Submitted to the Annals of Statistics
arXiv: math.PR/0000000

COMPUTATIONAL APPROACHES FOR EMPIRICAL BAYES METHODS AND BAYESIAN SENSITIVITY ANALYSIS

By Eugenia Buta and Hani Doss*

Yale University and University of Florida

We consider situations in Bayesian analysis where we have a family of priors ν_h on the parameter θ, where h varies continuously over a space $\mathcal{H}$, and we deal with two related problems. The first involves sensitivity analysis and is stated as follows. Suppose we fix a function f of θ. How do we efficiently estimate the posterior expectation of $f(\theta)$ simultaneously for all h in $\mathcal{H}$? The second problem is how do we identify subsets of $\mathcal{H}$ which give rise to reasonable choices of ν_h? We assume that we are able to generate Markov chain samples from the posterior for a finite number of the priors, and we develop a methodology, based on a combination of importance sampling and the use of control variates, for dealing with these two problems. The methodology applies very generally, and we show how it applies in particular to a commonly used model for variable selection in Bayesian linear regression, and give an illustration on the US crime data of Vandaele.

1. Introduction. In the Bayesian paradigm we have a data vector Y with density p_θ for some unknown $\theta \in \Theta$, and we wish to put a prior density on θ. The available family of prior densities is $\{\nu_h,\, h \in \mathcal{H}\}$, where h is called a hyperparameter. Typically, the hyperparameter is multivariate and choosing it can be difficult. But this choice is very important and can have a large impact on subsequent inference. There are two issues we wish to consider:

(A) Suppose we fix a quantity of interest, say $f(\theta)$, where f is a function. How do we assess how the posterior expectation of $f(\theta)$ changes as we vary h? More generally, how do we assess changes in the posterior distribution of $f(\theta)$ as we vary h?
(B) How do we determine if a given subset of $\mathcal{H}$ constitutes a class of reasonable choices?

The first issue is one of sensitivity analysis and the second is one of model selection.

*Supported by NSF Grant DMS-08-05860.

AMS 2000 subject classifications: Primary 62F15, 91-08; secondary 62F12.

Keywords and phrases: Bayes factors, control variates, ergodicity, hyperparameter selection, importance sampling, Markov chain Monte Carlo.

As an example of the kind of problem we wish to deal with, consider the problem of variable selection in Bayesian linear regression. Here, we have a response variable Y and a set of predictors $X_1, \ldots, X_q$, each a vector of length m. For every subset γ of $\{1, \ldots, q\}$ we have a potential model $\mathcal{M}_\gamma$ given by

$$Y = 1_m \beta_0 + X_\gamma \beta_\gamma + \epsilon,$$

where 1_m is the vector of m 1's, X_γ is the design matrix whose columns consist of the predictor vectors corresponding to the subset γ, β_γ is the vector of coefficients for that subset, and $\epsilon \sim \mathcal{N}_m(0, \sigma^2 I)$. Let q_γ denote the number of variables in the subset γ. The unknown parameter is $\theta = (\gamma, \sigma, \beta_0, \beta_\gamma)$, which includes the indicator of the subset of variables that go into the linear model. A very commonly used prior distribution on θ is given by a hierarchy in which we first choose the indicator γ from the "independence Bernoulli prior"—each variable goes into the model with a certain probability w, independently of all the other variables—and then choose the vector of regression coefficients corresponding to the selected variables. In more detail, the model is described as follows:

$$Y \sim \mathcal{N}_m(1_m \beta_0 + X_\gamma \beta_\gamma, \sigma^2 I), \tag{1.1a}$$

$$(\sigma^2, \beta_0) \sim p(\sigma^2, \beta_0) \propto 1/\sigma^2; \quad \text{given } \sigma, \ \beta_\gamma \sim \mathcal{N}_{q_\gamma}(0, g\sigma^2 (X_\gamma' X_\gamma)^{-1}), \tag{1.1b}$$

$$\gamma \sim w^{q_\gamma}(1-w)^{q-q_\gamma}. \tag{1.1c}$$

The prior on $(\sigma, \beta_0, \beta_\gamma)$ is Zellner's g-prior introduced in Zellner (1986), and is indexed by a hyperparameter g. Although this prior is improper, the resulting posterior distribution is proper.

Note that we have used the word "model" in two different ways: (i) a model is a specification of the hyperparameter h, and (ii) a model in regression is a list of variables to include. The meaning of the word will always be clear from context.

To summarize, the prior on the parameter $\theta = (\gamma, \sigma, \beta_0, \beta_\gamma)$ is given by the two-level hierarchy (1.1c) and (1.1b), and is indexed by $h = (w, g)$. Loosely speaking, when w is large and g is small, the prior encourages models with many variables and small coefficients, whereas when w is small and g is large, the prior concentrates its mass on parsimonious models with large coefficients. Therefore, the hyperparameter $h = (w, g)$ plays a very important role, and in effect determines the model that will be used to carry out variable selection.

A standard method for approaching model selection involves the use of Bayes factors. For each $h \in \mathcal{H}$, let $m_h(y)$ denote the marginal likelihood of the data under the prior ν_h, that is, $m_h(y) = \int p_\theta(y) \nu_h(\theta)\, d\theta$. We will write

m_h instead of $m_h(y)$. The Bayes factor of the model indexed by h_2 vs. the model indexed by h_1 is defined as the ratio of the marginal likelihoods of the data under the two models, m_{h_2}/m_{h_1}, and is denoted throughout by $B(h_2, h_1)$. Bayes factors are widely used as a criterion for comparing models in Bayesian analyses. For selecting models that are better than others from the family of models indexed by $h \in \mathcal{H}$, our strategy will be to compute and subsequently compare all the Bayes factors $B(h, h_1)$, for all $h \in \mathcal{H}$, and a fixed hyperparameter value h_1. We could then consider as good candidate models those with values of h that result in the largest Bayes factors.

Suppose now that we fix a particular function f of the parameter θ; for instance, in the example, this might be the indicator that variable 1 is included in the regression model. It is of general interest to determine the posterior expectation $E_h(f(\theta) \mid Y)$ as a function of h and to determine whether or not $E_h(f(\theta) \mid Y)$ is very sensitive to the value of h. If it is not, then two individuals using two different hyperparameters will reach approximately the same conclusions and the analysis will not be controversial. On the other hand, if for a function of interest the posterior expectation varies considerably as we change the hyperparameter, then we will want to know which aspects of the hyperparameter (e.g. which components of h) produce big changes and we may want to see a plot of the posterior expectations as we vary those aspects of the hyperparameter. Except for extremely simple cases, posterior expectations cannot be obtained in closed form, and are typically estimated via Markov chain Monte Carlo (MCMC). It is slow and inefficient to run Markov chains for every hyperparameter value h. Section 2 reviews an existing method for estimating $E_h(f(\theta) \mid Y)$ that bypasses the need to run a separate Markov chain for every h. The method has an analogue for the problem of estimating Bayes factors. Unfortunately, the method has severe limitations, which we also discuss.

In this paper we address the sensitivity analysis and model selection issues discussed above. Our approach involves running Markov chains corresponding to a few values of the hyperparameter, say $h_1, \dots, h_k$, and use these to estimate $E_h(f(\theta) \mid Y)$ for all $h \in \mathcal{H}$ and also the Bayes factors $B(h, h_1)$ for all $h \in \mathcal{H}$. The difficulty we face is that there is a severe computational burden caused by the requirement that we handle a very large number of values of h. Our approach for estimating large families of posterior expectations and Bayes factors is based on a combination of MCMC, importance sampling, and the use of control variates. The main contribution of this work is the development of theory to support the method. This theory can be used when dealing with implementation issues. The paper is organized as follows. In Section 2 we describe our methodology for estimating Bayes factors and posterior expectations, and give statements of theoretical results associated with the methodology. In Section 3 we discuss estimation of the variance and implementation issues. In Section 4 we return to the problem of variable selection in Bayesian linear regression, and show how our methodology applies in that model. The Appendix gives proofs of the theorems stated in the paper.

The idea of doing importance sampling using data streams from multiple densities has been investigated in several papers before. In Vardi (1985), Gill et al. (1988), Geyer (1994), Meng and Wong (1996), Kong et al. (2003), and Tan (2004), it is assumed that we have samples from each density and that each density is known except for a normalizing constant. The objective is to estimate all possible ratios of normalizing constants, and expectations of a given function with respect to each of the densities. The estimates in all these papers are identical, although the computational schemes to obtain them given in these papers are different. Gill et al. (1988) and Tan (2004) obtain the asymptotic distribution of the estimates when the samples are iid, and Geyer (1994) gives the asymptotic distribution when the samples are Markov chains satisfying certain regularity conditions.

Our Bayesian framework is the same as the framework described above. Let $\nu_{h,y}$ denote the posterior density of θ given $Y = y$ when the prior is ν_h. The posterior densities $\nu_{h_j,y}$ are given by $\nu_{h_j,y}(\theta) = p_\theta(y)\nu_{h_j}(\theta)/m_{h_j}$, where the functional form $p_\theta(y)\nu_{h_j}(\theta)$ is known, but the normalizing constant m_{h_j} is not. Our perspective is different from that of the previous authors in that we are interested in estimation of the ratios m_h/m_{h_1} and of posterior expectations $\int f(\theta)\nu_{h,y}(\theta)\,d\theta$ for a very large number of h's. Consequently, in addition to the obvious computational demands for handling many h's, we also have to deal with the fact that we will not have a sample from $\nu_{h,y}$ for every $h \in \mathcal{H}$, but only from $\nu_{h_j,y}$, $j = 1, \dots, k$. Thus, we are concerned with computational efficiency, in addition to statistical efficiency. These issues are discussed in detail in Section 2.

2. Estimation of Bayes Factors and Posterior Expectations.

Suppose that we have a sample $\theta_1, \dots, \theta_n$ (iid or ergodic Markov chain output) from the posterior density $\nu_{h_1,y}$ for a fixed h_1 and we are interested in the posterior expectation

$$E_h(f(\theta) \,|\, Y = y) = \int f(\theta) \frac{\nu_{h,y}(\theta)}{\nu_{h_1,y}(\theta)} \nu_{h_1,y}(\theta)\,d\theta \tag{2.1}$$

for different values of h. Using the fact that

$$\int \frac{p_\theta(y)\nu_h(\theta)/m_h}{p_\theta(y)\nu_{h_1}(\theta)/m_{h_1}} \nu_{h_1,y}(\theta)\,d\theta = 1,$$

we see that this expectation may be written as

$$\int f(\theta)\frac{p_\theta(y)\nu_h(\theta)/m_h}{p_\theta(y)\nu_{h_1}(\theta)/m_{h_1}}\nu_{h_1,y}(\theta)\,d\theta = \frac{\int f(\theta)\frac{\nu_h(\theta)}{\nu_{h_1}(\theta)}\nu_{h_1,y}(\theta)\,d\theta}{\int \frac{\nu_h(\theta)}{\nu_{h_1}(\theta)}\nu_{h_1,y}(\theta)\,d\theta}, \tag{2.2}$$

where the right side of (2.2) does not involve the ratio m_h/m_{h_1}. The idea to express $\int f(\theta)\nu_{h,y}(\theta)\,d\theta$ in this way was proposed in a different context by Hastings (1970). The right side of (2.2) is the ratio of two integrals with respect to $\nu_{h_1,y}$, each of which may be estimated from the sequence $\theta_1,\dots,\theta_n$. We may estimate the numerator and the denominator by

$$\frac{1}{n}\sum_{i=1}^{n} f(\theta_i)[\nu_h(\theta_i)/\nu_{h_1}(\theta_i)] \qquad \text{and} \qquad \frac{1}{n}\sum_{i=1}^{n}[\nu_h(\theta_i)/\nu_{h_1}(\theta_i)], \tag{2.3}$$

respectively, and $\int f(\theta)\nu_{h,y}(\theta)\,d\theta$ is estimated by the ratio of these two quantities.

