Scalable Empirical Bayes Inference and Bayesian Sensitivity Analysis

Hani Doss*
Department of Statistics
University of Florida
doss@stat.ufl.edu

Antonio Linero†
Department of Statistics and Data Science
University of Texas at Austin
antonio.linero@austin.utexas.edu

Abstract

Consider a Bayesian setup in which we observe $Y \sim p_\theta$, where $\theta \in \Theta$, and we have a parametric family $\{\nu_h,\, h \in \mathcal{H}\}$ of potential prior densities on Θ. A pervasive task is to select a member of that family, and in non-subjective Bayesian analyses this is typically done by choosing the value of the hyperparameter h that maximizes some criterion. Arguably the most common way of doing this is to let $m(h)$ be the marginal likelihood of h, i.e. $m(h) = \int p_\theta(y)\, \nu_h(\theta)\, d\theta$, and choose the value of h that maximizes $m(\cdot)$. Unfortunately, except for a handful of textbook examples, analytic evaluation of $\arg\max_h m(h)$ is not feasible. We review the literature on estimating it and find that all existing procedures are either potentially highly inaccurate or don't scale well with the dimension of h, the dimension of θ, or both. We present a method for estimating $\arg\max_h m(h)$, based on Markov chain Monte Carlo, that applies very generally and scales well with dimension. We provide theorems, based on empirical process theory, which enable us to obtain confidence sets for $\arg\max_h m(h)$, and to determine the Markov chain length needed to estimate $\arg\max_h m(h)$ to a preset level of accuracy. Let g be a real-valued function of θ, and let $I(h)$ be the posterior expectation of $g(\theta)$ when the prior is ν_h. As a byproduct of our approach, we show how to obtain point estimates and globally-valid confidence bands for the family $I(h)$, $h \in \mathcal{H}$. To illustrate the scope of our methodology we provide three detailed examples, having different characters (two are in the present paper and one is in a supplementary document).

Key words and phrases: Bayesian model selection, Donsker class, geometric ergodicity, hyperparameter selection, Markov chain Monte Carlo, regenerative simulation.

1 Introduction

Consider a Bayesian setup in which we observe $Y \sim p_\theta$, where $\theta \in \Theta$, and we have a family $\{\nu_h,\, h \in \mathcal{H}\}$ of potential prior densities on Θ. We observe $Y = y$, and after having selected a hyperparameter value h_0, Bayesian inference is based on the posterior distribution $\nu_{h_0,y}$. As is

*Research supported by NSF Grants DIIS-17-24174 and DMS-1854476
†Research supported by NSF Grant DMS-1712870

well known and we discuss later, inference can depend heavily on the hyperparameter value h that specifies the prior. The problem of selecting h in a principled manner comes up in a large number of situations, and below we describe two classes of nontrivial ones. Dirichlet process mixtures (DPM's, see Neal (2000) for a review) are Bayesian hierarchical models in which one of the levels of the hierarchy posits that a probability measure μ is distributed according to a Dirichlet process indexed by a base probability measure α and a precision parameter M. The base probability measure typically is taken from a parametric family $\{\alpha_a,\, a \in A\}$, so the hyperparameter for the DPM is $h = (M, a)$. As is well known, when M is large, the model based on a DPM behaves like a model in which the distribution μ is equal to α_a, i.e. the distribution of μ is a point mass at α_a, whereas when M is small, the distribution of μ is more diffuse, or "nonparametric." Thus, loosely speaking, the hyperparameter M controls the extent to which the model is parametric or nonparametric. Related to DPM's are hierarchical Dirichlet processes (Teh et al., 2006). For the most commonly-used version of these models, the one used for topic modelling (Blei et al., 2003), there are two levels in the hierarchy which involve Dirichlet processes. In this case, the hyperparameter h is the triple consisting of the two precision parameters and the parameter vector of the base probability measure at the bottom of the hierarchy. In each of these cases, the hyperparameter plays a crucial role, but the complex nature of the model makes its selection difficult; for example, for models based on hierarchical Dirichlet processes, there are currently no principled methods for dealing with the hyperparameters other than by placing a prior on them, an approach which has its own problems, as we discuss later.

The second situation involving selection of a hyperparameter involves a Bayesian formulation of the normal means problem introduced in a series of papers by Efron and Morris in the 1970's (see for example Efron and Morris (1973)). The setup is the hierarchical model

$$\begin{aligned} X_i \,|\, \theta_i &\overset{\text{ind}}{\sim} \mathcal{N}(\theta_i, 1), \quad i = 1, \ldots, p, \\ \theta_i &\overset{\text{iid}}{\sim} \mathcal{N}(\mu, \lambda), \quad i = 1, \ldots, p, \end{aligned}$$

where μ and λ are unknown. We observe $X \coloneqq (X_1, \ldots, X_p)$ and we wish to estimate $\theta \coloneqq (\theta_1, \ldots, \theta_p)$. The posterior mean of θ_i is

$$\hat{\theta}_i^{\lambda,\mu} = \frac{\lambda}{\lambda+1} X_i + \frac{1}{\lambda+1}\mu. \tag{1.1}$$

This gives a class of estimators of θ, indexed by (λ, μ), and the goal is to find the "optimal" value of (λ, μ). Let $R(\hat{\theta}^{\lambda,\mu}, \theta)$ denote the risk (under sum of squared errors as loss) of the estimate $\hat{\theta}^{\lambda,\mu} = (\hat{\theta}_1^{\lambda,\mu}, \ldots, \hat{\theta}_p^{\lambda,\mu})$. Note that $R(\hat{\theta}^{\lambda,\mu}, \theta)$ depends on (λ, μ) and also on θ, which is unknown. The optimal value of (λ, μ) is taken to be the value that minimizes $R(\hat{\theta}^{\lambda,\mu}, \theta)$, and this can't be obtained since we don't know θ. Stein's unbiased risk estimate (SURE, Stein (1981)), denoted $\mathrm{SURE}(\lambda, \mu)$, is a statistic with the property that $E_\theta[\mathrm{SURE}(\lambda, \mu)] = R(\hat{\theta}^{\lambda,\mu}, \theta)$, where the notation E_θ signifies that the expectation is taken under the assumption that $X_i \overset{\text{ind}}{\sim} \mathcal{N}(\theta_i, 1)$. Since we can't obtain $\arg\min_{\lambda,\mu} R(\hat{\theta}^{\lambda,\mu}, \theta)$, we compute $(\hat{\lambda}^{\text{opt}}, \hat{\mu}^{\text{opt}}) = \arg\min_{\lambda,\mu} \mathrm{SURE}(\lambda, \mu)$ and use $(\hat{\lambda}^{\text{opt}}, \hat{\mu}^{\text{opt}})$ in place of (λ, μ) in (1.1). The result is an estimator of θ which is very similar to the James-Stein estimate (James and Stein, 1961).

The approach of choosing the hyperparameter value by minimizing the SURE is an example of an empirical Bayes analysis: we use the data to select the hyperparameter of the prior. The approach is elegant and has been replicated in several situations that are more complex than the one described above. Unfortunately, it requires that we be able to find a closed-form expression for an unbiased risk estimate, and for complex models this may be extremely difficult or impossible. There is another approach, one that does not require us to specify a loss function, and which is much more commonly used, especially in the machine learning literature. It involves the marginal likelihood: observing $Y = y$ induces the marginal likelihood function $m(h) = \int \ell_y(\theta)\nu_h(\theta)\, d\theta$, where $\ell_y(\theta) = p_\theta(y)$ is the likelihood function, and h is chosen to be $\arg\max_h m(h)$. Unfortunately, analytic evaluation of $m(h)$ is not feasible except for a handful of textbook examples.

The literature gives several numerical methods for estimating the function $m(h)$ or its argmax. One of them is the EM algorithm, in which Y is viewed as observed data and θ is viewed as missing data. Typically, the "complete data likelihood" $f_h(\theta, y)$ is available in closed form, so the algorithm can in principle be applied. In most complex problems, however, the E-step in the algorithm is infeasible, because it requires calculating an expectation with respect to the intractable distribution $\nu_{h,y}$. Several variants of the EM algorithm have been proposed to deal with this difficulty, and these include Monte Carlo EM (MCEM), originally proposed in Wei and Tanner (1990), in which the E-step is approximated by a Monte Carlo estimate; and variational EM (VEM), see Beal and Ghahramani (2003) and also Blei et al. (2017), in which the E-step is approximated by an estimate produced through variational inference. Unfortunately, the EM algorithm can converge slowly, and therefore so can its variants, and this problem gets worse with "increasing missingness" (Liu, 1994; van Dyk and Meng, 2001). Thus, when the dimension of θ is large the rate of convergence can be very problematic, since here θ is what is missing. Also, if the marginal likelihood surface is multimodal, all EM-type algorithms can fail, with the user having no clue that the estimate of $\arg\max_h m(h)$ is only a local mode. Additionally, for both MCEM and VEM, because an approximation is used at every iteration of the EM algorithm, the theoretical basis for these two methods is weak. We elaborate on this point in Section 2.4.

Another approach for estimating $m(h)$ relies on importance sampling, and in its simplest form the approach is described as follows. Assume that all the priors in the family $\{\nu_h,\, h \in \mathcal{H}\}$ are mutually absolutely continuous, which entails that all the posteriors are also mutually absolutely continuous. Let $h_1 \in \mathcal{H}$, and suppose we are able to generate an ergodic Markov chain with invariant distribution $\nu_{h_1,y}$. Starting with the equation $\int [\nu_{h,y}(\theta)/\nu_{h_1,y}(\theta)]\nu_{h_1,y}(\theta)\, d\theta = 1$ and expressing the statement "the posterior is proportional to the likelihood times the prior" as $\nu_{h,y}(\theta) = (1/m(h))\ell_y(\theta)\nu_h(\theta)$, we trivially obtain the equation $\int [\nu_h(\theta)/\nu_{h_1}(\theta)]\nu_{h_1,y}(\theta)\, d\theta = m(h)/m(h_1)$. This is an interesting identity because it means that if $\theta_1, \theta_2, \ldots$ is an ergodic Markov chain with invariant distribution $\nu_{h_1,y}$, then

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\nu_h(\theta_i)}{\nu_{h_1}(\theta_i)} \xrightarrow{\text{a.s.}} \frac{m(h)}{m(h_1)}. \tag{1.2}$$

The significance of the convergence statement (1.2) is that, in principle, with a single Markov chain run, we can estimate the entire function $m(\cdot)$ up to a multiplicative constant. For the purpose of

estimating $\arg\max_h m(h)$, the information in the two functions $m(\cdot)$ and $m(\cdot)/m(h_1)$ is identical: the two functions have the same argmax, and the second derivative matrices of the logarithm of these two functions are identical. Therefore, the standard point estimates and confidence regions based on $m(\cdot)$ and $m(\cdot)/m(h_1)$ are identical.

Unfortunately, some of what is said above is too good to be true, and in reality the estimate on the left side of (1.2) has a serious defect: unless h is close to h_1, ν_h can be nearly singular with respect to ν_{h_1} over the region where the θ_i's are likely to be, resulting in a very unstable estimate. From a practical point of view, this means that there is effectively a "radius" around h_1 within which one can safely move, and there may not exist a single value of h_1 that gives rise to estimates that are stable for all $h \in \mathcal{H}$. One way of dealing with this problem is to select J fixed points $h_1, \ldots, h_J \in \mathcal{H}$ that "cover" $\mathcal{H}$ in the sense that for every $h \in \mathcal{H}$, ν_h is "close to" at least one of $\nu_{h_1}, \ldots, \nu_{h_J}$. We then develop an importance sampling expression analogous to the left side of (1.2), where the denominator is based on $\nu_{h_1}, \ldots, \nu_{h_J}$. This approach, which is not trivial, can give a great improvement over the simple importance sampling estimate, and Doss and Park (2018) developed theory for it. We explain the approach in detail in Section 2.4. Here, we mention only that it suffers from the curse of dimensionality: when $\dim(h)$ is large, it is necessary that J be huge in order for the points $h_1, \ldots, h_J$ to adequately cover $\mathcal{H}$, and in Section 2.4 we explain why this can cause the approach to fail when $\dim(h) \geq 3$ or even when $\dim(h) = 2$. (Note that here, dimension refers to the hyperparameter h, whereas for the EM variants discussed earlier, it is the dimension of θ that causes problems.)

The purpose of this paper is to review the literature on selection of the hyperparameter, and to present an approach for estimating $\arg\max_h m(h)$ which scales well with the dimensions of h and θ. The approach is described as follows. We consider a fully-Bayes model in which we put a prior distribution on the hyperparameter h. To make the discussion as simple as possible in this introductory description, we will take the prior on h to be the uniform distribution, although as we explain in Section 2, ultimately this is not the prior that we will use, and we will need to apply a minor adjustment to account for this. Let u denote the uniform distribution on h. This prior induces a joint distribution on (h, θ, Y), which we will denote by π. Let $\pi_{(h,\theta)\,|\,y}$ denote the posterior distribution of (h, θ) given $Y = y$, and let $\pi_{h\,|\,y}$ denote the marginal posterior distribution of h given $Y = y$. Regarding $\pi_{h\,|\,y}$, the statement "the posterior distribution is proportional to the likelihood times the prior" reads as $\pi_{h\,|\,y}(h) \propto m(h)u(h)$. Since $u \propto 1$, this may be rewritten as $\pi_{h\,|\,y} \propto m(h)$, so the mode of $\pi_{h\,|\,y}$ is $\arg\max_h m(h)$. Now, suppose that we can construct a geometrically ergodic Markov chain $(h_1, \theta_1), (h_2, \theta_2), \ldots$ whose invariant distribution is $\pi_{(h,\theta)\,|\,y}$. The marginal sequence $h_1, h_2, \ldots$ then has invariant distribution equal to $\pi_{h\,|\,y}$. Any method for estimating the mode of $\pi_{h\,|\,y}$ from the sequence $h_1, h_2, \ldots$ gives rise to an estimate of $\arg\max_h m(h)$. Generally speaking, estimation of the mode is a hard problem, and the optimal rates of convergence are worse than $n^{1/2}$. The theory is technical but a typical result states that if f is a density on $\mathbb{R}^k$, then under some regularity conditions, the optimal convergence rate for estimation of the mode of f based on an iid sample is $n^{1/(4+k)}$ (Tsybakov 1990, see also Donoho and Liu 1991), so even in the simplest case where $k = 1$, this is the very slow rate of $n^{1/5}$. However, these pessimistic results pertain only to the case where the only information we have about f is the Monte Carlo information in

the sample. In Bayesian problems, we typically also have some information concerning π. For example, the conditional density of h given θ and y may be available, and in this case, Rao-Blackwellization is possible: $\pi_{h\,|\,y}$ may be estimated by $\hat{\pi}_{h\,|\,y}(h) = (1/n)\sum_{i=1}^{n} \pi_{h\,|\,(\theta=\theta_i, Y=y)}(h)$. It turns out that for any fixed h, $\hat{\pi}_{h\,|\,y}(h)$ converges to $\pi_{h\,|\,y}(h)$ at the rate of $n^{1/2}$—after all, $\hat{\pi}_{h\,|\,y}(h)$ is simply an average. Using tools from empirical process theory, we show that we have $n^{1/2}$-convergence uniformly in h. More precisely, we show that $n^{1/2}\big(\hat{\pi}_{h\,|\,y}(\cdot) - \pi_{h\,|\,y}(\cdot)\big)$ converges to a mean-zero Gaussian process indexed by h. From this, we show that $\arg\max_h \hat{\pi}_{h\,|\,y}(h)$ converges to $\arg\max_h m(h)$, also at the rate of $n^{1/2}$, regardless of the dimension of h, and we show how confidence sets for $\arg\max_h m(h)$ can be constructed. Standard methods based on gradient-based approaches can be used to find $\arg\max_h \hat{\pi}_{h\,|\,y}(h)$ rapidly, even when $\dim(h)$ is moderately large.

As will be seen later, our approach may be used for hyperparameter selection problems when dealing with hierarchical Dirichlet processes and also problems for which it is very difficult or impossible to obtain an unbiased estimate of risk. Additionally, a by-product of our approach is a method for simultaneous estimation of a posterior expectation as the hyperparameter varies. More specifically, let g be some function of θ. In Bayesian sensitivity analysis, we are interested in $I(h) \coloneqq \int g(\theta)\nu_{h,y}(\theta)\,d\theta$ as h varies continuously. We show that from the single Markov chain $(h_1, \theta_1), (h_2, \theta_2), \ldots$ we can estimate the entire function $I(h)$ as h varies. More specifically, from the Markov chain we construct an estimator $\hat{I}(h)$ and show that $n^{1/2}\big(\hat{I}(\cdot) - I(\cdot)\big)$ converges to a mean-zero Gaussian process indexed by h. We also show how this result can be used to construct simultaneous confidence intervals for $I(h)$, as h varies continuously over $\mathcal{H}$.

