Bayesian Analysis (2004) **1**, Number 1, pp. 1–17

On Numerical Aspects of Bayesian Model Selection in High and Ultrahigh-dimensional Settings

Valen E. Johnson [*]

Abstract. This article examines the convergence properties of a Bayesian model selection procedure based on a non-local prior density in ultrahigh-dimensional settings. The performance of the model selection procedure is also compared to popular penalized likelihood methods. Coupling diagnostics are used to bound the total variation distance between iterates in an Markov chain Monte Carlo (MCMC) algorithm and the posterior distribution on the model space. In several simulation scenarios in which the number of observations exceeds 100, rapid convergence and high accuracy of the Bayesian procedure is demonstrated. Conversely, the coupling diagnostics are successful in diagnosing lack of convergence in several scenarios for which the number of observations is less than 100. The accuracy of the Bayesian model selection procedure in identifying high probability models is shown to be comparable to commonly used penalized likelihood methods, including extensions of smoothly clipped absolute deviations (SCAD) and least absolute shrinkage and selection operator (LASSO) procedures.

Keywords: MCMC algorithm, convergence diagnostic, coupling, SCAD, sure independence screening, penalized likelihood, variable selection

1 Introduction

Non-local prior densities were proposed by Johnson and Rossell (2012, JR12) as a mechanism to improve the performance of Bayesian model selection procedures in high-dimensional settings. The distinguishing feature of non-local prior densities on regression parameters is that such densities are identically zero whenever any component of the parameter vector equals its null value (typically 0). For variable selection in linear models, if p denotes the number of potential regressors and n the number of observations, JR12 showed that Bayesian model selection procedures based on non-local prior densities are model consistent if $p < n$ as n increases (i.e., the posterior probability of the true model converges to 1 as $n \to \infty$), provided that the design matrix satisfies certain regularity constraints. They also proposed a Markov chain Monte Carlo (MCMC) algorithm that could be used to sample from the posterior distribution on the model space. Further details regarding this method and the underlying model are provided in Section 2.

The MCMC algorithm proposed in JR12 proceeds by sequentially inserting or deleting individual regressors from the model based on comparisons of posterior model prob-

[*] Department of Statistics, Texas A&M University, vjohnson@stat.tamu.edu

abilities. Although the consistency results cited in JR12 do not extend to cases for which $p > n$, the MCMC algorithm proposed in JR12 can nevertheless be applied in such settings. It thus provides a potential alternative to the Sure Independence Screening (SIS) and Iterated SIS (ISIS) algorithms that are currently used in conjunction with penalized likelihood variable selection techniques (e.g., Fan and Lv (2008, 2010)). In contrast to the MCMC algorithm, SIS uses a screening procedure to identify subsets of components of the regression vector that contain fewer than n regressors. Penalized likelihood methods are then applied to these subsets to perform model selection. The ISIS algorithm is an iterative version of SIS in which subsets of regression components are iteratively considered for inclusion in a model that already contains regressors selected in previous SIS updates. Both SIS and ISIS methods have demonstrated substantial success in identifying important covariates in $p \gg n$ settings.

In this article, I examine the feasibility of extending the MCMC algorithm proposed in JR12 to ultrahigh dimensional settings. Two issues arise in making this extension: (i) evaluating the convergence properties of the resulting MCMC algorithms, and (ii) assessing the effectiveness of the algorithms in finding high probability models.

Convergence issues are of most concern when there are high correlations between columns of the design matrix. In practice, this problem tends to be most severe when there are groups of regressors that are highly correlated both with the response vector and other members of their group. Once one of these regressors is added to the current state of the MCMC chain, it can difficult for another regressor from the same group to also be included in the model. Because it is difficult for the chain to transition to a state that contains no members from the group, it then becomes difficult for the MCMC chain to transition between models that contain only one of the highly correlated regressors. To ameliorate this difficulty, I propose a modification of the MCMC algorithm proposed in JR12 that includes a "swap" step.

The convergence diagnostics studied in this article use coupling methods to obtain approximate bounds on the total variation distance (TVD) between the distribution of models sampled from the MCMC algorithm and the posterior distribution (e.g., Lindvall (1992); Johnson (1996, 1998)). Perhaps surprisingly, these diagnostics suggest that the distributions of iterates from a single MCMC chain often differ from the target distribution by less than 0.05 in TVD after only a few complete updates of the parameter vector. In fact, fewer than five updates are enough to achieve this level of convergence in several of the simulation studies considered below. In other settings, however, the coupling diagnostics show that the MCMC algorithm fails to converge even after several thousand updates of the parameter vector. Importantly, the proposed diagnostic provides a simple mechanism for determining whether a given chain is converging quickly, slowly, or at an intermediate rate. It also provides an estimate of how many updates are required to obtain what are essentially independent draws from the posterior distribution on the model space.

To assess the effectiveness of the proposed MCMC scheme and associated non-local prior specification for determining high probability models, several of the numerical experiments reported in Fan and Lv (2008, FL08) are repeated using Bayesian model selection procedures based on non-local prior densities in conjunction with this MCMC algorithm. In each of these experiments, the number of covariates p is much larger than the number of observations n. Because the numerical results obtained by Fan and Lv using sure independence screening-smoothly clipped absolute deviations (SIS-SCAD) and iterative sure independence screening-smoothly clipped absolute deviations (ISIS-SCAD) have proven to be among the most successful model selection procedures used in practice, they provide a useful benchmark for assessing model selection procedures in ultrahigh-dimensional settings. Simulation studies reported in Section 3 suggest that the Bayesian procedure, when implemented with the modified MCMC algorithm proposed in Section 2, is competitive with the SIS-SCAD and ISIS-SCAD procedures in identifying the correct model, at least for this set of simulation studies.

The remainder of this article is organized as follows. In the next section, I describe coupling algorithms that can be used to evaluate the convergence of MCMC algorithms for model selection. Section 3 illustrates the application of this coupling algorithm in several simulated data sets and provides a comparison of the performance of the Bayesian variable selection procedure proposed in JR12 to penalized likelihood methods based on SIS-SCAD and ISIS-SCAD in FL08. Concluding comments and discussion appear in Section 4.