The disappearance of the likelihood function on the right side of (2.2) is very convenient because its computation requires considerable effort in some cases (for example, when we have missing or censored data, the likelihood is a possibly high-dimensional integral). Note that the second average in (2.3) is an estimate of m_h/m_{h_1}, i.e. the Bayes factor $B(h,h_1)$. Ideally, we would like to use the estimates in (2.3) for multiple values of h using only a sample from the posterior distribution corresponding to the fixed hyperparameter value h_1. But, when the prior ν_h differs from ν_{h_1} greatly, the two estimates in (2.3) are unstable because of the potential that only a few observations will dominate the sums. Their ratio suffers the same defect.

A natural approach for dealing with the instability of these simple estimates is to choose k values $h_1,\dots,h_k \in \mathcal{H}$ and in (2.1) replace $\nu_{h_1,y}$ with a mixture $\sum_{s=1}^k a_s\nu_{h_s,y}$, where $a_s \geq 0$, for $s=1,\dots,k$, and $\sum_{s=1}^k a_s = 1$. For concreteness, consider the estimate of the Bayes factor. Let $\overline{\nu}_{\cdot y} = \sum_{s=1}^k a_s\nu_{h_s,y}$, and let $d_s = m_{h_s}/m_{h_1}$, $s=1,\dots,k$. Note that

$$B(h,h_1) = \int \frac{\nu_h(\theta)}{\sum_{s=1}^k a_s\nu_{h_s}(\theta)/d_s}\overline{\nu}_{\cdot y}(\theta)\,d\theta \tag{2.4}$$

and

$$\begin{aligned}\int f(\theta)\nu_{h,y}(\theta)\,d\theta &= (B(h,h_1))^{-1}\int f(\theta)\frac{\nu_h(\theta)}{\sum_{s=1}^k a_s\nu_{h_s}(\theta)/d_s}\overline{\nu}_{\cdot y}(\theta)\,d\theta \\ &= \frac{\int f(\theta)\frac{\nu_h(\theta)}{\sum_{s=1}^k a_s\nu_{h_s}(\theta)/d_s}\overline{\nu}_{\cdot y}(\theta)\,d\theta}{\int \frac{\nu_h(\theta)}{\sum_{s=1}^k a_s\nu_{h_s}(\theta)/d_s}\overline{\nu}_{\cdot y}(\theta)\,d\theta}. \qquad (2.5)\end{aligned}$$

(These two identities are valid under the condition that $\nu_h(\theta) = 0$ whenever $\nu_{h_s}(\theta) = 0$ for all s.) Suppose that for each $l = 1, \dots, k$ we have Markov chain samples $\theta_i^{(l)}$, $i = 1, \dots, n_l$ from the posterior density $\nu_{h_l,y}$. Letting $n = \sum_{s=1}^k n_s$, if $a_s = n_s/n$, then the pooled sample is a stratified sample from $\overline{\nu}_{.y}$. Doss (2010) considers the case where the vector $d = (d_2, \dots, d_k)'$ is known. In this situation, the right side of (2.4) is the integral of a known function with respect to the mixture density $\overline{\nu}_{.y}$. He shows that under certain regularity conditions, the estimate of $B(h, h_1)$ obtained by replacing the right side of (2.4) by its natural Monte Carlo estimate using the pooled sample is consistent and asymptotically normal.

In virtually all applications, the value of the vector d is unknown. The estimates of $B(h, h_1)$ and $\int f(\theta)\nu_{h,y}(\theta)\, d\theta$ that we consider in this paper are constructed by first forming an estimate $\hat{d}$ of d, and then using the natural Monte Carlo estimates of the integral in (2.4) and of the two integrals in (2.5) with $\hat{d}$ substituted for d. The MCMC scheme we will use involves the following two stages.

Stage 1 Generate samples $\theta_i^{(l)0}$, $i = 1, \dots, N_l$ from $\nu_{h_l,y}$, the posterior density of θ given $Y = y$, assuming that the prior is ν_{h_l}, for each $l = 1, \dots, k$, and use these $N = \sum_{l=1}^k N_l$ observations to form an estimate of d.

Stage 2 Independently of Stage 1, again generate samples $\theta_i^{(l)}$, $i = 1, \dots, n_l$ from $\nu_{h_l,y}$, for each $l = 1, \dots, k$, and construct the estimate of the Bayes factor $B(h, h_1)$ based on this second set of $n = \sum_{l=1}^k n_l$ observations and the estimate of d from Stage 1.

The estimate of d in Stage 1 is formed using a method introduced by Vardi (1985), and this estimate is discussed in the beginning of Section 2.1. From now on, for $l = 1, \dots, k$, we use the notations A_l and a_l to identify the ratios N_l/N and n_l/n, respectively.

It is natural to ask why we use two steps of sampling, instead of estimating the vector d and $B(h, h_1)$ from a single sample. The quantity considered in Doss (2010) is

$$\hat{B}(h, h_1, d) = \sum_{l=1}^{k} \sum_{i=1}^{n_l} \frac{\nu_h(\theta_i^{(l)})}{\sum_{s=1}^k n_s \nu_{h_s}(\theta_i^{(l)})/d_s}, \tag{2.6}$$

and it involves the vector d. The estimate considered in the present paper is $\hat{B}(h, h_1, \hat{d})$, where $\hat{d}$ is an estimate of d. The variance of $\hat{B}(h, h_1, \hat{d})$ turns out to be greater than that of $\hat{B}(h, h_1, d)$ (and this is true whether we use two steps of sampling or a single step). Thus, the variance decomposes as $\operatorname{Var}(\hat{B}(h, h_1, \hat{d})) = \operatorname{Var}(\hat{B}(h, h_1, d)) + V_d$, where V_d is the increase in variance resulting from using $\hat{d}$ instead of d. Because we wish to estimate $B(h, h_1)$ for a large number of h's and for each h the computational time needed is linear in the total sample size, this total sample size cannot be very large. On the other hand, d needs to be estimated only once. So if generating the chains is not computationally demanding, then one can use very long chains to estimate d and so greatly reduce the term V_d. A precise statement regarding the benefits of the two-stage scheme would have to take into account the cost of computing the typical term in (2.6) and the cost of generating a point in the chain, and no such statement can be made at the level of generality considered in this paper. However, in all the examples we have encountered, for fixed computational resources, the two-stage scheme gives estimates with considerably smaller variance. We mention here that our theoretical results are stated for the two-stage schemes, but these results have analogues for the case where a single sample is used to estimate both d and the family of Bayes factors $B(h, h_1)$, $h \in \mathcal{H}$, and these are given in Buta (2010).

A summary of the main contributions of the present work is as follows.

1. We develop a complete characterization of the asymptotic distribution of the estimate (2.6) and also of a variant involving the use of control variates developed by Doss (2010) for the realistic case where d is estimated from Stage 1 sampling. Included in our results is an explicit formula for the increase in variance resulting from using an estimate of d instead of d itself. (This contradicts statements in the literature to the effect that using a $\sqrt{n}$-consistent estimate of d rather than d itself does not inflate the variance; see our discussion in the Appendix.)
2. We develop an analogous theory for the problem of estimating a family of posterior expectations $E_h(f(\theta) \mid Y = y)$, $h \in \mathcal{H}$.
3. For any of our estimators, the variance is a sum of two components, and we discuss how each of these may be estimated. An important problem is how to properly select the skeleton points $h_1, \dots, h_k$, and ideally we would like to position these in such a way that the variance is minimized. We show how the variance estimates can be used to suggest good sets of skeleton points.
4. We apply the methodology to the problem of Bayesian variable selection discussed earlier. In particular, we show how our methods enable us to select good values of $h = (w, g)$ and to also see how the probability that a given variable is included in the regression varies with (w, g).

2.1. *Estimation of Bayes Factors.* Here, we analyze the asymptotic distributional properties of the estimator that results if in (2.6) we replace d with an estimate. Geyer (1994) proposes an estimator for d based on the "reverse logistic regression" method and Theorem 2 therein shows that this estimator is asymptotically normal when the samplers used satisfy certain regularity conditions. This estimator is obtained by maximizing with respect to $d_2, \dots, d_k$ the log quasi-likelihood

$$l_N(d) = \sum_{l=1}^{k} \sum_{i=1}^{N_l} \log\left(\frac{A_l \nu_{h_l}(\theta_i^{(l)0})/d_l}{\sum_{s=1}^{k} A_s \nu_{h_s}(\theta_i^{(l)0})/d_s} \right). \tag{2.7}$$

As was mentioned earlier, the estimate is the same as the estimates obtained by Vardi (1985), Meng and Wong (1996), and Kong et al. (2003). We assume that for all the Markov chains we use a Strong Law of Large Numbers (SLLN) holds for all integrable functions [for sufficient conditions see, e.g., Theorem 2 of Athreya et al. (1996)]. In the next theorem, we show that if $\hat{d}$ is the estimate produced by Geyer's (1994) method, or any of the equivalent estimates discussed above, then the estimate of the Bayes factor given by

$$\hat{B}(h, h_1, \hat{d}) = \sum_{l=1}^{k} \sum_{i=1}^{n_l} \frac{\nu_h(\theta_i^{(l)})}{\sum_{s=1}^{k} n_s \nu_{h_s}(\theta_i^{(l)})/\hat{d}_s} \tag{2.8}$$

is asymptotically normal if certain regularity conditions are met. In (2.8), $\hat{d}_1 = 1$.

Before we state the theorem, we need to define the expressions that appear in the asymptotic variance. For $l = 1, \dots, k$, $i = 1, \dots, n_l$, let

$$Y_{i,l} = \frac{\nu_h(\theta_i^{(l)})}{\sum_{s=1}^{k} a_s \nu_{h_s}(\theta_i^{(l)})/d_s} \tag{2.9}$$

(the $Y_{i,l}$'s depend on h, but this dependence is suppressed to lighten the notation), and let

$$\tau_l^2(h) = \mathrm{Var}(Y_{1,l}) + 2 \sum_{g=1}^{\infty} \mathrm{Cov}(Y_{1,l}, Y_{1+g,l}), \qquad \tau^2(h) = \sum_{l=1}^{k} a_l \tau_l^2(h).$$

Also, let $c(h)$ be the vector of length $k-1$ for which the $(j-1)^{\text{th}}$ coordinate is

$$[c(h)]_{j-1} = \frac{B(h, h_1)}{d_j^2} \int \frac{a_j \nu_{h_j}(\theta)}{\sum_{s=1}^{k} a_s \nu_{h_s}(\theta)/d_s} \cdot \nu_{h,y}(\theta)\, d\theta, \quad j = 2, \dots, k. \tag{2.10}$$

THEOREM 1. *Let $h \in \mathcal{H}$ be fixed. Suppose the chains in Stage 2 satisfy conditions A1 and A2 in Doss (2010):*

A1 *For each $l = 1, \ldots, k$, the chain $\{\theta_i^{(l)}\}_{i=1}^{\infty}$ is geometrically ergodic.*
A2 *For each $l = 1, \ldots, k$, there exists $\epsilon > 0$ such that*

$$E\left(\left|\frac{\nu_h(\theta_1^{(l)})}{\sum_{s=1}^k a_s \nu_{h_s}(\theta_1^{(l)})/d_s}\right|^{2+\epsilon}\right) < \infty. \tag{2.11}$$

In the expectation in (2.11), $\theta_1^{(l)} \sim \nu_{h_l,y}$. Assume also that the chains in Stage 1 satisfy the conditions in Theorem 2 of Geyer (1994) that imply $\sqrt{N}(\hat{d} - d) \xrightarrow{d} \mathcal{N}(0, \Sigma)$. In addition, suppose the total sample sizes for the two stages, N and n, satisfy $n \to \infty$, and $N \to \infty$ in such a way that $n/N \to q \in [0, \infty)$. Then

$$\sqrt{n}\big(\hat{B}(h, h_1, \hat{d}) - B(h, h_1)\big) \xrightarrow{d} \mathcal{N}\big(0, qc(h)'\Sigma c(h) + \tau^2(h)\big).$$

As alluded to earlier, there are two components to the expression for the variance. The first component arises from estimating d, and the second component is the variance that we would have if we had estimated the Bayes factor knowing what d is. As can be seen from the formula, the first component vanishes if $q = 0$, i.e., if the sample size for estimating the parameter d converges to infinity at a faster rate than does the sample size used to estimate the Bayes factor. In this case the Bayes factor estimator (2.8) using the estimate $\hat{d}$ has the same asymptotic distribution as the estimator in (2.6) which uses the true value of d. Otherwise, the variance of (2.8) is greater than that of (2.6), and the difference between the variances depends on the parameter q. This parameter is determined by the user and should be chosen in such a way as to minimize the variance given computer resources; this is discussed in Section 3.