The rest of the paper is organized as follows. Section 2 contains our methodological and theoretical results. There, we describe our approach in more detail, and deal with the situation where the distribution on $\mathcal{H}$ is not the uniform. We provide three theorems which give precise statements of the convergence results mentioned above. And we discuss modifications of our approach for cases where Rao-Blackwellization is not possible. Section 2 also contains our review of other approaches for selecting the hyperparameter and a comparison of our approach with these previous approaches. Section 3 provides two of three illustrations which are chosen to have very different characters in order to demonstrate the scope of our methodology. The first illustration involves an empirical Bayes approach to variable selection in regression. We consider an additive regression model in which each variable is fit using a regression spline. We show how our methodology can be used to select the significant knots for each predictor variable and, interestingly, also to eliminate variables which are not useful in the regression. The second illustration involves a Bayesian tree model for regression with many predictors. The model features a hyperparameter that controls sparsity. We show how our methodology can be used to select the sparsity parameter, and also show that with this adaptive choice, the model acts appropriately in sparse and nonsparse situations. The third illustration is in the supplement. In Section S-1 of Doss and Linero (2021) we discuss an example which involves a robust binary regression model. In this example, we use our methodology to select the "robustness" parameter, which is a parameter in the model, as opposed to the hyperparameter of the prior distribution. In Section 3 we also discuss the question of whether one should do a fully Bayes analysis instead of an empirical Bayes analysis. The proofs of our theoretical results are in Section S-2 of Doss and Linero (2021).

2 Estimation of the Marginal Likelihood Function and Its Argmax

This section consists of four parts. In the first, we explain in detail our approach for obtaining the maximizer of the marginal likelihood function $m(h)$; in the second, we show how the ideas underlying our approach can be used to construct simultaneous confidence bands for the posterior expectation of a function of θ as the hyperparameter of the prior varies continuously; in the third, we discuss the "Importance Weighted Marginal Density Method" of Chen (1994), which is an alternative to Rao-Blackwellization; and in the fourth, we describe existing methods for hyperparameter selection, discussing strengths and weaknesses, and compare our method with the existing methods (the reader who is interested in understanding and using our methodology but is not interested in a review and evaluation of other approaches can skip the fourth part without loss).

2.1 The Fully-Bayes Empirical Bayes Method

When h is random, we will use the letter π, with subscripts, to indicate distributions on the triple (h, θ, Y), in a self-explanatory manner. Thus, π_h will indicate a prior on h, $\pi_{h\,|\,y}$ will indicate the posterior distribution of h given $Y = y$, $\pi_{(h,\theta)\,|\,y}$ the joint conditional distribution of (h, θ) given $Y = y$, etc. (In the beginning of Section 1 we considered the model in which h is not random, but fixed, and we used the notation ν_h and $\nu_{h,y}$ to denote the prior on θ indexed by h, and the corresponding posterior, respectively. These correspond to $\pi_{\theta\,|\,h}$ and $\pi_{\theta\,|\,(h,y)}$, respectively.) The prior π_h may be something other than the uniform for several reasons. For example, certain forms for π_h may induce a partial conjugacy structure and thus enable the construction of MCMC algorithms in which the component updates exploit this conjugacy.

Recall that the marginal likelihood of h is $m(h) = \int \ell_y(\theta)\nu_h(\theta)\,d\theta$. The posterior distribution of h given $Y = y$ is $\pi_{h\,|\,y}(h) \propto m(h)\pi_h(h)$, which we rewrite as $\pi_{h\,|\,y}(h)/\pi_h(h) = cm(h)$, where c is the constant of proportionality, so we will need the maximizer of $\pi_{h\,|\,y}/\pi_h$, rather than the mode of $\pi_{h\,|\,y}$. Suppose that $(h_1, \theta_1), (h_2, \theta_2), \ldots$ is a Markov chain with invariant distribution equal to $\pi_{(h,\theta)\,|\,y}$. If the conditional density of h given θ and y is available, then from the chain we may form the Rao-Blackwellized estimate of $\pi_{h\,|\,y}$ given by

$$\hat{\pi}_{h\,|\,y}(h) = \frac{1}{n}\sum_{i=1}^{n} \pi_{h\,|\,(\theta=\theta_i, Y=y)}(h). \tag{2.1}$$

(Remark: notation of the sort $\pi_h(h)$ or $\pi_{h\,|\,y}(h)$ is unfortunately potentially confusing, with the lower case letter h playing two distinct roles: as a subscript, it indicates that π_h or $\pi_{h\,|\,y}$ is a density on the random variable h, while as an argument, it indicates that this density is evaluated at the point h. On occasion, we will use notation of the sort $\pi_h(h_*)$ to indicate the value of the density π_h at the point h_*, when this is necessary to avoid confusion.) For any fixed h, $\hat{\pi}_{h\,|\,y}(h)$ is an average. Suppose temporarily that $(h_1, \theta_1), (h_2, \theta_2), \ldots$ is a sequence of independently and identically distributed draws from $\pi_{(h,\theta)\,|\,y}$. If

$$E\big(\pi_{h\,|\,(\theta=\theta_1, Y=y)}(h)\big) = \pi_{h\,|\,y}(h) \tag{2.2}$$

and $E\big(\big[\pi_{h\,|\,(\theta=\theta_1,Y=y)}(h)\big]^2\big) < \infty$, then the estimate in (2.1) would be consistent, and also satisfy a central limit theorem. The second moment assumption is, of course, standard, but statement (2.2) actually requires an argument, which is given in Section S-2 of Doss and Linero (2021).

We now consider the Markov chain case. Consistency of $\hat{\pi}_{h\,|\,y}(h)$ is guaranteed under the minimal and easily checkable condition that the chain is Harris ergodic (that is, it is irreducible, aperiodic, and Harris recurrent; see chapter 17 of Meyn and Tweedie 1993 for definitions). We get a central limit theorem if

(i) the Markov chain mixes fast enough, and

(ii) the random variable $\pi_{h\,|\,(\theta=\theta_1,Y=y)}(h)$ has a high enough moment.

Weakening condition (i) requires strengthening condition (ii) and vice versa. There are many sets of conditions, and perhaps the most convenient is one given in Ibragimov and Linnik (1971) and that involves geometric ergodicity. (The definition of geometric ergodicity of a Markov chain $X_1, X_2, \ldots$ on the measurable space $(\mathsf{X}, \mathcal{B})$ and having Π as an invariant probability measure is as follows. Let $K^n(x, A)$ be the n-step Markov transition function. The chain is *geometrically ergodic* if there exist a constant $\rho \in [0, 1)$ and a function $M\colon \mathsf{X} \to [0, \infty)$ such that for $n = 1, 2, \ldots$, $\sup_{A\in\mathcal{B}} |K^n(x, A) - \Pi(A)| \leq M(x)\rho^n$ for all $x \in \mathsf{X}$.) The condition is that the chain is geometrically ergodic, and for some $\epsilon > 0$, $\pi_{h\,|\,(\theta=\theta_1,Y=y)}(h)$ has a moment of order $2 + \epsilon$, for some $\epsilon > 0$ (corollary to Theorem 18.5.3 of Ibragimov and Linnik 1971). Thus, under Harris and geometric ergodicity and the moment condition stated above, for each $h \in \mathcal{H}$,

$$\hat{\pi}_{h\,|\,y}(h) \xrightarrow{\text{a.s.}} \pi_{h\,|\,y}(h) \qquad \text{as } n \to \infty \tag{2.3}$$

and

$$n^{1/2}\big(\hat{\pi}_{h\,|\,y}(h) - \pi_{h\,|\,y}(h)\big) \xrightarrow{d} \mathcal{N}(0, \sigma^2(h)) \qquad \text{as } n \to \infty$$

for some $\sigma^2(h) < \infty$, and consequently, for each $h \in \mathcal{H}$, we have

$$\hat{\pi}_{h\,|\,y}(h)/\pi_h(h) \xrightarrow{\text{a.s.}} cm(h) \tag{2.4}$$

and

$$n^{1/2}\big(\hat{\pi}_{h\,|\,y}(h)/\pi_h(h) - cm(h)\big) \xrightarrow{d} \mathcal{N}\big(0, \sigma^2(h)/\pi_h^2(h)\big), \tag{2.5}$$

where c is the constant defined right above (2.1).

Our principal objectives are to show that

$$\arg\max_h [\hat{\pi}_{h\,|\,y}(h)/\pi_h(h)] \xrightarrow{\text{a.s.}} \arg\max_h m(h) \tag{2.6}$$

and, more importantly,

$$n^{1/2}\Big(\arg\max_h [\hat{\pi}_{h\,|\,y}(h)/\pi_h(h)] - \arg\max_h m(h)\Big) \xrightarrow{d} \mathcal{N}_{\dim(h)}(0, \Sigma) \tag{2.7}$$

where Σ can be estimated consistently, as this would enable us to construct confidence sets for $\arg\max_h m(h)$. A succinct summary of our methodology, including what to do if Rao-Blackwellization is not feasible, is given at the end of Section 2.3.

Note that the pointwise convergence statement that $\hat{\pi}_{h\,|\,y}(h)/\pi_h(h) \xrightarrow{\text{a.s.}} cm(h)$ for every h does not imply convergence of $\arg\max_h[\hat{\pi}_{h\,|\,y}(h)/\pi_h(h)]$ to $\arg\max_h m(h)$ in any sense at all. In fact, even in the very simple case where $\{f_n\}_{n=1}^{\infty}$ and f are deterministic functions defined on $[0,1]$, the statement $f_n(x) \to f(x)$ for every $x \in [0,1]$ does not imply that $\arg\max_x f_n(x)$ converges to $\arg\max_x f(x)$. (To get a counterexample, with $\phi_{\mu,v}$ denoting the density of the normal distribution with mean μ and variance v, consider $f_n(x) = \phi_{1/n,1/n}(x) + (x-.9)^2$, and $f(x) = (x-.9)^2$.) To obtain convergence of the argmax, it is necessary to have uniformity in the convergence of f_n to f. Theorem 1 below has two parts. The first gives a version of (2.3) that is uniform in h. Thus, the first part is a non-trivial generalization of what one gets from the usual strong law of large numbers in two directions: (i) it gives a convergence statement that is uniform in h, and (ii) it does this for the Markov chain case. (The first part also gives a version of (2.4) that is uniform in h, which is an immediate consequence of the version of (2.3) that is uniform in h.) The second part, which is really a simple consequence of the first part, is assertion (2.6). Theorem 2, which is the result of principal interest, uses the uniformity in Theorem 1 to arrive at (2.7), and the theorem is followed by the description of a method for obtaining confidence regions for $\arg\max_h m(h)$.

Before stating the theorems, we briefly review some basic facts regarding empirical process theory and regeneration sequences that form the underpinnings of the theorems. Suppose $(\Omega, \mathcal{A}, P)$ is a probability space, let $L_1(P) = L_1(\Omega, \mathcal{A}, P)$ be the usual space of measurable functions $f\colon \Omega \to \mathbb{R}$ satisfying $\int |f|\, dP < \infty$, and let $L_2(P)$ be the usual space of square-integrable functions. Assume that $X_1, X_2, \ldots \overset{\text{iid}}{\sim} P$. If $f \in L_1(P)$, then the strong law of large numbers (SLLN) states that $(1/n)\sum_{i=1}^n f(X_i) \xrightarrow{\text{a.s.}} E(f(X))$, and if $f \in L_2(P)$, the central limit theorem (CLT) states that $n^{1/2}\big(\big[(1/n)\sum_{i=1}^n f(X_i)\big] - E(f(X))\big) \xrightarrow{d} \mathcal{N}(0, \sigma^2(f))$, where $\sigma^2(f)$ is the variance of $f(X)$ under P. For example, if $\Omega = \mathbb{R}$ and $f_t(X) = I(X \le t)$, where I is the indicator function, then the SLLN states that

$$\left| \frac{1}{n}\sum_{i=1}^n f_t(X_i) - E(f_t(X)) \right| \xrightarrow{\text{a.s.}} 0, \tag{2.8}$$

and the CLT states that

$$n^{1/2}\left(\frac{1}{n}\sum_{i=1}^n f_t(X_i) - E(f_t(X)) \right) \xrightarrow{d} \mathcal{N}(0, \sigma^2(t)). \tag{2.9}$$

The classical Glivenko-Cantelli theorem asserts that convergence in (2.8) is uniform, i.e. $\sup_{t\in\mathbb{R}} \big|(1/n)\sum_{i=1}^n f_t(X_i) - E(f_t(X))\big| \xrightarrow{\text{a.s.}} 0$, and the classical Donsker theorem gives a uniform version of (2.9), namely $n^{1/2}\big(\big[\sum_{i=1}^n f_\cdot(X_i)\big]/n - E(f_\cdot(X))\big) \xrightarrow{d} W_0(F(\cdot))$, where $W_0(\cdot)$ is the Brownian bridge on $[0,1]$, and $F(t) = P(X \le t)$. In empirical process theory, the one-parameter class of functions $\{f_t,\, t \in \mathbb{R}\}$ is replaced with far more general classes (and Ω is not necessarily $\mathbb{R}$, but can be a more general space). There are two kinds of results. Glivenko-Cantelli results pertain to classes of functions $\mathcal{F} \subset L_1(P)$ and these assert statements of the form

$$\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n f(X_i) - E(f(X)) \right| \xrightarrow{\text{a.s.}} 0.$$

Donsker results pertain to classes of functions $\mathcal{F} \subset L_2(P)$ and these assert statements of the form

$$n^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - E(f(X))\right) \xrightarrow{d} W(f),$$

where W is a mean-zero Gaussian process indexed by $f \in \mathcal{F}$. A good review of empirical process theory is given in Kosorok (2008).

We will be interested in the case where $X = (h, \theta)$, P is the conditional distribution of (h, θ) given $Y = y$, and our class of functions is $\left\{\pi_{h\,|\,(\theta,y)}(h),\ h \in \mathcal{H}\right\}$. (Recall that these are the core functions of θ used to construct the Rao-Blackwellized estimate $\hat{\pi}_{h\,|\,y}(h)$ in (2.1), which is an average, and $\hat{\pi}_{h\,|\,y}(h)/\pi_h(h)$ is what features in (2.4) and (2.5).) Note that for fixed h, $\pi_{h\,|\,(\theta,y)}(h)$ is a random variable, with all the randomness given by θ (see the remark regarding the notation immediately following (2.1)), and we emphasize this in our notation by writing

$$f_h(\theta) = \pi_{h\,|\,(\theta,y)}(h). \tag{2.10}$$

Of course, f_h also depends on y, but we have conditioned on y, and this dependence is suppressed in the notation. Throughout, we make the benign (and checkable) assumption that $f_\cdot(\cdot)\colon \mathcal{H} \times \Theta \to \mathbb{R}$ is continuous in h for $[\pi_{\theta\,|\,y}]$-almost all θ. Glivenko-Cantelli results will give statements of the sort $\sup_{h\in\mathcal{H}} \left|(1/n)\sum_{i=1}^{n} f_h(\theta_i) - \pi_{h\,|\,y}(h)\right| \xrightarrow{\text{a.s.}} 0$, and hence

$$\sup_{h\in\mathcal{H}} \left|\frac{1}{n}\sum_{i=1}^{n} f_h(\theta_i)/\pi_h(h) - cm(h)\right| \xrightarrow{\text{a.s.}} 0,$$

which is precisely the uniformity that what we need in order to establish convergence of the estimator of the argmax; see (2.6). And Donsker theorems ultimately will give results of the sort (2.7).

Empirical process theory is fundamentally based on an iid assumption, whereas in our setting, typically the sequence $(h_1, \theta_1), (h_2, \theta_2), \ldots$ will be a Markov chain, and the component sequence $\theta_1, \theta_2, \ldots$ will also be a Markov chain. The best way to deal with the family of averages $(1/n)\sum_{i=1}^{n} f_h(\theta_i)$, $h \in \mathcal{H}$, is through the use of "regenerative simulation." A *regeneration* is a random time at which a stochastic process probabilistically restarts itself; therefore, the tours made by the process in between such random times are iid. Regeneration sequences are easy to construct and understand in the setting of Markov chains on a discrete state space. Suppose that α is a point to which the chain returns infinitely often with probability one. Assume we start the chain at α, and let $1 = \tau_0 < \tau_1 < \tau_2 < \cdots$ be the times of return to α. For each $h \in \mathcal{H}$ and $r = 1, 2, \ldots$, let

$$S_{h,r} = \sum_{i=\tau_{r-1}}^{\tau_r - 1} f_h(\theta_i) \qquad \text{and} \qquad N_r = \tau_r - \tau_{r-1}. \tag{2.11}$$

These are the sum of f_h over the r^{th} tour and the length of the r^{th} tour, respectively. By the Markov property, the pairs $\{(S_{h,r}, N_r)\}_{r=1}^{\infty}$ are iid, and we will show how the iid structure will enable us to convert Glivenko-Cantelli and Donsker theorems for the iid case to Glivenko-Cantelli and Donsker theorems for the case of Markov chains satisfying some regularity conditions.