2 Coupling diagnostics for Bayesian variable selection

To fix notation, let $\mathbf{k}$ denote a statistical model indexed by a p dimensional parameter vector $\boldsymbol{\beta}$. I assume that a component $\boldsymbol{\beta}_j$, $1 \leq j \leq p$, is excluded from a model if its value is 0, and I denote a model by $\mathbf{j} = \{j_1, \ldots, j_k\}$, $(1 \leq j_1 < \cdots < j_k \leq p)$ if and only if $\beta_{j_1} \neq 0, \cdots, \beta_{j_k} \neq 0$ and all other elements of $\boldsymbol{\beta}$ are 0. The number of non-zero components in model $\mathbf{j}$ is denoted by $|\mathbf{j}|$. The regression parameter associated with model $\mathbf{j}$ is denoted by $\boldsymbol{\beta}_{\mathbf{j}} = (\beta_{j_1}, \ldots, \beta_{j_{|\mathbf{j}|}})'$, and $\mathcal{K}$ denotes the set of 2^p possible models that are indexed by the p components of $\boldsymbol{\beta}$. Model $\mathbf{j} \cup i$ denotes model $\{j_1, \ldots, j_k\} \cup \{i\}$, the model obtained by adding the i^{th} component of the parameter vector $\boldsymbol{\beta}$ to $\boldsymbol{\beta}_{\mathbf{j}}$. Similarly, $\mathbf{j} \setminus i$ denotes the model obtained by excluding β_i from $\boldsymbol{\beta}_{\mathbf{j}}$.

The prior distribution on the model space is denoted by $\pi_{\mathcal{K}}$, and the posterior distribution on the model space, based on a data vector $\mathbf{y}$ (assumed to be generated by a "true" model $\mathbf{t} \in \mathcal{K}$), is denoted by $p_{\mathcal{K}}$. The marginal density of the data $\mathbf{y}$ under model $\mathbf{k}$ is denoted by $m_{\mathbf{k}}(\mathbf{y})$. In this article, a statistical model $\mathbf{k}$ refers to a sampling density for the data $\mathbf{y}$ and a proper prior density on the model parameter $\boldsymbol{\beta}_{\mathbf{k}}$.

The class of prior densities imposed on $\boldsymbol{\beta}$ in this article consists of the product moment (pMOM) densities, which are defined in JR12 by the equation

$$\pi(\boldsymbol{\beta} \,|\, \tau, \sigma^2, r) = d_p (2\pi)^{-p/2} (\tau\sigma^2)^{-rp-p/2} |\mathbf{A}_p|^{1/2} \exp\left[-\frac{1}{2\tau\sigma^2} \boldsymbol{\beta}' \mathbf{A}_p \boldsymbol{\beta}\right] \prod_{i=1}^{p} \boldsymbol{\beta}_i^{2r}, \qquad (1)$$

where $\tau > 0$, $\mathbf{A}_p$ is a $p \times p$ non-singular scale matrix, and $r = 1, 2, \ldots$. The normalizing constant d_p is independent of σ^2 and τ. The parameter r is called the order of the density, and in this manuscript is assigned the fixed value $r = 1$. Throughout this article, $\mathbf{A}_p$ is assumed to be the $p \times p$ identity matrix. Posterior model probabilities are calculated according to the formula

$$p_{\mathcal{K}}(\mathbf{k} \,|\, \mathbf{y}) = \frac{m_{\mathbf{k}}(\mathbf{y})\pi_{\mathcal{K}}(\mathbf{k})}{\sum_{\mathbf{j}\in\mathcal{K}} m_{\mathbf{j}}(\mathbf{y})\pi_{\mathcal{K}}(\mathbf{j})}.$$

Marginal densities $m_{\mathbf{j}}(\mathbf{y})$ are estimated using Laplace approximations. JR12 describe a Metropolis-Hastings (MH) algorithm that can be used to sample from the posterior distribution.

The coupling diagnostics proposed in this article are applied to the following modification of the MH algorithm proposed by JR12 for sampling from the model space:

Metropolis Hastings algorithm (MH)

1. Draw $\mathbf{k}^0$ from initialization distribution W on the model space $\mathcal{K}$, and set $t = 1$.
2. Set $\mathbf{k}^* = \mathbf{k}^{t-1}$. Draw $S_p^t = (h_1^t, \dots, h_p^t)$ as a random permutation of the integers $1, \dots, p$. For $j = h_1^t, \dots, h_p^t$,
   (a) Define $\mathbf{k}^{cand} = \mathbf{k}^* \setminus j$ if $\mathbf{k}^*$ includes β_j. Otherwise, define $\mathbf{k}^{cand} = \mathbf{k}^* \cup j$.
   (b) Draw $u \sim U(0,1)$ and define
   $$r = \frac{m_{\mathbf{k}^{cand}}(\mathbf{y})\pi(\mathbf{k}^{cand})}{m_{\mathbf{k}^*}(\mathbf{y})\pi(\mathbf{k}^*)}. \tag{2}$$
   (c) If $r > u$, set $\mathbf{k}^* = \mathbf{k}^{cand}$.
3. Swap step: If $t \bmod 5 = 0$,
   (a) For $j' = 1, \dots, p-1$ and for $k' = j+1, \dots, p$, let $j = h_{j'}^t$ and $k = h_{k'}^t$. Determine if exactly one of β_j and β_k is in the model. If so, define $\mathbf{k}^{cand}$ to be the model obtained by switching the inclusion status of β_j and β_k.
   (b) Draw $u \sim U(0,1)$ and define
   $$p = \frac{m_{\mathbf{k}^{cand}}(\mathbf{y})\pi(\mathbf{k}^{cand})}{m_{\mathbf{k}^{cand}}(\mathbf{y})\pi(\mathbf{k}^{cand}) + m_{\mathbf{k}^*}(\mathbf{y})\pi(\mathbf{k}^*)}. \tag{3}$$
   (c) If $p > u$, set $\mathbf{k}^* = \mathbf{k}^{cand}$.
4. Increment t and return to step 2.

Provided that it is possible to move between any two models in the discrete space $\mathcal{K}$, this sampler produces a chain of models $\mathbf{k}^0, \mathbf{k}^1, \ldots$ that converges to the posterior distribution on the model space. The use of a Gibbs rather than a Metropolis update in Step 3 of the algorithm is discussed below.

It is important to note that switching the inclusion status of variables in Step 3 of the algorithm requires $O(p \times |\mathbf{k}^*|)$ model updates. This number of updates may not be computationally feasible if the number of covariates included in the current model is large. It is for this reason that I recommend making this pass through the model space only infrequently within the MH algorithm, in this case only after every five repetitions of Step 2.

