Political Analysis Advance Access published August 19, 2007

doi:10.1093/pan/mpm019

Is Partial-Dimension Convergence a Problem for Inferences from MCMC Algorithms?

Jeff Gill
Center for Applied Statistics, Department of Political Science,
Washington University, One Brookings Drive, St Louis, MO 63130-4899
e-mail: jgill@wustl.edu

Increasingly, political science researchers are turning to Markov chain Monte Carlo methods to solve inferential problems with complex models and problematic data. This is an enormously powerful set of tools based on replacing difficult or impossible analytical work with simulated empirical draws from the distributions of interest. Although practitioners are generally aware of the importance of convergence of the Markov chain, many are not fully aware of the difficulties in fully assessing convergence across multiple dimensions. In most applied circumstances, every parameter dimension must be converged for the others to converge. The usual culprit is slow mixing of the Markov chain and therefore slow convergence towards the target distribution. This work demonstrates the partial convergence problem for the two dominant algorithms and illustrates these issues with empirical examples.

1 Introduction

We show here that users of Markov chain Monte Carlo (MCMC) algorithms must be careful to note whether the procedure has converged to the true distribution of interest for every parameter in the model, not just those of primary interest. Failing to do so leads to incorrect results statistically and substantively. This paper gives the theoretical basis for Markov chain convergence, which is measured imperfectly through single-dimensional diagnostics, and then gives detailed advice for practitioners who wish to increase the reliability of their empirical results. The findings here call for a higher level of caution than commonly observed in published and unpublished work in political science.

Bayesian applications appear increasingly often in political science research along with the concomitant MCMC estimation/marginalization process often required to get working inferences from complicated joint posterior distributions. Recent examples include Western (1998), Quinn, Martin, and Whitford (1999), Smith (1999), Jackman (2000a, 2000b, 2001), Hill and Kriesi (2001), Martin and Quinn (2002), and Gill and Walker (2005). It is (fortunately) now standard practice to "burn-in" the Markov chain by discarding some number observations at the beginning of the run under the assumption that they represent preconvergent samples and are therefore not representative of the distribution of interest. Regretfully, there exists no broad advice about how long the Markov chain should be run

in this way to ensure that it has converged to the distribution of interest before collecting empirical samples.

General issues of Markov chain mixing and convergence have greatly concerned statisticians working in this area since widespread use of MCMC began after 1990. Most of the progress here has been in the development of empirical convergence diagnostics that assess instability through various distributional tests, generally on a parameter-by-parameter basis. This information is used by practitioners to make decisions about how long to run the Markov chain since values from the preconvergent phase provide no useful inferential information. But these empirical diagnostics are explicit or implicit hypothesis tests of *nonconvergence*, meaning that they are capable of telling the user that a Markov chain has not converged but cannot directly assert that it has converged, nor that the chain is mixing well through the distribution of interest. Nonetheless, no indication of nonconvergence across several diagnostics run multiple times with varying parameter settings is safely considered sufficient evidence that the Markov chain has converged, even to conservative practitioners.

Importantly, very little attention has been given to the problem of evidential convergence amongst only subset of the dimensions of the chain (for an exception, see the casual guidance given in Kass et al. 1998). In fact, a few statisticians feel comfortable ignoring the problem if the nonconvergence is limited to nuisance parameters (see Chen, Shao, and Ibrahim 2000, p. 60), notably in statistical genetics where there can be thousands of parameters (e.g., Krishnan et al. 2004). Yet, the structure of at least one MCMC kernel (actually the most commonly used one) is built around marginal draws *conditional on other dimensions*. Therefore, even the parameters that appear to be converged are conditioned on nonconvergent parameters. The primary result shown in this work is that except for a small range of applications, evidence of nonconvergence in one dimension provides evidence of nonconvergence in all dimensions. So even with nuisance parameters, we need to worry about convergence since they can influence the mixing and convergence of parameters of primary interest.

In this paper, we address this lack of attention to convergence and mixing issues with an analysis of the problem focusing on mathematical properties of commonly used algorithms. A preliminary section gives an overview of the two dominant algorithms, followed by a section that gives an exact technical statement of the problem. The following sections explain how this problem of partial convergence differently affects the Metropolis-Hastings algorithm and the Gibbs sampler, and the final section provides specific examples to illustrate the problem.

2 Preliminaries

To establish notation, this section introduces the basic terms and concepts. For more leisurely theoretical discussions, see Nummelin (1984), Tierney (1994), Norris (1997), Liu (2001), or Robert and Casella (2004). The objective of MCMC work is to describe (marginalize) an unwieldy, usually high dimension, joint posterior distribution, $\pi(\boldsymbol{\theta})$ on $\Re^d$ known up to a scale factor. This problem naturally appears in realistic Bayesian model estimation in the social sciences. Two algorithms dominate applied work and are briefly described in the following subsections. The Gibbs sampler, however, is actually a special case of Metropolis-Hastings, and Metropolis-Hastings is sufficiently general that it has spawned a wide array of hybrid and specialized implementations (cf. Roberts and Rosenthal 1998a, section 4; Roberts and Rosenthal 1998b; or Chen, Shao, and Ibrahim 2000, chap. 2).