Before we do this, we remark that in virtually all cases that arise in Bayesian statistics, the state space is continuous, and there does not exist a point to which the chain returns infinitely often with probability one. The technique in Athreya and Ney (1978) allows us to construct a sequence of regeneration times $1 = \tau_0 < \tau_1 < \tau_2 < \cdots$ satisfying $E(\tau_r - \tau_{r-1}) < \infty$ in a very general setting that includes Markov chains on a continuous state spaces where the probability of visiting any particular point is always 0. Markov chains for which there exist such regeneration sequences are called regenerative. We discuss the regularity conditions needed for the Athreya and Ney (1978) construction to be feasible after the statements of the theorems.

Suppose now that our Markov chain is regenerative. If we run the chain for R regenerations, then the total number of cycles is given by $n = \sum_{r=1}^{R} N_r$. Also, $\sum_{i=1}^{n} f_h(\theta_i) = \sum_{r=1}^{R} S_{h,r}$. These two facts, which are true by definition, give rise to the key equation

$$\frac{\sum_{i=1}^{n} f_h(\theta_i)}{n} = \frac{\sum_{r=1}^{R} S_{h,r}}{\sum_{r=1}^{R} N_r} = \frac{\left(\sum_{r=1}^{R} S_{h,r}\right)/R}{\left(\sum_{r=1}^{R} N_r\right)/R}. \tag{2.12}$$

On the left are the averages of interest, but the θ_i's are not independent. On the right, the numerator is a class of averages of independent quantities, indexed by h, to which we can apply empirical process results. We have Glivenko-Cantelli theorems for the class of averages $\left(\sum_{r=1}^{R} S_{h,r}\right)/R$ and hence for the class of ratios on the right side of (2.12) (the denominator does not depend on h). And to obtain Donsker theorems for the class of ratios $\left[\left(\sum_{r=1}^{R} S_{h,r}\right)/R\right]\big/\left[\left(\sum_{r=1}^{R} N_r\right)/R\right]$, we apply the delta method to the function $t(x, y) = x/y$. There are, of course, many details that we have not discussed, but the present paragraph gives the big picture, and the details are dealt with in the proofs of the theorems.

Recall that f_h is defined by (2.10), and $S_{h,1}$ is defined by (2.11). Let $M(h) = cm(h)$ and let $\widehat{M}_n(h) = \hat{\pi}_{h\,|\,y}(h)/\pi_h(h)$ (see (2.4)). For a function $g\colon \mathcal{H} \to \mathbb{R}$, $\nabla_h g(h)$ denotes the gradient vector and $\nabla_h^2 g(h)$ denotes the Hessian matrix. We will refer to the regularity conditions below. They are discussed after the statements of Theorems 1 and 2.

- C1 The hyperparameter space $\mathcal{H}$ is a convex compact subset of $\mathbb{R}^k$.
- C2 The prior π_h is twice continuously differentiable on $\mathcal{H}$, and is positive on $\mathcal{H}$.
- C3 For every θ, $\nabla_h f_h(\theta)$ and $\nabla_h^2 f_h(\theta)$ exist and are continuous for all h.
- C4 The family $\{f_h,\, h \in \mathcal{H}\}$ is such that the interchange of order of integration and either first or second order differentiation is permissible: $\nabla_h \int f_h(\theta)\pi_{\theta\,|\,y}(\theta)\, d\theta = \int \nabla_h f_h(\theta)\pi_{\theta\,|\,y}(\theta)\, d\theta$ and $\nabla_h^2 \int f_h(\theta)\pi_{\theta\,|\,y}(\theta)\, d\theta = \int \nabla_h^2 f_h(\theta)\pi_{\theta\,|\,y}(\theta)\, d\theta$.
- C5 The function $m(\cdot)$ is twice continuously differentiable in $\mathcal{H}$, and the $k \times k$ Hessian matrix $\nabla_h^2 m(\arg\max_h m(h))$ is nonsingular.
- C6 Each of $\widehat{M}_n(\cdot)$, $n = 1, 2, \ldots$ and $m(\cdot)$ has a unique maximizer (thus it makes sense to talk about $\arg\max_h \widehat{M}_n(h)$ and $\arg\max_h m(h)$).
- C7 The sequence $\{\theta_i\}_{i=1}^{\infty}$ is a geometrically ergodic Markov chain with invariant distribution equal to $\pi_{\theta\,|\,y}$.

C8 For every $h \in \mathcal{H}$, there exists $\epsilon > 0$ such that $E\big(\|\nabla_h f_h(\theta)\|^{2+\epsilon}\big) < \infty$, where the expectation is with respect to $\pi_{\theta\,|\,y}$, and $\|\cdot\|$ is the Euclidean norm.

C9 $E(\sup_h S_{h,1}) < \infty$.

C10 $E\big(\sup_h |\nabla_h^2 S_{h,1}|\big) < \infty$.

The condition in C10 is of the form $E(|A|) < \infty$ where A is a $k \times k$ matrix, and the statement $E(|A|) < \infty$ should be taken to mean that the expected value is finite for every component of A.

Theorem 1 *Suppose that $\{\theta_i\}_{i=0}^{\infty}$ is a regenerative Harris ergodic Markov chain.*

1. *Under C2 and C9,* (2.3) *holds uniformly in h; consequently,* (2.4) *holds uniformly in h, i.e.*

$$\sup_{h\in\mathcal{H}} \big|\widehat{M}_n(h) - M(h)\big| \xrightarrow{\text{a.s.}} 0.$$

2. *Under C1, C2, C6, and C9,* $\arg\max_h \widehat{M}_n(h) \xrightarrow{\text{a.s.}} \arg\max_h M(h)$ *(which is equivalent to* (2.6)*).*

Theorem 2 *Suppose that $\{\theta_i\}_{i=0}^{\infty}$ is a regenerative Harris ergodic Markov chain. Under C1–C10,*

$$n^{1/2}\Big(\arg\max_h \widehat{M}_n(h) - \arg\max_h m(h)\Big) \xrightarrow{d} \mathcal{N}_k(0,\Sigma),$$

where Σ is a positive definite $k \times k$ matrix.

Batch-Based Estimation of Σ and Confidence Regions for $\arg\max_h m(h)$ The variance matrix Σ may be estimated in several ways. One way is to exploit the representation of $\widehat{M}_n(h)$ as the ratio of two averages, in which the numerator is an average of the iid quantities $S_{h,r}$, $r = 1,\ldots,R$ (see (2.12) and the proof of Theorem 1). We then produce an argument showing that $\arg\max_h \widehat{M}_n(h)$ inherits the representation in terms of averages, and apply standard methods for estimating the variance of an average.

Another way, which is much easier and is the one we recommend and we have implemented, is to use the method of batching, which is described as follows. The sequence $\theta_1,\ldots,\theta_n$ is broken up into B_n consecutive pieces of equal lengths called batches. Let $A_n^{[b]}$ denote the estimate of the argmax based on batch b, let A_n^{full} be the estimate based on the full sequence, and let A be short for $\arg\max_h m(h)$. Suppose that the number of batches and the batch length both go to infinity, i.e. $B_n \to \infty$ and $n/B_n \to \infty$ as $n \to \infty$. We note the following.

1. By Theorem 2, for $b = 1,\ldots,B_n$, the distribution of $(n/B_n)^{1/2}(A_n^{[b]} - A)$ is approximately $\mathcal{N}_k(0,\Sigma)$.
2. Under geometric ergodicity, for large n the variables $(n/B_n)^{1/2}A_n^{[1]},\ldots,(n/B_n)^{1/2}A_n^{[B_n]}$ are nearly independent.

If statements 1 and 2 above were exact, as opposed to approximations, then the batch-based estimate defined by

$$\widehat{\Sigma}_n = \frac{\sum_{b=1}^{B_n}(n/B_n)\big(A_n^{[b]} - A_n^{\text{full}}\big)\big(A_n^{[b]} - A_n^{\text{full}}\big)^{\top}}{B_n - 1}$$

would be a consistent estimator of Σ, since $\widehat{\Sigma}_n$ is (essentially) the sample variance based on the sequence $(n/B_n)^{1/2}A_n^{[1]}, \ldots, (n/B_n)^{1/2}A_n^{[B_n]}$.

In general terms, the literature shows that batch-based estimates are consistent under certain regularity assumptions. Jones et al. (2006) and Flegal and Jones (2010) establish strong consistency under the condition that $\{\theta_i\}_{i=1}^{\infty}$ is geometrically ergodic, some moment conditions, and stipulations regarding the rate at which $B_n \to \infty$ (Jones et al. 2006 recommend taking $B_n = n^{1/2}$, which is within the range of rates that Jones et al. 2006 and Flegal and Jones 2010 allow); see also Flegal et al. (2008). Their results pertain to the case where the statistic whose variance we need to estimate is an average, whereas our statistic is an argmax. However, in view of the fact that $\arg\max_h \widehat{M}_n(h)$ may be represented as an average plus a term that is asymptotically negligible (this fact is the crux of the proof of Theorem 2), we expect the batch-based estimate to be consistent in our situation also. An additional difference is that in our definition of $\widehat{\Sigma}_n$ we have used A_n^{full} as the centering value, instead of the traditional $A_n^{[\cdot]} := (1/B_n)\sum_{b=1}^{B_n} A_n^{[b]}$; but we do not think that making this change affects the theory.

If $\widehat{\Sigma}_n \xrightarrow{\text{a.s.}} \Sigma$, then $\widehat{\Sigma}_n$ is invertible for large n. Hence the ellipse $\mathcal{E}$ given by $\mathcal{E} = \big\{h : \big(\arg\max_h \widehat{M}_n(h) - h\big)^{\top}\widehat{\Sigma}_n^{-1}\big(\arg\max_h \widehat{M}_n(h) - h\big) \le \chi^2_{k,.95}/n\big\}$ is an asymptotic 95% confidence set for $\arg\max_h m(h)$. Here, $\chi^2_{k,.95}$ denotes the 0.95 quantile of the chi-square distribution with k degrees of freedom.

We now remark on the existence of regeneration sequences. First, we note that, at the theoretical level, it is a fact that for any chain satisfying our minimal regularity condition of Harris ergodicity, letting $K(x, A)$ denote its Markov transition function, there exists a $j \ge 1$ such that for the Markov chain driven by the j-step Markov transition function K^j, there is a regeneration sequence satisfying $E(\tau_r - \tau_{r-1}) < \infty$; see Meyn and Tweedie (1993, Theorem 5.2.2). At a more practical level, Mykland et al. (1995) have provided a very general method, the so-called distinguished point technique, for constructing regeneration sequences. When the method produces regeneration times that are reasonably short, estimation of standard errors of $(1/n)\sum_{i=1}^n f_h(\theta_i)$ becomes trivial, since by (2.12), the problem is reduced to estimating the standard error of ratios of averages of iid quantities. An often-heard criticism of the method is that it is very hard to tune, especially in high-dimensional problems. This criticism is irrelevant for us: for our theoretical development, we need only the *existence* of the construction, not a construction that is useful in the practical sense.

Conditions C1–C10 Conditions C1 and C3–C6 are standard. C2 is fairly benign and is likely to be satisfied by most priors in common use, possibly after a redefinition of $\mathcal{H}$. The geometric ergodicity condition C7 and the moment condition C8 are typical of the sort needed for proving CLT's for averages, and we recall that in the fixed h regime, the Rao-Blackwell estimate (2.1) is just that, a (single) average.

We now discuss C9 and C10. Given a family of functions $g_h \colon \Theta \to \mathbb{R}$, Glivenko-Cantelli theorems are uniform SLLN's, and Donsker theorems are "uniform CLT's," i.e. theorems that assert convergence to Gaussian processes indexed by $h \in \mathcal{H}$, as opposed to convergence to a multivariate normal random variable. The regularity conditions of substance are $E(\sup_h |g_h(\theta)|) < \infty$ and

$E\big(\sup_h |g_h(\theta)|^2\big) < \infty$, for Glivenko-Cantelli and Donsker theorems, respectively (there are other conditions, but they are very easily checked in Bayesian problems, where the prior is indexed by a finite-dimensional parameter). The conditions that these expectations are finite are *far* more stringent than the conditions $\sup_h E(|g_h(\theta)|) < \infty$ and $\sup_h E\big(|g_h(\theta)|^2\big) < \infty$. The difficulty is even more severe because in our situation, the functions we're dealing with are $\{S_{h,1},\, h \in \mathcal{H}\}$. For any h, $S_{h,1}$ is a sum of the functions $f_h(\theta_i)$ over a random number of indices i (see (2.11)), so verification of a condition of the form $E(\sup_h |S_{h,1}|) < \infty$ could potentially be extremely problematic, and we now address this issue. Consider the conditions below:

CC1 $\int f_h(\theta)\pi_{\theta\,|\,y}(\theta)\,d\theta < \infty$ for all $h \in \mathcal{H}$, and for some $d \geq 1$, there exist $h_1, \ldots, h_d \in \mathcal{H}$ and constants $c_1, \ldots, c_d$ such that $\sup_h f_h(\theta) \leq \sum_{j=1}^d c_j f_{h_j}(\theta)$ for all $\theta \in \Theta$.

CC2 $\int |\nabla_h f_h(\theta)|\pi_{\theta\,|\,y}(\theta)\,d\theta < \infty$ for all $h \in \mathcal{H}$, and for some $d \geq 1$, there exist $h_1, \ldots, h_d \in \mathcal{H}$ and constants $c_1, \ldots, c_d$ such that $\sup_h |\nabla_h f_h(\theta)| \leq \sum_{j=1}^d c_j |\nabla_h f_{h_j}(\theta)|$ for all $\theta \in \Theta$.

CC3 $\int |\nabla_h^2 f_h(\theta)|\,\pi_{\theta\,|\,y}(\theta)\,d\theta < \infty$ for all $h \in \mathcal{H}$, and for some $d \geq 1$, there exist $h_1, \ldots, h_d \in \mathcal{H}$ and constants $c_1, \ldots, c_d$ such that $\sup_h |\nabla_h^2 f_h(\theta)| \leq \sum_{j=1}^d c_j |\nabla_h^2 f_{h_j}(\theta)|$ for all $\theta \in \Theta$.

(In CC2 and CC3, the inequality signs are taken to mean component-wise inequalities.) CC2 is not relevant for Theorems 1 and 2, but we mention it because it will be needed later for Theorem 3.

Obviously, CC1 implies $E(\sup_h f_h(\theta)) < \infty$, CC2 implies $E(\sup_h |\nabla_h f_h(\theta)|) < \infty$, and CC3 implies $E\big(\sup_h |\nabla_h^2 f_h(\theta)|\big) < \infty$. But much more can be said: CC1 implies $E(\sup_h S_{h,1}) < \infty$, CC2 implies $E(\sup_h |\nabla_h S_{h,1}|) < \infty$, and CC3 implies $E\big(\sup_h |\nabla_h^2 S_{h,1}|\big) < \infty$. We will prove the first of these assertions. Let $\mathcal{T}_1$ denote the set of indices that comprise the first tour. We have

$$S_{h,1} = \sum_{i \in \mathcal{T}_1} f_h(\theta_i) \leq \sum_{i \in \mathcal{T}_1} \sum_{j=1}^d c_j f_{h_j}(\theta_i) = \sum_{j=1}^d c_j \sum_{i \in \mathcal{T}_1} f_{h_j}(\theta_i) \qquad \text{for any } h \in \mathcal{H}. \tag{2.13}$$

We may replace $S_{h,1}$ with $\sup_h S_{h,1}$ on the left side of (2.13), and then taking expectations, we obtain $E_P(\sup_h S_{h,1}) = \sum_{j=1}^d c_j E_P\big(\sum_{i \in \mathcal{T}_1} f_{h_j}(\theta_i)\big) = \sum_{j=1}^d c_j E_{\pi_{\theta\,|\,y}}(f_{h_j}(\theta)) E_P(N_1) < \infty$, where E_P denotes expectation with respect to the Markov chain. The other two assertions are proved in essentially the same way. To summarize: establishing C9–C10 reduces to checking CC1 and CC3, and these are typically not difficult to check using the compactness of $\mathcal{H}$. (And we will see later one of the principal regularity conditions required for Theorem 3 reduces to checking CC2.)