2.2. *Estimation of Bayes Factors Using Control Variates.* Recall that we have samples $\theta_i^{(l)}$, $i = 1, \ldots, n_l$ from $\nu_{h_l,y}$, $l = 1, \ldots, k$, with independence across samples (Stage 2 of sampling) and that, based on an independent set of preliminary MCMC runs (Stage 1 of sampling), we have estimated the constants $d_2, \ldots, d_k$. Also, $n_l/n = a_l$ and $n = \sum_{l=1}^k n_l$. Let

$$Y(\theta) = \frac{\nu_h(\theta)}{\sum_{s=1}^k a_s \nu_{h_s}(\theta)/d_s}. \tag{2.12}$$

Recalling that $\overline{\nu}_{.y} := \sum_{s=1}^k a_s \nu_{h_s,y}$, we have $E_{\overline{\nu}_{.y}}(Y(\theta)) = B(h, h_1)$, where the subscript $\overline{\nu}_{.y}$ to the expectation indicates that $\theta \sim \overline{\nu}_{.y}$. Also, for $j =$

$2, \dots, k$, let

$$Z^{(j)}(\theta) = \frac{\nu_{h_j}(\theta)/d_j - \nu_{h_1}(\theta)}{\sum_{s=1}^k a_s \nu_{h_s}(\theta)/d_s} \tag{2.13}$$

$$= \frac{\nu_{h_j,y}(\theta) - \nu_{h_1,y}(\theta)}{\sum_{s=1}^k a_s \nu_{h_s,y}(\theta)}. \tag{2.14}$$

Expression (2.14) shows that $E_{\overline{\nu}.y}(Z^{(j)}(\theta)) = 0$. This is true even if the priors ν_{h_j} and ν_{h_1} are improper, as long as the posteriors $\nu_{h_j,y}$ and $\nu_{h_1,y}$ are proper, exactly our situation in the Bayesian variable selection example of Section 1. On the other hand, the representation (2.13) shows that $Z^{(j)}(\theta)$ is computable if we know the d_j's—it involves the priors and not the posteriors. (A similar remark applies to (2.12).) Therefore, if as in Doss (2010) we define for $l = 1, \dots, k$, $i = 1, \dots, n_l$

$$Z_{i,l}^{(1)} = 1, \qquad Z_{i,l}^{(j)} = \frac{\nu_{h_j}(\theta_i^{(l)})/d_j - \nu_{h_1}(\theta_i^{(l)})}{\sum_{s=1}^k a_s \nu_{h_s}(\theta_i^{(l)})/d_s}, \quad j = 2, \dots, k, \tag{2.15}$$

then for any fixed $\boldsymbol{\beta} = (\beta_2, \dots, \beta_k)$,

$$\hat{I}_{\boldsymbol{\beta}}^d = \frac{1}{n} \sum_{l=1}^k \sum_{i=1}^{n_l} \left(Y_{i,l} - \textstyle\sum_{j=2}^k \beta_j Z_{i,l}^{(j)}\right), \tag{2.16}$$

is an unbiased estimate of $B(h, h_1)$. The value of $\boldsymbol{\beta}$ that minimizes the variance of $\hat{I}_{\boldsymbol{\beta}}^d$ is unknown. As is commonly done when one uses control variates, we use instead the estimate obtained by doing ordinary linear regression of the response $Y_{i,l}$ on the predictors $Z_{i,l}^{(j)}$, $j = 2, \dots, k$, and to emphasize that this estimate depends on d, we denote it by $\hat{\boldsymbol{\beta}}(d)$. Doss (2010) shows that $\hat{\boldsymbol{\beta}}(d)$ converges almost surely to a finite limit, $\boldsymbol{\beta}_{\text{lim}}$. His Theorem 1 states that the estimator $\hat{B}_{\text{reg}}(h, h_1) = \hat{I}_{\hat{\boldsymbol{\beta}}(d)}^d$, obtained under the assumption that we know the constants $d_2, \dots, d_k$, has an asymptotically normal distribution. As mentioned earlier, $d_2, \dots, d_k$ are typically unknown, and must be estimated. Let $\hat{d}_2, \dots, \hat{d}_k$ be estimates obtained from previous MCMC runs and let

$$\hat{I}_{\hat{\boldsymbol{\beta}}(\hat{d})}^{\hat{d}} = \frac{1}{n} \sum_{l=1}^k \sum_{i=1}^{n_l} \left(\hat{Y}_{i,l} - \sum_{j=2}^k \hat{\beta}_j(\hat{d}) \hat{Z}_{i,l}^{(j)}\right), \tag{2.17}$$

where $\hat{Y}_{i,l}$ and $\hat{Z}_{i,l}^{(j)}$ are as in (2.9) and (2.15), except using $\hat{d}$ for d, and $\hat{\boldsymbol{\beta}}(\hat{d})$ is the least squares regression estimator from regressing $\hat{Y}_{i,l}$ on predictors $\hat{Z}_{i,l}^{(j)}$, $j = 2, \dots, k$.

The next theorem gives the asymptotic distribution of this new estimator, and before we state it we introduce some notation. Let

$$U_{i,l} = Y_{i,l} - \sum_{j=2}^{k} \beta_{j,\lim} Z_{i,l}^{(j)}, \tag{2.18}$$

and let

$$\sigma_l^2(h) = \mathrm{Var}(U_{1,l}) + 2\sum_{g=1}^{\infty} \mathrm{Cov}(U_{1,l}, U_{1+g,l}), \quad \sigma^2(h) = \sum_{l=1}^{k} a_l \sigma_l^2(h). \tag{2.19}$$

Also, let $w(h)$ be the vector of length $k-1$ for which the $(t-1)^{\text{th}}$ coordinate $(t = 2, \dots, k)$ is

$$\begin{aligned} [w(h)]_{t-1} = {} & \frac{B(h,h_1)}{d_t^2} \int \frac{a_t \nu_{h_t}(\theta)}{\sum_{s=1}^{k} a_s \nu_{h_s}(\theta)/d_s} \cdot \nu_{h,y}(\theta)\, d\theta + \beta_{t,\lim} \frac{1}{d_t} \\ & + \sum_{j=2}^{k} \beta_{j,\lim} \int \frac{a_t \nu_{h_t}(\theta)}{d_t^2 \sum_{s=1}^{k} a_s \nu_{h_s}(\theta)/d_s} \cdot \big(\nu_{h_1,y}(\theta) - \nu_{h_j,y}(\theta)\big)\, d\theta. \end{aligned} \tag{2.20}$$

THEOREM 2. *Suppose all the conditions from Theorem 1 are satisfied. Moreover, assume that $\boldsymbol{R}$, the $k \times k$ matrix defined by*

$$R_{j,j'} = E\Big(\textstyle\sum_{l=1}^{k} a_l Z_{1,l}^{(j)} Z_{1,l}^{(j')}\Big), \qquad j, j' = 1, \dots, k,$$

is nonsingular. Then

$$\sqrt{n}\big(\hat{I}^{\hat{d}}_{\hat{\boldsymbol{\beta}}(\hat{d})} - B(h,h_1)\big) \xrightarrow{d} \mathcal{N}\big(0, q w(h)' \Sigma w(h) + \sigma^2(h)\big).$$

As mentioned above, for any $\boldsymbol{\beta}$, $\hat{I}^d_{\boldsymbol{\beta}}$ in (2.16) is an unbiased estimate of $B(h,h_1)$, which leads to the question of what is the optimal value of $\boldsymbol{\beta}$ to use. It is not difficult to see that when each of the sequences $\{\theta_i^{(l)}\}_{i=1}^{n_l}$ is iid, the value of $\boldsymbol{\beta}$ that minimizes the variance of $\hat{I}^d_{\boldsymbol{\beta}}$ is

$$\boldsymbol{\beta}_{\text{opt, iid}} := \mathrm{argmin}_{\boldsymbol{\beta}} \mathrm{Var}_{\overline{\nu}_{\cdot y}}\big(Y(\theta) - \sum_{j=2}^{k} \beta_j Z^{(j)}(\theta)\big),$$

i.e. the optimal value is the same whether we have a random sample from $\overline{\nu}_{\cdot y}$ or a stratified sample. It is natural to ask whether $\boldsymbol{\beta}_{\text{opt, iid}}$ is still optimal when the k sequences $\{\theta_i^{(l)}\}_{i=1}^{n_l}$ are Markov chains. It turns out that

(i) $\boldsymbol{\beta}_{\text{opt, iid}}$ is not optimal,
(ii) using $\boldsymbol{\beta}_{\text{opt, iid}}$ can actually increase the variance (when the Markov chains mix at significantly different rates, chains that are of the same length do not have the same "effective sample sizes," but $\boldsymbol{\beta}_{\text{opt, iid}}$ does not reflect this fact).

In our experience, using $\boldsymbol{\beta}_{\text{opt, iid}}$, or more precisely the least squares estimate (which in Doss (2010) was shown to converge almost surely to $\boldsymbol{\beta}_{\text{opt, iid}}$), typically gives a significant reduction in variance. Buta and Doss (2011) prove points (i) and (ii) above and also discuss an approach for estimating the value of $\boldsymbol{\beta}$ that is optimal in the Markov chain case.