2.1 *The Metropolis-Hastings Algorithm*

The Metropolis-Hastings algorithm comes from statistical physics (Metropolis et al. 1953) but has proved to be enormously useful in general statistical estimation (Barker 1965; Hastings 1970; Peskun 1973; Chib and Greenberg 1995). Suppose we want to simulate $\boldsymbol{\theta}$, a d-dimensional parameter vector, from the posterior distribution $\pi(\boldsymbol{\theta})$ with support known. At the tth step in the chain, we will draw θ'_j, $j = 1 : d$ from a multivariate *candidate generating distribution* (also called *jumping*, *proposal*, or *instrumental*) over this same support: $q_t(\boldsymbol{\theta}' \mid \boldsymbol{\theta})$. A common variant is the random walk chain where candidate jumping values are selected by an offset from the current state according to a simple additive scheme, $\theta' = \theta + \tau$, where τ is a random variable drawn from some convenient distribution.

It must be possible to determine $q_t(\boldsymbol{\theta} \mid \boldsymbol{\theta}')$ and $q_t(\boldsymbol{\theta}' \mid \boldsymbol{\theta})$. Under the original constraints of the Metropolis algorithm, these two conditionals needed to be equal (symmetrical Metropolis sampling), although we now know that this is not necessary, and have the more flexible restriction of reversibility. That is, the *detailed balance equation* must hold to ensure that $\pi(\boldsymbol{\theta})$ is the invariant (limiting) distribution:

$$K(\boldsymbol{\theta}, \boldsymbol{\theta}')\pi(\boldsymbol{\theta}) = K(\boldsymbol{\theta}', \boldsymbol{\theta})\pi(\boldsymbol{\theta}'), \tag{1}$$

where $K(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is the (probability) kernel of the Metropolis-Hastings algorithm going from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$. This means that if the chain were started in its invariant distribution, it has the same chance of starting at $\boldsymbol{\theta}$ and moving to $\boldsymbol{\theta}'$ as starting at $\boldsymbol{\theta}'$ and moving to $\boldsymbol{\theta}'$. Sometimes $K(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is labeled as $A(\boldsymbol{\theta}, \boldsymbol{\theta}')$ and called the *actual transaction function* from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$ to distinguish it from $a(\boldsymbol{\theta}, \boldsymbol{\theta}')$ below. The *acceptance ratio* is now defined with $\boldsymbol{\theta}^{[t]}$ as the current position and $\boldsymbol{\theta}'$ as the proposal:

$$a(\boldsymbol{\theta}, \boldsymbol{\theta}') = \frac{q_t(\boldsymbol{\theta}^{[t]} \mid \boldsymbol{\theta}')\ \pi(\boldsymbol{\theta}')}{q_t(\boldsymbol{\theta}' \mid \boldsymbol{\theta}^{[t]})\pi(\boldsymbol{\theta}^{[t]})}. \tag{2}$$

The subsequent decision that produces the $t + 1$st point in the chain is probabilistically determined according to:

$$\boldsymbol{\theta}^{[t+1]} = \begin{cases} \boldsymbol{\theta}' & \text{with probability} \quad \min(a(\boldsymbol{\theta}, \boldsymbol{\theta}'), 1) \\ \boldsymbol{\theta}^{[t]} & \text{with probability} \quad 1 - \min(a(\boldsymbol{\theta}, \boldsymbol{\theta}'), 1) \end{cases}. \tag{3}$$

In the case of symmetry in the candidate generating density, $q_t(\boldsymbol{\theta} \mid \boldsymbol{\theta}') = q_t(\boldsymbol{\theta}' \mid \boldsymbol{\theta})$, the acceptance criteria simplifies to a ratio of the posterior density values at the two points.

These steps can be summarized for a parameter vector of interest according to equation (3) with the following steps:

1. Sample $\boldsymbol{\theta}'$ from $q(\boldsymbol{\theta}' \mid \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the current location.
2. Sample u from $\mathcal{U}[0{:}1]$.
3. If $a(\boldsymbol{\theta}, \boldsymbol{\theta}') > u$, then accept $\boldsymbol{\theta}'$.
4. Otherwise keep $\boldsymbol{\theta}$ (the current value) as the new point.

The algorithm described by these steps also has desirable convergence properties to the distribution of interest (Roberts and Smith 1994), as discussed below.

2.2 *The Gibbs Sampler Algorithm*

The Gibbs sampler is a transition kernel created by a series of *full conditional distributions* that are updated based on cycling through these conditional probability statements.

