

Adaptive Markov Chain Monte Carlo for Bayesian Variable Selection

Chunlin Ji and Scott C. Schmidler*

Department of Statistical Science

Duke University

Abstract

We describe adaptive Markov chain Monte Carlo (MCMC) methods for sampling posterior distributions arising from Bayesian variable selection problems. Point mass mixture priors are commonly used in Bayesian variable selection problems in regression. However, for generalized linear and nonlinear models where the conditional densities cannot be obtained directly, the resulting mixture posterior may be difficult to sample using standard MCMC methods due to multimodality. We introduce an adaptive MCMC scheme which automatically tunes the parameters of a family of mixture proposal distributions during simulation. The resulting chain adapts to sample efficiently from multimodal target distributions. For variable selection problems point mass components are included in the mixture, and the associated weights adapt to approximate marginal posterior variable inclusion probabilities, while the remaining components approximate the posterior over non-zero values. The resulting sampler transitions efficiently between models, performing parameter estimation and variable selection simultaneously. Ergodicity and convergence are guaranteed by limiting the

* *Corresponding author:* Scott C. Schmidler, Department of Statistical Science, Duke University, Durham, NC 27708-0251. Tel: (919) 684-8064; Fax: (919) 684-8594; Email: schmidler@stat.duke.edu

adaptation based on recent theoretical results. The algorithm is demonstrated on a logistic regression model, a sparse kernel regression, and a random field model from statistical biophysics; in each case the adaptive algorithm dramatically outperforms traditional MH algorithms.

Keywords: adaptive Monte Carlo, Bayesian analysis, variable selection, kernel regression, graphical models

1 Introduction

Bayesian approaches to variable or feature selection often utilize prior distributions that assign non-zero probability to the event that a regression coefficient or other model parameter takes value exactly zero, leading to the removal of the corresponding variable from the model (Mitchell and Beauchamp, 1988). George and McCulloch (1997) and Clyde and George (2004) provide a thorough review of Bayesian variable selection methods and prior distributions, with particular emphasis on the Stochastic Search Variable Selection (SSVS) approach (George and McCulloch, 1993), which uses a Gibbs sampling algorithm to search for high posterior probability models. However with large numbers p of predictor variables under consideration, the search of 2^p candidate models is computationally challenging. Recently, Bayesian shrinkage regression has also attracted significant attention for solving the variable selection problem (Tipping, 2001; Figueiredo and Jain, 2001; Bae and Mallick, 2004; Griffin and Brown, 2005; Schmidler et al., 2007).

In many Bayesian variable selection approaches, the sparseness prior for inclusion can be written marginally or conditionally as a mixture of a parametric distribution and a point mass at zero:

$$\pi_0(\beta) = (1 - p)\delta(\beta) + p N(\beta|0, \sigma). \quad (1)$$

Variations on this theme include replacement of the normal component with uniform (the

'spike and slab' model of Mitchell and Beauchamp (1988)), or replacement of the point mass component with a mean-zero normal with high precision (George and McCulloch, 1993); such prior distributions have been well studied (Mitchell and Beauchamp, 1988; George and McCulloch, 1993; Geweke, 1996; George and McCulloch, 1997; West, 2003; Clyde and George, 2004). For concreteness in what follows we assume the form given in (1); the resulting posterior is a mixture of a point mass and a normal-likelihood product.

This point mass prior is especially popular for linear models, where use of conjugate priors enables conditional posterior distributions to be calculated in closed form for Gibbs sampling. For generalized linear and nonlinear models however, the conditional densities generally cannot be obtained explicitly. In such cases it is commonly assumed that one must resort to reversible-jump (Green, 1995) methods or approximation of marginal likelihoods. However, a simpler approach is to construct a Metropolis-Hastings algorithm (Hastings, 1970; Gilks et al., 1996), with the posterior distribution $\pi(\beta_1, \dots, \beta_r | X, Y)$ having density with respect to the product measure $\nu = (\mu + \delta)^r$ where μ is 1-dimensional Lebesgue measure, δ is the Dirac measure at 0, and r is the number of potential covariates. The drawback of such an approach is that the resulting mixture posterior can be difficult to sample using standard MCMC methods due to its inherent multimodality - the sampler fails to move easily between the zero and non-zero values components of the conditional posterior(s), leading to very slow convergence and extremely high autocorrelation and thus Monte Carlo variance.

In particular, the proposal distribution $q(\beta, \beta')$ must also have a density with respect to ν ; that is, must itself be a mixture of an atom at zero and a continuous component. However, mixing a point mass at zero with a standard random-walk Metropolis kernel performs poorly, tending to get stuck in the basin of attraction near zero. Instead, a Metropolized independence sampler (Hastings, 1970; Tierney, 1994) with proposals independent of the current state is desirable. The prior distribution is an obvious choice, and where priors are well informed this may be a viable alternative. However, when the posterior mass diverges significantly from the prior, proposing from the prior will also lead to unacceptably slow mix-

ing. Such cases tend to occur with larger sample sizes, where the non-zero component of the posterior will also be concentrated, making a diffuse proposal equally ineffective. A proposal distribution which better approximates the posterior distribution is therefore desirable, and the focus of this paper is an automatic approach to constructing such distributions.

We introduce adaptive Markov chain Monte Carlo (AMCMC) methods to sample efficiently from posteriors arising from point mass mixture priors. AMCMC has seen renewed interest in recent years due in part to the emergence of certain theoretical guarantees (Haario et al., 2001; Roberts and Rosenthal, 2007). With AMCMC algorithms, the entire sample history of process is used to tune parameters of the proposal density during simulation. A general framework for designing AMCMC algorithms is built around the adaptive Metropolized independence sampler (AMIS). To handle multimodality, we develop a simple but effective adaptation strategy using a family of mixture distribution proposals. For the Bayesian variable selection problem, we use a family of proposals containing both a point mass component and a Gaussian mixture. Under our adaptation strategy, the weight of the point mass component adapts to approximate (one minus) the posterior inclusion probability of its associated variable, while the Gaussian mixture distribution approximates the non-zero component of the coefficient’s posterior distribution. This mixture proposal enables efficient mixing between models with and without the variable included, and the resulting sampling scheme performs parameter estimation and variable selection simultaneously. The convergence and ergodicity of these algorithms is guaranteed by a careful design of the adaptation strategy.

Section 2 introduces the general framework of adaptive MCMC and AMIS. Section 2.2 describes the use of adaptive mixture proposals to sample from multimodal target distributions. Section 3 develops our adaptive MIS approach for sampling point mass mixture distribution and gives an illustrative example. Section 4 applies our approach to Bayesian variable selection in three realistic models: a logistic regression model, a sparse kernel regression model, and a random field model from statistical biophysics. We conclude with a discussion of the advantages and limitations of this approach.

2 Adaptive Metropolized Independence Samplers

Markov chain Monte Carlo methods are widely used to sample from analytically intractable probability distributions arising in statistics (Gilks et al., 1996; Robert and Casella, 1999). The efficiency of MCMC methods is of significant practical importance, and is determined by the convergence rate of the chain and asymptotic variance of ergodic averages, both of which are controlled by the spectral gap of the Markov kernel. Thus the efficiency of MCMC algorithms can depend significantly on the design of the Markov transition kernel; see e.g. (Hastings, 1970; Gelman et al., 1995; Mira, 2001; Roberts and Rosenthal, 2001). However choice of effective kernels and their associated tuning parameters is often difficult in precisely those problems where MCMC is most needed: high dimensional problems where we know little *a priori* about the shape of the (potentially multimodal) target posterior distribution.

Due to the difficulty of obtaining rapidly mixing Markov chains for simulating complicated target distributions, *adaptive* MCMC algorithms have been proposed which use the previous history of the chain to automatically tune or “learn” the proposal distribution parameters during simulation, with the goal of obtaining faster convergence or more efficient estimation (Gelfand and Sahu, 1994; Gilks et al., 1998). In adaptive MCMC, the proposal distribution is continually or periodically modified with the aim of improving efficiency. Although this idea is intuitively appealing, such algorithms generally fail to yield Markov chains, making design of adaptive MC schemes with theoretical convergence guarantees more challenging. Gilks et al. (1998) and Brockwell and Kadane (2005) approach this via regeneration times, at which the kernel may be modified while producing independent tours each generating correct ergodic averages. More recently, Haario et al. (2001) give an ergodic theorem for an adaptive Metropolis scheme based on the Robbins-Munro stochastic approximation algorithm (Robbins and Monro, 1951), and this result has led to significant renewed interest in adaptive algorithms and theory (Andrieu et al., 2005; Andrieu and Moulines, 2006; Erland, 2003; Roberts and Rosenthal, 2007, 2006). Recently, Roberts and Rosenthal (2007) provide a simple elegant proof and concise set of conditions under which ergodic theorems can be

obtained. One such condition requires that the magnitude of adaptation is continually decreasing in such a way that convergence of the chain to the target distribution in the limit is guaranteed; this kind of algorithms is referred as *diminishing adaptation* by Erland (2003). The other is a *bounded convergence* condition, which essentially guarantees that all transition kernels considered have bounded convergence time.