2.3. *Estimation of Posterior Expectations.* In this section we describe a method for estimating the posterior expectation of a function f when the prior is ν_h. Let us denote this quantity by

$$I^{[f]}(h) = \int f(\theta)\nu_{h,y}(\theta)\, d\theta.$$

Define

$$\begin{aligned} Y_{i,l}^{[f]} = \frac{f(\theta_i^{(l)})\nu_h(\theta_i^{(l)})}{\sum_{s=1}^k a_s \nu_{h_s}(\theta_i^{(l)})/d_s} &= \frac{f(\theta_i^{(l)})\nu_h(\theta_i^{(l)})/m_h}{\sum_{s=1}^k a_s \nu_{h_s}(\theta_i^{(l)})/m_{h_s}} \cdot \frac{m_h}{m_{h_1}} \\ &= \frac{f(\theta_i^{(l)})\nu_{h,y}(\theta_i^{(l)})}{\sum_{s=1}^k a_s \nu_{h_s,y}(\theta_i^{(l)})}\ B(h, h_1). \end{aligned}$$

With the view of applying identity (2.5), we note that, assuming a SLLN holds for the Markov chains $\theta_i^{(l)}$, $l = 1, \ldots, k$, $i = 1, \ldots, n_l$, we have

$$\begin{aligned} \frac{1}{n}\sum_{l=1}^k \sum_{i=1}^{n_l} Y_{i,l}^{[f]} &= \sum_{l=1}^k \frac{1}{n_l}\sum_{i=1}^{n_l} \frac{n_l}{n} Y_{i,l}^{[f]} \\ &\xrightarrow{\text{a.s.}} \int \frac{f(\theta)\nu_{h,y}(\theta)}{\sum_{s=1}^k a_s \nu_{h_s,y}(\theta)} \sum_{l=1}^k a_l \nu_{h_l,y}(\theta)\, d\theta \cdot B(h, h_1) = I^{[f]}(h) \cdot B(h, h_1) \end{aligned}$$

and

$$\frac{1}{n}\sum_{l=1}^k \sum_{i=1}^{n_l} Y_{i,l} \xrightarrow{\text{a.s.}} B(h, h_1).$$

(The $Y_{i,l}$'s are defined in (2.9); note that $Y_{i,l} = Y_{i,l}^{[f]}$ when $f \equiv 1$.) Letting

$$\hat{I}^{[f]}(h, d) = \frac{\sum_{l=1}^k \sum_{i=1}^{n_l} Y_{i,l}^{[f]}}{\sum_{l=1}^k \sum_{i=1}^{n_l} Y_{i,l}}, \tag{2.21}$$

we see that $\hat{I}^{[f]}(h,d) \xrightarrow{\text{a.s.}} I^{[f]}(h)$, and replacing d with the estimate $\hat{d}$ obtained from Stage 1 sampling, we form

$$\hat{I}^{[f]}(h,\hat{d}) = \frac{\displaystyle\sum_{l=1}^{k}\sum_{i=1}^{n_l} \frac{f(\theta_i^{(l)})\nu_h(\theta_i^{(l)})}{\sum_{s=1}^k a_s \nu_{h_s}(\theta_i^{(l)})/\hat{d}_s}}{\displaystyle\sum_{l=1}^{k}\sum_{i=1}^{n_l} \frac{\nu_h(\theta_i^{(l)})}{\sum_{s=1}^k a_s \nu_{h_s}(\theta_i^{(l)})/\hat{d}_s}}. \tag{2.22}$$

The following theorem concerns the asymptotic behavior of this estimator, and to state it we first define the expressions that appear in the asymptotic variance. Let

$$\begin{aligned}
\gamma_{11} &= \operatorname{Var}\bigl(Y_{1,l}^{[f]}\bigr) + 2\textstyle\sum_{g=1}^{\infty} \operatorname{Cov}\bigl(Y_{1,l}^{[f]}, Y_{1+g,l}^{[f]}\bigr),\\
\gamma_{12} = \gamma_{21} &= \operatorname{Cov}\bigl(Y_{1,l}^{[f]}, Y_{1,l}\bigr) + \textstyle\sum_{g=1}^{\infty}\bigl[\operatorname{Cov}\bigl(Y_{1,l}^{[f]}, Y_{1+g,l}\bigr) + \operatorname{Cov}\bigl(Y_{1,l}, Y_{1+g,l}^{[f]}\bigr)\bigr],\\
\gamma_{22} &= \operatorname{Var}\bigl(Y_{1,l}\bigr) + 2\textstyle\sum_{g=1}^{\infty} \operatorname{Cov}\bigl(Y_{1,l}, Y_{1+g,l}\bigr)
\end{aligned}$$

and

$$\Gamma_l(h) = \begin{pmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{pmatrix}, \qquad \Gamma(h) = \sum_{l=1}^{k} a_l \Gamma_l(h). \tag{2.23}$$

Since (2.21) and (2.22) are ratios to which we will apply the delta method, we will consider the function $g(u,v) = u/v$, whose gradient is $\nabla g(u,v) = (1/v, -u/v^2)'$. Let

$$\rho(h) = \nabla g\bigl(I^{[f]}(h)B(h,h_1), B(h,h_1)\bigr)' \cdot \Gamma(h) \cdot \nabla g\bigl(I^{[f]}(h)B(h,h_1), B(h,h_1)\bigr). \tag{2.24}$$

Finally, let $v(h)$ be the vector of length $k-1$ for which the $(j-1)^{\text{th}}$ coordinate is

$$[v(h)]_{j-1} = \int \frac{[f(\theta) - I^{[f]}(h)]a_j \nu_{h_j}(\theta)/d_j^2}{\sum_{s=1}^k a_s \nu_{h_s}(\theta)/d_s} \nu_{h,y}(\theta)\, d\theta, \quad j = 2, \dots, k. \tag{2.25}$$

THEOREM 3. *Suppose the conditions stated in Theorem 1 are satisfied and, in addition, for each $l = 1, \dots, k$, there exists an $\epsilon > 0$ such that*

$$E\bigl(\bigl|Y_{1,l}^{[f]}\bigr|^{2+\epsilon}\bigr) < \infty. \tag{2.26}$$

Then

$$\sqrt{n}\bigl(\hat{I}^{[f]}(h,\hat{d}) - I^{[f]}(h)\bigr) \xrightarrow{d} \mathcal{N}\bigl(0, q v(h)'\Sigma v(h) + \rho(h)\bigr).$$

The numerator of $\hat{I}^{[f]}(h, \hat{d})$ is an estimate of $I^{[f]}(h)\, B(h, h_1)$ and the denominator is an estimate of $B(h, h_1)$. It is possible to adjust both the numerator and denominator through the use of control variates and thus arrive at a variant of $\hat{I}^{[f]}(h, \hat{d})$; the theory for this is developed in Buta (2010). As for the case of estimating the Bayes factors, the variant is not guaranteed to give an improvement, but a large improvement is often noted.

3. Variance Estimation and Selection of the Skeleton Points. Estimation of the variance of our estimates is important for several reasons. In addition to the usual need for providing error margins for our point estimates, variance estimates are of great help in selecting the skeleton points. The main approaches for estimation of the variance are (i) spectral methods, (ii) methods based on batching, and (iii) methods based on regeneration; see Flegal and Jones (2010) and Mykland et al. (1995) for a review. Methods based on batching are difficult to use in our framework because of two complications, namely that we are dealing with multiple chains, and we have a two-stage scheme; and procedures based on regeneration are often difficult to implement. Here we describe a way of estimating the variance using spectral methods.

For the sake of concreteness, consider $\hat{B}(h, h_1, \hat{d})$, whose asymptotic variance is the expression $\kappa^2(h) = qc(h)'\Sigma c(h) + \tau^2(h)$ (see Theorem 1). The term $\tau^2(h)$ is the asymptotic variance of the quantity $\hat{B}(h, h_1, d)$ in (2.6), and since the k Markov chains are independent, $\tau^2(h) = \sum_{l=1}^{k} a_l \tau_l^2(h)$, where $\tau_l^2(h)$ is the asymptotic variance of

$$\frac{1}{n_l} \sum_{i=1}^{n_l} \frac{\nu_h(\theta_i^{(l)})}{\sum_{s=1}^{k} a_s \nu_{h_s}(\theta_i^{(l)})/d_s}. \tag{3.1}$$

Now for each l we will estimate $\tau_l^2(h)$ by the asymptotic variance of

$$\frac{1}{n_l} \sum_{i=1}^{n_l} \frac{\nu_h(\theta_i^{(l)})}{\sum_{s=1}^{k} a_s \nu_{h_s}(\theta_i^{(l)})/\hat{d}_s}, \tag{3.2}$$

where $\hat{d}$ is formed from Stage 1 runs. It is not too difficult to show that under our asymptotic regime where $n/N \to q \in [0, \infty)$, standard consistent spectral estimates of the asymptotic variance of (3.2) are also consistent estimates of the asymptotic variance of (3.1); details are given in Buta and Doss (2011). Geyer (1994) gives an expression for Σ that is explicit enough to enable us to estimate it via standard spectral methods. Now, $c(h)$ is a vector each of whose components is an integral with respect to the posterior $\nu_{h,y}$

(see (2.10)). The estimate derived in Section 2.3 (see (2.22)) is designed precisely to estimate such posterior expectations. Combining, we arrive at an overall estimate of $\kappa^2(h)$, and the asymptotic variances of our other estimates are handled similarly.

Selection of the Skeleton Points. The asymptotic variances of any of our estimates depend on the choice of the points $h_1, \dots, h_k$. For concreteness, consider $\hat{B}(h, h_1, \hat{d})$, and to emphasize this dependence, let $V(h, h_1, \dots, h_k)$ denote the asymptotic variance of $\hat{B}(h, h_1, \hat{d})$. For fixed $h_1, \dots, h_k$, identifying the set of h's for which $V(h, h_1, \dots, h_k)$ is *finite* is typically a feasible problem. For instance, Doss (1994) considered the pump data example discussed in Tierney (1994), for which the hyperparameter h has dimension 3, and determined this set for the case $k = 1$. He showed that one can go as far away from h_1 as one wants in certain directions, but in other directions the range is limited. (The calculation can be extended to any k.) Suppose now that we fix a range $\mathcal{H}$ over which h is to vary. A necessary first step is to select $h_1, \dots, h_k$ such that $V(h, h_1, \dots, h_k) < \infty$ for all $h \in \mathcal{H}$. Typically however, we will want more, and we will face the problem below.

Design Problem Find the values of the skeleton points $h_1, \dots, h_k$ that minimize $\max_{h \in \mathcal{H}} V(h, h_1, \dots, h_k)$.

Unfortunately, except for extremely simple cases, it is not possible to calculate $V(h, h_1, \dots, h_k)$ analytically (even if $k = 1$, $V(h, h_1)$ is an infinite sum each of whose terms depends on the Markov transition function in a complicated way), and maximizing it over $h \in \mathcal{H}$ would present additional difficulties. Furthermore, even if we were able to calculate $\max_{h \in \mathcal{H}} V(h, h_1, \dots, h_k)$, the design problem would involve the minimization of a function of $k \times \dim(\mathcal{H})$ variables, and in general, solving the design problem is hopeless.

In our experience, we have found that the following method works reasonably well. Having specified the range $\mathcal{H}$, we select trial values $h_1, \dots, h_k$ and plot the estimated variance as a function of h, using one of the methods described above. If we find a region in $\mathcal{H}$ where this variance is unacceptably large, we "cover" this region by moving some h_l's closer to the region, or by simply adding new h_l's in that region, which increases k. This is illustrated in the example in Section 4.

The Relative Lengths of the Stage 1 and Stage 2 Chains. The parameter q affects the performance of any of the methods, and the optimal value involves a tradeoff between time spent calculating density ratios in Stage 2 and time spent generating the chains in Stage 1. Consider for instance the estimate (2.8), whose asymptotic variance is given by Theorem 1 and which we will write as $\kappa^2(h) = qv_1(h) + v_2(h)$. In the discussion below, we assume that we have run a small pilot experiment that has enabled us to adequately estimate the components $v_1(h)$ and $v_2(h)$, and we assume that the total sample sizes n and N are both large. The discussion is heuristic in that we assume that $v_1(h)$ and $v_2(h)$ are nearly constant in h. Let t_1 denote the time it typically takes to generate a single step in a chain, let t_2 denote the time it takes to compute the typical term in (2.8), and let g denote the number of values in $\mathcal{H}$ for which we wish to compute the estimate (2.8). Suppose we are given a computational budget of T units of time. For any $q \in (0, \infty)$, the time it takes to compute (2.8) for g values of h is $t(q) = (n/q)t_1 + nt_1 + ngt_2$, and setting this equal to T determines n to be $qT/\big((q+1)t_1 + qgt_2\big)$. The variance of the estimate is then $V(q) = T^{-1}\big(v_1(h) + v_2(h)/q\big)\big((q+1)t_1 + qgt_2\big)$. Clearly $V(q)$ is unbounded as $q \to 0$ or $q \to \infty$. The function has a unique minimum, which occurs at $q_{\text{opt}} = \sqrt{[v_2(h)t_1]/[v_1(h)(t_1 + gt_2)]}$. This last formula expresses in a usable manner the intuitive notion that if g is large, or if the cost of evaluating the density ratios in (2.8) is high relative to the cost of running the chains, then a small value of q should be used.