2.2 Simultaneous Estimation of a Family of Posterior Expectations

Let g be a function of θ, and let $I(h) = \int g(\theta)\nu_{h,y}(\theta)\,d\theta = \int g(\theta)\pi_{\theta\,|\,(h,y)}(\theta)\,d\theta$ be the posterior expectation of $g(\theta)$ when the prior on θ is ν_h. Suppose we are interested in the family $I(h)$ as h

varies continuously. For any $h \in \mathcal{H}$, we have

$$\int g(\theta)\pi_{\theta\,|\,(h,y)}(\theta)\,d\theta = \frac{\pi_{h\,|\,y}(h)\int g(\theta)\pi_{\theta\,|\,(h,y)}(\theta)\,d\theta}{\pi_{h\,|\,y}(h)} \tag{2.14a}$$

$$= \frac{\int g(\theta)\pi_{\theta\,|\,(h,y)}(\theta)\pi_{h\,|\,y}(h)\,d\theta}{\pi_{h\,|\,y}(h)} \tag{2.14b}$$

$$= \frac{\int g(\theta)\pi_{(\theta,h)\,|\,y}(\theta,h)\,d\theta}{\pi_{h\,|\,y}(h)} \tag{2.14c}$$

$$= \frac{\int [g(\theta)\pi_{h\,|\,(\theta,y)}(h)]\pi_{\theta\,|\,y}(\theta)\,d\theta}{\pi_{h\,|\,y}(h)} =: \frac{N(h)}{\pi_{h\,|\,y}(h)}. \tag{2.14d}$$

Now suppose, as we did before, that we can construct a geometrically ergodic Markov chain $(h_1,\theta_1),(h_2,\theta_2),\ldots$ whose invariant distribution is $\pi_{(h,\theta)\,|\,y}$; and also suppose, as before, that the conditional density of h given θ and y is available. The numerator, $N(h)$, and denominator, $\pi_{h\,|\,y}(h)$, of (2.14d) may be estimated (using only the θ-component of the chain) by

$$\widehat{N}_n(h) \coloneqq \frac{1}{n}\sum_{i=1}^{n} g(\theta_i)\pi_{h\,|\,(\theta=\theta_i,y)}(h) \qquad \text{and} \qquad \hat{\pi}_{h\,|\,y}(h) = \frac{1}{n}\sum_{i=1}^{n}\pi_{h\,|\,(\theta=\theta_i,y)}(h), \tag{2.15}$$

respectively, where $\hat{\pi}_{h\,|\,y}(h)$ was defined earlier (see (2.1)). Therefore, we may estimate $I(h)$ by the ratio of $\widehat{N}_n(h)$ and $\hat{\pi}_{h\,|\,y}(h)$, i.e. estimate $I(h)$ via

$$\hat{I}_n(h) = \frac{\sum_{i=1}^{n} g(\theta_i)\pi_{h\,|\,(\theta=\theta_i,y)}(h)}{\sum_{i=1}^{n}\pi_{h\,|\,(\theta=\theta_i,y)}(h)}. \tag{2.16}$$

It is interesting to note that if we let $w_{h,i} = \big[\pi_{h\,|\,(\theta=\theta_i,y)}(h)\big/\sum_{j=1}^{n}\pi_{h\,|\,(\theta=\theta_j,y)}(h)\big]$, then the $w_{h,i}$'s are weights, and $\hat{I}_n(h) = \sum_{i=1}^{n} g(\theta_i)w_{h,i}$, i.e. $\hat{I}_n(h)$ has the interpretation as a weighted average of the $g(\theta_i)$'s, with weights given by the $w_{h,i}$'s.

For any fixed h, $\hat{I}_n(h) \xrightarrow{\text{a.s.}} I(h)$. To see this informally, note that $\widehat{N}_n(h) \xrightarrow{\text{a.s.}} N(h)$ and $\hat{\pi}_{h\,|\,y}(h) \xrightarrow{\text{a.s.}} \pi_{h\,|\,y}(h)$, because $\widehat{N}_n(h)$ and $\hat{\pi}_{h\,|\,y}(h)$ are averages over an ergodic Markov chain. Also, for fixed h, $n^{1/2}\big(\hat{I}_n(h) - I(h)\big)$ is asymptotically normal. Informally, this is because under regularity conditions on the mixing rate of the Markov chain and some moment conditions, the bivariate vector $n^{1/2}\big(\widehat{N}_n(h) - N(h), \hat{\pi}_{h\,|\,y}(h) - \pi_{h\,|\,y}(h)\big)$ is asymptotically jointly bivariate normal, by the Markov chain CLT. Asymptotic normality of $n^{1/2}\big(\hat{I}_n(h) - I(h)\big)$ follows from the delta method applied to the function $t(x,y) = x/y$. However, we will be interested in versions of these consistency and asymptotic normality statements that are uniform in h. For example, if $\dim(h) = 1$, in order to form simultaneous confidence bands for $I(h)$ (or regions, if $\dim(h) > 1$), we need the asymptotic distribution of $\hat{I}_n(h)$, viewed as a process in h. The uniform versions of these convergence statements are given by Theorem 3, whose proof uses results from empirical process theory.

Before stating the theorem, we give some definitions and state the assumptions we will need. Let $C(\mathcal{H})$ be the space of all continuous functions $x\colon \mathcal{H} \to \mathbb{R}$, with the topology induced by the sup norm metric ρ: for $x, y \in C(\mathcal{H})$, $\rho(x,y) = \|x - y\|_\infty = \sup_h |x(h) - y(h)|$. For a

regenerative Markov chain $\theta_1, \theta_2, \ldots$ with regeneration times $1 = \tau_0 < \tau_1 < \tau_2 < \cdots$ satisfying $E(\tau_r - \tau_{r-1}) < \infty$, the sequence $S_{h,r}$, $r = 1, 2, \ldots$ was defined by (2.11), and in analogy, we define

$$T_{h,r} = \sum_{i=\tau_{r-1}}^{\tau_r - 1} g(\theta_i) f_h(\theta_i), \qquad \text{for } h \in \mathcal{H},\ r = 1, 2, \ldots$$

We will refer to the following assumptions.

D1 $\pi_{h\,|\,y}(\cdot)$ is continuous and positive on $\mathcal{H}$.

D2 For every θ, $\nabla_h f_h(\theta)$ exists and is continuous for all h.

D3 The families $\{f_h,\ h \in \mathcal{H}\}$ and $\{g f_h,\ h \in \mathcal{H}\}$ are such that the order of integration and differentiation can be interchanged, i.e. $\nabla_h \int f_h(\theta)\pi_{\theta\,|\,y}(\theta)\,d\theta = \int \nabla_h f_h(\theta)\pi_{\theta\,|\,y}(\theta)\,d\theta$, and $\nabla_h \int g(\theta) f_h(\theta)\pi_{\theta\,|\,y}(\theta)\,d\theta = \int \nabla_h [g(\theta) f_h(\theta)]\pi_{\theta\,|\,y}(\theta)\,d\theta$.

D4 For every $h \in \mathcal{H}$, there exists $\epsilon > 0$ such that $E\big(f_h^{2+\epsilon}(\theta)\big) < \infty$ and $E\big((g(\theta) f_h(\theta))^{2+\epsilon}\big) < \infty$, where the expectations are with respect to $\pi_{\theta\,|\,y}$.

D5 For every $h \in \mathcal{H}$, there exists an $\epsilon > 0$ such that $E\big(\|\nabla_h f_h(\theta)\|^{2+\epsilon}\big) < \infty$ and $E\big(\|\nabla_h [g(\theta) f_h(\theta)]\|^{2+\epsilon}\big) < \infty$, where the expectations are with respect to $\pi_{\theta\,|\,y}$.

D6 $E\big(\sup_h \|\nabla_h S_{h,1}\|^2\big) < \infty$.

D7 $E\big(\sup_h \|\nabla_h T_{h,1}\|^2\big) < \infty$.

For Theorem 3, recall that $\hat{\pi}_{h\,|\,y}(\cdot)$, $N(\cdot)$, and $\widehat{N}_n(\cdot)$ are defined by (2.1), (2.14d), and (2.15), respectively, and that $\hat{I}_n(\cdot) = \widehat{N}_n(\cdot)/\hat{\pi}_{h\,|\,y}(\cdot)$.

Theorem 3 *Suppose that $\{\theta_i\}_{i=0}^{\infty}$ is a regenerative Harris ergodic Markov chain, and assume C1, C7, and D1–D7. Then:*

1.
$$\sup_{h \in \mathcal{H}} \big|\hat{I}_n(h) - I(h)\big| \xrightarrow{\text{a.s.}} 0.$$

2.
$$n^{1/2}\big(\hat{\pi}_{h\,|\,y}(h) - \pi_{h\,|\,y}(h)\big) \xrightarrow{d} \mathbb{P}(\cdot), \tag{2.17}$$

$$n^{1/2}\big(\widehat{N}_n(\cdot) - N(\cdot)\big) \xrightarrow{d} \mathbb{N}(\cdot), \tag{2.18}$$

$$n^{1/2}\big(\hat{I}_n(\cdot) - I(\cdot)\big) \xrightarrow{d} \mathbb{I}(\cdot), \tag{2.19}$$

where $\mathbb{P}$, $\mathbb{N}$, and $\mathbb{I}$ are mean 0 Gaussian processes indexed by $\mathcal{H}$, and the convergence takes place in $C(\mathcal{H})$.

In Part 2 of the theorem, (2.17) and (2.18) may be viewed as lemmas that are needed to prove (2.19), which is the result of principal interest. As will emerge in the proof of the theorem, it is possible to give explicit expressions for the covariance function of $\mathbb{I}(\cdot)$ in terms of first and second moments of certain random variables and, in principle, it is possible to estimate these moments and hence the covariance function of $\mathbb{I}(\cdot)$. However, to use this covariance function to

form simultaneous confidence bands (or sets) for $I(h)$, we would also need the distribution of $\sup_{h\in\mathcal{H}} |\mathbb{I}(h)|$, which is extremely complicated, even for the simplest parametric models. A convenient alternative way to form simultaneous confidence bands for $I(h)$ is through the method of batching, which we describe in the paragraph below.

Batch-Based Simultaneous Confidence Bands for the Family $\{I(h),\, h \in \mathcal{H}\}$ The map $K\colon C(\mathcal{H}) \to [0,\infty)$ defined by $K(x) = \sup_{h\in\mathcal{H}} |x(h)|$ is continuous, so by (2.19) in Theorem 3, $\sup_h n^{1/2}|\hat{I}_n(h) - I(h)| \xrightarrow{d} \sup_h |\mathbb{I}(h)|$. Suppose that the distribution of $\sup_h |\mathbb{I}(h)|$ is continuous. For $\alpha \in (0,1)$, let q_α be such that $P(\sup_h |\mathbb{I}(h)| \le q_\alpha) = 1-\alpha$. If q_α was known, then $P\big(\sup_h n^{1/2}|\hat{I}_n(h) - I(h)| \le q_\alpha\big) \to P(\sup_h |\mathbb{I}(h)| \le q_\alpha) = 1-\alpha$, i.e.

$$P\left(\hat{I}_n(h) - \frac{q_\alpha}{n^{1/2}} \le I(h) \le \hat{I}_n(h) + \frac{q_\alpha}{n^{1/2}} \text{ for all } h \in \mathcal{H}\right) \to 1-\alpha.$$

The difficulty is that the distribution of $\sup_h |\mathbb{I}(h)|$ is analytically intractable, so q_α is not known. The method of batching can be used to estimate it. As before, the sequence $\theta_1,\ldots,\theta_n$ is broken up into B_n consecutive batches of equal lengths. Let $\hat{I}^{(m)}(h)$ be the estimate of $I(h)$ formed from the b^{th} batch and, as before, suppose that as $n\to\infty$, $B_n \to \infty$ and $n/B_n \to \infty$. We will write B instead of B_n. For $b = 1,\ldots,B$, let $S_b = \sup_h (n/B)^{1/2}|\hat{I}_b(h) - I(h)|$. Then, because the batch length is large, the distribution of S_b is approximately equal to that of $\sup_h |\mathbb{I}(h)|$. Therefore, we may estimate q_α by the $(1-\alpha)$-quantile of the sequence $S_1,\ldots,S_B$. Unfortunately, the S_b's are not available, because they involve $I(h)$, which is unknown. So instead we use $\mathcal{S}_b = \sup_h (n/B)^{1/2}|\hat{I}_b(h) - \hat{I}(h)|$, in which we have substituted $\hat{I}(h)$ for $I(h)$. To conclude, let $\mathcal{S}_{[1]} \le \mathcal{S}_{[2]} \le \cdots \le \mathcal{S}_{[B]}$ denote the ordered values of the sequence $\mathcal{S}_1,\ldots,\mathcal{S}_B$. We estimate q_α via $\mathcal{S}_{[(1-\alpha)B]}$, and our simultaneous $(1-\alpha)$-level confidence band for $\{I(h),\, h\in\mathcal{H}\}$ is $\big\{\hat{I}(h) \pm \mathcal{S}_{[(1-\alpha)B]}/n^{1/2},\, h\in\mathcal{H}\big\}$.

Remark 1 There is a highly-developed theory on the consistency of estimates based on batching, including results on the optimal rate at which $B_n \to \infty$. However, this theory is focused primarily on the case where we are estimating a variance, whereas in the present situation, we are estimating the quantile of the distribution of the supremum of a stochastic process. There are some differences; for example, theoretically we need to show that substitution of $\hat{I}(h)$ for $I(h)$ in going from S_m to $\mathcal{S}_m$ does not cause any difficulties, and on the practical side, the rate at which B_n goes to infinity may differ because here we are estimating a moderately large quantile. Establishing the theoretical validity of the batch-based simultaneous confidence bands and providing theoretical and empirical results on the optimal rate at which $B_n \to \infty$ are interesting open problems.

Remark 2 Result (2.17) is of interest in its own right: together with the method of batching, it enables us to form simultaneous confidence bands for the marginal posterior density of h. Moreover, there is nothing intrinsic about h being a hyperparameter in the model. Result (2.17) and the construction of the batch-based confidence band apply in any situation where we have a multivariate parameter $\theta = (\theta_1,\ldots,\theta_k)$ and we wish to form a simultaneous confidence band for the

posterior density of θ_j for some j. The conditions needed are those stated in Theorem 3 (except for those that involve the function g), and we need a proof that the simultaneous confidence bands are asymptotically valid.

2.3 An Alternative to Rao-Blackwellization

Suppose that $(U_1, V_1), (U_2, V_2), \ldots$ is a Markov chain with invariant density $f_{U,V}$ on a space $\mathsf{U} \times \mathsf{V}$ where V is Euclidean. For the purpose of estimating the marginal density f_V, Chen (1994) introduced the so-called Importance Weighted Marginal Density Estimation (IWMDE) method, a very general procedure which can be applied in cases where Rao-Blackwellization is not feasible. In our context, in which θ corresponds to U, h corresponds to V, and our Markov chain is $(h_1, \theta_1), (h_2, \theta_2), \ldots$, the method is described as follows. Let $\{w_\theta(\cdot),\, \theta \in \Theta\}$ be a family of densities on $\mathcal{H}$. To estimate $\pi_{h\,|\,y}$ we use the estimator $\hat{\pi}^{\text{iwmde}}_{h\,|\,y}$ whose value at h_* is given by

$$\hat{\pi}^{\text{iwmde}}_{h\,|\,y}(h_*) = \frac{1}{n}\sum_{i=1}^{n} w_{\theta_i}(h_i)\frac{\pi_{(\theta,h)\,|\,y}(\theta_i, h_*)}{\pi_{(\theta,h)\,|\,y}(\theta_i, h_i)}. \tag{2.20}$$

Note that to calculate (2.20), we need only that $\pi_{(\theta,h)\,|\,y}$ is known up to a normalizing constant, and this is typically the case in Bayesian problems, where the posterior is proportional to the likelihood times the prior. Chen (1994) required that the family $\{w_\theta(\cdot),\, \theta \in \Theta\}$ correspond to a joint distribution on (θ, h); more precisely, he required that there exist a joint density $W_{\theta,h}$ on (θ, h), and that $w_\theta(h) = W_{h\,|\,\theta}(h)$, in self-explanatory notation. Actually, no such stipulation is needed, and in Section S-2 of Doss and Linero (2021) we show that under the minimal condition that for each θ, w_θ is a density on $\mathcal{H}$, $\hat{\pi}^{\text{iwmde}}_{h\,|\,y}(\cdot)$ is an unbiased estimate of $\pi_{h\,|\,y}(\cdot)$.

In principle, any family $\{w_\theta(\cdot),\, \theta \in \Theta\}$ of densities can be used in (2.20), but Chen (1994) showed that the choice $w_\theta(\cdot) = \pi_{h\,|\,(\theta,y)}(\cdot)$ is optimal in the sense of minimizing the asymptotic variance and, moreover, for this choice the estimator reduces to the Rao-Blackwell estimate $\hat{\pi}^{\text{RB}}_{h\,|\,y}(h_*) = (1/n)\sum_{i=1}^{n} \pi_{h\,|\,(\theta_i,y)}(h_*)$. This leads to the heuristic that the family $\{w_\theta(\cdot),\, \theta \in \Theta\}$ should be taken to be as close to $\pi_{h\,|\,(\theta,y)}(\cdot)$ as possible.