In this paper we describe a general approach to the design of adaptive MCMC algorithms which utilizes a mixture distribution for the proposal kernel, and adapts the parameters of this proposal distribution to minimize Kullback-Leibler divergence from the target distribution. We illustrate our approach using a Metropolized independence sampler (MIS) (Hastings, 1970; Tierney, 1994), a special case of the Metropolis algorithm where the proposal is independent of the current state. (The method described here utilizes the stochastic approximation approach of Ji (2006). Andrieu and Moulines (2006) have proposed a closely related method for adapting MIS mixtures using KL divergence, although to our knowledge it has not been applied to the variable selection problems studied here. See also Gasemyr (2003) and Holden et al. (2009) for additional work on adaptive MIS samplers. Andrieu and Thoms (2008) and Craiu et al. (2009), which appeared while our paper was under review, also use adaptive mixture distributions similar to our intermediate algorithm (Algorithm 2).)

Performance of MIS samplers is strongly dependent on the proposal distribution selected. Our adaptation strategy tunes the parametrized proposal distribution to approximate the target distribution in the sense of minimizing Kullback-Leibler (KL)-divergence. Thus for independence proposal density $q(x; \psi)$ with parameters ψ , and target distribution $\pi(x)$, we wish to find the optimal parameters ψ^* which minimize $\mathcal{D}[\pi(x) \parallel q(x; \psi)] = \mathbb{E}_\pi \left[\log \frac{\pi(x)}{q(x; \psi)} \right]$, or equivalently maximize the negative cross-entropy $\mathcal{H}(\pi(x), q(x; \psi)) = \int \pi(x) \log q(x; \psi) \nu(dx)$. Then ψ^* is obtained as a root of the derivative of $\mathcal{H}(\pi(x), q(x; \psi))$:

$$h(\psi) = \int \frac{\pi(x)}{q(x; \psi)} \frac{\partial}{\partial \psi} q(x; \psi) = 0 \quad (2)$$

where we assume the integrands on both sides are continuous. Exact solution of the integral equation (2) is generally intractable, as $h(\psi)$ involves an integral with respect to the target distribution $\pi(x)$ which cannot be calculated directly. However, denoting $f(x, \psi) = \frac{\partial}{\partial \psi} [\log \frac{\pi(x)}{q(x; \psi)}]$ and assuming $f(x, \psi) \in L_2(\pi)$, we can approximate $h(\psi)$ by Monte Carlo integration $h(\psi) \approx \frac{1}{K} \sum_{k=1}^K f(X^{(k)}, \psi)$ where $X^{(k)} \sim \pi(x)$.

When $q(x; \psi)$ is in the exponential family, so $q(x; \psi) = c(x) \exp(t(x)' \psi - A(\psi))$ in canonical form with natural parameter ψ , we obtain $\int \pi(x) t(x) = \frac{\partial}{\partial \psi} A(\psi)$, which says that we should match the expected sufficient statistics under π to the moments of q . However this expectation $E_\pi(t(x))$ is an integral with respect to the MCMC target distribution $\pi(x)$, and as such is assumed to be analytically intractable. Instead, we adaptively match the moments of q to a Monte Carlo approximation of $E_\pi(t(x))$ based on the current sample history.

Let $\hat{h}(X^{(1:K)}; \psi)$ denote the estimate of $h(\psi)$ based on the previous sample path $X^{(1:K)}$ from $\pi(x)$, which can be therefore viewed as a noisy ‘observation’ of $h(\psi)$. A common approach to obtaining roots of $h(\psi) = 0$ when only noisy evaluations of $h(\psi)$ are available is the Stochastic Approximation (SA) algorithm (Robbins and Monro, 1951; Kushner and Yin, 1997). Stochastic approximation is an iterative algorithm expressed as

$$\begin{aligned} \psi_{n+1} &= \psi_n - r_{n+1}(0 - (h(\psi_n) + \xi_{n+1})) \\ &= \psi_n + r_{n+1} \hat{h}(X_n^{(1:K)}; \psi_n) \end{aligned} \tag{3}$$

where $X_n^{(1:K)} \sim \pi(x)$ are samples generated in our case by Metropolized independence sampling with proposal distribution $q(x; \psi_n)$, $\{\xi_n\}$ is a sequence of ‘noise’, and $\{r_n\}$ is a sequence of decreasing step-sizes satisfying $\sum_n r_n = \infty$ and $\sum_n r_n^2 < \infty$.

In our case, SA can be viewed as performing an iterative gradient descent, with Monte Carlo approximation of the gradient at each iteration. When q is an exponential family $\mathcal{D}[\pi(x) \parallel q(x; \psi)]$ is convex, and the sequence $\{\psi_n\}$ defined by equation (3) converges to the unique root of equation (2) under mild conditions on $\{\xi_n\}$ and $\{r_n\}$ (Andrieu et al., 2005).

However Andrieu and Moulines (2006) also show that an adaptive proposal $q(x; \psi)$ for MIS with ψ unrestricted does not guarantee convergence of the algorithm. A straightforward solution due to Haario et al. (2001) is to use an additional fixed mixture component $q(x; \zeta)$ which is not modified during the adaptive updating; in what follows we take $q(x; \zeta) = N(x; \tilde{\mu}, \tilde{\Sigma})$ for some fixed $(\tilde{\mu}, \tilde{\Sigma})$. Note that an MIS chain will be geometrically ergodic (have a spectral gap) if and only if $\text{ess sup}_x \frac{\pi(x)}{q(x)} < \infty$; thus it may be desirable to use at least one mixture component with heavy tails, e.g. to replace $N(x; \tilde{\mu}, \tilde{\Sigma})$ with a t -distribution.

2.1 Adaptive MIS

As a simple illustrative example, choosing the adaptive proposal distributions $q(x; \psi)$ to be normal $N(x; \mu, \Sigma)$ with parameters $\psi = (\mu, \Sigma)$ yields a simple AMIS algorithm:

Algorithm 1: Initialize $\psi_0 = (\mu_0, \Sigma_0)$. Then iteration $n + 1$ becomes (see Appendix):

1. Simulate $X_{n+1}^{(1:K)}$ by MIS with proposal distribution $q_n = \lambda N(x; \tilde{\mu}, \tilde{\Sigma}) + (1-\lambda)N(x; \mu_n, \Sigma_n)$
2. Update the parameters of adaptive proposal q_n by

$$\mu_{n+1} = \mu_n + r_{n+1} \left[\frac{1}{K} \sum_{k=1}^K (X_{n+1}^{(k)} - \mu_n) \right] \quad (4)$$

$$\Sigma_{n+1} = \Sigma_n + r_{n+1} \left[\frac{1}{K} \sum_{k=1}^K (X_{n+1}^{(k)} - \mu_n) (X_{n+1}^{(k)} - \mu_n)^T - \Sigma_n \right] \quad (5)$$

where r_{n+1} is the step-size of the SA algorithm.

The covariance update of Step 2 is similar to that of (Haario et al., 2001), but as we use an independence proposal rather than a random walk proposal the mean is also approximated. The above adaptive MCMC algorithm satisfies the *diminishing adaptation* condition of Roberts and Rosenthal (2007) as long as the step-size sequence $r_n \rightarrow 0$. It will also satisfy the *bounded convergence* condition for $\lambda > 0$ as long as the non-adaptive component $q(x; \tilde{\mu}, \tilde{\Sigma})$ has sufficiently heavy tails (or \mathcal{X} is compact) as mentioned above, since $\text{ess sup}_x \frac{\pi(x)}{q(x)}$

is independent of ψ . Together these two conditions ensure asymptotic convergence and a weak law of large numbers for this algorithm (Roberts and Rosenthal, 2007).