To apply coupling diagnostics to this MH algorithm, introduce a second chain, say $\mathbf{j}^0, \mathbf{j}^1, \dots$ that is updated synchronously with the first chain. That is, suppose $\mathbf{k}^T$ is the model sampled after T updates from algorithm above. In the coupling version of the MH algorithm, both $\{\mathbf{j}^s\}_{s=0}^S$ and $\{\mathbf{k}^{T+s}\}_{s=0}^S$ are updated according to the following modification of the MH algorithm.

Coupled MH algorithm (CMH)

1. Draw $\mathbf{j}^0$ from W, and set $s = 1$.
2. Set $\mathbf{k}^* = \mathbf{k}^{T+s-1}$ and $\mathbf{j}^* = \mathbf{j}^{s-1}$. Draw $S_p^t = (h_1^t, \dots, h_p^t)$ as a random permutation of the integers $1, \dots, p$. For $j = h_1^t, \dots, h_p^t$,
   (a) Define $\mathbf{k}^{cand} = \mathbf{k}^* \setminus j$ if $\mathbf{k}^*$ includes β_j. Otherwise define $\mathbf{k}^{cand} = \mathbf{k}^* \cup j$. Similarly, if $\mathbf{j}^*$ includes β_j, define $\mathbf{j}^{cand} = \mathbf{k}^* \setminus j$. Otherwise define $\mathbf{j}^{cand} = \mathbf{j}^* \cup j$.

   (b) Draw $u_1 \sim U(0, 1)$. If $\mathbf{j}^*$ and $\mathbf{k}^*$ either both contain β_j or both exclude β_j, set $u_2 = u_1$. Otherwise, define $u_2 = 1 - u_1$.

   (c) Define
   $$r_1 = \frac{m_{\mathbf{k}^{cand}}(\mathbf{y})\pi(\mathbf{k}^{cand})}{m_{\mathbf{k}^*}(\mathbf{y})\pi(\mathbf{k}^*)} \tag{4}$$
   and
   $$r_2 = \frac{m_{\mathbf{j}^{cand}}(\mathbf{y})\pi(\mathbf{j}^{cand})}{m_{\mathbf{j}^*}(\mathbf{y})\pi(\mathbf{j}^*)}. \tag{5}$$

   (d) If $r_1 > u_1$, set $\mathbf{k}^* = \mathbf{k}^{cand}$. If $r_2 > u_2$, set $\mathbf{j}^* = \mathbf{j}^{cand}$.
3. Swap step: If $t \bmod 5 = 0$,
   (a) For $j' = 1, \dots, p-1$ and for $k' = j+1, \dots, p$, let $j = h_{j'}^t$ and $k = h_{k'}^t$. Determine if exactly one of β_j and β_k is in model $\mathbf{k}^*$. If so, define $\mathbf{k}^{cand}$ to be the model obtained by switching the inclusion status of β_j and β_k. Otherwise, do not update $\mathbf{k}^*$. Similarly, determine if exactly one of β_j and β_k is in model $\mathbf{j}^*$. If so, define $\mathbf{j}^{cand}$ to be the model obtained by switching the inclusion status of β_j and β_k. Otherwise, do not update $\mathbf{j}^*$.

   (b) Draw $u_1 \sim U(0, 1)$. If both $\mathbf{k}^*$ and $\mathbf{j}^*$ are being updated and agree on the inclusion of β_j, or if only $\mathbf{j}^*$ is being updated, define $u_2 = u_1$. Otherwise, define $u_2 = 1 - u_1$. Define
   $$p_1 = \frac{m_{\mathbf{k}^{cand}}(\mathbf{y})\pi(\mathbf{k}^{cand})}{m_{\mathbf{k}^{cand}}(\mathbf{y})\pi(\mathbf{k}^{cand}) + m_{\mathbf{k}^*}(\mathbf{y})\pi(\mathbf{k}^*)} \tag{6}$$

and

$$p_2 = \frac{m_{\mathbf{j}^{cand}}(\mathbf{y})\pi(\mathbf{j}^{cand})}{m_{\mathbf{j}^{cand}}(\mathbf{y})\pi(\mathbf{j}^{cand}) + m_{\mathbf{j}^*}(\mathbf{y})\pi(\mathbf{j}^*)}. \tag{7}$$

(c) If $p_1 > u_1$, set $\mathbf{k}^* = \mathbf{k}^{cand}$. If $p_2 > u_2$, set $\mathbf{j}^* = \mathbf{j}^{cand}$.

4. If $\mathbf{k}^{t+s} = \mathbf{j}^s$, set $S = s$ and exit. Otherwise, set $\mathbf{k}^{T+s} = \mathbf{k}^*$, $\mathbf{j}^s = \mathbf{j}^*$, increment s and return to step 2.

The operation of this coupling scheme, which represents a Metropolis-Hastings version of the coupling scheme proposed in Johnson (1998), can be understood by examining the outcomes of component updates when $\mathbf{k}^{cand}$ and $\mathbf{j}^{cand}$ agree or differ in their inclusion of the parameter β_j being updated.

If the two chains agree on the inclusion or exclusion of β_j, then the use of a common uniform deviate u_1 to update both chains is likely to result in both chains either accepting the candidate draw or both chains rejecting it. Indeed, if the chains are identical before the update, they will remain so afterwards and are said to have coupled. The number of parameter updates required to obtain coupled chains, S, provides a convergence diagnostic for the chain.

Conversely, if the two chains disagree on the inclusion or exclusion of β_j, then the use of u_1 and $u_2 = 1 - u_1$ in the acceptance step of the algorithm will increase the probability that one chain rejects the candidate draw while the other accepts it. If this event occurs, then the chains will agree in their inclusion or exclusion of β_j after the update. This coupled update procedure thus encourages the two chains to move closer to a coupled state.

The swap step requires a slight modification of the acceptance probability in (7) to encourage this type of coupling. To see why, note that if the probability that the two model configurations obtained by switching the inclusion status of two variables in one chain are exactly equal, then variables will always be switched according to (7). Thus, if the primary and secondary chains disagree before the swap step, they will disagree afterwards. It is for this reason that a Gibbs update, rather than a standard Metropolis-Hastings update, is performed in this step of the algorithm. If a Gibbs update is proposed and the primary and secondary chains agree on the inclusion of the variables prior to the proposed swap, then the same uniform deviate is used to update both chains. That is, if $u \sim U(0,1)$, then the variables in each chain are swapped if $p_1 > u$ and $p_2 > u$. If the two chains disagree on the inclusion status of the swapped variables before the proposed switch, then the variables in the second chain are swapped if $p_2 > 1 - u$. This represents the maximal coupling procedure for this step in the algorithm (Lindvall (1992)).