4. Illustration on Variable Selection in Bayesian Linear Regression.

There exist many classes of problems in Bayesian analysis in which the sensitivity analysis and model selection issues discussed earlier arise; see Section 5. Here we give an illustration involving the hierarchical prior used in variable selection in the Bayesian linear regression model discussed in Section 1. For this model, the parameter is the vector $\theta = (\gamma, \sigma, \beta_0, \beta_\gamma)$, and the prior on θ is given by the hierarchy (1.1c) and (1.1b). There exist several MCMC-based methods for estimating the posterior distribution of θ given $Y = y$, and the algorithm we use here is based on the Gibbs sampler of Smith and Kohn (1996), which runs on the space of model indicators. Our algorithm, developed in Buta (2010), is a Markov chain on θ that is uniformly ergodic and also computationally efficient (it avoids the need for repeated time-consuming matrix inversion). It is implemented in the R package `bvslr`, available from `http://www.stat.ufl.edu/~ebuta/BVSLR`.

In Sections 1 and 2, ν_h and $\nu_{h,y}$ refer to the prior and posterior *densities*, and all estimates in Section 2 involve ratios of these prior densities. In the Bayesian linear regression model that we are considering here, the priors ν_h on $(\gamma, \sigma, \beta_0, \beta_\gamma)$ are actually probability measures on $\{0,1\}^q \times (0,\infty) \times \mathbb{R}^{q+1}$, which in fact are not absolutely continuous with respect to the product of counting measure on $\{0,1\}^q$ and Lebesgue measure on $(0,\infty) \times \mathbb{R}^{q+1}$. For $h_1 = (w_1, g_1)$ and $h_2 = (w_2, g_2)$, the Radon-Nikodym derivative of ν_{h_1} with respect to ν_{h_2} is given by

(4.1)
$$\left[\frac{d\nu_{h_1}}{d\nu_{h_2}}\right](\gamma,\sigma,\beta_0,\beta_\gamma) = \left(\frac{w_1}{w_2}\right)^{q_\gamma}\left(\frac{1-w_1}{1-w_2}\right)^{q-q_\gamma} \times \frac{\phi_{q_\gamma}(\beta_\gamma;0,g_1\sigma^2(X'_\gamma X_\gamma)^{-1})}{\phi_{q_\gamma}(\beta_\gamma;0,g_2\sigma^2(X'_\gamma X_\gamma)^{-1})},$$

where $\phi_{q_\gamma}(u;a,V)$ is the density of the q_γ-dimensional normal distribution with mean a and covariance V, evaluated at u (Doss (2007)). It is immediate that all formulas in Section 2 remain valid if ratios of the form $\nu_h(\theta)/\nu_{h_1}(\theta)$ (see, e.g., equation (2.3)) are replaced by the Radon-Nikodym derivative $[d\nu_h/d\nu_{h_1}](\theta)$. Fortunately, evaluation of (4.1) requires neither matrix inversion nor calculation of a determinant, so can be done very quickly. Note that in view of (4.1), it is not enough to have Markov chains running on the γ's and we need Markov chains running on the θ's (or at least $(\gamma,\sigma,\beta_\gamma)$).

There is a large literature on dealing with the hyperparameter in models involving Zellner's g-prior (with or without the variable inclusion line (1.1c)). Some of the proposals involve putting a prior on g, or on both g and w. Liang et al. (2008) propose and discuss priors on g; priors on w are generally taken to be beta distributions. Other proposals give g as a deterministic function of m and q (e.g., $g = \max\{m, q^2\}$ in Fernandez et al. (2001)). Liang et al. (2008) contains an extensive and critical review of the recommendations given in this literature. The most common deterministic choice for w is $w = 1/2$. George and Foster (2000) recommend the empirical Bayes (EB) approach for estimating the pair (w,g): the marginal likelihood of (w,g) is computed over a grid, and the value of (w,g) that maximizes it is taken as the estimate of (w,g). As with many likelihood-based methods, special care needs to be taken when the maximizing value is at the boundary. Cui and George (2008) give evidence that the EB method outperforms fully Bayes methods in this problem. Unfortunately, the EB method is in general computationally demanding because the likelihood is a sum over all 2^q models γ, so it is practically feasible only for relatively small values of q. Our methodology handles this problem by estimating ratios of marginal likelihoods, i.e. Bayes factors and, besides giving the maximizing values of w and g, gives a plot which shows the behavior of the Bayes factors for a wide range of other values of w and g.

We illustrate our methods on the US crime data of Vandaele (1978), which can be found in the R library `MASS` under the name `UScrime`. This data set seems ideal, because it has been studied in several papers already so we can compare our results with previous analyses, and also because its modest size enables a closed-form calculation of the marginal likelihood m_h, so we can compare our estimates with the gold standard. The data set gives, for each of $m = 47$ states of the USA, the crime rate, defined as number of offenses per

100,000 individuals (the response variable), and $q = 15$ predictors measuring different characteristics of the population, such as average number of years of schooling, average income, unemployment rate, etc.

To be consistent with what is done in the literature, we applied a log transformation to all variables, except the indicator variable. We took the baseline hyperparameter to be $h_1 = (w_1, g_1) = (.5, 15)$, and our goal was to estimate $B(h, h_1)$ for the 924 values of h obtained when w ranges from 0.1 to 0.91 by increments of 0.03, and g ranges from 4 to 100 by increments of 3. We used (2.17) and this estimate was based on 16 chains each of length 10,000, corresponding to the skeleton grid of hyperparameter values

$$(w, g) \in \{.3, .5, .6, .8\} \times \{15, 50, 100, 225\} \tag{4.2}$$

for the Stage 1 samples, and 16 new chains, each of length 1000, corresponding to the same hyperparameter values, for the Stage 2 samples. The plots in Figure 1 give graphs of the estimate (2.17) as w and g vary, from two different angles. These indicate that values for w around 0.65 and for g around 20 seem appropriate, while values of w less than .3 and values of g greater than 60 should be avoided. A side calculation showed that, interestingly, for $g = \max\{m, q^2\}$ $(= 225)$, the estimate of $B\big((w, g), (.65, 20)\big)$ is less than .008 regardless of the value of w, so this choice should not be used for this data set. With the long chains used and the estimate that uses control variates, the Bayes factor estimates in Figure 1 are extremely accurate—root mean squared errors are less than 0.04 uniformly over the entire domain of the plot and considerably less in the convex hull of the skeleton grid (our calculation of the root mean squared errors used the closed-form expression for the Bayes factors based on complete enumeration). The figure took about a half hour to generate on an Intel 2.8 GHz Q9550 running Linux. (The accuracy we obtained is overkill and the figure can be created in a few minutes if we use more typical Markov chain lengths.)

Table 1 gives the posterior inclusion probabilities for each of the fifteen predictors, i.e. $P(\gamma_i = 1 \,|\, y)$ for $i = 1, \ldots, 15$, under several models. Line 2 gives the inclusion probabilities when we use model (1.1) with the values $w = .65$ and $g = 20$, which are the values at which the graph in Figure 1 attains its maximum. Line 4 gives the inclusion probabilities when the hyper-g prior "HG3" in Liang et al. (2008) is used. As can be seen, the inclusion probabilities we obtained under the EB model are comparable to, but somewhat larger than, the probabilities when the HG3 prior is used. This is not surprising since our model allows w to be chosen, and the data-driven choice gives a value (.65) greater than the value $w = .5$ used in Liang et al. (2008). (Table 2 of Liang et al. (2008) gives a comparison of posterior inclu-

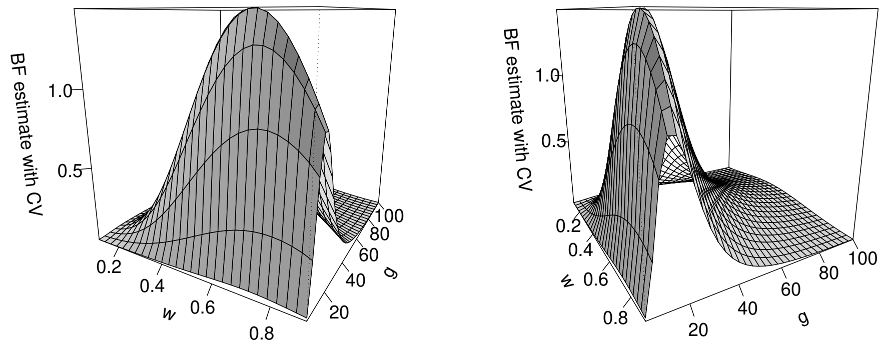

FIG 1. *Estimates of Bayes factors for the US crime data. The plots give two different views of the graph of the Bayes factor as a function of w and g when the baseline value of the hyperparameter is given by $w = 0.5$ and $g = 15$. The estimate is (2.17), which uses control variates.*

sion probabilities for a total of ten models taken from the literature.) Line 3 of Table 1 gives the inclusion probabilities under model (1.1) when we use $w = .5$ and the value of g that maximizes the likelihood with w constrained to be .5. It is interesting to note that the inclusion probabilities are then strikingly close to those under the HG3 model.

	Age	S	Ed	Ex0	Ex1	LF	M	N	NW	U1	U2	W	X	Prison	Time
EB(.65, 20)	.93	.39	.99	.70	.51	.34	.35	.52	.83	.40	.76	.55	1.00	.96	.55
EB(.5, 20)	.85	.29	.97	.67	.45	.22	.22	.38	.70	.27	.62	.38	1.00	.90	.39
HG3	.84	.29	.97	.66	.47	.23	.23	.39	.69	.27	.61	.38	.99	.89	.38

TABLE 1

Posterior inclusion probabilities for the fifteen predictor variables in the US crime data set, under three models. Names of the variables are as in Table 2 of Liang et al. (2008) (but all variables except for the binary variable S have been log transformed).

Buta (2010) uses the estimates in Section 2.3 to produce plots of posterior inclusion probabilities for several of the predictors, as w and g vary. The plots enable one to read the posterior inclusion probabilities under various choices for g and w proposed in the literature, and also show that the extent to which these probabilities change with the choices is striking.

Selection of the skeleton points was discussed at the end of Section 3, and we now return to this issue. Consider the Bayes factor estimate based on the skeleton (4.2), which was chosen in an ad-hoc manner. The left panel in Figure 2 gives a plot of the variance of this estimate, as a function of h. As can be seen from the plot, the variance is greatest in the region where g is small and w is large. We changed the skeleton from (4.2) to

$$(w, g) \in \{.5, .7, .8, .9\} \times \{10, 15, 50, 100\} \tag{4.3}$$

and reran the algorithm. The variance for the estimate based on (4.3) is given by the right panel of Figure 2, from which we see that the maximum variance has been reduced by a factor of about 9.