2.2 Adaptive MIS with Mixture Proposal Distribution

When the above AMIS algorithm is applied to sample from a multimodal target distribution, it will generally perform poorly due to the difficulty in approximating the posterior with a unimodal q distribution. An alternative is to take q to be a mixture distribution:

$$q(x) = \lambda q_0(x; \tilde{\psi}) + (1 - \lambda) \sum_{m=1}^M w_m q_m(x; \psi_m)$$

and adapt the mixture component parameters $\psi = (w_{1:M}, \psi_{1:M})$ to approximate the multimodal target distribution by minimizing KL-divergence. The number of components M required is problem-dependent but should be chosen relatively large to enable adequate coverage in the event of multiple modes. Then the adaptation strategy is easily derived (see Appendix); for example taking all q_i 's to be normal gives:

Algorithm 2: Initialize $(\mathbf{w}_0, \mu_0, \Sigma_0) = \{w_{i,0}, \mu_{i,0}, \Sigma_{i,0}\}_{i=1}^M$. At iteration $n + 1$,

1. Draw X_{n+1} by MIS with proposal $q_n(x) = \lambda N(x; \tilde{\mu}, \tilde{\Sigma}) + (1 - \lambda) \sum_{m=1}^M w_m N(x; \mu_{m,n}, \Sigma_{m,n})$
2. Update the parameters $(\mathbf{w}_{n+1}, \mu_{n+1}, \Sigma_{n+1})$ by

$$w_{i,n+1} = w_{i,n} + r_{n+1}(O_i(X_{n+1}) - \bar{O}) \tag{6}$$

$$\mu_{i,n+1} = \mu_{i,n} + \kappa_{i,n+1}(X_{n+1} - \mu_{i,n}) \tag{7}$$

$$\Sigma_{i,n+1} = \Sigma_{i,n} + \kappa_{i,n+1} \left[(X_{n+1} - \mu_{i,n})(X_{n+1} - \mu_{i,n})^T - \Sigma_{i,n} \right] \tag{8}$$

where $\kappa_{i,n+1} = r_{n+1} w_{i,n} O_i(X_{n+1})$, $\bar{O} = \frac{1}{M} \sum_{i=1}^M O_i(X_{n+1})$, and $O_i(X_{n+1}) = \frac{\phi(X_{n+1}; \mu_{i,n}, \Sigma_{i,n})}{\sum_{m=1}^M w_{m,n} \phi(X_{n+1}; \mu_{m,n}, \Sigma_{m,n})}$.

Here $\phi(X; \mu, \Sigma)$ denotes the multivariate normal density with mean μ and covariance Σ .

For notational simplicity the parameter updates (6-8) show the use of a single sample X_{n+1} , but as discussed previously using $K > 1$ samples $X_{n+1}^{(1:K)}$ will enable the SA algorithm to converge more smoothly. (Examples in Section 4 use $K = 20$). A derivation for general q_i 's is given in the Appendix; for example, Section 4.2 utilizes a mixture of gamma distributions for adaptive sampling of scale parameters.

3 Adaptive MCMC with Point Mass Mixture Proposal

When performing Bayesian variable selection using priors of the form (1) as described in Section 1, the resulting conditional posterior is a mixture of point mass and an normal-likelihood product. When this conditional distribution is not available in closed form (e.g. due to nonlinearity or non-conjugacy) so that Gibbs samplers are not available, sampling from the posterior via MCMC can be difficult as described in Section 1. In particular, random-walk Metropolis can converge very slowly due to multimodality, and an MIS sampler will perform poorly unless the proposal distribution can be chosen in advance to closely approximate the target distribution. However, the adaptive mixture MIS algorithm described in the previous section can successfully handle both of these difficulties. We need simply modify the family of proposal mixture distributions to include both point mass and continuous components:

$$q(x) = \lambda q_0(x; \tilde{\psi}) + (1 - \lambda) \left[w_0 \delta(x) + \sum_{m=1}^M w_m q_m(x; \psi_m) \right]$$

where the parameters $\psi = (w_{0:M}, \psi_{1:M})$ can be tuned using an adaptive scheme similar to that of the previous section. For example, taking the q_i 's to be normal gives:

Algorithm 3: Adaptive MCMC with Point Mass Mixture Proposal

Initialize $\{\mathbf{w}_0, \mu_0, \Sigma_0\}$. At iteration $n + 1$:

1. Sample $Y \sim \lambda N(x; \tilde{\mu}, \tilde{\Sigma}) + (1 - \lambda) \left[w_{0,n} \delta(x) + \sum_{m=1}^M w_{m,n} N(x; \mu_{m,n}, \Sigma_{m,n}) \right]$, and set

$X_{n+1} = Y$ with acceptance probability $\alpha(X_n, Y) = \min(1, \eta)$ where:

$$\eta = \begin{cases} 1 & \text{if } X_n = Y = 0 \\ \frac{\pi(Y)}{\pi(X_n)} \frac{s_n(X_n)}{(1-\lambda)w_{0,n}} & \text{if } X_n \neq 0, Y = 0 \\ \frac{\pi(Y)}{\pi(X_n)} \frac{(1-\lambda)w_{0,n}}{s_n(Y)} & \text{if } X_n = 0, Y \neq 0 \\ \frac{\pi(Y)}{\pi(X_n)} \frac{s_n(X_n)}{s_n(Y)} & \text{if } X_n \neq 0, Y \neq 0 \end{cases}$$

where $s_n(x) = \lambda N(x, \tilde{\mu}, \tilde{\Sigma}) + (1 - \lambda) \sum_{m=1}^M w_{m,n} N(x; \mu_{m,n}, \Sigma_{m,n})$.

- Update the parameters $\{w_{i,n+1}, \mu_{i,n+1}, \Sigma_{i,n+1}\}$ via equations (6-8) of the previous section, except now $(\mu_{i,n+1}, \Sigma_{i,n+1}) = (\mu_{i,n}, \Sigma_{i,n})$ when $X_{n+1} = 0$, and we have a slight modification in the calculation of $O_i(X_{n+1})$:

$$O_i(X_{n+1}) = \begin{cases} \frac{1}{w_{0,n}} & \text{for } X_{n+1} = 0; i = 0 \\ \frac{\phi(X_{n+1}; \mu_{i,n}, \Sigma_{i,n})}{\sum_{m=1}^M w_{m,n} \phi(X_{n+1}; \mu_{m,n}, \Sigma_{m,n})} & \text{for } X_{n+1} \neq 0; i = 1, \dots, M \\ 0 & \text{otherwise} \end{cases}$$

and now $\bar{O}(X) = \frac{1}{M+1} \sum_{m=0}^M O_i(X)$.

3.1 Example

We begin with a simple concrete example to illustrate the performance of our adaptive MIS with point mass mixture proposal. Suppose we consider inclusion or exclusion of a single parameter, with posterior distribution given by point mass mixture $\pi(x) = 0.3\delta(x) + 0.7N(x; 5, 1)$. We apply the adaptive MIS Algorithm 3 to sample from this target distribution $\pi(x)$. We set $M = 1$, making the proposal distribution of the form $w_1\delta + w_2N(\mu, \sigma)$. Therefore for this illustrative example, q and π are of the same parametric (w_1, μ, σ) family, and it is expected that q will converge to π .

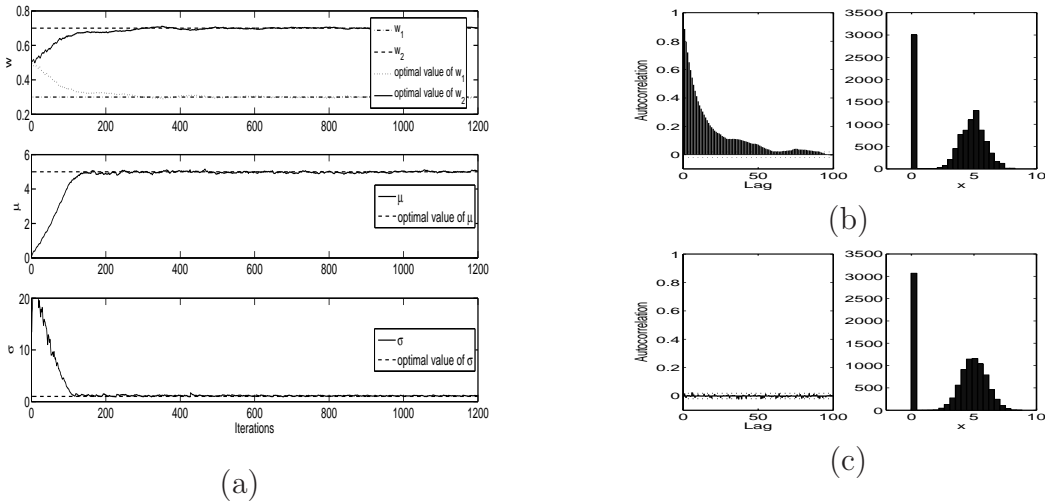


Figure 1: (a) Traceplots of proposal distribution parameters for the adaptive MIS algorithm applied to toy example described in text. Proposal parameters converge to their respective optimal values: $w = [0.3, 0.7]$, $\mu^* = 5$, and $\sigma^* = 1$. (b) Autocorrelation plots and posterior histograms for toy example, obtained by non-adaptive MIS algorithm, versus (c) adaptive MIS algorithm with point mass mixture proposal.