Convergence diagnostics for the MH algorithm follow from the coupling inequality (e.g., Lindvall (1992)), which states that

$$\mathbf{P}[s > S] \geq ||\mathcal{L}(\mathbf{j}^S) - \mathcal{L}(\mathbf{k}^{T+S})||, \tag{8}$$

where $||\mathcal{L}(\mathbf{a}) - \mathcal{L}(\mathbf{b})||$ denotes the TVD between the distributions of the random vectors $\mathbf{a}$ and $\mathbf{b}$.

If the MH algorithm is run for a sufficiently large number of updates, then $\mathbf{k}^T$ and $\mathbf{k}^{T+S}$ represent draws from the posterior distribution on the model space. Provided that T is large enough, this means that (8) provides a bound on the TVD distance between the S^{th} iterate in the chain (initialized from distribution W) and the posterior distribution.

In practice, the use of the coupling inequality to obtain a bound on the TVD between iterates in the chain and the stationary distribution requires an estimate of the number of updates T for the first chain $\{\mathbf{k}^t\}$ to reach its stationary distribution, as well as an empirical estimate of $\mathbf{P}[s > S]$, the coupling time distribution. I propose the following lead-in procedure to obtain preliminary estimates of both quantities.

First, an initial estimate of the burn-in period T is specified, and a preliminary run of the CMH algorithm is performed for that number of iterations. As soon as the two chains $\mathbf{k}$ and $\mathbf{j}$ couple, the “auxiliary” chain $\mathbf{j}$ is restarted at a random draw from distribution W. In implementing this procedure, it is important to define W so that it represents a disperse distribution on the model space $\mathcal{K}$, meaning that random draws from W are likely to fall within the domains of attraction of different posterior modes when multiple modes are likely to exist. During the preliminary run of the CMH algorithm, both the number of couplings and the lengths of time required to obtain each coupling are recorded and used to plan the inferential run of the MCMC algorithm.

In the examples that follow, W was defined so that each regressor had a small, independent probability of being included in the initial models. To insure both that neither the null model nor inappropriately large models were assigned too much prior probability, the prior probability that each variable was sampled was arbitrarily set to $q = 8/p$. The resulting probability mass function of W can be expressed as

$$w(\mathbf{k}) = q^{|\mathbf{k}|}(1-q)^{p-|\mathbf{k}|}.$$

If a large number of chains initialized from this distribution have converged to the same primary chain $\mathbf{k}$ when it is run for T iterations, then it is likely that the primary chain has reached the stationary distribution (Johnson 1996). Conversely, if numerous couplings have not occurred, then it is probable that convergence has not occurred within T iterations. In this case, a larger value of T should be selected and the lead-in procedure should be repeated.

At the completion of the lead-in procedure, a “restart” interval I is defined to be some multiple of the maximum coupling time observed during lead-in. A factor of three was used in the numerical studies reported in Section 3. Using this restart interval, a second run of the CMH algorithm is performed, but in the second run the auxiliary chain $\mathbf{j}$ is re-initialized after every I updates of the primary chain, whether or not it has coupled with the primary chain $\mathbf{k}$. If the auxiliary chain has not coupled with the primary chain, then a right-censored coupling time of I is recorded. It is important to choose the interval I so that coupling occurs with high probability within I updates.

Otherwise, the bounds on the TVD between the target distribution and iterates in the MCMC algorithm prescribed in (8) will not be sufficiently tight for practical use, and there will tend to be a high correlation between the coupling times used to estimate $P(S > s)$.

By re-initializing the secondary chain at fixed intervals, biases in the distribution of the coupling distribution that might result from more frequent restarts when the primary chain was near the posterior mode (or other high probability states) can be avoided.

The coupling times obtained using the fixed restart intervals are then used to obtain empirical estimates of the coupling survival function $P(S > s)$. This estimate of the survival function provides an estimate of the bound on the TVD distance between the s^{th} iterate in a chain initialized from distribution W and the stationary distribution based on equation (8).

The quantity $P(S > s)$ also provides an estimate the number of updates required to obtain what are essentially independent draws of the target distribution after the burn-in period. Letting $p_{\mathcal{K}} \times p_{\mathcal{K}}$ denote the distribution of two independent draws from the stationary distribution of the chain and $q^T \times q^{T+r}$ the distribution of two updates separated by r updates in the MH algorithm, it follows that (Johnson 1998)

$$\mathbf{P}(S > r) \geq ||p_{\mathcal{K}} \times p_{\mathcal{K}} - q^T \times q^{T+r}||. \tag{9}$$

By choosing r large enough so that $\mathbf{P}(S > r)$ is small, this inequality provides a bound on the number of updates in the MCMC algorithm that are required to obtain draws that would, with high probability, be accepted as independent draws from the posterior distribution in an acceptance sampling scheme.

3 Applications

In order to evaluate the performance of coupling diagnostics for assessing convergence of the MH algorithm, they were used to assess the convergence of this algorithm for variable selection based on non-local prior densities. To facilitate comparisons with the SIS and ISIS algorithms, the diagnostics were applied to a number of simulation studies presented in FL08. These studies focused on variable selection for linear models having sampling densities of the form

$$\mathbf{y} \,|\, \boldsymbol{\beta}_{\mathbf{k}}, \sigma^2 \sim N_n(\mathbf{X}_{\mathbf{k}}\boldsymbol{\beta}_k, \sigma^2\mathbf{I}_n). \tag{10}$$

Throughout the remainder of this article, the prior densities for the parameters appearing in (10) are assumed to be expressible as

$$\sigma^2 \sim IG(10^{-3}, 10^{-3}), \qquad \pi(\mathbf{k}) \propto B(k+a, p-k+b), \tag{11}$$

$$\pi(\boldsymbol{\beta}_{\mathbf{k}} \,|\, \tau, \sigma^2) = (2\pi)^{-k/2}(\tau\sigma^2)^{-3k/2} \exp\left[-\frac{1}{2\tau\sigma^2}\boldsymbol{\beta}'_{\mathbf{k}}\boldsymbol{\beta}_{\mathbf{k}}\right] \prod_{i=1}^{k} \boldsymbol{\beta}^2_{\mathbf{k}_i}. \tag{12}$$