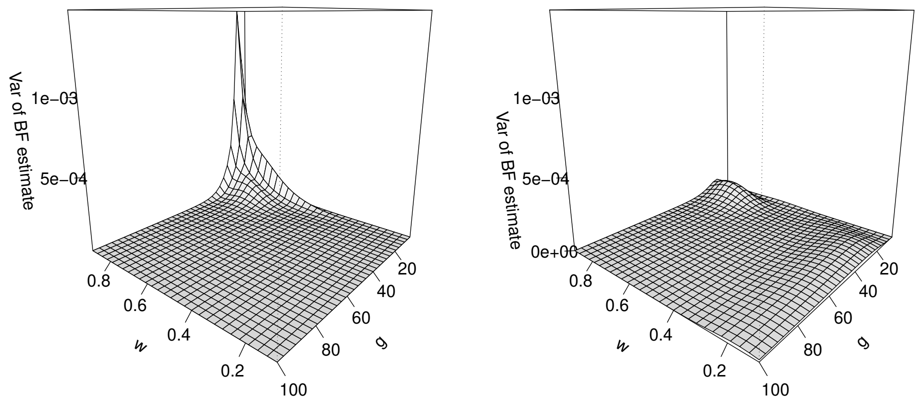

FIG 2. *Variance functions for two versions of $\hat{I}^{\hat{d}}_{\hat{\beta}(\hat{d})}$. The left panel is for the estimate based on the skeleton (4.2). The points in this skeleton were shifted to better cover the problematic region near the back of the plot (g small and w large), creating the skeleton (4.3). The maximum variance is then reduced by a factor of 9 (right panel).*

5. Discussion. The following fact is obvious, but it may be worthwhile to state it explicitly. If h_1 is fixed, maximizing $B(h, h_1)$ and maximizing the marginal likelihood m_h are equivalent. Choosing the value of h that maximizes m_h is by definition the empirical Bayes method. Thus, the development in Section 2 can be used to implement empirical Bayes methods.

Our methodology for dealing with the sensitivity analysis and model selection problems discussed in Section 1 can be applied to many classes of Bayesian models. In addition to the usual parametric models, we mention also Bayesian nonparametric models involving mixtures of Dirichlet processes (Antoniak (1974)), in which one of the hyperparameters is the so-called total mass parameter—very briefly, this hyperparameter controls the extent to which the nonparametric model differs from a purely parametric model. (Among the many papers that use such models, we mention in particular Burr and Doss (2005), who give a more detailed discussion of the role of the total mass parameter.) The approach developed in Sections 2.1 and 2.2 can be used to select this parameter.

When the dimension of h is low, it will be possible to plot $B(h, h_1)$, or at least plot it as h varies along some of its dimensions. Empirical Bayes methods are notoriously difficult to implement when the dimension of the hyperparameter h is high. In this case, it is possible to use the methods developed in Sections 2.1 and 2.2 to enable approaches based on stochastic search algorithms. These require the calculation of the gradient $\partial B(h, h_1)/\partial h$. We note that the same methodology used to estimate $B(h, h_1)$ can also be used to estimate its gradient. For example, in (2.8), $\nu_h(\theta_i^{(l)})$ is simply replaced by $\partial \nu_h(\theta_i^{(l)})/\partial h$.

APPENDIX

Proof of Theorem 1. We begin by writing

$$\text{(A.1)}\quad \begin{aligned} &\sqrt{n}(\hat{B}(h, h_1, \hat{d}) - B(h, h_1)) \\ &\quad = \sqrt{n}(\hat{B}(h, h_1, \hat{d}) - \hat{B}(h, h_1, d)) + \sqrt{n}(\hat{B}(h, h_1, d) - B(h, h_1)). \end{aligned}$$

The second term on the right side of the equation in (A.1) involves randomness coming only from the second stage of sampling. This term was analyzed by Doss (2010), who showed that it is asymptotically normal, with mean 0 and variance $\tau^2(h)$. The first term ostensibly involves randomness from both Stage 1 and Stage 2 sampling. However, as will emerge from our proof, the randomness from Stage 2 is of lower order, and effectively all the randomness is from Stage 1. This randomness is non-negligible. We mention here the often-cited work of Geyer (1994) (whose nice results we use in the present paper). In the context of a setup very similar to ours, his Theorem 4 states that using an estimated d and using the true d results in the same asymptotic variance. From our proof [refer also to the extension of our Theorem 1 to the case of a simple sample given in Buta (2010)], we see that this statement is not correct.

To analyze the first term on the right side of (A.1), define the function $F(u) = \hat{B}(h, h_1, u)$, where $u = (u_2, \dots, u_k)'$ is a real vector with $u_l > 0$, $l = 2, \dots, k$. Then, by the Taylor series expansion of F about d, we get

$$\text{(A.2)}\quad \begin{aligned} \sqrt{n}(\hat{B}(h, h_1, \hat{d}) - \hat{B}(h, h_1, d)) &= \sqrt{n}(F(\hat{d}) - F(d)) \\ &= \sqrt{n}\nabla F(d)'(\hat{d} - d) + \frac{\sqrt{n}}{2}(\hat{d} - d)'\nabla^2 F(d^*)(\hat{d} - d), \end{aligned}$$

where d^* is between d and $\hat{d}$.

First, we show that the gradient $\nabla F(d) = (\partial F(d)/\partial d_2, \dots, \partial F(d)/\partial d_k)'$ converges almost surely to a finite constant. Recall that $c(h)$ is defined in (2.10). For $j = 2, \dots, k$, the $(j-1)^{\text{th}}$ component of $\nabla F(d)$ converges almost surely since, with the SLLN assumed to hold for the Markov chains used, we have

$$[\nabla F(d)]_{j-1} = \sum_{l=1}^{k} \frac{1}{n_l} \sum_{i=1}^{n_l} \frac{a_j a_l \nu_h(\theta_i^{(l)}) \nu_{h_j}(\theta_i^{(l)})}{d_j^2 \big(\sum_{s=1}^{k} a_s \nu_{h_s}(\theta_i^{(l)})/d_s\big)^2} \xrightarrow{\text{a.s.}} [c(h)]_{j-1}.$$

Next, we show that the random Hessian matrix $\nabla^2 F(d^*)$ of second-order derivatives of F evaluated at d^* is bounded in probability. To this end, it suffices to show that each element of this matrix, say $[\nabla^2 F(d^*)]_{t-1,j-1}$, where $t, j \in \{2, \dots, k\}$, is $O_p(1)$. Since $\|d^* - d\| \le \|\hat{d} - d\| \xrightarrow{p} 0$, it follows that $d^* \xrightarrow{p} d$.

Let $\epsilon \in (0, \min(d_2, \dots, d_k))$. Then we have $P(\|d^* - d\| \le \epsilon) \to 1$. We now show that, on the set $\{\|d^* - d\| \le \epsilon\}$, $\nabla^2 F(d^*)$ is bounded in probability. Let

$$\mathcal{I} = I(\|d^* - d\| \le \epsilon).$$

For $t \neq j$, we have

$$\begin{aligned}
\big|[\nabla^2 F(d^*)]_{t-1,j-1}\big| \cdot \mathcal{I} &= \sum_{l=1}^{k} \frac{2}{n_l} \sum_{i=1}^{n_l} \frac{a_j a_l a_t \nu_h(\theta_i^{(l)}) \nu_{h_j}(\theta_i^{(l)}) \nu_{h_t}(\theta_i^{(l)})}{d_j^{*2} d_t^{*2} \big(\sum_{s=1}^{k} a_s \nu_{h_s}(\theta_i^{(l)})/d_s^*\big)^3} \cdot \mathcal{I} \\
&\le \sum_{l=1}^{k} \frac{2}{n_l} \sum_{i=1}^{n_l} \frac{a_j a_l a_t \nu_h(\theta_i^{(l)}) \nu_{h_j}(\theta_i^{(l)}) \nu_{h_t}(\theta_i^{(l)})}{(d_j - \epsilon)^2 (d_t - \epsilon)^2 \big[\sum_{s=1}^{k} a_s \nu_{h_s}(\theta_i^{(l)})/(d_s + \epsilon)\big]^3} \\
&\xrightarrow{\text{a.s.}} \sum_{l=1}^{k} B(h, h_l) \int \left\{ \frac{a_j a_l a_t \nu_{h_j}(\theta) \nu_{h_t}(\theta) \nu_{h_l}(\theta)}{\big[\sum_{s=1}^{k} a_s \nu_{h_s}(\theta)/(d_s + \epsilon)\big]^3} \right\} \nu_{h,y}(\theta)\, d\theta \\
&\qquad \times \frac{2}{(d_j - \epsilon)^2 (d_t - \epsilon)^2}.
\end{aligned} \tag{A.3}$$

Note that the expression inside the braces in (A.3) is clearly bounded above by a constant, so expression (A.3) is finite. Similarly, for $t = j$, we can show that $\big|[\nabla^2 F(d^*)]_{j-1,j-1}\big| \cdot \mathcal{I}$ is $O_p(1)$. Since $P(\|d^* - d\| \le \epsilon) \to 1$, it follows that $\nabla^2 F(d^*)$ is bounded in probability. Now, by combining (A.1) and (A.2), we obtain

$$\begin{aligned}
\sqrt{n}\big(\hat{B}(h, h_1, \hat{d}) - B(h, h_1)\big) &= \sqrt{\frac{n}{N}} \nabla F(d)' \sqrt{N}(\hat{d} - d) \\
&\quad + \frac{1}{2\sqrt{N}} \sqrt{\frac{n}{N}} \big[\sqrt{N}(\hat{d} - d)\big]' \nabla^2 F(d^*) \big[\sqrt{N}(\hat{d} - d)\big] \\
&\quad + \sqrt{n}\big(\hat{B}(h, h_1, d) - B(h, h_1)\big) \\
&= \sqrt{q}\, c(h)' \sqrt{N}(\hat{d} - d) + \sqrt{n}\big(\hat{B}(h, h_1, d) - B(h, h_1)\big) + o_p(1),
\end{aligned}$$

where the last line follows from the fact that $\nabla F(d) \xrightarrow{\text{a.s.}} c(h)$ established earlier, the assumptions of Theorem 1 that $\sqrt{n/N} \to \sqrt{q}$ and that $\sqrt{N}(\hat{d} - d)$ converges in distribution (hence is $O_p(1)$). Because the two sampling stages (for estimating d and $B(h, h_1)$) are assumed to be independent, using the assumption that $\sqrt{N}(\hat{d} - d) \xrightarrow{d} \mathcal{N}(0, \Sigma)$ in conjunction with the result $\sqrt{n}(\hat{B}(h, h_1, d) - B(h, h_1)) \xrightarrow{d} \mathcal{N}(0, \tau^2(h))$ established in Theorem 1 of Doss (2010) under conditions A1 and A2, we conclude that

$$\sqrt{n}\big(\hat{B}(h, h_1, \hat{d}) - B(h, h_1)\big) \xrightarrow{d} \mathcal{N}\big(0, qc(h)'\Sigma c(h) + \tau^2(h)\big).$$

$\square$

Proof of Theorem 2. We begin by writing

$$\sqrt{n}\big(\hat{I}^{\hat{d}}_{\hat{\boldsymbol{\beta}}(\hat{d})} - B(h, h_1)\big) = \sqrt{n}\big(\hat{I}^{\hat{d}}_{\hat{\boldsymbol{\beta}}(\hat{d})} - \hat{I}^{d}_{\hat{\boldsymbol{\beta}}(d)}\big) + \sqrt{n}\big(\hat{I}^{d}_{\hat{\boldsymbol{\beta}}(d)} - B(h, h_1)\big), \tag{A.4}$$

where the second term on the right side of (A.4) was analyzed by Doss (2010) who showed that it is asymptotically normal, with mean 0 and variance $\sigma^2(h)$. Our plan is to show that $\hat{\boldsymbol{\beta}}(d)$ and $\hat{\boldsymbol{\beta}}(\hat{d})$ converge in probability to the same limit, which we denote $\boldsymbol{\beta}_{\text{lim}}$. We then expand the first term on the right side of (A.4) by writing

$$\sqrt{n}\big(\hat{I}^{\hat{d}}_{\hat{\boldsymbol{\beta}}(\hat{d})} - \hat{I}^{d}_{\hat{\boldsymbol{\beta}}(d)}\big) = \sqrt{n}\big(\hat{I}^{\hat{d}}_{\hat{\boldsymbol{\beta}}(\hat{d})} - \hat{I}^{\hat{d}}_{\boldsymbol{\beta}_{\text{lim}}}\big) + \sqrt{n}\big(\hat{I}^{\hat{d}}_{\boldsymbol{\beta}_{\text{lim}}} - \hat{I}^{d}_{\boldsymbol{\beta}_{\text{lim}}}\big) + \sqrt{n}\big(\hat{I}^{d}_{\boldsymbol{\beta}_{\text{lim}}} - \hat{I}^{d}_{\hat{\boldsymbol{\beta}}(d)}\big). \tag{A.5}$$

Our proof is organized as follows:

- We note that the third term on the right side of (A.5) was shown to converge to 0 in probability by Doss (2010).
- We will show that the first term on the right side of (A.5) also converges to 0 in probability.
- The second term on the right side of (A.5) involves randomness from both Stage 1 and Stage 2. However, we will show that the randomness from Stage 2 is asymptotically negligible, and that this term is asymptotically equivalent to an expression of the form $w(h)'(\hat{d} - d)$, where $w(h)$ is a deterministic vector. This will show that the second term is asymptotically normal.