Figure 1 shows results on this simple example, using initial parameter values $w_1 = 0.5$, $w_2 = 0.5$, $\mu = 0$, and $\sigma = 10$, and SA step-size $r_n = 0.1/n$. Proposal distribution parameters (w_1, w_2, μ, σ) are seen to converge to their respective optimal values $w_1^* = 0.3$, $w_2^* = 0.7$, $\mu^* = 5$, and $\sigma^* = 1$; thus the proposal distribution converges to the target distribution. Figure 1b,c compares the performance of this adaptive scheme with a non-adaptive Metropolized independence sampler using fixed proposal $0.5\delta + 0.5N(0, 100)$, via posterior histograms and autocorrelation plots. The adaptive algorithm is seen to perform significantly better.

4 Applications

We now evaluate the performance of the adaptive MC variable selection algorithm on several realistic statistical models: Bayesian variable selection in generalized linear models; a sparse Bayesian kernel regression problem; and a model selection problem for a Gibbs random field model from statistical biophysics. The logistic regression model example can be viewed as a demonstration of the adaptive MCMC approach, as (approximate) methods for this model

are available. The latter two examples represent complex models from the recent literature, where traditional MCMC methods such as the Gibbs sampler or random walk Metropolis algorithm perform very poorly. We show that the adaptive MCMC methods of this paper can be applied in all cases to obtain significant improvements in sampling efficiency.

4.1 Logistic Regression

We begin with Bayesian analysis of a standard logistic regression. We first apply the adaptive mixture MIS (Algorithm 2) of Section 2.2 to sample the posterior on coefficients for a model with fixed covariates. We then consider Bayesian variable selection in this model using spike-and-slab priors for the coefficients, and apply the adaptive MCMC with point mass mixture proposal (Algorithm 3) of Section 3. In these examples the mixture approximation is applied to each parameter independently, thus adapting to the marginal posteriors.

Consider the Bayesian logistic regression model, $y_i | x_i, \beta \sim \text{Bernoulli}(g^{-1}(x_i\beta))$ where $y_i \in \{0, 1\}$ for $i = 1, \dots, n$ is a binary response variable for a collection of n subjects, each with p associated covariates $x_i = (x_{i1}, \dots, x_{ip})$, $g(u)$ is the logistic link function, and β is a $(p \times 1)$ column vector of regression coefficients (including intercept) with prior distribution $\pi_0(\beta)$. We wish to sample from the posterior distribution

$$\pi(\beta | X, Y) \propto \pi_0(\beta) \prod_{i=1}^n (g^{-1}(x_i\beta))^{y_i} (1 - g^{-1}(x_i\beta))^{1-y_i}. \quad (9)$$

As closed form conditional distributions for Gibbs sampling are unavailable, it is standard to use Metropolis-Hastings. Recently Holmes and Held (2006) extended the data augmentation method of Albert and Chib (1993), which provides closed form conditionals for probit regression, to obtain an Gibbs sampler which approximates (9). This is done by expressing the logistic model via auxiliary variables z_i as $y_i = 1$ if $z_i > 0$ and 0 otherwise, with $z_i = x_i\beta + \epsilon_i$ where $\epsilon_i \sim N(0, \lambda_i)$, $\lambda_i = (2\psi_i)^2$, $\psi_i \sim KS$, and $\psi_i, i = 1, \dots, n$, are independent random variables with Kolmogorov-Smirnov (KS) distribution (Devroye, 1986). Then ϵ_i has

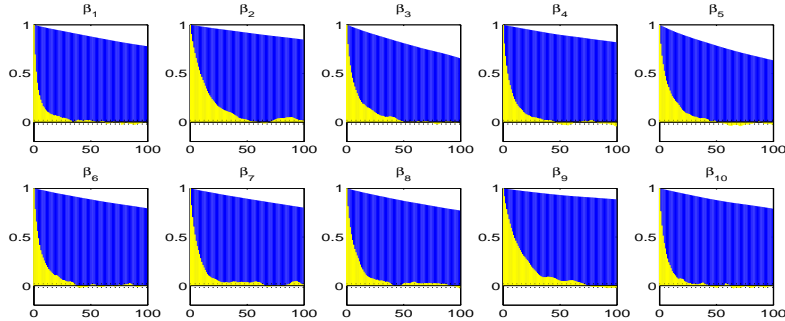


Figure 2: Autocorrelation plots for logistic model coefficients $\beta_{1:10}$ using the Gibbs sampling method of Holmes and Held (2006) (dark) versus the adaptive MCMC algorithm (light). The adaptive algorithm is seen to mix significantly faster.

distribution given by a scale mixture of normals which is marginally a standard logistic distribution (Andrews and Mallows, 1974), giving the marginal likelihood of the original logistic regression model. However in practice the KS distribution has an infinite representation and is intractable to sample from. Instead we may take λ_i to have Gamma distribution and ψ_i a T-distribution, which can approximate the logistic distribution with appropriate choice of the T d.o.f. (Albert and Chib, 1993).

In contrast, the adaptive MCMC proposed in this work requires no such approximation, and enables efficient sampling directly from (9). In addition, the adaptation provides significant efficiency improvements over the approximate auxiliary variable Gibbs sampler. To illustrate, we applied our adaptive MCMC to the posterior distribution (9) of the logistic regression model, and compared with the approximate auxiliary variable model. We simulated a dataset of 200 data points directly from the model, using $r = 10$ covariates and with coefficients $\beta_{1:10} = [-0.01, -1.50, 0.15, 0.50, -0.15, -0.20, -0.60, 0.25, 1.50, -0.05]$. Figure 2 shows autocorrelation plots for the MCMC samples of the ten regression coefficients. Blue autocorrelation plots show slow mixing of the auxiliary variable scheme, due to the strong posterior dependency between the regression and auxiliary variables (Holmes and Held, 2006); in contrast the red autocorrelation plots demonstrate that the adaptive MCMC algorithm performs significantly better in this scenario.

Extending this example to the Bayesian variable selection context, we place spike-and-

slab prior distributions $\pi(\beta_i)$ on the coefficients $\pi_0(\beta_i) = 0.5 \delta(\beta_i) + 0.5 N(\beta_i | 0, \sigma^2)$ taking $\sigma = 100$. We applied the adaptive MCMC with point mass mixture proposal (Algorithm 3) of Section 3 to sample from the joint posterior $\pi(\beta | X, Y) \propto \mathcal{L}(Y|\beta) \prod_{i=1}^p \pi_0(\beta_i)$, where $\mathcal{L}(Y|\beta)$ denotes the logistic likelihood function. We simulated 200 data points using $r = 10$ covariates with coefficients $\beta_{1:4} = [1.0, 4.0, 2.0, -2.0]$ and $\beta_{5:10} = 0$. Figure 3 shows posterior histograms obtained for the coefficients; we see that the algorithm correctly selects the relevant predictor variables. Figure 4 shows sample autocorrelation plots for the regression coefficients using the adaptive point mass method, compared to a non-adaptive random-walk Metropolis which would be the standard choice for this problem, and again the adaptive algorithm shows dramatic speedup. Table 1 shows the estimated Monte Carlo standard errors of the parameters, where we see that the decrease in autocorrelation of the adaptive scheme yields effective sample sizes 50-150 \times larger. Thus the adaptive algorithm computes the Bayesian variable selection solution for the logistic regression model with spike-and-slab priors, approximately 100 \times faster than the standard approach.

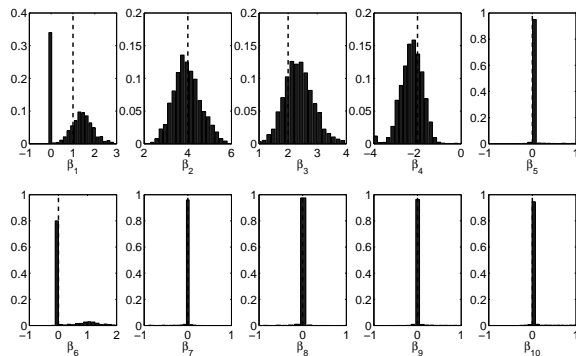


Figure 3: Posterior histogram obtained via point mass mixture AMCMC for coefficients $\beta_{1:10}$ of logistic model, with corresponding true values (dashed line). Posteriors concentrate at correct zero ($\beta_{5:10}$) and non-zero ($\beta_{1:4}$) values.