Defining the posterior distribution of interest as $\pi(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a d-length parameter vector, the set of full conditional distributions for $\boldsymbol{\theta}$ are expressed as $\pi(\theta_j \mid \boldsymbol{\theta}_{-j})$ for $j = 1, \ldots, d$, where the notation $\boldsymbol{\theta}_{-j}$ indicates a specific conditional parametric form with θ_j absent on the right-hand-side of the conditional. The Gibbs sampler is then performed by the following steps:

1. Choose starting values: $\boldsymbol{\theta}^{[0]} = [\theta_1^{[0]}, \theta_2^{[0]}, \ldots, \theta_d^{[0]}]$.
2. At the tth step, starting at $t = 1$, complete the single cycle by drawing values from the d conditional distributions given by:

$$\begin{aligned}
\theta_1^{[t]} &\sim \pi(\theta_1 |\ \theta_2^{[t-1]}, \theta_3^{[t-1]}, \ldots, \theta_{d-1}^{[t-1]}, \theta_d^{[t-1]}) \\
\theta_2^{[t]} &\sim \pi(\theta_2 |\ \theta_1^{[t]}, \theta_3^{[t-1]}, \ldots, \theta_{d-1}^{[t-1]}, \theta_d^{[t-1]}) \\
&\vdots \sim \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad . \\
\theta_{d-1}^{[t]} &\sim \pi(\theta_{d-1} |\ \theta_1^{[t]}, \theta_2^{[t]}, \theta_3^{[t]} \ldots, \theta_d^{[t-1]}) \\
\theta_d^{[t]} &\sim \pi(\theta_d |\ \theta_1^{[t]}, \theta_2^{[t]}, \theta_3^{[t]} \ldots, \theta_{d-1}^{[t]})
\end{aligned}$$

3. Increment t and repeat step 2.

Since the Gibbs sampler conditions only on values from its previous iteration, it clearly has the Markovian property The Gibbs sampler is a homogeneous Markov chain: the consecutive probabilities are independent of t, the current length of the chain. This should be apparent from the algorithm above as there is nothing in drawing from the full conditionals that is dependent on the absolute value of t. The Gibbs sampler has the true posterior distribution of parameter vector as its limiting distribution, meaning that it uses these conditional distributions to produce empirical draws from the marginal distributions of interest.

Credit for introducing the Gibbs sampler on finite state spaces is usually given to Geman and Geman (1984), but Ulf Grenander (a student of Harald Cramér) actually applied it to Bayesian modeling in a well known but unpublished paper in 1983. Early restricted versions, typically labeled as the *heatbath algorithm* in statistical physics can be found in Creutz (1979), Ripley (1979), and Creutz, Jacobs, and Rebbi (1983).

3 Statement of the Problem

In applied work there are two practical questions that a user of MCMC algorithms must ask: (1) how long should I run the chain before I can claim that it has converged to its invariant (stationary) distribution, and (2) how long do I need to run the chain in stationarity before it has sufficiently mixed throughout the target distribution? The key factor driving both these questions is the *rate* at which the Markov chain is mixing through the parameter space: slow mixing means that the definition of "long" gets considerably worsened. This work addresses an additional question of equal importance but insufficient attention: What does it mean when a univariate MCMC convergence diagnostic, applied to each dimension of a multidimensional Markov chain, provides evidence that some, but not all, of the dimensions have not converged? This question is important because while convergence is a multidimensional concept, as explained in the following paragraphs, the convergence diagnostics commonly used in practice all operate on a single dimension at a time.

There is also a difference between *being* in the state of convergence and *measuring* the state of convergence. A Markov chain has converged at time t to its invariant distribution

(the posterior distribution of interest for correctly setup Bayesian applications) when the transition kernel produces draws arbitrarily close to this distribution and the process therefore generates only legitimate values from a distribution in proportion to the actual target density. For a given measure of "closeness" (i.e., for some specified threshold, see below), a Markov chain is either in its invariant distribution or it is not. There is no such thing as "somewhat converged" or "approaching convergence." In fact, Rosenthal (1995a) gives an example Markov chain that converges in exactly one step. So diagnostics and mathematical proofs that make claims about convergence are thus analyzing only a two-state world.

To put more precision and clarity on such statements, start with the nonempty outcome space, Ω, that has an associated σ-algebra, and an arbitrary state space, denoted S, that define the possible realizations of the Markov chain of interest. Now define the vector $\boldsymbol{\theta}_t \in S \subseteq \Re^d$ ($d \gg 1$) as the tth empirical draw (reached point) from the chain $\mathrm{MC}(\boldsymbol{\theta}_t, t \geq 0)$, operating on the d-dimensional measurable space (S, Ω), having the transition operator f defined on the Banach space of bounded measurable functions,[1] and having $\pi(\boldsymbol{\theta})$ as its invariant distribution. Invariance in this context means that π is a probability measure on (S, Ω), such that $\pi(s) = \int f(\boldsymbol{\theta}, s)\pi(d\boldsymbol{\theta}), \forall s \in \Omega$. So far, all we have done is notate the measure space for the Markov chain and require that the applicable probability function on this space is well behaved.

Define A (and later B) as an arbitrary and finite subset of S. We will use these subsets to make statements about the behavior of the Markov chain as it traverses the state space. In the discrete case, this can be a single event or a collection of events. In the continuous case, this is a defined subregion of S. The transition kernel of the Markov chain, $f()$ (generalizing K above), is the mechanism that maps $S \times \Omega \rightarrow [0, 1]$ such that for every $A \in \Omega$, the function $f(\cdot, A)$ is measurable and for every $\boldsymbol{\theta} \in S$ the function $f(\theta, \cdot)$ is a valid probability function. So, all this says that we can define the measure space for the Markov chain and properly identify the probability mechanism (function) that governs movement through this space.