In these equations, $IG(\cdot,\cdot)$ denotes an inverse gamma distribution, $B(\cdot,\cdot)$ denotes the beta function, $\mathbf{X_k}$ denotes the design matrix containing those covariates that correspond to $\boldsymbol{\beta}_\mathbf{k}$, $\mathbf{I}_n$ is an $n \times n$ identity matrix, and τ is a prior hyperparameter. The prior model assumed for $\boldsymbol{\beta}_k$ is a particular case of the first-order pMOM prior proposed in JR12, and the prior on the model space is the beta-binomial prior proposed by Scott and Berger (2010). The default values of a and b recommended by Scott and Berger are $a = b = 1$. However, these hyperparameter values assign the same prior mass to the null and saturated models when $p = n$, and lead to the same paradox confronting empirical Bayes methods when the marginal density of the data is maximized at the saturated model (c.f. Lemma 4.1, Scott and Berger (2010)). To avoid this paradox, the values of a and b were fixed at 1 and 20, respectively, in the examples that follow. Loosely speaking, this means that each variable was assigned a 1/20 chance of being included in a model *a priori*. The value of τ was fixed at 2.85, which corresponds to the assignment of 0.05 prior mass to values of each regression coefficient that are less than σ in magnitude in settings for which the columns of the design matrix have been standardized. This value of τ is larger than the value of τ suggested in JR12 for $p < n$ settings, in which standardized regression coefficients greater than 0.2 in magnitude were sought. The modes of the corresponding standardized pMOM prior density (i.e., when $\sigma^2 = 1$) occur at ± 2.39, which means that the posterior mean of regression coefficients will tend to be shrunk toward one of these modes. In practice, the identification of the maximum *a posteriori* (MAP) model does not seem to be highly sensitive to the choice of τ, although the choice of τ does affect the bias of the posterior mean, particularly for small n.

3.1 Study 1

The first set of simulation studies was patterned after the simulations presented in Section 3 of FL08 and involved five parameter settings (i.e., simulation truths). In the first two settings, the elements of the design matrix were generated as independent standard normal variables. For the final three settings, the columns of the design matrix $\mathbf{X}$ were correlated.

Two sets of simulation parameters were tested for the independent design. In the first case, $n = 200$, $p = 1000$, and the dimension of the true model t was 8. The values of the non-zero regression coefficients were independently set to $(-1)^u(c\log(n)/\sqrt{n} + |z|)$, where u was a Bernoulli random variable with success probability 0.4, z was a standard normal deviate, and $c = 4$. The second case was defined by taking $n = 800$, $p = 20000$, $t = 18$, and $c = 5$. The observational variance was chosen to be $\sigma^2 = 1.5^2$ in both cases.

For the dependent designs, three additional scenarios were defined by taking $(n, p, t, c, \sigma) = (200, 1000, 5, 2, 1.0)$, (200,1000,8,4,1.5), and (800,20000,14,4, 2.0), respectively. To generate correlated design matrices, the columns of $\mathbf{X}$ were generated by first simulating a $t \times t$ symmetric matrix $\mathbf{A}$ with condition number $n^{1/2}\log(n)$ and a sample of t predictors $X_1, \ldots, X_t \sim N(0, \mathbf{A})$. Next, $p - t$ vectors were simulated according to $Z_{t+1}, \ldots, Z_p \sim N(0, \mathbf{I}_{p-t})$; these vectors were then used to define the remaining predictors as $X_i = Z_i + rX_{i-t}$, $i = t+1, \ldots, 2t$, and $X_i = Z_i + (1-r)X_1$, $i = 2t+1, \ldots, p$, where $r = 1 - d\log(n)/n$. The values of d used in the three scenarios were $d = 4$, 4, and

5, respectively (FL08).

It is important to note that the columns of the design matrix **X** were not standardized in the dependent scenarios. Ideally, this standardization should be performed so that τ can be interpreted in terms of standardized regression coefficients. However, the columns of the design matrix were not standardized in FL08, even though a common tuning parameter λ was used for all regression coefficients. To maintain comparability across studies, the columns of the design matrix were therefore not standardized here, either.

Two hundred data sets were generated under each of the five scenarios. Before conducting the coupling experiment for each data set, the CMH algorithm was run for 100 iterations. During this burn-in period, the auxiliary chain $\mathbf{j}$ was reinitialized immediately after coupling with the primary chain $\mathbf{k}$. The coupling times observed during burn-in were then used to establish the re-initialization period I for the coupling experiment performed on each data set. These re-initialization periods were defined to be the maximum of 150 and three times the longest coupling times observed during burn-in.

Following burn-in, the CMH algorithm was used to obtain 20 coupling times for each data set. The primary chain was initialized with its value at the end of the burn-in phase. The auxiliary chains were reinitialized every I iterations from W by independently including each regressor in the chain with probability $8/p$. If a chain did not couple before I iterations, a censored coupling time was recorded.

The survival functions for the coupling times obtained under the five simulation scenarios are displayed in Figure 1. Under all five scenarios, coupling occurred within 5 iterations with probability exceeding 0.995. In practical terms, this means that 5 iterations are enough for a chain initialized from W to reach the target distribution in these simulation scenarios, and that essentially independent draws from the posterior distribution are obtained after every 5 iterations.

The speed with which coupling occurred in these experiments is somewhat surprising, and may be be attributed to the fact that the posterior distribution is relatively concentrated for values of $n = 200$ or 800. The next two studies illustrate scenarios in which convergence was substantially slower.

The accuracy of the Bayesian variable selection algorithm in identifying the true model compared well with the penalized likelihood methods proposed by FL08. Estimation errors for the Bayesian variable selection routine, along with the SIS-SCAD and ISIS-SCAD procedures from FL08, are reported in Table 1. In all five scenarios, the median model size identified by each method is listed in the row labeled “t”. The median model size estimated by the pMOM procedure matched the true model size in each scenario. In fact, the MAP model estimated by the pMOM procedure corresponded to the simulation truth with probability 0.99 or greater under all five scenarios. The rows labeled “MEE” provide the median estimation error for each method; standard errors of estimation based on the 200 simulated data sets are provided in parentheses. The MEE values listed for the “LSE-Truth” correspond to the least squares estimates obtained