4.2 Kernel Regression

Kernel models have been used extensively in machine learning for classification and regression problems (Poggio and Girosi, 1990; Vapnik, 1998; Scholkopf and Smola, 2001; Shawe-Taylor

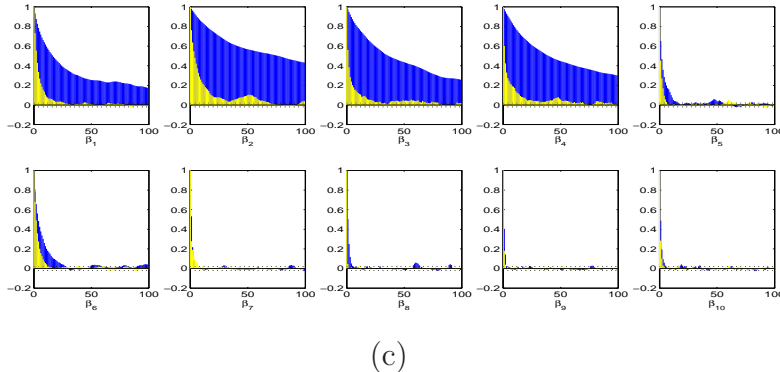


Figure 4: Autocorrelation plots for the logistic model coefficients $\beta_{1:10}$ using random walk non-adaptive MCMC (dark) and the point-mass mixture adaptive MCMC algorithm described in text (light). The adaptive algorithm is seen to mix significantly faster.

| | Metropolis | | Adaptive | | Eff. sample size |
|-----------|-----------------|-----------|-----------------|-----------|--|
| | $\hat{\beta}_i$ | std error | $\hat{\beta}_i$ | std error | $\sigma_{\text{MCMC}}^2 / \sigma_{\text{AMCMC}}^2$ |
| β_1 | 1.59 | 1.31 | 0.95 | 0.108 | 147.1 |
| β_2 | 6.55 | 0.59 | 3.97 | 0.052 | 127.5 |
| β_3 | 2.82 | 0.76 | 2.37 | 0.063 | 146.5 |
| β_4 | -3.70 | 0.05 | -2.27 | 0.007 | 50.8 |

Table 1: Parameter estimates for the Bayesian logistic regression model selection. The adaptive MCMC shows dramatic improvements in Monte Carlo variance of parameter estimates, with effective sample sizes 50-150 \times larger than the standard Metropolis scheme.

and Cristianini, 2004). Variable selection technologies for kernel models are of great interest, especially in situations where the number of potential variables is comparable to or larger than the number of observations. Recently Liang et al. (2006) develop a MCMC sampling procedure for treating the Relevance Vector Machine (RVM) kernel regression model from a fully Bayesian perspective. However the MCMC algorithm provided there mixes slowly; here we apply our adaptive MCMC variable selection scheme to this model.

Suppose we have n observations (x_i, y_i) with p explanatory variables $x_i \in \mathbb{R}^p$ and corresponding responses $y_i \in \{0, 1\}$, on which we wish to train a classifier for future observations. If p is large relative to n , we wish to include only salient features in order to reduce classifier variance and improve predictive performance. The model uses a probit link, with latent variable $z_i \sim N(\mu_i, 1)$ introduced for each observation, where μ_i is a predictor and $z_i > 0$

iff $y_i = 1$, so $P(y_i = 1) = \Phi(\mu_i)$. The regression model is specified through a *kernel* matrix $\mu = w_0 + Kw$ where $\mu = (\mu_1, \dots, \mu_n)'$, w_0 is an intercept term, w is a vector of regression coefficients, and K is an $n \times n$ kernel matrix. Then we have

$$\mu_i = w_0 + \sum_{j=1}^n K(x_i, x_j)w_j \quad \text{for } i = 1, \dots, n$$

Any Mercer (continuous, symmetric, positive definite) kernel K may be used; common choices include a radial basis function (RBF) kernel $K(x, x^*) = \exp\{-\sum_{k=1}^p \rho_k(x_k - x_k^*)^2\}$ for scale parameters for $\rho \in \mathbb{R}^p$, and the linear kernel $K(x, x^*) = \sum_{k=1}^p \rho_k x_k x_k^*$. These kernels measure similarity between two data points $x, x^* \in \mathbb{R}^p$, accounting for possible differences in significance of the various dimensions. Large ρ_k 's contribute to defining similarity and represent salient explanatory variables, while small ρ_k 's represent insignificant variables.

Kernel classification and regression methods typically utilize fixed parameters (for example, fixed ρ in the RBF kernel) and do not perform variable selection directly. The Bayesian approach of Liang et al. (2006) estimates the kernel function parameters simultaneously with the regression coefficients, allowing for feature selection during model fitting.

The probit likelihood is of the form

$$\prod_{j=1}^n \Phi(\mu_j)^{y_j} (1 - \Phi(\mu_j))^{1-y_j}$$

and parameter prior distributions are specified hierarchically: $\rho_k \sim (1-\gamma)\delta + \gamma, \text{Gamma}(a_\rho, a_\rho s)$ for $k = 1, \dots, p$, $s \sim \text{Exp}(a_s)$ and $\gamma \sim \text{Beta}(a_\gamma, b_\gamma)$ where $a_\rho, a_s, a_\gamma, b_\gamma$ are all hyperparameters that are prespecified.

Posterior inference for ρ is of particular interest here, as ρ represents the significance of various dimensions. Liang et al. (2006) construct a special-purpose Metropolis scheme using a proposal distribution for ρ with a mixture of a global and local moves, but the chain still mixes somewhat slowly. Here we implement the general-purpose adaptive MCMC procedure of Section 3 (Algorithm 3) with independent proposal of the ρ_k 's. Since the prior distribution

of ρ_k is a mixture of a point mass and a Gamma distribution, we also take the proposal for each ρ_k as a mixture of point mass at 0 and $M = 5$ Gamma distributions:

$$q(\rho) = \lambda \mathcal{G}(\rho; 1, 0.1) + (1 - \lambda) \left[w_0 \delta(\rho) + \sum_{m=1}^4 w_m \mathcal{G}(\rho; \alpha_m, \beta_m) \right].$$

Parameter updates for Algorithm 3 using a mixture of Gammas proposal are given by (see Appendix):

$$\alpha_{i,k,n+1} = \alpha_{i,k,n} + \kappa_{i,n+1} (\log(\beta_{i,k,n}) - \psi(\alpha_{i,k,n}) + \log(\rho)) \quad (10)$$

$$\beta_{i,k,n+1} = \beta_{i,k,n} + \kappa_{i,n+1} \left(\frac{\alpha_{i,k,n}}{\beta_{i,k,n}} - \rho \right) \quad (11)$$

where subscripts i, k, n index the proposal mixture component, the covariate, and the iteration, respectively; $\psi(\alpha)$ is the digamma function; and $\kappa_{i,n+1}$ are defined similarly to Section 2 with Gamma densities in place of normals. Thus $\rho_{k,n+1}$ is proposed from the adaptive mixture distribution with parameters obtained from the previous iteration. Sampling of all other parameters remains unchanged from the original method (Liang et al., 2006).

We evaluated our adaptive MCMC algorithm for this model using the synthetic data set studied in Liang et al. (2006). As our emphasis here is solely on the sampling algorithm rather than the larger model, we report here only results for the parameters ρ of interest for variable selection; further details of the model applied to this data set can be found in the original paper (Liang et al., 2006). Figure 5 shows the sample autocorrelation of the ρ 's obtained by simulation using the proposed MCMC method of Liang et al. (2006), compared with the adaptive MCMC method developed in this paper, where again the adaptive algorithm shows significantly better mixing. Estimated exclusion probabilities estimates are shown in Table 2 along with those of Liang et al. (2006); where it can be seen that the decrease in Monte Carlo variance obtained via the adaptive algorithm increases the accuracy (only variables ρ_1 and ρ_2 are nonzero in the synthetic data set).