An *aperiodic* chain is one that has no deterministically repeating sequence: for an irreducible Markov chain (all substates communicate) on a finite or continuous state space, the greatest length for a repeating cycle is the trivial value of 1. This is easy to visualize in the case of finite state spaces since these are discrete values or positions. For continuous state spaces, these are subspaces or regions, broadly defined as done in the last paragraph. For Markov chains operating on a finite state space, *recurrence* is the property that the chain visits every state infinitely often in the limit. This is implied if the chain is *positive*: the expected time to return to any state having started there is finite (Robert and Casella 2004, chap. 6). For continuous state spaces, positive Harris recurrence for a Markov chain (sometimes called ϕ-recurrence) means that there is a σ-finite measure ϕ on (S, Ω) such that for every definable subspace A in Ω with $\phi(A) > 0$ the probability that the chain reaches A infinitely often in the limit is one for all starting points: $P(\boldsymbol{\theta}_t \in A \; i.o.) = 1, \; \forall A \in S$ and all $\boldsymbol{\theta}_0$ (in our specific case ϕ is π). Colloquially, Harris recurrence assures us that the chain is not impeded from reaching every arbitrary subspace from any given point, even wildly misguided starting points: "there is no measure-theoretic pathology" to worry about (Chan and Geyer 1994). All Markov chains discussed from this point on are assumed to meet the conditions of aperiodicity and (appropriate)

[1] A normed vector space is called a Banach space if it is complete under this metric. Completeness means that for a given probability measure space (typically given as (Ω, $\mathcal{F}$, P) denoting a space, a field (class of subsets of Ω), and a probability measure on $\mathcal{F}$), if $A \subset B, B \in \mathcal{F}, P(B) = 0$, then $A \in \mathcal{F}$ and $P(A) = 0$. This condition for f allows us to ignore a set of measure problems that can otherwise occur. It also provides results on a general state space as well as the easier case of a finite countable state space.

recurrence, although we will occasionally issue a reminder. See Athreya, Doss, and Sethuraman (1996) for additional technical details on theoretical issues beyond those discussed here.

If $\text{MC}(\boldsymbol{\theta}_t, t \geq 0)$ is an aperiodic positive Harris recurrent Markov chain (such as the Gibbs sampler and common variants of the Metropolis-Hastings algorithm), then we call it *ergodic* (Tweedie 1975). Ergodic Markov chains have the important property that:

$$\lim_{t\to\infty} \frac{1}{t}\sum f(\boldsymbol{\theta}_t) = \int_{\Theta} f(\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}, \tag{4}$$

proven originally by Doeblin (1940). This result means that empirical averages for the function $f()$ converge to a probabilistic average of the function over the limiting distribution. In fact, it is this principle that underlies and justifies all MCMC for Bayesian stochastic simulation; it is exactly the link between "Markov chain" and "Monte Carlo." Since we are often interested in the posterior mean of the distribution of interest from a Bayesian model, equation (4) then implies that $\frac{1}{t}\sum f(\boldsymbol{\theta}_t) \approx \int_{\boldsymbol{\Theta}} \boldsymbol{\theta}\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$, for sufficiently large t.

A chain that is positive recurrent (finite state space) or positive Harris recurrent (continuous state space) and aperiodic is also α-mixing in t, meaning that:

$$\alpha(t) = \sup_{A,B} \mid P(\boldsymbol{\theta}_t \in B, \boldsymbol{\theta}_0 \in A) - P(\boldsymbol{\theta}_t \in B)P(\boldsymbol{\theta}_0 \in A) \mid \underset{t\to\infty}{\to} 0 \tag{5}$$

(Rosenblatt 1971). This means that for the subspaces of S, A, and B that produce the largest difference [the designation $\sup_{A,B}$ in equation (5)], the joint probability of starting at some point in A and ending at some point in B at time t converges to the product of the individual probabilities. In other words, these events are asymptotically (in t) independent for any definable subspaces. This second result from ergodicity justifies our treatment of Markov chain iterations as independent and identically distributed samples (Chan 1993).

There are actually some additional measure-theoretic nuances and extensions of these properties, such as the implied assumption that $\pi(\boldsymbol{\theta})$ is not concentrated on a single point as in a Dirac delta function (a "distribution" only in one technical sense), but the definitions given here are sufficient for the present purposes. Also, it is important to remember that ergodicity is just one way to assert convergence. It turns out, for instance, that a *periodic* Markov chain can also converge under a different and more complicated set of assumptions (Meyn and Tweedie 1993, chap. 13), and we can even define ergodicity without an invariant distribution (Athreya and Ney 1978). All this theory leads now to a rigorous definition of convergence that we can use to understand partial results.