Now we prove that the first term on the right side of (A.5) is $o_p(1)$, and to do this we begin by showing that $\hat{\boldsymbol{\beta}}(d)$ and $\hat{\boldsymbol{\beta}}(\hat{d})$ converge in probability to the same limit. Let $\boldsymbol{Z}$ be the $n \times k$ matrix whose transpose is

$$\text{(A.6)} \quad \boldsymbol{Z}' = \begin{pmatrix} 1 & \dots & 1 & 1 & \dots & 1 & \dots & 1 & \dots & 1 \\ Z_{1,1}^{(2)} & \dots & Z_{n_1,1}^{(2)} & Z_{1,2}^{(2)} & \dots & Z_{n_2,2}^{(2)} & \dots & Z_{1,k}^{(2)} & \dots & Z_{n_k,k}^{(2)} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ Z_{1,1}^{(k)} & \dots & Z_{n_1,1}^{(k)} & Z_{1,2}^{(k)} & \dots & Z_{n_2,2}^{(k)} & \dots & Z_{1,k}^{(k)} & \dots & Z_{n_k,k}^{(k)} \end{pmatrix}$$

and let $\boldsymbol{Y}$ be the vector

$$\text{(A.7)} \qquad \boldsymbol{Y} = \big(Y_{1,1}, \dots, Y_{n_1,1}, Y_{1,2}, \dots, Y_{n_2,2}, \dots, Y_{1,k}, \dots, Y_{n_k,k}\big)'.$$

Let $\hat{\boldsymbol{Z}}$ be the $n \times k$ matrix corresponding to $\boldsymbol{Z}$ when we replace d by $\hat{d}$. Similarly, $\hat{\boldsymbol{Y}}$ is like $\boldsymbol{Y}$, but using $\hat{d}$ for d.

For fixed $j, j' \in \{2, \dots, k\}$, consider the function

$$\text{(A.8)} \quad G(u) = \frac{1}{n} \sum_{l=1}^{k} \sum_{i=1}^{n_l} \frac{\nu_{h_j}(\theta_i^{(l)})/u_j - \nu_{h_1}(\theta_i^{(l)})}{\sum_{s=1}^k a_s \nu_{h_s}(\theta_i^{(l)})/u_s} \cdot \frac{\nu_{h_{j'}}(\theta_i^{(l)})/u_{j'} - \nu_{h_1}(\theta_i^{(l)})}{\sum_{s=1}^k a_s \nu_{h_s}(\theta_i^{(l)})/u_s},$$

where $u = (u_2, \dots, u_k)'$ and $u_l > 0$, for $l = 2, \dots, k$. (On the right side of (A.8), u_1 is taken to be 1.) Note that setting $u = d$ gives $G(d) = \frac{1}{n} \sum_{l=1}^k \sum_{i=1}^{n_l} Z_{i,l}^{(j)} Z_{i,l}^{(j')}$. By the Mean Value Theorem, we know that there exists a d^* between d and $\hat{d}$ such that

$$G(\hat{d}) = G(d) + \nabla G(d^*)'(\hat{d} - d) = \boldsymbol{R}_{j,j'} + \nabla G(d^*)'(\hat{d} - d) + o_p(1).$$

Note that the last equality above comes from applying the SLLN. An argument similar to that used in Theorem 1 to show that $\nabla^2 F(d^*) = O_p(1)$ can now be applied to show that $\nabla G(d^*) = O_p(1)$.

Therefore

$$G(\hat{d}) = \boldsymbol{R}_{j,j'} + \nabla G(d^*)'(\hat{d} - d) + o_p(1) = \boldsymbol{R}_{j,j'} + O_p(1) o_p(1) + o_p(1) \xrightarrow{p} \boldsymbol{R}_{j,j'}.$$

Similar arguments extend to the case $j = 1$ or $j' = 1$. By the fact that $\boldsymbol{R}$ is assumed invertible, we have

$$\text{(A.9)} \qquad n(\hat{\boldsymbol{Z}}' \hat{\boldsymbol{Z}})^{-1} \xrightarrow{p} \boldsymbol{R}^{-1}.$$

In a similar way, it can be shown that

$$\text{(A.10)} \qquad \hat{\boldsymbol{Z}}' \hat{\boldsymbol{Y}}/n \xrightarrow{p} \boldsymbol{v},$$

where $\boldsymbol{v}$ is the same limit vector to which $\boldsymbol{Z}'\boldsymbol{Y}/n$ has been proved to converge in Doss (2010). Combining (A.9) and (A.10) we have

$$(\hat{\beta}_0(\hat{d}), \hat{\boldsymbol{\beta}}(\hat{d})) = [n(\hat{\boldsymbol{Z}}'\hat{\boldsymbol{Z}})^{-1}][\hat{\boldsymbol{Z}}'\hat{\boldsymbol{Y}}/n] \xrightarrow{p} (\beta_{0,\mathrm{lim}}, \boldsymbol{\beta}_{\mathrm{lim}}) = \boldsymbol{R}^{-1}\boldsymbol{v}.$$

Let $e(j,l) = E(Z_{1,l}^{(j)})$. We now have

$$\begin{aligned}\sqrt{n}(\hat{I}^{\hat{d}}_{\hat{\boldsymbol{\beta}}(\hat{d})} - \hat{I}^{\hat{d}}_{\boldsymbol{\beta}_{\mathrm{lim}}}) &= \sum_{j=2}^{k}(\beta_{j,\mathrm{lim}} - \hat{\beta}_j(\hat{d}))\left(\sum_{l=1}^{k} a_l\, n^{1/2}\sum_{i=1}^{n_l}\left(\frac{\hat{Z}_{i,l}^{(j)} - e(j,l)}{n_l}\right)\right)\\ &= \sum_{j=2}^{k} o_p(1)\left(\sum_{l=1}^{k} a_l\, n^{1/2}\sum_{i=1}^{n_l}\left(\frac{\hat{Z}_{i,l}^{(j)} - e(j,l)}{n_l}\right)\right). \qquad \text{(A.11)}\end{aligned}$$

To show that (A.11) converges to 0 in probability it suffices to show that for each l and j

$$n_l^{1/2}\sum_{i=1}^{n_l}\left(\frac{\hat{Z}_{i,l}^{(j)} - e(j,l)}{n_l}\right) = O_p(1). \qquad \text{(A.12)}$$

For fixed $j \in \{2,\dots,k\}$ and $l \in \{1,\dots,k\}$, define

$$H(u) = n_l^{-1/2}\sum_{i=1}^{n_l}\frac{\nu_{h_j}(\theta_i^{(l)})/u_j - \nu_{h_1}(\theta_i^{(l)})}{\sum_{s=1}^{k} a_s \nu_{h_s}(\theta_i^{(l)})/u_s}$$

for $u = (u_2,\dots,u_k)'$ with $u_l > 0$, $l = 2,\dots,k$, $u_1 = 1$. Note that $H(d) = n_l^{-1/2}\sum_{i=1}^{n_l} Z_{i,l}^{(j)}$. To see why (A.12) is true we begin by writing

$$\begin{aligned}n_l^{1/2}\sum_{i=1}^{n_l}\left(\frac{\hat{Z}_{i,l}^{(j)} - e(j,l)}{n_l}\right) &= n_l^{1/2}\sum_{i=1}^{n_l}\left(\frac{\hat{Z}_{i,l}^{(j)} - Z_{i,l}^{(j)}}{n_l}\right) + n_l^{1/2}\sum_{i=1}^{n_l}\left(\frac{Z_{i,l}^{(j)} - e(j,l)}{n_l}\right)\\ &= H(\hat{d}) - H(d) + O_p(1). \qquad \text{(A.13)}\end{aligned}$$

Note that the fact that $n_l^{1/2}\sum_{i=1}^{n_l}([Z_{i,l}^{(j)} - e(j,l)]/n_l) = O_p(1)$, which was used to establish the second equality in (A.13), is proved in Doss (2010). Now, applying the Mean Value Theorem to the function H, we know that there exists a point d^* between d and $\hat{d}$ such that (A.13) becomes

$$\begin{aligned}n_l^{1/2}\sum_{i=1}^{n_l}\left(\frac{\hat{Z}_{i,l}^{(j)} - e(j,l)}{n_l}\right) &= \nabla H(d^*)'(\hat{d} - d) + O_p(1)\\ &= \sqrt{a_l}\sqrt{\frac{n}{N}}\, n_l^{-1/2}\nabla H(d^*)'\sqrt{N}(\hat{d} - d) + O_p(1), \qquad \text{(A.14)}\end{aligned}$$

so that the right side of (A.14) is $O_p(1)$. We now consider $\sqrt{n}(\hat{I}^{\hat{d}}_{\boldsymbol{\beta}_{\text{lim}}} - \hat{I}^{d}_{\boldsymbol{\beta}_{\text{lim}}})$, the middle term in (A.5). Define

$$K(u) = \frac{1}{n}\sum_{l=1}^{k}\sum_{i=1}^{n_l}\left(\frac{\nu_h(\theta_i^{(l)})}{\sum_{s=1}^k a_s \nu_{h_s}(\theta_i^{(l)})/u_s} - \sum_{j=2}^{k}\beta_{j,\text{lim}}\frac{\nu_{h_j}(\theta_i^{(l)})/u_j - \nu_{h_1}(\theta_i^{(l)})}{\sum_{s=1}^k a_s \nu_{h_s}(\theta_i^{(l)})/u_s}\right),$$

where $u = (u_2, \dots, u_k)'$, and $u_l > 0$ for $l = 2, \dots, k$. By Taylor series expansion, we have

(A.15)
$$\sqrt{n}(\hat{I}^{\hat{d}}_{\boldsymbol{\beta}_{\text{lim}}} - \hat{I}^{d}_{\boldsymbol{\beta}_{\text{lim}}}) = \sqrt{n}\nabla K(d)'(\hat{d} - d) + \sqrt{n}\frac{1}{2}(\hat{d} - d)'\nabla^2 K(d^*)(\hat{d} - d),$$

where d^* is between $\hat{d}$ and d. We now consider $\nabla K(d)$. For $t = 2, \dots, k$ we have

$$[\nabla K(d)]_{t-1} \xrightarrow{\text{a.s.}} [w(h)]_{t-1},$$

where $[w(h)]_{t-1}$ was defined in (2.20). The Hessian matrix $\nabla^2 K(d^*)$ can be shown to be bounded in probability, using an argument similar to the one used in the proof of Theorem 1. Therefore, using the fact that $\nabla^2 K(d^*)$ is bounded in probability, we can now rewrite (A.15) as

$$\begin{aligned}
\sqrt{n}(\hat{I}^{\hat{d}}_{\boldsymbol{\beta}_{\text{lim}}} - \hat{I}^{d}_{\boldsymbol{\beta}_{\text{lim}}}) &= \sqrt{\frac{n}{N}}w(h)'\sqrt{N}(\hat{d} - d) \\
&\quad + \sqrt{\frac{n}{N}}\frac{1}{2\sqrt{N}}\sqrt{N}(\hat{d} - d)'O_p(1)\sqrt{N}(\hat{d} - d) \\
&= \sqrt{q}w(h)'\sqrt{N}(\hat{d} - d) + o_p(1).
\end{aligned}$$