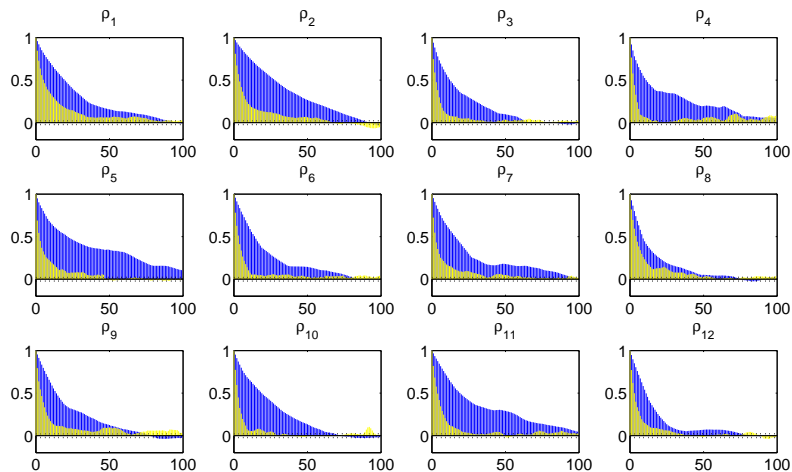


Figure 5: Autocorrelation plots for kernel regression scale parameters (ρ_k 's) obtained using the MCMC algorithm of Liang et al. (2006) (dark) and point-mass mixture AMCMC (light).

| Algorithm | ρ_1 | ρ_2 | ρ_3 | ρ_4 | ρ_5 | ρ_6 | ρ_7 | ρ_8 | ρ_9 | ρ_{10} | ρ_{11} | ρ_{12} |
|--------------|----------|----------|----------|----------|----------|----------|----------|----------|----------|-------------|-------------|-------------|
| Non-adaptive | 0.33 | 0.30 | 0.61 | 0.62 | 0.61 | 0.59 | 0.61 | 0.61 | 0.65 | 0.63 | 0.60 | 0.61 |
| Adaptive | 0.01 | 0.03 | 0.82 | 0.88 | 0.90 | 0.89 | 0.79 | 0.83 | 0.86 | 0.91 | 0.76 | 0.77 |

Table 2: Estimated exclusion probabilities for kernel regression scale parameters (ρ_k 's) obtained via the MCMC algorithm of Liang et al. (2006) versus the point-mass mixture AMCMC algorithm. Rapid convergence of the AMCMC estimates yields more accurate inclusion ($\rho_{1:2}$) and exclusion ($\rho_{3:12}$) probabilities.

4.3 Model Selection in Gibbs Random Fields

Schmidler et al. (2007) describe Bayesian estimation and model selection for the parameters of a statistical mechanical model arising in bioinformatics and molecular biophysics. The helix-coil model describes equilibrium behavior of a short polypeptide which adopts a helical shape in solution (Poland and Scheraga, 1970); in recent years this model has seen extensive study and has been generalized to model the effects of peptide sequence on this equilibrium (see Schmidler et al. (2007) and references therein). Let $R = (R_1, \dots, R_l)$ denote a sequence of categorical variables specifying the amino acid sequence of a peptide, with each R_i taking values in the set of 20 amino acids. Let $X = (x_1, \dots, x_l)$ be an associated vector of binary indicators with $x_i = 1$ if the i^{th} amino acid is in helical conformation and 0 otherwise.

The model is a Gibbs random field with short-range neighborhood interactions, with potential $U(X, R)$ given by

$$U(X, R) = \sum_{i=1}^l x_i \alpha_{R_i} + \sum_{i=1}^{l-3} x_{i:i+3} \beta_{R_i R_{i+3}} + \sum_{i=1}^{l-4} x_{i:i+4} \gamma_{R_i R_{i+4}} \quad (12)$$

where $x_{i:k} = \prod_{j=i}^k x_j$ indicate contiguous stretches of helical amino acids. Here the α_i 's are "free energy" parameters which quantify the differing tendency of distinct amino acids to adopt helical conformations, and the β_{ij} 's and γ_{ij} 's denote interaction contributions involving positions at lags 3 and 4, respectively. (The absence of lags 1 and 2 are predetermined by the physical characteristics of the molecules being modeled). More detailed description of these parameters, and the many other parameters which fully specify the model, is given in Schmidler et al. (2007).

The resulting Gibbs distribution is $P(X \in \mathcal{X} \mid R) = Z^{-1} e^{-\frac{1}{kT} U(X, R)}$ where Z is the normalizing constant or partition function involving a sum over all configurations $X \in \mathcal{X} = \mathbb{Z}_2^l$. The *helicity* of a peptide is then given by the expectation or ensemble average $\mathcal{A}(R) = \sum_{X \in \mathcal{X}} h(X) P(X \mid R)$ where $h(X) = l^{-1} \sum_{i=1}^l x_i$, and it is this ensemble average quantity for which data can be measured via experimental methods such as circular dichroism.

Bayesian estimation of the parameter vectors (α, β, γ) , as well as other parameters in the model, is described in Schmidler et al. (2007). Denoting by θ all model parameters, a simple additive noise model $\tilde{h}_R = \mathcal{A}(R, \theta) + \epsilon$ with $\epsilon \sim N(0, \sigma^2)$ yields posterior distribution

$$\pi(\theta \mid D) = \frac{\sum_{\mathbf{x}} P(\mathbf{R}, \mathbf{x}, \mathbf{h} \mid \theta) P(\theta)}{\sum_{\mathbf{x}} P(\mathbf{R}, \mathbf{x}, \mathbf{h})} \propto \pi_0(\theta) (2\pi\sigma^2)^{-\frac{n}{2}} \prod_{i=1}^n e^{-\frac{1}{2\sigma^2} g(h_i - \mathcal{A}(R_i, \theta))^2}$$

where $\mathcal{A}(R)$ arises as calculation of the marginal likelihood of observations h_i .

Of particular interest is the reduction of the 800+ parameters contained in the vectors β and γ by selection of only those which are supported by the experimental data. In addition to stabilizing the model by reducing posterior and predictive variance, application of model selection to these interaction energies also yields important scientific insight into

which molecular interactions are important. Applying point-mass model selection priors of the form (1) to the individual β_{ij} 's and γ_{ij} 's achieves model selection in the graph of possible interactions. However, a random-walk Metropolis algorithm design for this purpose mixes rather slowly (Lucas, 2006).

Figure 6 compares autocorrelation plots for a representative subset of the interaction parameters using the adaptive MCMC with point mass mixture and the MCMC algorithm proposed in Lucas (2006). Although the previous algorithm uses a precomputing strategy to construct a good proposal distribution, the adaptive algorithm nevertheless shows significant improvement in mixing for many of the parameters. Posterior densities estimated from the two methods shown in Figure 7 indicate that at convergence the two samplers produce the same answer as expected; however the adaptive algorithm converges significantly faster and yields significantly lower Monte Carlo variance.

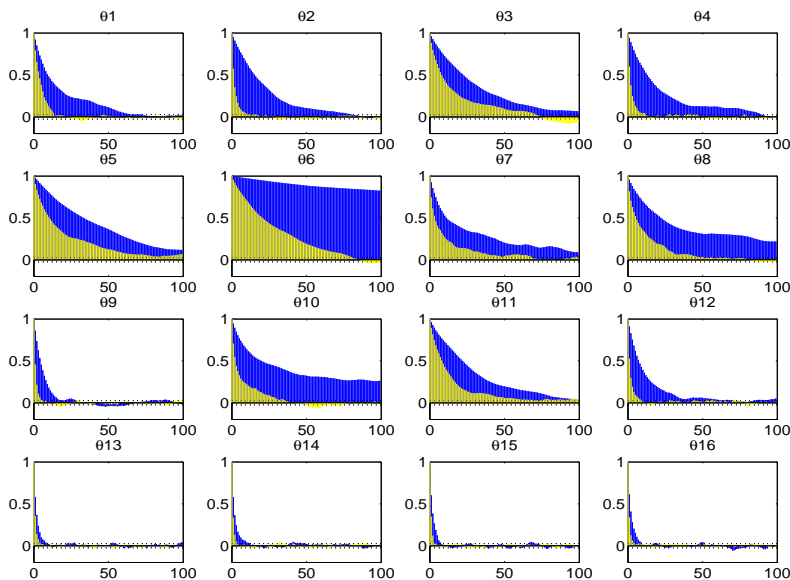


Figure 6: Autocorrelation plots for energy parameters of the helix-coil Gibbs random field model, using MCMC samples generated by the algorithm of Lucas (2006) (dark) versus the point-mass mixture adaptive MCMC algorithm (light).

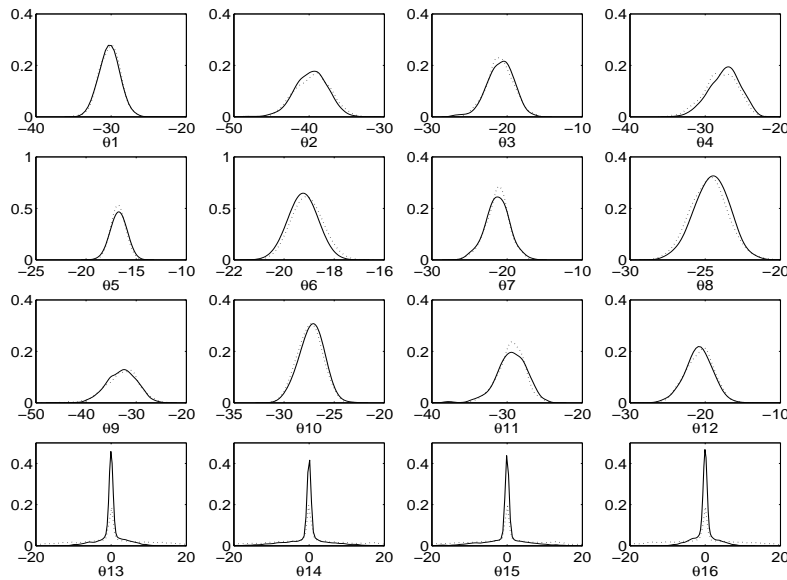


Figure 7: Posterior distributions of the helix-coil model parameters of Figure 6, obtained from the MCMC algorithm of Lucas (2006) (dashed line) and the point-mass mixture adaptive MCMC algorithm (solid line). Both samplers converge to the same distribution; however the adaptive algorithm mixes significantly faster (Figure 6).