For every fixed $h_* \in \mathcal{H}$, consider the function $f^{\text{iwmde}}_{h_*} \colon \Theta \times \mathcal{H} \to \mathbb{R}$ defined by

$$f^{\text{iwmde}}_{h_*}(\theta, h) = w_\theta(h)\frac{\pi_{(\theta,h)\,|\,y}(\theta, h_*)}{\pi_{(\theta,h)\,|\,y}(\theta, h)}. \tag{2.21}$$

The IWMDE is $\hat{\pi}^{\text{iwmde}}_{h\,|\,y}(h_*) = (1/n)\sum_{i=1}^{n} f^{\text{iwmde}}_{h_*}(\theta_i, h_i)$. Our statement that $\hat{\pi}^{\text{iwmde}}_{h\,|\,y}(\cdot)$ is an unbiased estimate of $\pi_{h\,|\,y}(\cdot)$ may be written as $E_{\pi_{(\theta,h)\,|\,y}}(f^{\text{iwmde}}_{h_*}(\theta, h)) = \pi_{h\,|\,y}(h_*)$, and Theorems 1, 2 and 3 hold for the IWMDE: we simply replace $f_{h_*}(\theta) = \pi_{h\,|\,(\theta,y)}(h_*)$ (see (2.10)) with $f^{\text{iwmde}}_{h_*}(\theta, h)$ defined by (2.21), and definitions and assumptions involving f_{h_*} are now taken to refer to the function $f^{\text{iwmde}}_{h_*}$. Additionally, the requirement that $\theta_1, \theta_2, \ldots$ is a geometrically ergodic Markov chain with invariant distribution equal to $\pi_{\theta\,|\,y}$ now is replaced by the requirement that $(h_1, \theta_1), (h_2, \theta_2), \ldots$ is a geometrically ergodic Markov chain whose invariant distribution is $\pi_{(h,\theta)\,|\,y}$.

Our methodology is summarized as follows.

1. Choose a prior for h. The methodology is invariant to this choice, and different priors give rise to the same answers in the limit, so the selection should be based on convenience, for example exploiting any conjugacy or partial conjugacy in the problem.
2. Generate a suitably ergodic Markov chain on (h, θ) with invariant distribution equal to the posterior distribution of (h, θ) given $Y = y$.
3. If the conditional distributions needed for Rao-Blackwellization are available, then calculate the Rao-Blackwellized estimate 2.1. If the conditionals are not available, then use the IWMDE (whose variance is greater than that of the Rao-Blackwellized estimate).
4. Adjust the Rao-Blackwellized estimate or the IWMDE via division by $\pi_h(\cdot)$, and find the argmax of this ratio, which is an estimate of $\arg\max_h m(h)$.
5. Form confidence sets for $\arg\max_h m(h)$ via the method of batching.

2.4 Comparison with Other Methods

Here we discuss current schemes for estimating $\arg\max_h m(h)$, which are schemes for implementing the empirical Bayes method. The number of such schemes is very large, so we limit our review to the methods that are the most competitive, and also to those that are frequently used in the machine learning literature, whether or not these are competitive. Our discussion spans three groups.

Direct Monte Carlo Estimation of the Marginal Likelihood Function In approaches from this group, for each h over a fine grid in $\mathcal{H}$, we run a Monte Carlo experiment to form an estimate $\widehat{m}(h)$ of $m(h)$; we do this separately for each h, and we estimate $\arg\max_h m(h)$ via $\arg\max_h \widehat{m}(h)$. Papers that proceed in this way include Chib (1995), Chib and Jeliazkov (2001), and Newton and Raftery (1994) which introduced the "harmonic mean estimator." These approaches do not scale well with $\dim(h)$, because the size of the grid needed to cover $\mathcal{H}$ grows exponentially with $\dim(h)$. We mention them here simply because the machine learning literature frequently uses them, and this includes the harmonic mean estimator even though, typically, for each h, the harmonic mean estimator of $m(h)$ converges at a rate that is much slower than $n^{1/2}$ (Wolpert and Schmidler, 2012).

EM-Based Approaches As mentioned in the Introduction, the basic EM algorithm is rarely feasible, except in simple problems, and the principal variants are MCEM and VEM. In MCEM, the E-step is replaced by a Monte Carlo estimate, so an error is introduced at every iteration, and there is no reason to expect that the algorithm will converge at all, let alone to the true maximizer of the likelihood. In fact, Wei and Tanner (1990) recognized this problem and suggested that the Markov chain length be increased at every iteration of the EM algorithm. Let m_k denote the Markov chain length at the k^{th} iteration. Fort and Moulines (2003) showed that a minimal condition for convergence is that $m_k \to \infty$ at the rate of k^a, for some $a > 1$ (they do not give guidelines for choosing a). MCEM has been shown to perform very poorly in some cases. For example, George and Doss (2018), who deal with latent Dirichlet allocation (which is used in topic modeling and where $\dim(\theta)$ is very high), showed poor performance of MCEM even when a is taken to be 2.

The estimate of $\arg\max_h m(h)$ produced by VEM is obtained as follows. If h_k is the current value of h, the E-step of the EM algorithm is to calculate $E_{h_k}\big(\log(p_h(\theta, y))\big)$, where $p_h(\theta, y)$ is the joint distribution of (θ, y) under the model indexed by h, and the subscript to the expectation indicates that the expectation is taken with respect to $\nu_{h_k,y}$ (recall that the ν-notation refers to the model where h is not random, and $\nu_{h_k,y}$ is the posterior distribution of θ given $Y = y$ when the prior on θ is ν_{h_k}). This step is infeasible because $\nu_{h_k,y}$ is analytically intractable. We consider $\{q_\psi,\ \psi \in \Psi\}$, a (finite-dimensional) parametric family of analytically tractable distributions on θ, and within this family, we find the distribution, say q_{ψ_k}, which is closest to $\nu_{h_k,y}$ in the sense of minimizing the Kullback-Leibler (KL) divergence: $\psi_k = \arg\min_\psi \mathrm{KL}(q_\psi \| \nu_{h_k,y})$. Let $Q(h)$ be the expected value of $\log(p_h(\theta, y))$ with respect to q_{ψ_k}. We view $Q(h)$ as a surrogate for $E_{h_k}\big(\log(p_h(\theta, y))\big)$, and the M-step is then to maximize $Q(h)$ with respect to h, to produce h_{k+1}. The maximization is done analytically.

Suppose that $\theta = (\theta_1, \ldots, \theta_p)$ for some p. In mean-field variational inference, the version of variational inference that is most commonly used, the distributions q_ψ are all products of marginal densities, i.e. under q_ψ, $\theta_1, \ldots, \theta_p$ are independent. At each iteration of VEM, the minimization step is carried out through an iterative scheme. At convergence of this scheme, what is obtained is a member, q_{ψ_k}, of the parametric family. Let $\mathcal{P}_\Theta$ be the space of distributions on Θ (endowed with some topology, which will not concern us here), let ρ be any given metric on $\mathcal{P}_\Theta$, and let $h_{\text{true}} = \arg\max_h m(h)$ denote the target of the EM scheme. Also, let $\delta = \inf_{\psi\in\Psi} \rho(\nu_{h_{\text{true}},y}, q_\psi)$. Unless $\nu_{h_{\text{true}},y}$ corresponds to a product measure, we necessarily have $\delta > 0$. Variational inference is very useful because it is fast and can handle very large datasets (and stochastic variational inference (Hoffman et al., 2013) can scale variational inference to massive data). On the other hand, even if at each outer iteration of VEM the inner iterative scheme was run long enough for convergence to take place, we would still have $\rho(\nu_{h_{\text{true}},y}, q_{\psi_k}) \geq \delta > 0$, and therefore, there is no reason to think that the sequence $h_1, h_2, \ldots$ will converge to h_{true}. George and Doss (2018) have documented cases (in the latent Dirichlet allocation model) where the sequence produced by VEM converges, but to a value that is far from h_{true}, and that the predictive accuracy of the model that results from VEM is worse than that of the scheme that uses a combination of importance sampling and serial tempering MCMC.

Importance Sampling Through Serial Tempering MCMC As discussed in Section 1, to construct the "naive" estimate of $m(h)$ (up to a constant), we select a point $h_1 \in \mathcal{H}$, generate an ergodic Markov chain $\theta_1, \theta_2, \ldots$ with invariant distribution $\nu_{h_1,y}$, and form the estimate on the left of (1.2). The estimate has high variance if ν_h and ν_{h_1} are nearly mutually singular. In serial tempering MCMC (Marinari and Parisi, 1992; Geyer and Thompson, 1995), we select hyperparameter points $h_1, \ldots, h_J \in \mathcal{H}$ that "cover" $\mathcal{H}$ well, and the goal is to generate a Markov chain whose invariant distribution is a mixture of the $\nu_{h_j,y}$'s. The updates will sample different components of this mixture, with jumps from one component to another. Define $\mathcal{L} = \{1, \ldots, J\}$; the elements of $\mathcal{L}$ will be called "labels." We will sometimes write ν_j instead of ν_{h_j}. This is a slight abuse of notation, but we use do it in order to avoid having double and triple subscripts.

The serial tempering chain is really a data augmentation chain that runs on the product space

$\mathcal{L} \times \Theta$, and we now describe it. Let $\Gamma(j, \cdot)$ be a Markov transition function on $\mathcal{L}$. We typically take $\Gamma(j, \cdot)$ to be the uniform distribution on $\mathcal{N}_j$, where $\mathcal{N}_j$ is a set consisting of the indices of the h_l's which are close to h_j. For each $j \in \mathcal{L}$, let Φ_j be a Markov transition function on Θ with invariant distribution $\nu_{h_j,y}$. Also, let $\zeta_1, \ldots, \zeta_J > 0$. These are tuning parameters which we discuss shortly. The serial tempering chain can be viewed as a two-block Metropolis-Hastings (i.e. Metropolis-within-Gibbs) algorithm, and is run as follows. Suppose that the current state of the chain is (L_{i-1}, θ_{i-1}).

- A new value $j \sim \Gamma(L_{i-1}, \cdot)$ is proposed. We set $L_i = j$ with the Metropolis probability

$$\rho_\zeta = \min\left\{1, \frac{\Gamma(j, L_{i-1})}{\Gamma(L_{i-1}, j)} \frac{\nu_j(\theta_{i-1})/\zeta_j}{\nu_{L_{i-1}}(\theta_{i-1})/\zeta_{L_{i-1}}}\right\}, \tag{2.22}$$

  and with the remaining probability we set $L_i = L_{i-1}$.

- Generate $\theta_i \sim \Phi_{L_i}(\theta_{i-1}, \cdot)$.

By standard arguments, the density p_ζ given by $p_\zeta(j, \theta) \propto \ell_y(\theta)\nu_j(\theta)/\zeta_j$ is an invariant density for the serial tempering chain. The θ-marginal of p_ζ is

$$f_\zeta(\theta) = (1/c_\zeta) \sum_{j=1}^{J} \ell_y(\theta)\nu_j(\theta)/\zeta_j, \qquad \text{where} \qquad c_\zeta = \sum_{j=1}^{J} m(h_j)/\zeta_j.$$

Suppose that $(L_1, \theta_1), (L_2, \theta_2), \ldots$ is a serial tempering chain, as described above, and suppose that we have established that it is ergodic. To estimate $m(h)$ up to a multiplicative constant, consider

$$\widehat{M}_\zeta(h) = \frac{1}{n} \sum_{i=1}^{n} \frac{\nu_h(\theta_i)}{(1/J)\sum_{j=1}^{J} \nu_j(\theta_i)/\zeta_j}.$$

For any $h \in \mathcal{H}$, we have

$$\widehat{M}_\zeta(h) \xrightarrow{\text{a.s.}} \int \frac{\nu_h(\theta)}{(1/J)\sum_{j=1}^{J} \nu_j(\theta)/\zeta_j} \frac{\sum_{j=1}^{J} \ell_y(\theta)\nu_j(\theta)/\zeta_j}{c_\zeta}\, d\theta = \int \frac{\ell_y(\theta)\nu_h(\theta)}{c_\zeta/J}\, d\theta = \frac{m(h)}{c_\zeta/J} \tag{2.23}$$

This means that for any vector ζ, the family $\left\{\widehat{M}_\zeta(h),\, h \in \mathcal{H}\right\}$ can be used to estimate the family $\{m(h),\, h \in \mathcal{H}\}$, up to a single multiplicative constant.

We now discuss the choice of ζ. Ideally, we would take $\zeta_j = cm(h_j)$, where c is a constant, for then f_ζ would be the desired mixture $f_\zeta(\theta) = (1/J)\sum_{j=1}^{J} \nu_{h_j,y}(\theta)$. But, of course, the $m(h_j)$'s are unknown. The convergence statement (2.23) enables us to use an iterative scheme for selecting ζ: if $\zeta^{(t)}$ is the current value of ζ, for each $j = 1, \ldots, J$, $\widehat{M}_{\zeta^{(t)}}(h_j) \xrightarrow{\text{a.s.}} (J/c_{\zeta^{(t)}})m(h_j) =: \zeta_j^{(t+1)}$, and $(\zeta_1^{(t+1)}, \ldots \zeta_J^{(t+1)})$ is close to a multiplicative constant times $(m(h_1), \ldots, m(h_j))$, as desired. Details on the iterative scheme, including how to assess its convergence are given in the review paper Geyer (2011).

Serial tempering is a member of a wider family of MCMC schemes called simulated tempering. These, as well as the closely related "umbrella sampling," all involve selecting J grid points

$h_1, \ldots, h_J \in \mathcal{H}$, and considering the priors $\nu_{h_1}, \ldots, \nu_{h_J}$ and the corresponding posteriors. All these schemes are reviewed in Geyer (2011). When compared to the simple importance sampling method discussed in Section 1 (see (1.2)), importance sampling via serial tempering and its relatives enable us to reliably estimate the marginal likelihood (up to a constant) for a much greater range of hyperparameter values. These methods have been used successfully in Geyer and Thompson (1995), and also in Buta and Doss (2011), Roy (2014), and Roy et al. (2018), among many others papers. Nevertheless, importance sampling via serial tempering suffers from two significant deficiencies. One is caused by the dimension of h: for the method to work properly, the number of hyperparameter points J must grow exponentially with $\dim(h)$.

A second difficulty is caused by the fact that, in some situations, the number of latent variables grows with the data sample size. A problem can then occur even if $\dim(h) = 1$. If the data sample size is large, ν_h and ν_{h_1} are distributions on a high-dimensional parameter, and they can be nearly singular with respect to each other even if h and h_1 are close. (As a simple analogy, consider the case of the $\mathcal{N}(h, 1)$ family: if h is close to h_1, then $\mathcal{N}(h, 1)$ is close to $\mathcal{N}(h_1, 1)$, but for large m, the m-fold product of $\mathcal{N}(h, 1)$ is not close to the m-fold product of $\mathcal{N}(h_1, 1)$.) An acute version of this situation was encountered by George and Doss (2018) in their work on the latent Dirichlet allocation model, where the dimension of the parameter is often in the millions. In the robit regression illustration of Section S-1 of Doss and Linero (2021), we provide a different kind of example of a situation where h and h_1 are close, but ν_h and ν_{h_1} are not. When ν_{h_i} and ν_{h_j} are nearly singular with respect to each other even when h_i and h_j are close, serial tempering does not work. This is because there are no good values for J, the number of grid points. If J is small, the distributions ν_{h_i} and ν_{h_j} are far apart, and in the serial tempering chain, the proposal $j \sim \Gamma(L_{i-1}, \cdot)$ is almost always rejected (see (2.22)), so the chain does not mix well. On the other hand, if J is taken to be large enough so that the distributions ν_{h_i} and ν_{h_j} are close enough that the label variable has a reasonable chance of changing, it becomes difficult to traverse the entire label space: to go from any given label to a label that is distant, it is necessary to go through many intermediate labels, again causing the chain to not mix well. In the robit regression illustration, the fully-Bayes empirical Bayes method works very well, but none of the other methods do.

We have discussed the shortcomings of serial tempering, but the two issues we have mentioned (namely the near mutual singularity of ν_{h_i} and ν_{h_j} for $i \neq j$, and the requirement that J grows exponentially with $\dim(h)$) apply to other forms of simulated tempering, as well as to umbrella sampling.