A Markov chain that has converged has the property that repeated applications of the transition kernel produce an identical distribution: $\pi f = \pi$. By far, the most commonly used method[2] of measuring such convergence is the *total variation norm*:

$$\| f(\boldsymbol{\theta}_t) - \pi(\boldsymbol{\theta}) \| = \frac{1}{2}\sup_{\boldsymbol{\theta}\in A} A \int_{\Theta} \mid f(\boldsymbol{\theta}_t) - \pi(\boldsymbol{\theta}) \mid d\boldsymbol{\theta}, \tag{6}$$

which is also half of the L_1 distance (the 1/2 is sometimes omitted by authors). The 1/2 comes from limit theory: it turns out that as $t \to \infty$, the total variation norm for A converges to twice the empirical difference for all such subspaces (Meyn and Tweedie 1993, p. 311).

[2]The L_2 "chi-square" distance is also useful; see Diaconis and Saloff-Coste (1996). Another suggestion is the infinity norm $\|f\|_\infty = \sup_{\theta\in\Re} \|f(\theta)\|$ (Roberts and Polson 1994). Geometric ergodicity can be asserted with the V-norm as shown in Meyn and Tweedie (1993, chap. 16). Zellner and Min (1995) propose three potentially useful alternatives as well.

Here $\boldsymbol{\theta}$ is a d-dimensional random variable lying within A, the subspace that makes the difference within the integral as great as possible. When we integrate over $\boldsymbol{\theta} \in A$, it produces a supremum over the measurable subspace A for the set of all measurable functions on A (a set that includes $f(\boldsymbol{\theta}_t)$ and $\pi(\boldsymbol{\theta})$). So there are two important operations occurring in the statement of equation (6): (1) selection of a subspace that makes the resulting quantity as large as possible and (2) integration of the distributional difference over this subspace. In this way, we get the most pessimistic view of the difference between $f(\boldsymbol{\theta}_t)$ and $\pi(\boldsymbol{\theta})$ possible. Another common way to write the total variation norm (Meyn and Tweedie 1993, p. 311) defines $\mu(A)$ as a signed measure on the state space S for the subspace A. For our purposes, $\mu(A)$ is the integrated difference of two distributional statements over all of A. Then the total variation norm can be expressed as:

$$\| \mu \| = \sup_{A \in S} \mu(A) - \inf_{A \in S} \mu(A), \tag{7}$$

which shows the same principle as equation (6) due to the explicit statement of the integral.

Thus, if $\| f(\boldsymbol{\theta}_t) - \pi(\boldsymbol{\theta}) \| \to 0$ as $t \to \infty$, $\boldsymbol{\theta}$ converges to a random variable from $\pi(\boldsymbol{\theta})$, and this convergence is actually stronger than standard convergence in distribution (i.e., convergence of cumulative distribution functions [CDFs]). When $\| f(\boldsymbol{\theta}_t) - \pi(\boldsymbol{\theta}) \|$ reaches a value close to 0 (say δ for now), we are willing to assert convergence. The problem, of course, is that $\pi(\boldsymbol{\theta})$ is a difficult form analytically, which is why we are using MCMC in the first place. Theoretical work that puts explicit bounds on convergence includes Lawler and Sokal (1988), Sinclair and Jerrum (1988), Frieze, Kannan, and Polson (1993), Ingrassia (1994), Robert (1995), Rosenthal (1995a), Liu (1996), Mengersen and Tweedie (1996), as well as Roberts and Tweedie (1996). For discrete problems, it turns out that the converge rate can be established in proportion to the absolute value of the second eigenvalue of the transition matrix (kernel) (Sinclair and Jerrum 1989; Diaconis and Stroock 1991; Fill 1991; Fulman and Wilmer 1999), but this can also be quite difficult to produce for realistic problems (Frigessi et al. 1993). For examples where these approaches work in practice, see Goodman and Sokal (1989), Amit (1991), Amit and Grenander (1991), Meyn and Tweedie (1994), Rosenthal (1995b, 1996), Polson (1996), Cowles and Rosenthal (1998), Roberts and Rosenthal (1999), and Mira and Tierney (2001). Usually, these solutions are particularistic to the form of the kernel and can also produce widely varying or impractical bounds.

Markov chain mixing is a related, but different, concern than convergence. Mixing is the rate at which a Markov chain moves about the parameter space, before or after reaching the stationary distribution. Thus, slow mixing causes two problems: it retards the advance towards the target distribution, and once there, it makes full exploration of this distribution take longer. Both these considerations are critical to providing valid inferences since the preconvergence distribution does not describe the desired marginals and failing to mix through regions of the final distribution biases summary statistics. Mixing problems come from high correlations between model parameters, weakly identified model specifications, and are often more pronounced for model precision parameters. Although we are more concerned with the convergence implications of poor mixing here, recommendations for identifying and solving such problems are given later.