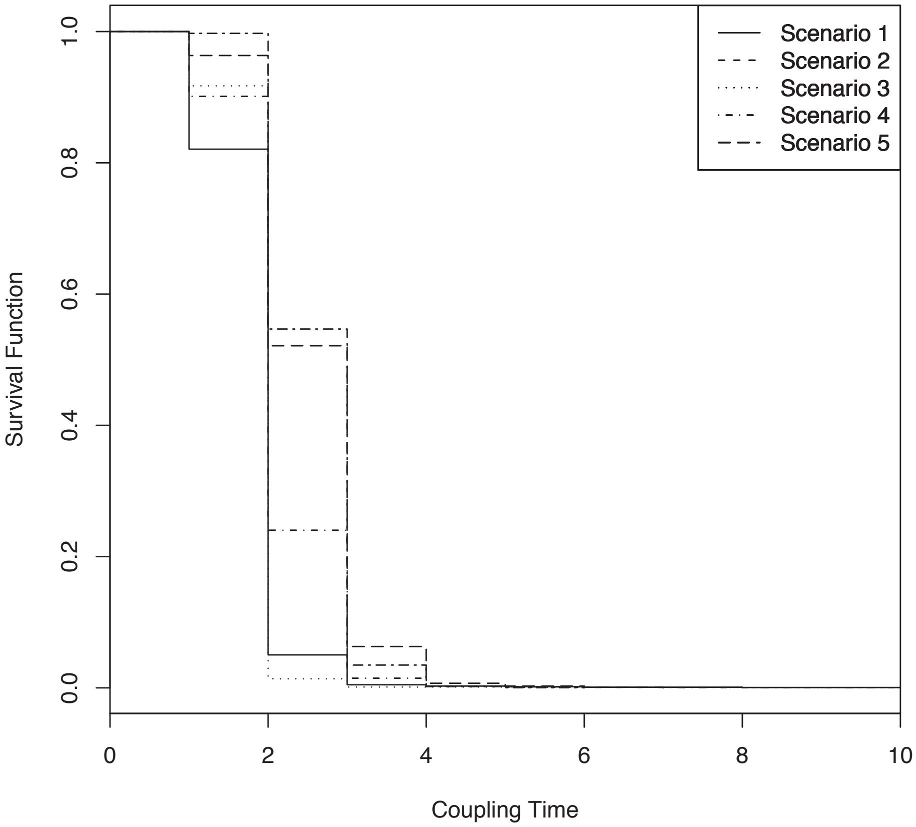

Figure 1: Survival function of coupling times for simulation scenarios in Study 1.

under the true model. Because of the high probability assigned to the true model by the pMOM procedure, there is close agreement between the LSE-Truth values and that procedure. In general, the MEE reported in FL08 for SIS-SCAD and ISIS-SCAD appear to be comparable in magnitude to the values obtained using the Bayesian procedure. From the standard deviations of the MEE values reported for the MAP model in this table, it is clear that substantially larger simulation studies would be required to determine which of the methods provided the smallest MEE for these simulation scenarios.

3.2 Study 2

In the next simulation study, linear models of the form (10) were again considered, but now the true model size was fixed at 3 and the regression coefficient was set to $\boldsymbol{\beta} = (5, 5, 5, 0, \ldots, 0)'$. Each row of the design matrix $\mathbf{X}$ was independently drawn from a multivariate normal distribution having mean 0 and covariance matrix Σ with

Scenario		1	2	3	4	5
	n	200	800	200	200	800
	p	1000	20,000	1000	1000	20000
	t	8	18	5	8	14
LSE-Truth	MEE	0.30 (0.08)	0.22 (0.04)	0.21 (0.07)	0.38 (0.10)	0.40 (0.08)
pMOM	med(t)	8	18	5	8	14
	MEE	0.30 (0.08)	0.22 (0.04)	0.19 (0.08)	0.39 (0.10)	0.37 (0.09)
SIS-SCAD	med(t)	15	37	21	18	36
	MEE	0.37	0.29	0.33	0.46	0.37
ISIS-SCAD	med(t)	13	31	11	13.5	27
	MEE	0.33	0.25	0.22	0.37	0.32

Table 1: Summary statistics for the MAP models and penalized likelihood methods in ultrahigh dimensional settings for Study 1. The median model sizes, med(t), listed for the Bayesian model correspond to the true model sizes in all studies. MEE denotes median estimation error, or the median of $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||$ across datasets. For the penalized likelihood methods, $\hat{\boldsymbol{\beta}}$ represents the penalized MLE; for the Bayesian methods it represents the posterior mean of $\boldsymbol{\beta}$ obtained under the MAP model. Summaries for SIS-SCAD and ISIS-SCAD are taken from FL08; standard errors of estimation are not available for these summaries but are likely similar to those reported for the pMOM estimates.

diagonal entries equal to 1 and off-diagonal entries equal to 0.5. Five combinations of (n, p) were considered: (20,100), (50,100), (20,1000), (50,1000), and (70,1000). These scenarios mimic simulation studies reported in Section 4.2.1 of FL08. Two hundred data sets were simulated under each scenario.

The lead-in experiments conducted for Scenario 3 indicated that coupling of the auxiliary chain with the primary chain did not occur within 100 parameter updates for a majority of simulated data sets. For this reason, the CMH algorithm described in Study 1 was modified so that 2,000 burn-in iterations were performed for each simulated data set under this scenario. Following burn-in of the primary chain, the restart interval for the auxiliary chains was set to 500 iterations, and 20 auxiliary chains were run for each simulated data set. The coupling procedures described in Study 1 were applied to all other scenarios.

Figure 2 displays the estimates of the survival distributions obtained for the coupling times under each of the five scenarios. As is evident from this figure, the coupling times obtained under Scenario 3 suggest that convergence of the MH algorithm for these data sets required far more than 500 iterations. Indeed, by extrapolating the extreme tail of its survival curve, it appears that more than 5,000 parameter updates are likely to be needed to achieve burn-in of the MH algorithm under this scenario.

The convergence of the MH algorithm under Scenario 1, the other scenario with a sample size of 20, was also substantially slower than it was under Scenarios 2, 4, and 5, which had sample sizes of 50 or 70. This suggests that the convergence of the MH algorithm was driven more by sample size than it was by the number of possible covariates, at least in this study. Evidently, larger sample sizes provide more concentrated posterior distributions, and thus more rapid convergence to high probability models. For sample sizes of 50 and 70, convergence of the MH algorithm was again quite rapid, requiring fewer than 10 iterations to obtain iterates that differed from the true posterior distribution by less than 0.025 in TVD.