3 Illustrations

We provide three illustrations of our methodology, with three different purposes. The first involves many hyperparameters, and we show that because our fully-Bayes empirical Bayes (FBEB) approach can handle a high-dimensional hyperparameter, it actually gives rise to new statistical methodology, namely a new way to do variable selection in nonparametric regression. The second illustration deals with the Dirichlet additive trees model mentioned in Section 1. Although it in-

volves only a single hyperparameter, none of the existing methods (see Section 2.4) can be used to estimate it, because of the inherent complexity of this model. In the third illustration, which is in Section S-1 of Doss and Linero (2021), we use our FBEB methodology to select the likelihood function, as opposed to a hyperparameter of the prior, and this shows that in some cases our methodology can be used, effectively, to do model selection. Before proceeding, we mention two points regarding computational considerations.

Computation of the Argmax When $\dim(h)$ is 1 or 2, we can simply evaluate our estimate of $\pi_{h\,|\,y}(\cdot)$ over a fine grid and find the maximizer via a grid search; and we can even plot the estimate and inspect it visually. When $\dim(h) > 2$, suppose first that Rao-Blackwellization is possible, in which case our estimate of $\pi_{h\,|\,y}(\cdot)$ is given by (2.1) (we're temporarily assuming that the prior on h is the uniform). Thus, we seek $\arg\max_h (1/n)\sum_{i=1}^n \pi_{h\,|\,(\theta=\theta_i,y)}(h)$. Now, if the function $\pi_{h\,|\,(\theta,y)}(\cdot)$ is available in closed form, then so is its derivative, and therefore the derivative of $\hat{\pi}_{h\,|\,y}(\cdot)$ in (2.1) is available in closed form. This means that all gradient-based optimization methods are available to us. (If the prior on h is not the uniform, we make the obvious adjustment.) If Rao-Blackwellization is not feasible, then we need to use Chen's (1994) method, but the comments above still apply.

Construction of a Markov Chain on (θ, h) In many situations, for the model in which h is a fixed constant, there will already exist a Markov transition function $\Phi_h(\cdot,\cdot)$ on Θ-space with $\pi_{\theta\,|\,(h,y)}$ as invariant density. In this case, we may be able to use Hamiltonian Monte Carlo (HMC, see, e.g. Neal 2011 for a review) to construct a Markov transition function $\Psi_\theta(\cdot,\cdot)$ on $\mathcal{H}$-space with $\pi_{h\,|\,(\theta,y)}$ as invariant density. The only requirement that we need in order to implement HMC is that we know $\nabla_h \log(\pi_{h\,|\,(\theta,y)}(h))$. Now typically, $\pi_{h\,|\,(\theta,y)}$ is available in closed form, except for a normalizing constant that may involve θ and y, but does not involve h. So, except for pathological models in which there are non-differentiability issues, $\nabla_h \log(\pi_{h\,|\,(\theta,y)}(\cdot))$ exists and is available in closed form. The chain that alternates between Φ_h draws and Ψ_θ draws then has $\pi_{(\theta,h)\,|\,y}$ as invariant density.

3.1 Variable Selection in Bayesian Nonparametric Additive Regression

Bayesian approaches to model selection provide a natural way of simultaneously treating both model and parameter uncertainty. In a linear regression situation, we have a response variable Y and a set of predictors $X_1, \ldots, X_p$, each a vector of length m. For $\gamma \subseteq \{1, \ldots, p\}$ we have a potential model 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 $\epsilon \sim \mathcal{N}_m(0, \sigma^2 I)$. We view γ as a binary vector of length p, whose j^{th} component, γ_j, is 1 if variable j is in the model, and 0 otherwise. 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.

In a common formulation of the Bayesian approach, the prior on θ is given by a hierarchy in which we first choose the indicator γ from some distribution, a "non-informative prior" is given to (σ^2, β_0), and given γ and σ, we choose β_γ from some proper distribution. The distribution for

γ is the so-called independence Bernoulli prior—each variable goes into the model with a certain probability w, independently of all the other variables—and the distribution for β_γ is taken to be Zellner's g-prior (see (3.1b) below). Let $p_\gamma = \sum_{j=1}^{p} \gamma_j$ denote the number of variables in the model indexed by γ. In detail, this common formulation is described as follows:

$$Y \sim \mathcal{N}_m(1_m\beta_0 + X_\gamma\beta_\gamma, \sigma^2 I), \tag{3.1a}$$

$$\beta_\gamma \sim \mathcal{N}_{p_\gamma}\big(0, g\sigma^2(X_\gamma^\top X_\gamma)^{-1}\big), \tag{3.1b}$$

$$(\sigma^2, \beta_0) \sim p(\beta_0, \sigma^2) \propto 1/\sigma^2, \tag{3.1c}$$

$$\gamma \sim w^{p_\gamma}(1-w)^{p-p_\gamma}. \tag{3.1d}$$

In (3.1), each line is understood to be a distributional statement conditional on all the variables specified in the lines below it. Zellner's g prior is indexed by a hyperparameter g, which plays an important role in variable selection: generally speaking, when g is large the prior is concentrated on models with few variables and large regression coefficients, and when g is small the prior is concentrated on large models with small coefficients. The hyperparameter w in (3.1d) has the opposite effect, with small w favoring models with only a few variables, and large w favoring models with many variables. The interplay between w and g is not well understood. The prior is improper because the prior on (σ^2, β_0) is improper (see (3.1c)); however, the posterior distribution of θ is proper.

The posterior distribution of θ given Y may be estimated by MCMC schemes which run on the variable $\gamma = (\gamma_1, \ldots, \gamma_p)$, with $(\beta_0, \beta_\gamma, \sigma)$ integrated out. Because the state space for γ is finite, the Markov chains are uniformly ergodic. The original papers which develop such schemes are Madigan and York (1995), Smith and Kohn (1996), Clyde et al. (1996), and Raftery et al. (1997), and there have been many enhancements since. Model (3.1) was considered by Liang et al. (2008), who showed that g-priors with a fixed g give rise to posteriors with paradoxical (and highly undesirable) properties. They propose to use mixtures of g-priors; specifically, they advocate "hyper-g" priors (which we discuss shortly), and show that if we use them, the paradoxes do not arise.

Smith and Kohn (1996) considered the case where some variables need to be treated nonlinearly, and so considered the additive model $Y_l = \beta_0 + \sum_{j=1}^{p} f_j(x_{lj}) + \epsilon_l$, $l = 1, \ldots, m$, in which the f_j's are represented by regression splines: $f_j(x) = \sum_{k=1}^{K} \beta_{jk} B_{jk}(x)$, where $B_{j1}, \ldots, B_{jK}$, are cubic regression splines with evenly-spaced knots. The model may then be expressed as

$$Y = 1_m\beta_0 + \sum_{j=1}^{p} f_j + \epsilon, \quad \text{where} \quad f_j = B_j\beta_j, \tag{3.2}$$

where B_j is an $m \times K$ matrix with $(\ell, k)^{\text{th}}$ entry $B_{jk}(x_{j\ell})$ and $\beta_j = (\beta_{j1}, \ldots, \beta_{jK})^\top$. We can also write $Y = 1_m\beta_0 + B\beta + \epsilon$, where $B = [B_1, \ldots, B_p]$, and $\beta = (\beta_1^\top, \ldots, \beta_p^\top)^\top$, this last representation fitting in with the usual formulation of a linear model.

When estimating the parameters of the spline model (3.2), one option is to use a relatively small number of knots, positioned at evenly spaced points along the predictor axes, and to choose

the number of knots via cross-validation (or some model selection criterion). A second option is to use a large number of knots but then apply ℓ_1 penalization on the regression coefficients. This has the advantage that it captures local curvature better. Using ℓ_1 penalization sets some of the knot coefficients to 0 and therefore leads to knot selection. The approach of Smith and Kohn (1996) is, effectively, a Bayesian version of the second option. If in model (3.1) we replace (3.1b) with the statement that β_γ is distributed according to a hyper-g prior, then the model is indexed by the hyperparameter w. Smith and Kohn (1996) use a fixed value for g ($g = 100$) and specify $w = 1/2$.

We will take the approach of Smith and Kohn (1996), modified so that for each $j = 1, \ldots, p$, there is a separate inclusion probability w_j for the knots corresponding to variable j, and a hyper-g prior is used for g. The hyper-g prior is given by $\mu(g) \propto (1+g)^{-a/2}$ for $g > 0$, and is indexed by the parameter $a > 2$. Following Liang et al. (2008), we take $a = 3$. The reason for having separate inclusion probabilities for the p variables is that this allows the different f_j's to a-priori have different curvatures but, as we will see, allowing separate inclusion probabilities has interesting ramifications for variable selection. Let $w = (w_1, \ldots, w_p)$. Also, let $\gamma_{j1}, \ldots, \gamma_{jK}$ be the knot-inclusion indicators for variable j, define $\gamma_{j\cdot} = \sum_{k=1}^{K} \gamma_{jk}$, and denote $\gamma_{[j]} = (\gamma_{j1}, \ldots, \gamma_{jK})$ and $\gamma = (\gamma_{[1]}, \ldots, \gamma_{[p]})$. Lines (3.1d) and (3.1b) of Model (3.1) now need to be changed to

$$\gamma_{[j]} \overset{\text{indep}}{\sim} w_j^{\gamma_{j\cdot}}(1-w_j)^{K-\gamma_{j\cdot}}, \quad j = 1, \ldots, p,$$

$$g \sim \mu, \qquad \text{and} \qquad \text{given } g, \;\; \beta_{\gamma_{[j]}} \overset{\text{indep}}{\sim} \mathcal{N}_{p_{\gamma_{[j]}}}\big(0, g\sigma^2(B_{j,\gamma_{[j]}}^\top B_{j,\gamma_{[j]}})^{-1}\big), \quad j = 1, \ldots, p,$$

respectively. We then estimate the hyperparameter w using the methods of this paper.

None of the procedures described in Section 2.4 works here: EM-based approaches give estimates which converge, but to incorrect values, and serial tempering MCMC gives estimates which are extremely unstable, as discussed earlier. We now describe how our FBEB approach may be implemented. We take $w_j \overset{\text{iid}}{\sim} \text{Unif}(0,1)$. Our Markov chain will run over (γ, w, g), with $(\beta_0, \beta_\gamma, \sigma)$ integrated out. It will be driven by a Markov transition function (MTF) which consists of the composition of three MTF's, of which the first updates γ, the second updates w, and the third updates g. Our first MTF modifies the proposal of Yang et al. (2016) to take into account that $\gamma_{[1]}, \ldots, \gamma_{[p]}$ constitute p groups; it applies one of the two changes below, each with probability $1/2$. (1) Flip a randomly selected γ_{jk}. (2) For each $j = 1, \ldots, p$, randomly select two indicators from the set $\{\gamma_{j1}, \ldots, \gamma_{jK}\}$, one of which is a 1 and the other a 0, and swap their values (if $\gamma_{j1}, \ldots, \gamma_{jK}$ are all 1 or all 0, then we do nothing). In either case, the move is accepted or rejected based on the Metropolis-Hastings acceptance probability. This update leaves the conditional distribution of γ given (w, g, Y) invariant.

The second MTF generates w according to its conditional distribution given (γ, g, Y). From the hierarchical nature of the model, it is easy to see that

$$\pi(w \,|\, \gamma, g, Y) = \pi(w \,|\, \gamma) \propto \prod_{j=1}^{p} \prod_{k=1}^{K} w_j^{\gamma_{jk}}(1-w_j)^{1-\gamma_{jk}} \propto \prod_{j=1}^{p} \text{beta}(w_j; 1+\gamma_{j\cdot}, 1+K-\gamma_{j\cdot}),$$

where $\text{beta}(w; a, b)$ denotes the $\text{beta}(a, b)$ density evaluated at w. This update leaves the conditional distribution of w given γ, g, and Y invariant. The equation above gives rise to the Rao-

Blackwellized estimate

$$M_n(w) = \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{p}\text{beta}(w_j; 1+\gamma_{j\bullet}^{(i)}, 1+K-\gamma_{j\bullet}^{(i)}), \tag{3.3}$$

of the marginal likelihood of w (up to a constant), where $i = 1, \ldots, n$ indexes the iterations of the Markov chain. In our illustrations, the empirical Bayes estimate of w is obtained by maximizing (3.3) through the `optim` function in R using the `L-BFGS-B` algorithm (Byrd et al., 1995).

The final MTF updates g, and this is done through the slice sampler (Neal, 2011). We remark that Liang et al. (2008) show that we can actually modify the first two MTF's by marginalizing out g, thereby eliminating the need to construct an update for g; we include an update for g to allow for the possibility of including g in our empirical Bayes analysis, if this is desired. The composition of the three updates leaves the conditional distribution of (γ, w, g) given Y invariant.

It is notable that some of the components of $\arg\max_w m(w)$ can be zero (and if $\gamma_{j\bullet}^{(i)} = 0$ for the n iterations of the Markov chain, then the j^{th} component of the maximizer of $M_n(w)$ in (3.3) is zero, i.e. w_j is estimated to be zero.) In this case, it is not just some knots that are excluded; rather variable j in its entirety is eliminated from the model. As is the case for many likelihood-based methods, some issues arise when the maximizing value is at the boundary of the parameter space, and this was noted in the context of Bayesian variable selection by Scott and Berger (2010). (The situation here is similar to that where we have $Y \sim \text{binomial}(m, p)$: if we observe $Y = 0$, then not only is the maximum likelihood estimate of p equal to 0, but the associated standard error estimate is also 0, and the naive Wald-type confidence interval for p is the singleton $\{0\}$.) As is the case for the simple binomial example, if uncertainty quantification regarding variable selection is required, a fully-Bayes approach can be used.

We apply our empirical Bayes approach to the ragweed dataset of Stark et al. (1997). This dataset consists of measurements of the ragweed pollen for 335 days in Kalamazoo, Michigan, along with several meteorological predictors: the day number of the ragweed pollen season, the temperature, the wind speed, and whether it rained. The first three predictors are numeric, and the fourth is binary, and the effect of the three numeric predictors is modeled using cubic regression splines with at most $K = 50$ knots for each predictor. The binary predictor contributes an additive effect $\gamma_j\beta_j$ where $\gamma_j \sim \text{Bernoulli}(w_j)$.

Before proceeding, we check that the empirical Bayes procedure gives reasonable results by comparing the fit it gives to an additive fit using the `gam` function in the R package `mgcv`. This experiment and all the others below are based on a Markov chain of length $55{,}000$, with the first $5{,}000$ cycles discarded as burn-in and the remaining cycles thinned by 10, giving $5{,}000$ cycles. The results are displayed in Figure 1. We see that for the day in season and temperature variables the fit from `gam` is quite similar to the one using our empirical Bayes approach, although the fit for the wind speed variable is somewhat different due to the preference of `gam` to favor linear terms before including non-linear terms (our approach can be easily extended to also favor linear terms).

To evaluate the variable selection performance of our model, we added noise variables as follows. Let D_l, T_l, W_l, and R_l be the day, temperature, wind, and rain variables for observation l ($l = 1, \ldots, 335$), and let π be a random permutation of the integers $1, \ldots, 335$. With a single per-

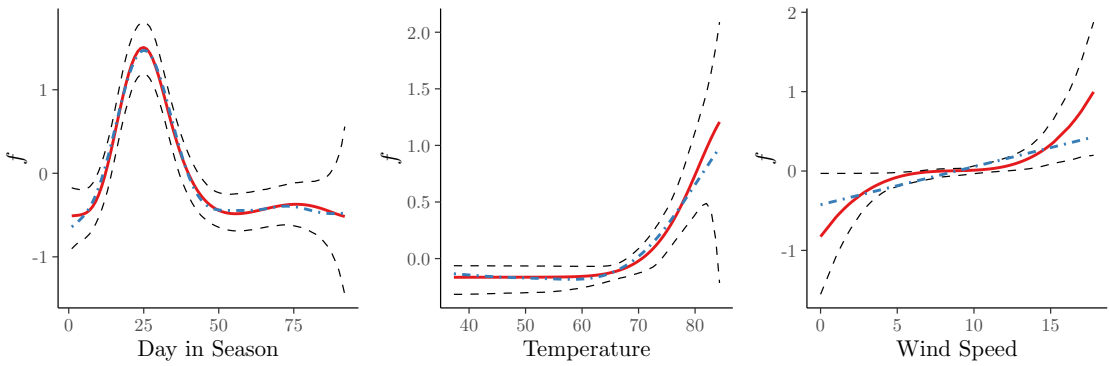

Figure 1: Estimates of the f_j's for the ragweed data. Solid line (—) gives the posterior mean, dashed lines (– – –) give 95% pointwise credible intervals, and dot-dashed line (– · –) gives the estimate obtained from the `gam` function in the R package `mgcv`.

mutation, the augmented data set is $(Y_l, D_l, T_l, W_l, D_{\pi(l)}, T_{\pi(l)}, W_{\pi(l)}, R_l)$, $l = 1, \ldots, 335$. We then modeled each of the continuous predictors using our spline basis function expansion. The reason for creating noise variables in this way is that the correlation structure for the added variables is identical to that in the original variables.