5 Conclusions

The adaptive MIS algorithm with point mass mixture works well on all the examples we have tried. In several cases, where multimodality in the target distribution leads to slow mixing of random-walk Metropolis samplers, the adaptive algorithm yields dramatic speedups. For Bayesian variable selection problems, the algorithm appears to handle easily the multimodality commonly encountered using point mass priors. We have demonstrated this here using mixtures of normal distributions for location parameters, mixtures of gamma distributions for scale parameters, using varying numbers of components, and on small, moderate, and high-dimensional model selection problems.

Although we did not encounter this in our examples, as with any component-wise Metropolis or Gibbs sampler the adaptive algorithm may have difficulty in the presence of significant posterior covariance among the coefficients arising from correlated predictors. The multimodality addressed by the current adaptive algorithm lies in the conditional posterior

distributions and is due to the point mass mixture priors. When this effect is encountered in combination with correlated predictors, the *joint* conditional posteriors of pairs or subsets of coefficients may be multimodal. For extreme such cases, no component-wise MCMC algorithm will mix well, including the adaptive one given here. Instead, joint updates of dependent parameters must be introduced. It would be relatively straightforward to extend the current algorithm to handle such situations. Currently the algorithm builds a proposal distribution which approximates the posterior by a product of univariate mixture distributions for each parameter. If significant posterior covariance exists among parameters, an alternative is to selectively replace component-wise proposals with joint proposals, using mixtures of bivariate or multivariate normals and tables of joint inclusion probabilities. The need for such additional terms can be identified by examining sample estimates of joint inclusion odds ratios $\hat{p}_{ij}/\hat{p}_i\hat{p}_j$ where $\hat{p}_{ij} = \frac{1}{n} \sum_{k=1}^n \delta_i^{(k)} \delta_j^{(k)}$, as well as estimated posterior covariances for non-zero values of the β_i and β_j 's. Note that this should be based on multiple MCMC runs from widely dispersed starting points (Gelman and Rubin, 1992), as lack of covariance in a single run may be misleading. However if significant multivariate posterior covariance exists, inclusion of higher dimensional proposal distributions will require an explosion of adaptive proposal parameters. In such cases orthogonalization of the predictor variables is to be preferred when possible. Nevertheless, extension of the adaptive sampling algorithm described here to handle such cases automatically is of interest.

Finally, it should be emphasized that all our results, while convincing, are empirical. The theory for adaptive MCMC algorithms still lags far behind that of even non-adaptive MCMC algorithms. In particular, recent results in this area (Andrieu et al., 2005; Andrieu and Moulines, 2006; Erland, 2003; Roberts and Rosenthal, 2007, 2006) provide weak laws of large numbers, but often do not yet provide central limit theorems or any information about convergence rates. Indeed, adaptive schemes may potentially *slow* convergence by adapting prematurely to local characteristic of the target distribution. For example, Schmieder and Woodard (2009) prove that several adaptive algorithms proposed in the literature

can converge no faster than their non-adaptive counterparts, at least in a formal computational complexity sense. Significant work remains to be done on *rates* of convergence for these adaptive schemes. Nevertheless, empirical results such as reported here for Bayesian variable selection demonstrate that adaptive MCMC algorithms have the potential to be a powerful addition to the Monte Carlo toolbox available to the applied Bayesian statistician.

Acknowledgments

SCS was partially supported by NSF grant DMS-0204690 (SCS). We thank Kai Mao for help with the kernel regression example.

References

- Albert, J. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88:669–679.
- Andrews, D. F. and Mallows, C. L. (1974). Scale mixtures of normal distributions. *J. Roy. Stat. Soc. B*, 36(1):99–102.
- Andrieu, C. and Moulines, E. (2006). On the ergodicity properties of some adaptive MCMC algorithms. *Annals of Applied Probability*, 16:1462–1505.
- Andrieu, C., Moulines, E., and Priouret, P. (2005). Stability of stochastic approximation under verifiable conditions. *SIAM Journal on Control and Optimization*, 44:283–312.
- Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive MCMC. *Statistics and Computing*, 18:343–373.
- Bae, K. and Mallick, B. (2004). Gene selection using a two level hierarchical Bayesian model. *Bioinformatics*, 20:3423–3430.

- Brockwell, A. E. and Kadane, J. B. (2005). Identification of regeneration times in MCMC, with application to adaptive schemes. *J. Comp. Graph. Stat.*, 14(2):436–458.
- Clyde, M. and George, E. I. (2004). Model uncertainty. *Statistical Science*, 19:81–94.
- Craiu, R. V., Rosenthal, J., and Yang, C. (2009). Learn from thy neighbor: Parallel-chain and regional adaptive mcmc. *J. Amer. Statist. Assoc.*, 104(488):1454–1466.
- Devroye, L. (1986). *Non-Uniform Random Variate Generation*. New York: Springer.
- Erland, S. (2003). *On Eigen-Decompositions and Adaptivity of Markov Chains*. PhD thesis, Norwegian University of Science and Technology.
- Figueiredo, M. and Jain, A. (2001). Bayesian learning of sparse classifiers. In *Proc. IEEE Int. Conference on Computer Vision and Pattern Recognition*.
- Gasemyr, J. (2003). On an adaptive version of the Metropolis-Hastings algorithm with independent proposal distribution. *Scand J. Stat.*
- Gelfand, A. E. and Sahu, S. K. (1994). On Markov chain Monte Carlo acceleration. *J. Comp. Graph. Stat.*, 3(3):261–276.
- Gelman, A., Roberts, G. O., and Gilks, W. R. (1995). Efficient Metropolis jumping rules. In Bernardo, J. M., Berger, J., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics 5*, pages 599–607, Oxford. Oxford University Press.
- Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Stat. Sci.*, 7:457–511.
- George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.*, 88:881–889.
- George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, 7:339–373.

- Geweke, J. (1996). *Variable selection and model comparison in regression*. Oxford Univ. Press, New York.
- Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., editors (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall.
- Gilks, W. R., Roberts, G. O., and Sahu, S. K. (1998). Adaptive Markov chain Monte Carlo through regeneration. *J. Amer. Statist. Assoc.*, 93(443):1045–1054.
- Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–32.
- Griffin, J. E. and Brown, P. J. (2005). Alternative prior distributions for variable selection with very many more variables than observations. Technical report, Dept. of Statistics, University of Warwick.
- Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli*, 7:223–242.
- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109.
- Holden, L., Hauge, R., and Holden, M. (2009). Adaptive independent Metropolis-Hastings. (to appear).
- Holmes, C. and Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1:145–168.
- Ji, C. (2006). Adaptive Monte Carlo methods for Bayesian inference. Master’s thesis, University of Cambridge, UK.
- Kushner, H. J. and Yin, G. G. (1997). *Stochastic approximation algorithms and applications*. Springer-Verlag, New York.

- Liang, F., Mao, K., Mukherjee, S., Liao, M., and West, M. (2006). Nonparametric Bayesian kernel models. Technical report, Duke University.
- Lucas, J. E. (2006). *Sparsity Modeling for High Dimensional Systems: Applications in Genomics and Structural Biology*. PhD thesis, Duke University.
- Mira, A. (2001). Ordering and improving the performance of Monte Carlo Markov chains. *Stat. Sci.*, 16(4):340–350.
- Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression (with discussion). *J. Amer. Statist. Assoc.*, 83:1023–1036.
- Poggio, T. and Girosi, F. (1990). Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, 247:978–982.
- Poland, D. and Scheraga, H. A. (1970). *Theory of Helix-Coil Transitions in Biopolymers: Statistical Mechanical Theory of Order-Disorder Transitions in Biological Macromolecules*. Academic Press.
- Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Ann. Math. Stat.*, 22:400–407.
- Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods*. New York: Springer-Verlag.
- Roberts, G. and Rosenthal, J. (2006). Examples of adaptive MCMC. *Preprint*. Preprint.
- Roberts, G. O. and Rosenthal, J. S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Stat. Sci.*, 16(4):351–367.
- Roberts, G. O. and Rosenthal, J. S. (2007). Coupling and ergodicity of adaptive MCMC. *J. Appl. Prob.*, 44:458–475.