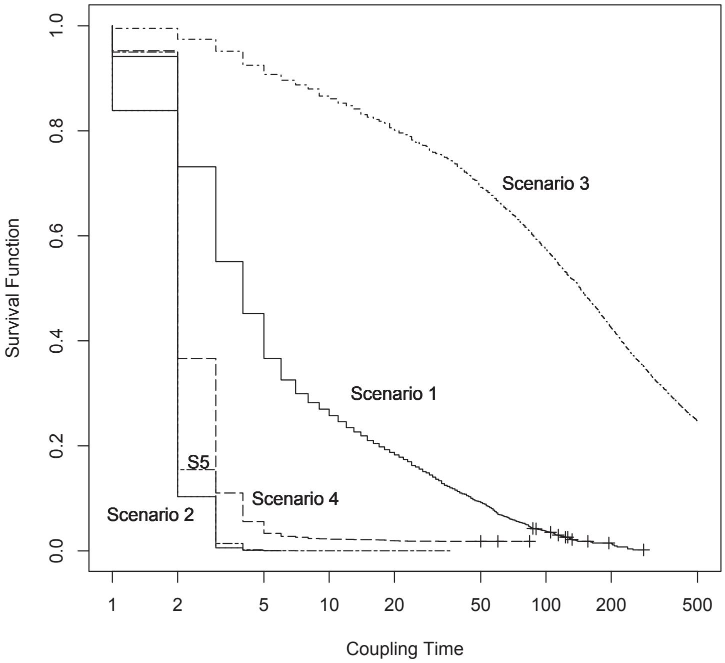

Figure 2: Survival function of coupling times for simulation scenarios in Study 2.

Table 2 provides a comparison of the performance of the Bayesian model selection procedure against SIS-SCAD and ISIS-SCAD. Numerical estimates of the summary statistics for the penalized likelihood methods are taken from FL08. For the SIS and ISIS procedures, n variables were (iteratively) preselected for inclusion in the final model, and Fan and Lv retained $n - 1$ variables in the final models in order to make comparisons between SIS, ISIS, and LASSO more commensurate. For this reason, the median model sizes reported for these methods overestimate the median model sizes that might otherwise have been obtained for each of these methods. The inclusion probabilities reported for these methods refer to the proportion of final models that contained the true model.

The summary statistics reported for the Bayesian methods refer to the maximum a posteriori (MAP) model identified through the CMH algorithm. It is important to

Scenario		1	2	3	4	5
	n	20	50	20	50	70
	p	100	100	1000	1000	1000
	t	3	3	3	3	3
pMOM	med(t)	3	3	3	3	3
	inclusion prob.	1.0	1.0	1.0	1.0	1.0
SIS-SCAD	med(t)	19	49	19	49	69
	inclusion prob	0.69	1.0	0.15	0.87	0.97
ISIS-SCAD	med(t)	19	49	19	49	69
	inclusion prob	1.0	1.0	1.0	1.0	1.0

Table 2: Summary statistics for the MAP models and penalized likelihood methods in ultrahigh dimensional settings for Study 2. The median model sizes (t) listed for the Bayesian model correspond to the true model sizes in all studies. For the Bayesian selection based on the pMOM prior, inclusion probabilities refer to the probability that the MAP model was identical to the true model. For the penalized likelihood methods, the inclusion probabilities refer to the probability that a final model of dimension $n-1$ included the true model. The MAP model identified by the Bayesian procedure always identified the (exactly) true model. Summaries for SIS-SCAD and ISIS-SCAD are taken from FL08.

note that the search for the MAP model utilized states visited by both the primary and auxiliary (coupling) chains. For all but the third scenario, the MAP model identified by both chains always corresponded to the true model. However in the third scenario, the MAP model was often only identified by one or more of the 20 auxiliary chains that were reinitialized from W every 500 updates. This is an important observation, because it shows that the use of the coupling diagnostics not only provided information regarding the convergence of the primary chain, but also led to better coverage of the posterior distribution on the model space. This confirms the assertion made by many researchers (e.g., Gelman and Rubin (1992)) that it is often advantageous to use multiple runs of an MCMC algorithm to explore highly multimodal target distributions.

Interestingly, the Bayesian model selection procedure identified the (exactly) correct model in all 200 data sets under every simulation scenario examined in this study.

3.3 Study 3

The scenarios studied in the final simulation study were very similar to those examined in Study 2, except that the true model was augmented by an additional non-zero regression coefficient whose magnitude was equal to $-15\sqrt{\rho} \approx -10.607$. The value of this coefficient was selected so that the marginal correlation between the fourth explanatory variable and the response variable was 0, which means that this variable is unlikely to be included in a regression model that does not contain at least one of the other three variables. With the addition of the fourth variable, linear models of the form (10) were again considered, but now the true model size was 4 and the regression coefficient was set to $\boldsymbol{\beta} = (5, 5, 5, -15\sqrt{\rho}, 0, \ldots, 0)'$, with $\rho = 0.5$. As in Study 2, each row of the design matrix $\mathbf{X}$ was independently drawn from a multivariate normal distribution having mean 0 and covariance matrix Σ with diagonal entries equal to 1 and off-diagonal entries equal to 0.5. Six scenarios were considered: $(n, p) = (20, 100), (50, 100), (70, 100), (20, 1000), (50, 1000)$, and (70,1000). These scenarios mimic simulation studies reported in Section 4.2.2 of FL08. Two hundred data sets were simulated under each scenario.

The lead-in experiments conducted for Scenarios 1 and 4 produced only a few couplings during a burn-in period of 100 updates. As a consequence, the CMH algorithm was again modified so that 2,000 burn-in iterations were performed for each simulated data set under these scenarios. Following burn-in of the primary chains, the restart interval for the auxiliary chains was set to 500 iterations, and 20 auxiliary chains were run for each simulated data set. The coupling procedures described in Study 1 were applied to all other scenarios.

Figure 3 displays the estimates of the survival distributions for the coupling times under the six scenarios. As is evident from this figure, the coupling times obtained under Scenarios 1 and 4 suggest that the MH algorithm has not converged within 500 iterations. Extrapolating the survival distributions suggests that in excess of 2,000 updates would be required to achieve convergence for data sets simulated under Scenario 1, and that more than 5,000 updates are required to achieve convergence under Scenario 4. As in Study 2, the slowest convergence occurs with the smallest sample size (i.e., $n = 20$), suggesting again that sample size plays a more critical role in determining the convergence of the chains than p does. Of course, this observation should be qualified by noting that one iteration of the chain was defined as an update of the inclusion status of all p variables, which means that the magnitude of p has a direct effect on the number of MCMC acceptance steps that must be performed.

Convergence to the target distribution was achieved with high probability within 500 updates for each of the other scenarios.

Table 3 provides a comparison of the performance of the Bayesian model selection procedure against SIS-SCAD and ISIS-SCAD. The entries in this table have the same interpretation as the entries in Table 2. The summary statistics for the penalized likelihood methods were again taken from FL08. As in Study 2, the Bayesian model procedure (exactly) identified the correct model for every dataset simulated under each scenario.