Together with (A.4), this gives

$$\begin{aligned}
\sqrt{n}(\hat{I}^{\hat{d}}_{\hat{\boldsymbol{\beta}}(\hat{d})} - B(h, h_1)) &= \sqrt{q}w(h)'\sqrt{N}(\hat{d} - d) + \sqrt{n}(\hat{I}^{d}_{\hat{\boldsymbol{\beta}}(\boldsymbol{d})} - B(h, h_1)) + o_p(1) \\
&\xrightarrow{d} \mathcal{N}(0, qw(h)'\Sigma w(h) + \sigma^2(h)),
\end{aligned}$$

by the independence of the two stages of sampling, the assumption that $\sqrt{N}(\hat{d} - d)$ is asymptotically normal with mean 0 and variance Σ, and the result from Doss (2010) that $\sqrt{n}(\hat{I}^{d}_{\hat{\boldsymbol{\beta}}(d)} - B(h, h_1))$ is asymptotically normal with mean 0 and variance $\sigma^2(h)$. □

Proof of Theorem 3. First, we note that

(A.16)
$$\sqrt{n}(\hat{I}^{[f]}(h, \hat{d}) - I^{[f]}(h)) = \sqrt{n}(\hat{I}^{[f]}(h, \hat{d}) - \hat{I}^{[f]}(h, d)) + \sqrt{n}(\hat{I}^{[f]}(h, d) - I^{[f]}(h)).$$

We begin by analyzing the second term on the right side of (A.16), which only involves randomness from the second stage of sampling, and show that it is asymptotically normal. As for the first term, a closer examination reveals that it is also asymptotically normal, with all its randomness coming from Stage 1. The asymptotic normality of the sum of these two terms then follows immediately from the independence of the two stages of sampling.

Note that $\sum_{l=1}^k a_l E(Y_{1,l}^{[f]}) = I^{[f]}(h) \cdot B(h, h_1)$, and in particular, when $f \equiv 1$, this gives $\sum_{l=1}^k a_l E(Y_{1,l}) = B(h, h_1)$. Also, we have

$$
\begin{aligned}
\text{(A.17)}\quad n^{1/2} & \begin{pmatrix} \frac{1}{n}\sum_{l=1}^k \sum_{i=1}^{n_l} Y_{i,l}^{[f]} - I^{[f]}(h)\cdot B(h,h_1) \\ \frac{1}{n}\sum_{l=1}^k \sum_{i=1}^{n_l} Y_{i,l} - B(h,h_1) \end{pmatrix} \\
&= n^{1/2} \begin{pmatrix} \frac{1}{n}\sum_{l=1}^k \sum_{i=1}^{n_l} Y_{i,l}^{[f]} - \sum_{l=1}^k a_l E(Y_{1,l}^{[f]}) \\ \frac{1}{n}\sum_{l=1}^k \sum_{i=1}^{n_l} Y_{i,l} - \sum_{l=1}^k a_l E(Y_{1,l}) \end{pmatrix} \\
&= \sum_{l=1}^k a_l^{1/2} \cdot \frac{1}{n_l^{1/2}} \sum_{i=1}^{n_l} \left[\begin{pmatrix} Y_{i,l}^{[f]} \\ Y_{i,l} \end{pmatrix} - \begin{pmatrix} E(Y_{1,l}^{[f]}) \\ E(Y_{1,l}) \end{pmatrix} \right].
\end{aligned}
$$

By condition (2.26), assumption A2 of Theorem 1, and the assumed geometric ergodicity and independence of the k Markov chains used, the vector in (A.17) converges in distribution to a normal random vector with mean 0 and covariance matrix $\Gamma(h)$ where $\Gamma(h)$ is defined in (2.23). Since $\hat{I}^{[f]}(h,d)$ is given by the ratio (2.21), in view of (A.17), its asymptotic distribution may be obtained by applying the delta method to the function $g(u,v) = u/v$. This gives $\sqrt{n}(\hat{I}^{[f]}(h,d) - I^{[f]}(h)) \xrightarrow{d} \mathcal{N}(0, \rho(h))$, where $\rho(h)$ is given in (2.24).

We now consider the first term on the right side of (A.16). Define

$$
L(u) = \frac{\displaystyle\sum_{l=1}^k \sum_{i=1}^{n_l} \frac{f(\theta_i^{(l)})\nu_h(\theta_i^{(l)})}{\sum_{s=1}^k a_s \nu_{h_s}(\theta_i^{(l)})/u_s}}{\displaystyle\sum_{l=1}^k \sum_{i=1}^{n_l} \frac{\nu_h(\theta_i^{(l)})}{\sum_{s=1}^k a_s \nu_{h_s}(\theta_i^{(l)})/u_s}},
$$

for $u = (u_2, \dots, u_k)'$ with $u_l > 0$ for $l = 2, \dots, k$. Then

$$
L(d) = \hat{I}^{[f]}(h,d) = \frac{\sum_{l=1}^k \sum_{i=1}^{n_l} Y_{i,l}^{[f]}}{\sum_{l=1}^k \sum_{i=1}^{n_l} Y_{i,l}},
$$

and $\sqrt{n}(\hat{I}^{[f]}(h,\hat{d}) - \hat{I}^{[f]}(h,d)) = \sqrt{n}(L(\hat{d}) - L(d))$. Now, by the Taylor series expansion of L about d we get

$$
\sqrt{n}(\hat{I}^{[f]}(h,\hat{d}) - \hat{I}^{[f]}(h,d)) = \sqrt{n}\nabla L(d)'(\hat{d} - d) + \frac{\sqrt{n}}{2}(\hat{d} - d)'\nabla^2 L(d^*)(\hat{d} - d),
$$

where d^* is between d and $\hat{d}$. First, we show that the gradient $\nabla L(d)$ converges almost surely to a finite constant vector by proving that each one of its components, $[L(d)]_{j-1}$, $j = 2, \ldots, k$, converges almost surely. We have

$$[\nabla L(d)]_{j-1} \xrightarrow{\text{a.s.}} [v(h)]_{j-1}, \qquad j = 2, \ldots, k,$$

where $[v(h)]_{j-1}$ is given in (2.25). As in the proof of Theorem 1, it can be shown that each element of the second-derivative matrix $\nabla^2 L(d^*)$ is $O_p(1)$. Now, we can rewrite (A.16) as

$$\begin{aligned}\sqrt{n}\big(\hat{I}^{[f]}(h,\hat{d}) - I^{[f]}(h)\big) &= \sqrt{\frac{n}{N}}\nabla L(d)'\sqrt{N}(\hat{d}-d) + \sqrt{n}\big(\hat{I}^{[f]}(h,d) - I^{[f]}(h)\big) \\ &\quad + \frac{1}{2\sqrt{N}}\sqrt{\frac{n}{N}}\big[\sqrt{N}(\hat{d}-d)\big]'\nabla^2 L(d^*)\big[\sqrt{N}(\hat{d}-d)\big] \\ &= \sqrt{q}v(h)'\sqrt{N}(\hat{d}-d) + \sqrt{n}\big(\hat{I}^{[f]}(h,d) - I^{[f]}(h)\big) + o_p(1).\end{aligned}$$

Since the two sampling stages are assumed to be independent, we conclude that

$$\sqrt{n}\big(\hat{I}^{[f]}(h,\hat{d}) - I^{[f]}(h)\big) \xrightarrow{d} \mathcal{N}\big(0, qv(h)'\Sigma v(h) + \rho(h)\big).$$

□

ACKNOWLEDGEMENTS

We thank the reviewers for their careful reading and helpful comments. We are especially grateful to the associate editor for a very thorough report and for suggestions which led to several improvements in the paper.

SUPPLEMENTARY MATERIAL

Additional technical details
(doi: ???; .pdf). We show that when estimating the Bayes factors using control variates, the estimate that is optimal when the samples are iid sequences is no longer optimal when the samples are Markov chains. We also give technical arguments regarding the consistency of spectral estimates of the variance of our estimators.

REFERENCES

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics* **2** 1152–1174.

Athreya, K. B., Doss, H. and Sethuraman, J. (1996). On the convergence of the Markov chain simulation method. *The Annals of Statistics* **24** 69–100.

Burr, D. and Doss, H. (2005). A Bayesian semiparametric model for random-effects meta-analysis. *Journal of the American Statistical Association* **100** 242–251.

Buta, E. (2010). *Computational Approaches for Empirical Bayes Methods and Bayesian Sensitivity Analysis*. Ph.D. thesis, University of Florida.

Buta, E. and Doss, H. (2011). Supplement to "Computational approaches for empirical Bayes methods and Bayesian sensitivity analysis".

Cui, W. and George, E. (2008). Empirical Bayes vs. fully Bayes variable selection. *Journal of Statistical Planning and Inference* **138** 888–900.

Doss, H. (1994). Comment on "Markov chains for exploring posterior distributions" by Luke Tierney. *The Annals of Statistics* **22** 1728–1734.

Doss, H. (2007). Bayesian model selection: Some thoughts on future directions. *Statistica Sinica* **17** 413–421.

Doss, H. (2010). Estimation of large families of Bayes factors from Markov chain output. *Statistica Sinica* **20** 537–560.

Fernandez, C., Ley, E. and Steel, M. F. J. (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics* **100** 381–427.

Flegal, J. M. and Jones, G. L. (2010). Batch means and spectral variance estimators in Markov chain Monte Carlo. *The Annals of Statistics* **38** 1034–1070.

George, E. I. and Foster, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87** 731–747.

Geyer, C. J. (1994). Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo. Tech. Rep. 568r, Department of Statistics, University of Minnesota.

Gill, R. D., Vardi, Y. and Wellner, J. A. (1988). Large sample theory of empirical distributions in biased sampling models. *The Annals of Statistics* **16** 1069–1112.

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

Kong, A., McCullagh, P., Meng, X.-L., Nicolae, D. and Tan, Z. (2003). A theory of statistical models for Monte Carlo integration (with discussion). *Journal of the Royal Statistical Society, Series B* **65** 585–618.

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

Meng, X.-L. and Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica* **6** 831–860.

Mykland, P., Tierney, L. and Yu, B. (1995). Regeneration in Markov chain samplers. *Journal of the American Statistical Association* **90** 233–241.

Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics* **75** 317–343.

Tan, Z. (2004). On a likelihood approach for Monte Carlo integration. *Journal of the American Statistical Association* **99** 1027–1036.

Tierney, L. (1994). Markov chains for exploring posterior distributions (Disc: p1728–1762). *The Annals of Statistics* **22** 1701–1728.

Vandaele, W. (1978). Participation in illegitimate activities: Ehrlich revisited. In *Deterrence and Incapacitation*. US National Academy of Sciences, Washington DC.

Vardi, Y. (1985). Empirical distributions in selection bias models. *The Annals of Statistics* **13** 178–203.

Zellner, A. (1986). 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* (P. K. Goel and A. Zellner, eds.). Elsevier, New York.

Department of Epidemiology and Public Health
Yale University
New Haven, CT 06510
E-mail: eugenia.buta@yale.edu

Department of Statistics
University of Florida
Gainesville, FL 32611
E-mail: doss@stat.ufl.edu