Figure 2 shows the empirical Bayes estimate of w obtained from our procedure when we used two permutations to generate noise variables. We see that our approach correctly removes all the noise variables and includes all the original variables. Maximizing (3.3) numerically confirms that the empirical Bayes estimators of the w's corresponding to the noise variables are all 0. To get more detail, we would like to plot the marginal likelihood function $m(\cdot)$, but it is not possible to do so because $m(\cdot)$ is a function of 10 variables. Instead, for each $j = 1, \ldots, 10$, we can take $(w_1, \ldots, w_{j-1}, w_{j+1}, \ldots, w_{10}) \overset{\text{iid}}{\sim} \text{Unif}(0, 1)$ and plot the marginal likelihood function for w_j, with all the other w's integrated out. Figure 3 presents the plots for variables $1, \ldots, 9$. Here, w_1, w_2, and w_3 are associated with the original continuous variables, and $w_4, \ldots, w_9$ are associated with the noise variables. The figure additionally gives pointwise and uniform confidence bands for these marginal likelihoods. This marginal analysis also suggests that these predictors do not appear in the model selected by the empirical Bayes method.

We now make a brief comparison of our empirical Bayes approach to penalized regression using the commonly-used composite minimax concave penalty (cMCP) method proposed by Breheny and Huang (2009). The cMCP method carries out bi-level selection by performing both group-level (predictor) and within-group level (basis function) selection, which is similar to what our method is doing. We fit the cMCP regression using the `grpreg` package in R and selected tuning parameters using cross-validation. In addition to the dataset augmented with two permutations, we considered a second dataset which used six permutations. The two methods were evaluated using the number of true positives (regarding the four original predictors as positives) and the number of false positives. Table 1 gives the results. From the table, we see that for both datasets, each method correctly identified all the original predictors; however, we also see that cMCP is prone to false positives when we increase the number of noise variables. Of course, there

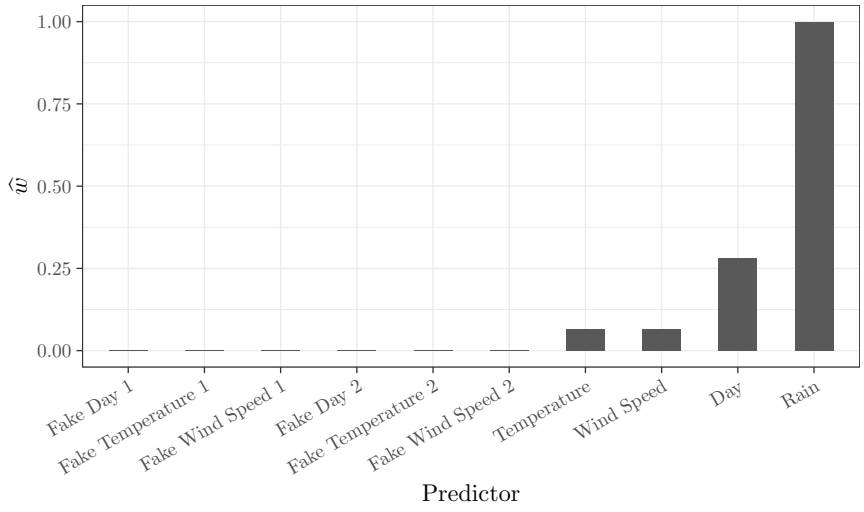

Figure 2: Empirical Bayes estimates of the w_j's for the ragweed data augmented with noise predictor variables.

are several other procedures for doing bi-level variable selection, e.g. group bridge (Huang et al., 2009) and `spikeSlabGAM` (Scheipl et al., 2012). A thorough comparison of our empirical Bayes approach with all other methods is beyond the scope of this paper. Our goal in this first illustration is only to demonstrate the potential of the FBEB method for enabling the implementation of empirical Bayes approaches in models where these would be useful but no existing procedure for estimating $\arg\max_h m(h)$ works. (Note that there is a distinction between FBEB, which is a Monte Carlo method for obtaining an estimate of $\arg\max_h m(h)$, and the resulting empirical Bayes scheme, which is a statistical procedure.)

3.2 Choosing the Sparsity Parameter in the Dirichlet Additive Regression Trees Model

Consider the nonparametric regression model $Y_l = f(x_l) + \epsilon_l$, $l = 1, \ldots, m$, where $\epsilon_l \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ and $x_l \in \mathbb{R}^p$. An increasingly popular strategy for estimating f is to take a Bayesian approach in which f is modeled as a sum of random decision trees (see Chipman et al. 2013 and Linero 2017 for reviews). The most popular such approach is to use a Bayesian additive regression trees (BART, Chipman et al. 2010) model, which sets $f(x) = \sum_{t=1}^{T} g(x; \mathcal{T}_t, \mathcal{M}_t)$, where the $\mathcal{T}_t$'s are regression trees and the $\mathcal{M}_t$'s are the corresponding vectors of terminal node parameters. Here, $g(x; \mathcal{T}_t, \mathcal{M}_t) = \mu_{t\ell}$ if x goes to terminal node ℓ of tree t.

In the generative tree-construction process, there is a variable, $s = (s_1, \ldots, s_p)$, where s_j is the probability that the variable chosen for a split is variable j. Chipman et al. (2010) take s to be deterministic: $s = (p^{-1}, \ldots, p^{-1})$. Linero (2018) argues that when dealing with regression that is potentially sparse, it is better to allow s to be random. He specifies that s is drawn from the Dirichlet distribution $\text{Dir}_p(\alpha/p, \ldots, \alpha/p)$, in which α is a hyperparameter; we refer to this model

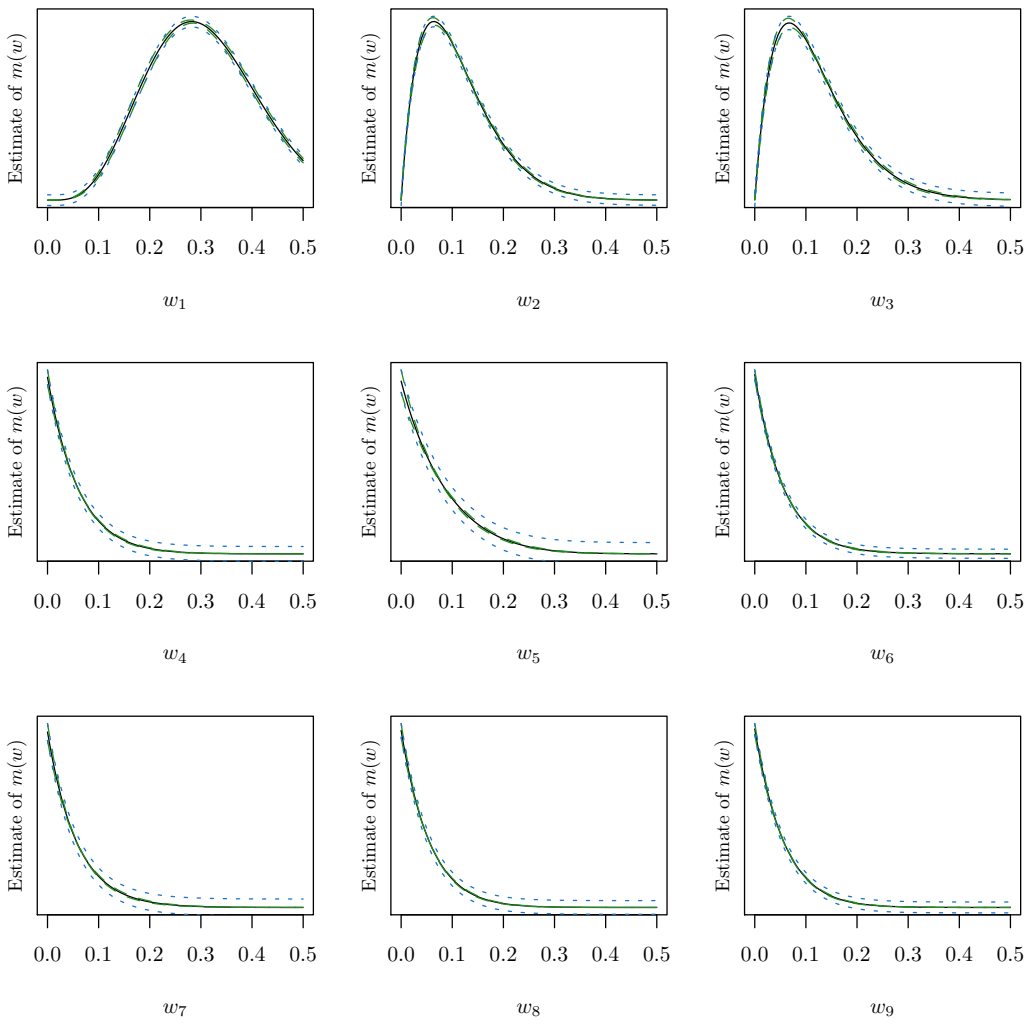

Figure 3: Estimates of the marginal likelihood of w_j when treating all other w's as $\mathrm{Unif}(0, 1)$, for the synthetic ragweed dataset. Estimated marginal likelihood is given by solid lines (—), uniform 95% confidence bands are given by short-dashed lines (- - -), and pointwise 95% confidence bands are given by long-dashed lines (– – –).

as "Dirichlet additive regression trees" (DART) for the purpose of comparison with BART. The hyperparameter α plays a key role related to sparsity. Suppose $U \sim \mathrm{Dir}_p(a, \ldots, a)$. As is well known, if a is small, there is a tendency for most of the components of U to be near zero, and in the limit where $a \to 0$, one component is 1, the rest are zero, and the position of the nonzero component is uniformly distributed on $\{1, \ldots, p\}$. On the other hand, when a increases to infinity, U is nearly the vector $(p^{-1}, \ldots, p^{-1})$. In the generative tree-construction process, the variable s is chosen once, and then is applied to all T trees. As a consequence, when α is small, only a few of the predictor variables are chosen for splits in the entire ensemble; and when α is large, many predictors are involved, and DART reverts to BART. (See Linero 2018 and the supplement to Linero 2018 for a precise version of this statement.) This technique has also been used by Linero

Replications = 2		
Method	TP	FP
cMCP	4	1
EB	4	0

Replications = 6		
Method	TP	FP
cMCP	4	6
EB	4	1

Table 1: Comparison of performance of the empirical Bayes and cMCP methods, on the augmented ragweed data. TP denotes the number of true positives and FP denotes the number of false positives. Replications denotes the number of permutations used to construct the noise variables.

and Yang (2018).

In Linero (2018), α is given a prior of the form $\rho_\eta(\alpha) = \eta \cdot [2\alpha^{1/2}(\alpha+\eta)^{3/2}]^{-1}$, where the scale parameter η is taken to be p by default. This prior is unbounded near 0, which enables sparsity, and it has a Cauchy-like right tail, which allows DART to revert to BART. Unfortunately, these two features of the prior can result in poor mixing of the Markov chain used to estimate the posterior distribution. An alternative to putting a prior on α is to select α by maximum marginal likelihood, i.e. via the empirical Bayes method, for which the methodology of the present paper is designed. In short, the empirical Bayes method will enable us to determine whether to use BART ($\arg\max_\alpha m(\alpha)$ is large) or DART ($\arg\max_\alpha m(\alpha)$ is not large). In the rest of this section we will explain how our approach may be implemented (we will present our MCMC algorithm and discuss how the IWMDE method of Chen (1994) may be carried out), and then illustrate, on real and artificial data, how DART implemented through the empirical Bayes approach acts appropriately in both sparse and non-sparse situations. (The only hyperparameter we will allow to vary is α, all the others being set to the default values recommended by Chipman et al. 2010.)

Our methodology provides an asymptotically exact estimate of $\arg\max_\alpha m(\alpha)$ regardless of our choice of prior on α, so we will take $\pi_\alpha = \text{gam}(a_0, b_0)$. This choice enables a very convenient data augmentation scheme that takes advantage of some conjugacy in the problem. Here, $\text{gam}(x; a, b)$ is the density proportional to $x^{a-1}\exp(-xb)$. We introduce the augmentation variable $\lambda \sim \text{gam}(\alpha, 1)$, where λ is independent of s, and define $Z = \lambda s$. Then $Z_j \overset{\text{iid}}{\sim} \text{gam}(\alpha/p, 1)$, $j = 1, \ldots, p$ follows from routine distribution theory (see, e.g., Devroye, 1986, Chapter 11, Theorem 4.1).

Denote $Y = (Y_1, \ldots, Y_m)$, $y = (y_1, \ldots, y_m)$, and let $\theta = (s, \sigma^2, \lambda, \{\mathcal{T}_t, \mathcal{M}_t, t = 1, \ldots, T\})$. We will generate a Markov chain on (θ, α) whose invariant distribution is the distribution of (θ, α) given $Y = y$. The MTF for doing so will consist of the composition of four MTF's. The first updates $(\sigma^2, \{\mathcal{T}_t, \mathcal{M}_t, t = 1, \ldots, T\})$ using the Bayesian backfitting algorithm of Chipman et al. (2010). The second MTF is the update from Linero (2018), which updates s. The third MTF draws λ from its conditional distribution given all the other parameters and $Y = y$, which is $\text{gam}(\alpha, 1)$. The combination of these MTF's leaves the full-conditional distribution of θ invariant.

The final MTF updates α. The conditional distribution of α given θ and $Y = y$ is given by $\pi_{\alpha\,|\,(\theta,y)}(\alpha) \propto \Gamma(\alpha/p)^{-p}\exp\left\{(\alpha/p)\sum_{j=1}^p \log(Z_j)\right\}\alpha^{a_0-1}\exp(-b_0\alpha)$. This depends only on α and the Z_j's because the only factors of the joint distribution of (θ, α, Y) in which α appears are $\pi_\alpha, \pi_{s\,|\,\alpha}$, and $\pi_{\lambda\,|\,\alpha}$. It is easy to see that this distribution is also equal to the posterior distribution

of α in the model where $\alpha \sim \text{gam}(a_0, b_0)$ and $Z_j \overset{\text{iid}}{\sim} \text{gam}(\alpha/p, 1)$ for $j = 1, \ldots, p$. Unfortunately, there is no analytic expression for the posterior distribution. Miller (2019) considered a setup which includes precisely this situation. He showed that the posterior distribution of α is very well approximated by $\text{gam}(A(Z), B(Z))$, if $A(Z)$ and $B(Z)$ are chosen appropriately. Ideally, the parameters $A(Z)$ and $B(Z)$ are chosen by matching the first and second derivatives of $\log(p_{\alpha \mid Z})$ to those of $\log(\text{gam}(A(Z), B(Z)))$ at the mean of $\text{gam}(A(Z), B(Z))$. Unfortunately, the mean of $\text{gam}(A(Z), B(Z))$ is not known, because $A(Z)$ and $B(Z)$ are not known. So a trial value for the mean is initially used, leading to an improved estimate, and this scheme is iterated. Miller (2019) states that convergence occurs within four iterations in all the situations he has seen, and that the final approximation is excellent.

Our update of α will be a Metropolis-Hastings step that uses the Miller (2019) approximation as a proposal density (as suggested by Miller 2019). This gives rise to an MTF for which the invariant density is $\pi_{\alpha \mid (\theta, y)}$. Because the approximation is excellent, the acceptance probability is nearly one, and we are essentially sampling from $\pi_{\alpha \mid (\theta, y)}$. For estimating $\pi_{\alpha \mid y}$, Rao-Blackwellization is not feasible, and we must use the IWMDE method of Chen (1994). The Miller (2019) approximation provides a very convenient family $\{w_\theta, \theta \in \Theta\}$ of densities on α: we take $w_\theta = \text{gam}(A(Z), B(Z))$. Because the approximation of $\pi_{\alpha \mid (\theta, y)}$ by $\text{gam}(A(Z), B(Z))$ is very good, the IWMDE method is essentially as good as Rao-Blackwellization.

We now illustrate our methodology on real and simulated data. Our goals are limited: recalling that small values of α encourage sparse models and that as $\alpha \to \infty$ DART reverts to BART, we establish that the empirical Bayes choice of α reflects the sparsity in the regression. A broad comparison of DART, with α chosen via the empirical Bayes method (DART-EB), to other regression methods is beyond our scope. We consider three datasets: (1) the Waste Isolation Pilot Plant (WIPP) data, described in Storlie and Helton (2008), for which there are $m = 300$ observations and $p = 31$ variables; (2) the Triazines data, available from the UCI Machine Learning repository, and for which $n = 186$ and $p = 60$; and (3) the Blood Brain Barrier data (BBB), available and described in the `caret` package in R, and for which $n = 208$ and $p = 134$.