4 Discussion

The coupling diagnostics proposed in this article provide a convenient mechanism for evaluating the convergence of many MCMC algorithms for model selection. The basic requirement for their implementation is that updates within an MCMC algorithm be "synchronizable" in the sense that the same regression components in two chains can

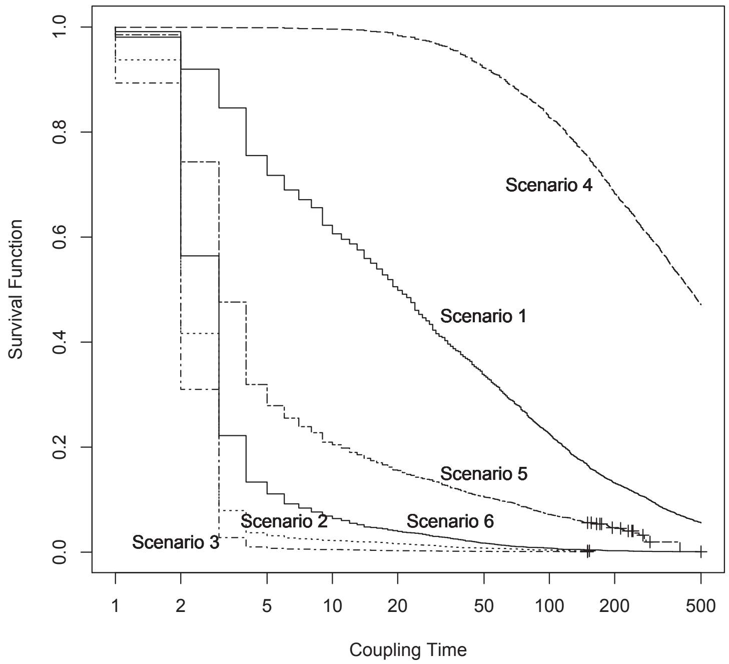

Figure 3: Survival function of coupling times for simulation scenarios in Study 3.

be updated at the same time, without altering the update procedure of either chain.

From a numerical standpoint, the cost of implementing the coupling diagnostics is approximately a two-fold increase in computation time. This cost is offset, however, by the fact that the auxiliary chains provide additional information about the posterior distribution. In several of the simulation scenarios considered in this article, they were critical in identifying high probability models when it was not feasible to run an MH algorithm long enough to thoroughly explore the target distribution.

There is one additional aspect of the simulation studies that deserves comment. Namely, the convergence properties of an MH algorithm often depend on the particular design matrix and observation vector specific to a problem. There is often substantial variation between the coupling-time distributions obtained for data simulated under similar scenarios. For this reason, it is important to apply coupling diagnostics to individual data sets, rather than relying on summary properties obtained from data sets generated under similar conditions.

The Bayesian selection procedure, based on the specification of non-local prior densities on regression coefficients, identified the correct model with an accuracy that

Scenario		1	2	3	4	5	6
	n	20	50	70	20	50	70
	p	100	100	1000	1000	1000	1000
	t	4	4	4	4	4	4
pMOM	med(t)	4	4	4	4	4	4
	prob.	1.0	1.0	1.0	1.0	1.0	1.0
SIS-SCAD	med(t)	19	49	69	19	49	69
	prob	0.025	0.49	0.74	0.00	0.00	0.00
ISIS-SCAD	med(t)	19	49	69	19	49	69
	prob	1.0	1.0	1.0	1.0	1.0	1.0

Table 3: Summary statistics for the MAP models and penalized likelihood methods in ultrahigh dimensional settings for Study 3. The median model sizes (t) listed for the Bayesian model correspond to the true model sizes in all studies. For the Bayesian selection based on the pMOM prior, inclusion probabilities refer to the probability that the MAP model was identical to the true model. For the penalized likelihood methods, the inclusion probabilities refer to the probability that a final model of dimension $n-1$ included the true model. The MAP model identified by the Bayesian procedure always identified the (exactly) true model. Summaries for SIS-SCAD and ISIS-SCAD are taken from FL08.

matched or exceeded the ISIS-SCAD and SIS-SCAD algorithms described in FL08 in all of the simulation examples considered. Furthermore, the availability of coupling diagnostics for the Bayesian procedure is important for providing some assurance that the selection procedure has adequately probed the model space in order to identify high probability models. Of course, these advantages do not come without cost: the Bayesian procedure requires substantially more computational resources to implement than do penalized likelihood methods. For very large p (i.e., $p > O(10^6)$), computational costs may preclude the use of the Bayesian selection procedure without specialized software and/or computational infrastructure.

Finally, the simulation studies conducted in this article suggest that convergence properties of model selection algorithms depend critically on the sample size and concentration of the posterior distribution. Even for settings in which $p \gg n$, the MCMC algorithm converged quickly to the target distribution–often in fewer than 10 iterations–provided that the sample size exceeded about 100. Conversely, great caution should be exercised in attempting to apply Bayesian methods with this MCMC algorithm when n is very small ($n \approx 30$). In such settings, the convergence of the MCMC algorithm can be exceedingly slow, and it may be necessary to perform long runs of multiple chains in order to adequately probe the posterior distribution on the model space.

References

Fan, J. and Lv, J. (2008). "Sure independence screening of ultrahigh dimensional feature space." *Journal of the Royal Statistical Society, Series B*, 70: 849–911.

— (2010). "A selective overview of variable selection in high dimensional feature space." *Statistica Sinica*, 20: 101–148.

Gelman, A. and Rubin, D. (1992). "Inference from iterative simulation using multiple sequences." *Statistical Sciences*, 7: 457–472.

Johnson, V. (1996). "Studying convergence of Markov chain Monte Carlo algorithms using coupled sample paths." *Journal of the American Statistical Association*, 91: 154–166.

— (1998). "A cooupling-regeneration scheme for diagnosing convergence in Markov chain Monte algorithms." *Journal of the American Statistical Association*, 93: 238–248.

Johnson, V. and Rossell, D. (2012). "Bayesian variable selection in high dimensional settings." *Journal of the American Statistical Association*, 107: 649–660.

Lindvall, T. (1992). *Lectures on the coupling method*. NY: Wiley.

Scott, J. and Berger, J. (2010). "Bayes and empirical-Bayes multiplicity adjustment in variable selection problems." *Annals of Statistics*, 38: 2587–2619.

Acknowledgments

The author thanks the Associate Editor and two referees for a number of helpful comments and suggestions. The author was supported by Award Number R01 CA158113 from the National Cancer Institute. Article content is solely the responsibility of the author and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health.