- Schmidler, S. C., Lucas, J., and Oas, T. G. (2007). Statistical estimation in statistical mechanical models: Helix-coil theory and peptide helicity prediction. *J. Comp. Biol.*, 14(10):1287–1310.
- Schmidler, S. C. and Woodard, D. (2009). Lower bounds on the convergence of adaptive MCMC estimators. (submitted to *Annals of Statistics*).
- Scholkopf, B. and Smola, A. J. (2001). *Learning with Kernels: Support vector machines, regularization, optimization, and beyond*. Cambridge: The MIT Press.
- Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge: Cambridge Univ. Press.
- Tierney, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist.*, 22(4):1701–1728.
- Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.*, 1:211–244.
- Vapnik, V. (1998). *Statistical Learning Theory*. New York: Wiley.
- West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian Statistics*, 7:723–732.

APPDX: Derivation of Adaptive Updates for Mixtures

For mixture proposal distribution $q(x; \mathbf{w}, \Psi) = \sum_{m=1}^M w_m q_m(x; \psi_m)$ with $\mathbf{w} = \{w_1, \dots, w_M\}$ and $\Psi = \{\psi_1, \dots, \psi_M\}$ the parameters to be optimized, we wish to find optimal parameters (\mathbf{w}^*, Ψ^*) that minimize the KL-divergence $\mathcal{D}[\pi(x)||q(x; \mathbf{w}, \Psi)]$, or equivalently maximize the negative cross-entropy $\mathcal{H}(\mathbf{w}, \Psi) = \int \pi(x) \log q(x; \mathbf{w}, \Psi) \nu(dx)$. Taking partial derivatives with

respect to the w_i 's gives:

$$h_{w_i}(\mathbf{w}, \Psi) = \frac{\partial}{\partial w_i} \left[\int \pi(x) \log(q(x; \mathbf{w}, \Psi)) dx + \lambda \left(\sum_{i=1}^M w_i - 1 \right) \right] = \int \pi(x) \frac{q_i(x; \psi_i)}{q(X; \mathbf{w}, \Psi)} + \lambda$$

for Lagrange multiplier λ . As before, h_{w_i} involves an intractable integration w.r.t. $\pi(x)$, but can be estimated from the previous sample path X by $\hat{h}_{w_i} = \frac{1}{K} \sum_{k=1}^K H_{w_i}(X^{(k)}; \mathbf{w}, \Psi)$ where

$$H_{w_i}(X; \mathbf{w}, \Psi) = \frac{q_i(X; \psi_i)}{q(X; \mathbf{w}, \Psi)} + \lambda$$

Then using $\sum_{i=1}^M \hat{h}_{w_i}(x; \mathbf{w}^*, \Psi^*) = 0$ yields $\lambda = -\frac{1}{MK} \sum_{m=1}^M \sum_{k=1}^K \frac{q_m(X_{n+1}^{(k)}; \psi_m)}{q(X_{n+1}^{(k)}; \mathbf{w}, \Psi)}$, and for $K = 1$ we obtain the stochastic approximation recursive update for w_i :

$$w_{i,n+1} = w_{i,n} + r_{n+1} (O_i(X_{n+1}) - \bar{O}(X_{n+1}))$$

where X_{n+1} is the sample path at iteration $n + 1$, r_{n+1} is the step-size in the stochastic approximation algorithm, $O_i(X) = \frac{q_i(X; \psi_i)}{q(X; \mathbf{w}, \Psi)}$ and $\bar{O}(X) = \frac{1}{M} \sum_{i=1}^M O_i(X)$.

In Algorithm 3 the mixture also contains a point mass, in which case we instead get $\bar{O}(X) = \frac{1}{M+1} \sum_{m=0}^M O_i(X)$ where

$$O_i(X) = \begin{cases} \frac{1}{w_0} & \text{for } X = 0; \quad i = 0 \\ \frac{q_i(X; \psi_i)}{\sum_{m=1}^M w_m q_m(X; \psi_m)} & \text{for } X \neq 0; \quad i = 1, \dots, M \\ 0 & \text{otherwise} \end{cases}$$

Note these updates enforce $\sum_m w_m = 1$ but not $w_m \geq 0$. Rather than add slack variables to satisfy the Karush-Kuhn-Tucker conditions, we simply project back onto the unit simplex if weights become negative, as common in stochastic approximation (Kushner and Yin, 1997).

The partial derivatives of $\mathcal{H}(\mathbf{w}, \Psi)$ with respect to component parameters ψ_i are similar

$$\begin{aligned} h_{\psi_i}(\mathbf{w}, \Psi) &= \frac{\partial}{\partial \psi_i} \left[\int \pi(x) \log q(x; \mathbf{w}, \Psi) dx \right] \\ &= \int \pi(x) \frac{w_i \frac{\partial}{\partial \psi_i} [q_i(x; \psi_i)]}{q(X; \mathbf{w}, \Psi)} \\ &= \int \pi(x) \frac{w_i q_i(x; \psi_i)}{q(X; \mathbf{w}, \Psi)} \frac{\partial}{\partial \psi_i} \log q_i(x; \psi_i) = \int \pi(x) H_{\psi_i}(x; \mathbf{w}, \Psi) \end{aligned}$$

with associated sample path Monte Carlo estimate $\hat{h}_{\psi_i} = \frac{1}{K} \sum_{k=1}^K H_{\psi_i}(X^{(k)}; \mathbf{w}, \Psi)$.

For the specific case of q a normal mixture, we have $q_i(x; \psi_i) = \phi(x; \mu_i, \Sigma_i)$ and

$$\begin{aligned} H_{\mu_i}(X; \mathbf{w}, \Psi) &= \frac{w_i \phi(X; \psi_i)}{\sum_{m=1}^M w_m \phi(X; \psi_m)} \Sigma_i^{-1} (X - \mu_i) \\ H_{\Sigma_i}(X; \mathbf{w}, \Psi) &= \frac{w_i \phi(X; \psi_i)}{\sum_{m=1}^M w_m \phi(X; \psi_m)} \frac{1}{2} \left(\Sigma_i^{-1} (X - \mu_i) (\Sigma_i^{-1} (X - \mu_i))^T - \Sigma_i^{-1} \right) \end{aligned}$$

But since Σ_i is positive definite, the roots of H_{μ_i} and H_{Σ_i} are the same as those of

$$\begin{aligned} H'_{\mu_i}(X; \mathbf{w}, \Psi) &= \frac{w_i \phi(X; \psi_i)}{\sum_{m=1}^M w_m \phi(X; \psi_m)} (X - \mu_i) = 0 \\ H'_{\Sigma_i}(X; \mathbf{w}, \Psi) &= \frac{w_i \phi(X; \psi_i)}{\sum_{m=1}^M w_m \phi(X; \psi_m)} \left[(X - \mu_i) (X - \mu_i)^T - \Sigma_i \right] = 0 \end{aligned}$$

with the latter yielding the recursive updates (6-8) for $\psi_i = [\mu_i, \Sigma_i]$ given in Algorithm 2.

Taking $M = 1$ yields the updates (4-5) given in Algorithm 1. When including a point mass component, the same derivation gives all H_{ψ_i} multiplied by indicator $\mathbf{1}_{\{X \neq 0\}}$.

Similarly, when q is a mixture of gamma distributions (Section 4.2) we have $q_i(\rho; \psi_i) = \mathcal{G}(\rho; \alpha_i, \beta_i) = \frac{\beta_i^\alpha}{\Gamma(\alpha)} \rho^{\alpha-1} e^{-\beta_i \rho}$, and the associated derivatives become:

$$\begin{aligned} H_{\alpha_i}(\rho; \mathbf{w}, \Psi) &= \frac{w_i \mathcal{G}(\rho; \alpha_i, \beta_i)}{\sum_{m=1}^M w_m \mathcal{G}(\rho; \alpha_m, \beta_m)} (\log(\beta_i) - \psi(\alpha_i) + \log(\rho)) \\ H_{\beta_i}(\rho; \mathbf{w}, \Psi) &= \frac{w_i \mathcal{G}(\rho; \alpha_i, \beta_i)}{\sum_{m=1}^M w_m \mathcal{G}(\rho; \alpha_m, \beta_m)} \left(\frac{\alpha_i}{\beta_i} - \rho \right) \end{aligned}$$

where $\psi(\alpha)$ denotes the digamma function, yielding the updates (10-11) of Section 4.2.