4 A Convergence Finding for the Metropolis-Hastings Algorithm

As the number of dimensions increases, the sensitivity (and complexity) of the Metropolis-Hastings algorithm increases dramatically since the measure space (S, Ω) is defined such

that each abstract point in S is d-dimensional and the σ-field of subsets of Ω is generated by a countable collection of sets on $\Re^d$. An ergodic Markov chain $\text{MC}(\boldsymbol{\theta}_t, t \geq 0)$ has the property:

$$\| f(\boldsymbol{\theta}_t) - \pi(\boldsymbol{\theta}) \| \rightarrow 0 \quad \text{as} \ \ t \rightarrow \infty, \ \forall \boldsymbol{\theta} \in \Omega, \tag{8}$$

but the size of d is critical in determining the rate since each step is d-dimensional. The primary complexity introduced by dimensionality here has to do with the strictness by which we apply $f(\boldsymbol{\theta}_t) - \pi(\boldsymbol{\theta})$. Suppose now that there is a subset of these dimensions $e < d$ that are of primary interest and the remaining $d - e$ are essentially a result of nuisance parameters. Is it then reasonable to require only evidence of *partial convergence*? That is, at time t for some small δ:

$$\| f(\boldsymbol{\theta}_t^*) - \pi(\boldsymbol{\theta}^*) \| \approx \delta, \quad \forall \boldsymbol{\theta}^* \in \Omega^e \tag{9}$$

but,

$$\| f(\boldsymbol{\theta}_t^\dagger) - \pi(\boldsymbol{\theta}^\dagger) \| \gg \delta \quad \forall \boldsymbol{\theta}^\dagger \in \Omega^{d-e}, \tag{10}$$

where decisions are made one at a time for each of these dimensions using standard empirical diagnostics of stationarity. Even though our evidence is derived from these diagnostics, it is important to note that they measure Markov chain *stability* rather than actual convergence, and so the sum of each of these across dimensions is used to just *assert* convergence (convergence in total variation norm gives stationarity but the converse is not similarly guaranteed).

The term *partial convergence* (or *incomplete convergence*) is introduced here since there appears to be no scholarly concern to the problem by probabilists or statisticians and therefore no vernacular to adopt. Partial convergence comes from a similar, but distinct, problem of convergence for contact processes on connected graphs (see Salzano and Schonmann 1997, 1999).[3]

There is one very important distinction to be made here. The Markov chain $\text{MC}(\boldsymbol{\theta}_t, t \geq 0)$ is assumed ergodic over all of S and is thus guaranteed to *eventually* converge across all of $\Re^d$. What we see by observing equations (9) and (10) at time point t is a lack of evidence to say that there is full dimensional convergence. Can a Markov chain operating in d dimensions be drawing from the true invariant distribution in e subdimensions but not in $d - e$ subdimensions? Recall that the standard empirical diagnostics in WinBUGS, coda, and boa (Brooks, Gelman, and Rubin [BGR]; Geweke, Heidelberger, and Welch; Raftery and Louis, plus graphical methods), as well as others used in practice (Brooks, Dellaportas, and Roberts [1997] develop one based on the total variation norm discussed above), provide a series of parameter-by-parameter tests of nonconvergence.[4] Hence, they indicate when a single dimension chain is sufficiently trending as to violate specific distributional assumptions that reflect stability, but they do assert convergence in the opposite case.

These diagnostics operate on marginal distributions individually since the output of the MCMC process is a set of *marginal* empirical draws. Unfortunately, the total variation norm given above only shows us if the chain has converged simultaneously across every dimension, providing a strong disconnect between theoreticians who derive convergence

[3]There is also a neat parallel with Duncan Macintosh's (1994) theory that all empirically adequate scientific theories about nature partially converge to the true theory.

[4]Good overviews and critiques of these diagnostics can be found in Cowles and Carlin (1996), Brooks and Roberts (1998), and at a more basic level in Gill (2007, chap. 12).

properties for specific chains under specific circumstances and the masses who want to run simple empirical diagnostics in easy-to-use software environments. To see this disconnect more clearly, consider the right-hand-side of equation (6) written out with more detail:

$$\frac{1}{2}\sup_{\theta\in A}A\left|\int_{\theta_1}\cdots\int_{\theta_e}\int_{\theta_{e+1}}\cdots\int_{\theta_d}[f(\theta_1,\ldots,\theta_d)_t - \pi(\theta_1,\ldots,\theta_d)]d\theta_1\cdots d\theta_e d\theta_{e+1}\cdots d\theta_d\right|. \tag{11}$$

If $f()$ and $\pi()$ were able to be expressed as independent products [i.e., $f(x, y) = f(x)f(y)$] for the θ_i, then this would be a straightforward integration process. As stated, this cannot be true for π since we are doing MCMC for this very reason. But what about $f()$? Consider the *actual* transition probability for the Metropolis-Hastings algorithm from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$:

$$A(\boldsymbol{\theta},\boldsymbol{\theta}') = \min\left\{\frac{\pi(\boldsymbol{\theta}')g(\boldsymbol{\theta}\mid\boldsymbol{\theta}')}{f(\boldsymbol{\theta})g(\boldsymbol{\theta}'\mid\boldsymbol{\theta})},1\right\}g(\boldsymbol{\theta}'|\boldsymbol{\theta}) + (1-r(\boldsymbol{\theta}))\delta_{\boldsymbol{\theta}}(\boldsymbol{\theta}'), \tag{12}$$

where $g()$ is the proposal distribution,

$$r(\boldsymbol{\theta}) = \int \min\left\{\frac{f(\boldsymbol{\theta}')g(\boldsymbol{\theta}\mid\boldsymbol{\theta}')}{f(\boldsymbol{\theta})g(\boldsymbol{\theta}'\mid\boldsymbol{\theta})},1\right\}g(\boldsymbol{\theta}'\mid\boldsymbol{\theta})d\boldsymbol{\theta}',$$

and

$$\delta_{\boldsymbol{\theta}}(\boldsymbol{\theta}') = 1 \quad \text{if } \boldsymbol{\theta} = \boldsymbol{\theta}' \text{ and } 0 \text{ otherwise}$$