We set $a_0 = 1$, $b_0 = 1/50$. We used our MTF to generate $120{,}000$ samples of θ, and discarded the first $20{,}000$ as burn-in. We then estimated the marginal likelihood (up to a constant) using the IWMDE procedure, and also estimated the posterior density of α, using $\hat{\pi}^{\text{iwmde}}_{h \mid y}$ (see (2.20)). We also formed pointwise and simultaneous confidence bands for both the marginal likelihood function and the posterior density. The pointwise bands were constructed via standard batching methodology, with the number of batches set using the values recommended by Jones et al. (2006). The simultaneous confidence bands were constructed using the procedure described after the statement of Theorem 3. The Markov chain lengths were chosen to make the simultaneous confidence bands acceptably narrow. Plots for all three data sets are given in Figure 4. The results are as follows.

WIPP It is a priori known that the regression for this data set is sparse; for example, Storlie et al. (2011) report that only 8 of the 31 variables are informative. The top-right panel of Figure 4 correctly reflects this sparsity, and indicates that the marginal likelihood of BART vs. DART with $\alpha = 5$ is essentially 0.

Triazines Linero (2018) noted that for this dataset, the performance of DART (with the default prior for α) was about the same as that of BART, under a criterion of predictive error. The top-middle panel of Figure 4 confirms this: the marginal likelihood of BART is higher than that of DART for all α.

BBB Linero (2018) noted that for this data set, BART and DART have similar performance in terms of prediction error, but that DART used fewer variables; hence DART obtained a more parsimonious fit to the data with the same predictive accuracy. The top-left panel of Figure 4 confirms this: the marginal likelihood of DART at its argmax is only slightly higher than the marginal likelihood of BART.

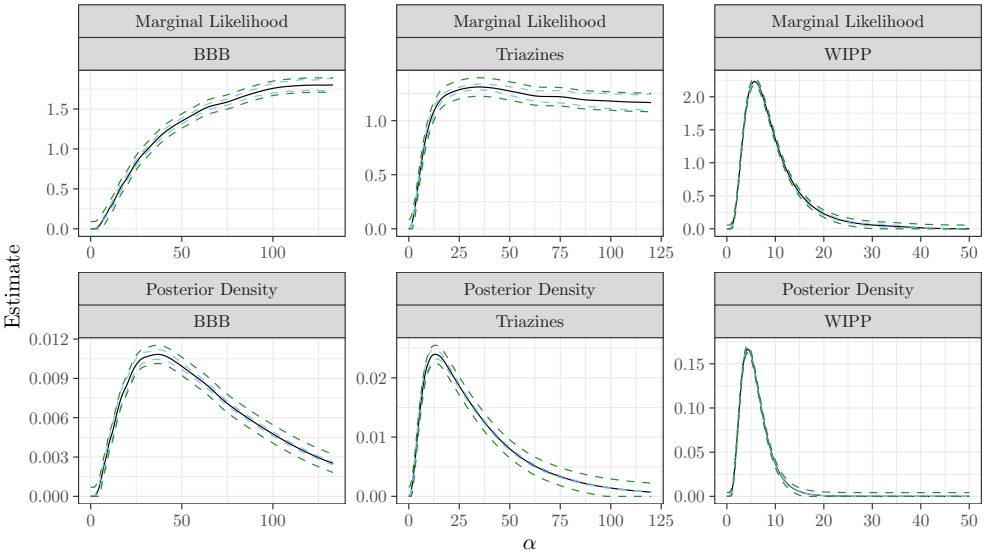

Figure 4: Estimates of the marginal likelihood (top) and posterior distribution (bottom) of α for the DART model. Left: the BBB dataset; middle: the Triazines dataset; right: the WIPP dataset. The estimated marginal likelihood is given by solid lines (—), simultaneous 95% confidence bands are given by long-dashed lines (– – –), and pointwise 95% confidence bands are given by short-dashed lines (- - -).

Next, we show that DART-EB behaves appropriately in a simple simulation study in which the true underlying regression function is $f_0(x) = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5$. This function was introduced by Friedman (1991) and has been used many times in the context of BART (see, e.g., Chipman et al., 2010 and Linero and Yang, 2018). Here, $f_0(x)$ depends on only the first five predictors, $x_6, x_7, \ldots, x_p$ being irrelevant noise variables. We consider 200 independent replicates of the following experiment. First, we sample $X_l \overset{\text{iid}}{\sim} \text{Unif}([0,1]^p)$, $l = 1, \ldots, 250$, and draw $Y_l \overset{\text{indep}}{\sim} \mathcal{N}(f_0(X_l), \sigma^2)$. Then, based on (X_l, Y_l), $l = 1, \ldots, 100$, we fit (i) the usual BART model and (ii) DART-EB, using $\text{gam}(1, 1/20)$ as the prior on α. For each replication and each method, we consider the median probability model, which is defined to be the model that includes all variables that occur in at least half of the samples from the posterior distribution (i.e.

	$\sigma = 1, p = 100$		$\sigma = 1, p = 500$		$\sigma = 5, p = 100$		$\sigma = 5, p = 500$	
Method	RMSE	F_1	RMSE	F_1	RMSE	F_1	RMSE	F_1
DART-EB	1.00	1.00	1.00	1.00	1.00	0.903	1.00	0.877
BART	1.75	0.48	2.26	0.76	1.25	0.251	1.29	0.719
Avg($\hat{\alpha}_{\text{opt}}$)	1.31		1.27		4.52		5.06	

Table 2: DART simulation results. RMSE is relative to the RMSE of DART-EB so that DART-EB by definition has an RMSE of 1. Avg($\hat{\alpha}_{\text{opt}}$) is the the estimate of $\arg\max_\alpha m(\alpha)$, averaged over the 200 replications of the simulation study. The Monte Carlo standard error of Avg($\hat{\alpha}_{\text{opt}}$) is less than 0.21 throughout.

the variables for which the marginal inclusion probability is estimated to be at least $1/2$). The integrated root-mean-squared error is given by rmse $= \left\{\int (\hat{f}(x) - f_0(x))^2\, dx\right\}^{1/2}$, where $\hat{f}(x)$ is the Bayes estimate of $f_0(x)$. For each method and replication we approximate this by simple Monte Carlo and we form RMSE, which is rmse averaged over the 200 replicates. Furthermore, for each method and replication we compute the *precision* and *recall*, defined by precision $= \mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$ and recall $= \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$, where TP, FP, and FN denote the number of true positives, false positives, and false negatives in carrying out variable selection. The results are given in Table 2 for $p \in \{100, 500\}$ and $\sigma \in \{1, 5\}$. We report the F_1 score, which is the harmonic mean of the precision and recall, averaged over the replications, as well as RMSE relative to the RMSE of DART-EB. From the table we see that DART-EB significantly outperforms BART. It behaves appropriately as the number of irrelevant predictors increases. Specifically, for both the $\sigma = 1$ and $\sigma = 5$ cases, as p increases, the estimated value of α remains roughly constant, so that in the statement $s \sim \mathrm{Dir}_p(\alpha/p, \ldots, \alpha/p)$, the shape parameter of the Dirichlet decreases roughly in proportion with p, correctly reflecting the increase in sparsity.

We note that importance sampling via serial tempering will not work here because it is difficult to obtain an analytic expression for the ratio of densities $\nu_{\alpha_1}/\nu_{\alpha_2}$ due to the complexity of the parameter θ, and even if such an expression was available, unless α_1 and α_2 are extremely close, ν_{α_1} and ν_{α_2} will be nearly singular because of the high dimension of θ. VEM is currently not an option either, as there are no variational algorithms for BART.

A Recapitulation of the FBEB Method In the FBEB method, the prior on h does not affect the final inference. Any prior leads to an asymptotically exact estimate of $\arg\max_h m(h)$, our empirical Bayes estimate of h. So the prior may be taken to be any convenient choice, for example one that takes advantage of any conjugacy that exists in the model. In order to obtain our empirical Bayes estimate of h, we need to run a Markov chain on the augmented space $\mathcal{H} \times \Theta$, so it is perhaps natural to ask why do we not stop there? That is, why should we use an empirical Bayes approach instead of using a fully-Bayes approach, and what is the value of our methodology? Aside from the fact that our approach enables sensitivity analysis as a by-product, we mention the following. A fully-Bayes approach can certainly be very useful; however, it requires a choice of prior on

h. Proper priors generally involve a subjective choice, which must be justified, since different priors give different conclusions, particularly in small sample situations. (Consider, for example, the Triazines and BBB data sets in Section 3.2—see Figure 4. For these data sets, the marginal likelihood is not very informative, so a proper prior has a large influence on the posterior.) On the other hand, objective priors are usually improper, and these can lead to improper posteriors (for example, if $m(\alpha)$ is bounded away from 0 as $\alpha \to \infty$, as is the case for the Triazines and BBB data sets, then any prior giving infinite mass to the interval $[1, \infty)$ necessarily results in an improper posterior). In this case, insidiously, if we use Gibbs sampling to estimate the posterior, it is possible that all conditionals needed to implement the sampler are proper; but Hobert and Casella (1996) have shown that the Gibbs sampler output may not give a clue that there is a problem.

As mentioned above, the empirical Bayes approach avoids the prior specification issue. In broad terms, the general interest in empirical Bayes methods arises in part from a desire to select a specific value of the hyperparameter vector because this gives a model that is more parsimonious and interpretable. These points are discussed more fully in George and Foster (2000) and Robert (2001, Chapter 7). The question of whether in general one should use empirical Bayes or fully-Bayes methods has been around for decades and is unlikely to be settled soon. Our purpose here is not to advocate an empirical Bayes over a fully-Bayes approach at a philosophical level, but rather to provide a methodology that gives the user the option of using an empirical Bayes approach.

References

Athreya, K. B. and Ney, P. (1978). A new approach to the limit theory of recurrent Markov chains. *Transactions of the American Mathematical Society* **245** 493–501.

Beal, M. and Ghahramani, Z. (2003). The variational Bayesian EM algorithm for incomplete data: With application to scoring graphical model structures. In *Bayesian Statistics 7—Proceedings of the Sixth Valencia International Meeting* (J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, M. West et al., eds.). Oxford University Press.

Blei, D. M., Kucukelbir, A. and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association* **112** 859–877.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research* **3** 993–1022.

Breheny, P. and Huang, J. (2009). Penalized methods for bi-level variable selection. *Statistics and Its Interface* **2** 369–380.

Buta, E. and Doss, H. (2011). Computational approaches for empirical Bayes methods and Bayesian sensitivity analysis. *The Annals of Statistics* **39** 2658–2685.

Byrd, R. H., Lu, P., Nocedal, J. and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing* **16** 1190–1208.

Chen, M.-H. (1994). Importance-weighted marginal Bayesian posterior density estimation. *Journal of the American Statistical Association* **89** 818–824.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* **90** 1313–1321.

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association* **96** 270–281.

Chipman, H., George, E. I., Gramacy, R. B. and McCulloch, R. (2013). Bayesian treed response surface models. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **3** 298–305.

Chipman, H. A., George, E. I. and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics* **4** 266–298.

Clyde, M., DeSimone, H. and Parmigiani, G. (1996). Prediction via orthogonalized model mixing. *Journal of the American Statistical Association* **91** 1197–1208.

Devroye, L. (1986). *Nonuniform Random Variate Generation*. Springer-Verlag, New York.

Donoho, D. L. and Liu, R. C. (1991). Geometrizing rates of convergence, II. *The Annals of Statistics* **19** 633–667.

Doss, H. and Linero, A. (2021). Supplement to "Scalable empirical Bayes inference and Bayesian sensitivity analysis".

Doss, H. and Park, Y. (2018). An MCMC approach to empirical Bayes inference and Bayesian sensitivity analysis via empirical processes. *The Annals of Statistics* **46** 1630–1663.

Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors: An empirical Bayes approach. *Journal of the American Statistical Association* **68** 117–130.

Flegal, J. M., Haran, M. and Jones, G. L. (2008). Markov chain Monte Carlo: Can we trust the third significant figure? *Statistical Science* **23** 250–260.

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.

Fort, G. and Moulines, E. (2003). Convergence of the Monte Carlo expectation maximization for curved exponential families. *The Annals of Statistics* **31** 1220–1259.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics* **19** 1–67.

George, C. P. and Doss, H. (2018). Principled selection of hyperparameters in the latent Dirichlet allocation model. *Journal of Machine Learning Research* **18** 1–38.

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

Geyer, C. J. (2011). Importance sampling, simulated tempering, and umbrella sampling. In *Handbook of Markov Chain Monte Carlo* (S. P. Brooks, A. E. Gelman, G. L. Jones and X. L. Meng, eds.). Chapman & Hall/CRC, Boca Raton, 295–311.

Geyer, C. J. and Thompson, E. A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *Journal of the American Statistical Association* **90** 909–920.

Hobert, J. P. and Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association* **91** 1461–1473.

Hoffman, M. D., Blei, D. M., Wang, C. and Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research* **14** 1303–1347.

Huang, J., Ma, S., Xie, H. and Zhang, C.-H. (2009). A group bridge approach for variable selection. *Biometrika* **96** 339–355.

Ibragimov, I. A. and Linnik, Y. V. (1971). *Independent and Stationary Sequences of Random Variables*. Wolters-Noordhoff, Groningen.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press.

Jones, G. L., Haran, M., Caffo, B. S. and Neath, R. (2006). Fixed-width output analysis for Markov chain Monte Carlo. *Journal of the American Statistical Association* **101** 1537–1547.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York.

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.

Linero, A. R. (2017). A review of tree-based Bayesian methods. *Communications for Statistical Applications and Methods* **24** 543–559.

Linero, A. R. (2018). Bayesian regression trees for high-dimensional prediction and variable selection. *Journal of the American Statistical Association* **113** 626–636.

Linero, A. R. and Yang, Y. (2018). Bayesian regression tree ensembles that adapt to smoothness and sparsity. *Journal of the Royal Statistical Society,* Series B **80** 1087–1110.

Liu, J. S. (1994). The fraction of missing information and convergence rate for data augmentation. In *Computationally Intensive Statistical Methods: Proceedings of the 26th Symposium Interface* (J. Sall and A. Lehman, eds.). Interface Foundation, Fairfax Station, VA.

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

Marinari, E. and Parisi, G. (1992). Simulated tempering: A new Monte Carlo scheme. *Europhysics Letters* **19** 451–458.

Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer-Verlag, New York.

Miller, J. W. (2019). Fast and accurate approximation of the full conditional for gamma shape parameters. *Journal of Computational and Graphical Statistics* **28** 476–480.

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

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **9** 249–265.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo* (S. P. Brooks, A. Gelman, G. L. Jones and X.-L. Meng, eds.). CRC Press, Boca Raton, 113–162.

Newton, M. and Raftery, A. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion). *Journal of the Royal Statistical Society,* Series B **56** 3–48.

Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* **92** 179–191.

Robert, C. P. (2001). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer-Verlag, New York.

Roy, V. (2014). Efficient estimation of the link function parameter in a robust Bayesian binary regression model. *Computational Statistics & Data Analysis* **73** 87–102.

Roy, V., Tan, A. and Flegal, J. (2018). Estimating standard errors for importance sampling estimators with multiple Markov chains. *Statistica Sinica* **28** 1079–1101.

Scheipl, F., Fahrmeir, L. and Kneib, T. (2012). Spike-and-slab priors for function selection in structured additive regression models. *Journal of the American Statistical Association* **107** 1518–1532.

Scott, J. G. and Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics* **38** 2587–2619.

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

Stark, P. C., Ryan, L. M., McDonald, J. L. and Burge, H. A. (1997). Using meteorologic data to model and predict daily ragweed pollen levels. *Aerobiologia* **13** 177–184.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics* **9** 1135–1151.

Storlie, C. and Helton, J. (2008). Multiple predictor smoothing methods for sensitivity analysis: Example results. *Reliability Engineering and System Safety* **93** 55–77.

Storlie, C. B., Bondell, H. D., Reich, B. J. and Zhang, H. H. (2011). Surface estimation, variable selection, and the nonparametric oracle property. *Statistica Sinica* **21** 679–705.

Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association* **101** 1566–1581.

Tsybakov, A. B. (1990). Recursive estimation of the mode of a multivariate distribution. *Problemy Peredachi Informatsii* **26** 38–45.

van Dyk, D. A. and Meng, X.-L. (2001). The art of data augmentation (with discussion). *Journal of Computational and Graphical Statistics* **10** 1–50.

Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association* **85** 699–704.

Wolpert, R. L. and Schmidler, S. C. (2012). α-stable limit laws for harmonic mean estimators of marginal likelihoods. *Statistica Sinica* **22** 1233–1251.

Yang, Y., Wainwright, M. J. and Jordan, M. I. (2016). On the computational complexity of high-dimensional Bayesian variable selection. *The Annals of Statistics* **44** 2497–2532.