(the Dirac delta function). It is pretty obvious from looking at equation (12) that we cannot generally disentangle dimensions. In particular, note the conditionals that exist across $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$. What this means is that decisions to jump to a proposal point in d-space (S, Ω) are made based on the current position in every dimension for the Metropolis-Hastings algorithm. So if the chain has not converged in the ith dimension, $i \in [e+1:d]$, its current placement affects the single acceptance ratio and therefore the probability of making a complete d-dimensional jump. And this is all under the assumption of ergodicity or better: geometric ergodicity or uniform ergodicity. Geometric ergodicity requires that:

$$\| f(\boldsymbol{\theta}_t) - \pi(\boldsymbol{\theta}) \| \leq m(\boldsymbol{\theta})\rho^t, \quad \forall\boldsymbol{\theta},\ 0<\rho<1, \tag{13}$$

where $m(\boldsymbol{\theta})$ is any finite, nonnegative function. Under these conditions, the tth step transition probability converges to the invariant distribution at a geometric rate, which can be very quick depending on the value of ρ. If instead of specifying the function $m(\boldsymbol{\theta})$, we find a constant m such that:

$$\| f(\boldsymbol{\theta}_t) - \pi(\boldsymbol{\theta}) \| \leq m\rho^t, \quad \forall\boldsymbol{\theta},\ 0<\rho<1, \tag{14}$$

then the chain is uniformly ergodic, which means it converges even faster.

So now that we know that nonconvergence in at least one dimension affects decisions to move in all dimensions, the natural question is how does this work? A Metropolis-Hastings chain dimension that has not converged is producing on average lower density contributions in the acceptance ratio. Therefore, in cases where the conditionality on the current state is explicit (all general forms except the independence chain Metropolis-Hastings where jumping values are selected from a convenient form as in the random walk chain, but ignoring the current position completely: $g(\boldsymbol{\theta}' \mid \boldsymbol{\theta}) = f(\boldsymbol{\theta}')$), it retards the mixing of the

whole chain. Because the chain is ergodic, it is α-mixing ($\sup_{A,B}|P(\boldsymbol{\theta}_t \in B, \boldsymbol{\theta}_0 \in A) - P(\boldsymbol{\theta}_t \in B)P(\boldsymbol{\theta}_0 \in A)|$ goes to 0 in the limit), but inefficiently so (slowly) since non-convergence for the $d - e$ dimensions implies poorer mixing and greater distance between $P(\boldsymbol{\theta}_t \in B, \boldsymbol{\theta}_0 \in A)$ and $P(\boldsymbol{\theta}_t \in B)P(\boldsymbol{\theta}_0 \in A)$.

It is no secret that a chain that is completely in its invariance distribution mixes better (Robert and Casella 2004, chap. 12). So even in the case where θ_i, the unconverged dimension here is not a parent node in the model specified by π, there is a negative effect: a Markov chain that has not sufficiently mixed through the target distribution produces biased empirical summaries because collected chain values will be incomplete, having had insufficient time to fully explore the target.

5 A Convergence Finding for the Gibbs Sampler

Robert and Richardson (1998) show that when a Markov chain, MC(θ_t, $t \geq 0$), is derived from another Markov chain, MC(ϕ_t, $t \geq 0$), by simulating from a distribution according to $\pi(\theta \mid \phi_t)$, the properties of the first chain inherit that of the conditional. Critically, this conditionality defines new subspaces of (S, Ω) with new measure properties.

For our purposes, the important point is that if MC(ϕ_t, $t \geq 0$) is geometrically ergodic, then MC(θ_t, $t \geq 0$) is as well, which is easy to demonstrate using the data augmentation principle. The marginal distribution for the geometrically ergodic chain at time t is $f_t(\phi)$ with invariant distribution for $\pi(\phi)$. We can now express the invariant distribution of θ in conditional terms:

$$\pi(\theta) = \int_{\phi} \pi(\theta \mid \phi)\pi(\phi)d\phi, \tag{15}$$

with the marginal distribution at time t:

$$f_t(\theta) = \int_{\phi} \pi(\theta \mid \phi)f_t(\phi)d\phi. \tag{16}$$

These integrals over ϕ result from a standard data augmentation process and currently come before the integration of $\theta \in A$ in the calculation of the total variation norm (we will switch this order below). Note that $\pi(\theta \mid \phi)$ appears in the second expression without reference to time since θ_t is simulated *at* each step from $\pi(\theta \mid \phi)$. These define the total variation norm for θ_t:

$$\begin{aligned} \| f_t(\theta) - \pi(\theta) \| &= \frac{1}{2}\sup_{\theta \in A} A\left| \int_{\theta}\int_{\phi} \pi(\theta|\phi)f_t(\phi)d\phi d\theta - \int_{\theta}\int_{\phi} \pi(\theta|\phi)\pi(\phi)d\phi d\theta \right| \\ &= \frac{1}{2}\sup_{\theta \in A}\left| \int_{\theta}\int_{\phi} \pi(\theta|\phi)[f_t(\phi) - \pi(\phi)]d\phi d\theta \right| \\ &\leq \| f_t(\phi) - \pi(\phi) \|, \end{aligned} \tag{17}$$

where the inequality comes from $\pi(\theta \mid \phi) \leq 1$ by the integration of a probability function over the measure space for ϕ (the rate ρ carries through as well). This says that the total variation norm of θ can never be higher than the total variation norm of ϕ, the random

quantity it is conditioned on. Switching the order of integration in the last line comes from stated regularity conditions on probability functions. Note that this process in equation (17) is related to, but distinct from, so-called Rao-Blackwellization where intentional conditioning is imposed to reduce the variance of computed expectations or marginals (Casella and Robert 1996). The result in equation (17) is that a nonconvergent dimension to the Gibbs sampler (just θ here) "pushes" the others (just θ here) away from stationarity as well, even if these pass an empirical diagnostic for convergence. Observe also, that if the two integrals in the first line of equation (17) were calculated-out at that point, it would just be a way to calculate the total variation norm for θ using data augmentation (which might or might not be easier). Although we may want to perform these calculations in practice, such a step does not help in producing the inequality result.

One utility of this result is that if we can intentionally augment a target chain with a simple form that is known to be geometrically ergodic, then we can impose this property even though we increase the dimension of (S, Ω) (Fill 1991; Diaconis and Saloff-Coste 1993). Robert and Richardson (1998) point out that this is particularly useful when a target chain of unknown convergence characteristics is conditioned on a simple discrete Markov chain known to be geometrically ergodic with specific ρ and $m(\theta)$ (alternately m). Also, if the chain that is conditioned on is α-mixing, the target chain will be as well.

The big point comes from the structure of the Gibbs sampler (the default engine of WinBUGS). Since the kernel is an iteration of full conditionals, $\pi(\theta_j \mid \boldsymbol{\theta}_{-j})$ for $j = 1, \ldots, d$, then according to the logic just discussed, either the Gibbs sampler is geometrically ergodic in every dimension or it is not geometrically ergodic in any dimension. Importantly, since the subchains share the same geometric rate of convergence, ρ, one should be cautious with empirical diagnostics since they provide evidence of nonconvergence not evidence of convergence (see Asmussen, Glynn, and Thorisson 1992, for a detailed discussion on this point). Recall that at any given time t, a Markov chain is either converged to its invariant distribution or it has not: there is a specific time when $\| f(\boldsymbol{\theta}_t) - \pi(\boldsymbol{\theta}) \| < \delta$ for some chosen δ, we just do not necessarily know the moment.

Suppose for a two-parameter Gibbs sampler θ_1 passes some empirical diagnostic and θ_2 does not. Since they share the same rate of convergence, for θ_1 to be in its invariant distribution while θ_2 is not means that you are testing the chain for convergence during the very small number of intervals where the differing results are due to probabilistic features of the chain or the test. Given the standard number iterations expected for MCMC work in political science (generally tens or hundreds of thousands), the probability that you have stumbled up this exact interval is essentially 0. Conversely, the test fails for θ_2 because the Markov chain for this dimension is either not yet in stationarity or is in stationarity but has failed to sufficiently explore the target distribution to produce a stable summary statistic for the chosen diagnostic. The latter condition only exists for a relatively short period of time, even with poor mixing. Moreover, the faster the convergence rate (i.e., geometric or uniform), the smaller the numerator in the calculation of this probability making it even less likely that the user caught the interval of intermediate results using the Gibbs samplers with listed properties above, where the size of this effect is notably a function of $m(\theta)$ (or m) and ρ. Therefore for the Gibbs sampler, evidence of nonconvergence in any dimension is evidence of nonconvergence in all dimensions. In textbook, Gibbs sampling every parameter is conditional on every other parameter (actually in Bayesian hierarchical models most parameters are not conditional on every other parameter). Returning to the notation of equation (17), we note that θ and ϕ would both be conditional on each other in a Gibbs sampler. If θ cannot converge faster than ϕ and ϕ cannot converge faster than θ, then they converge at the same rate regardless of what various empirical tests are telling us. So for

users of the usual diagnostic packages, coda and boa, the standard for multidimensional convergence needs to be high. For a relatively small number of parameters, one would expect broad consensus across diagnostics. However, for large numbers of model parameters, we need to be aware that the diagnostics are built on formal hypothesis tests at selected α levels and therefore $1 - \alpha$ tests for large numbers of parameters will fail for about α proportion of dimensions, even in full convergence.

Note that this same logic applies to Metropolis-Hastings MCMC for parameters with conditions formed by hierarchies, which are a natural and common feature of Bayesian model specifications. This inheritance of convergence properties does not necessarily occur, however, for every parameter as in the perfectly symmetric case of the Gibbs sampler. It also is not reciprocal in that the conditions in a Bayesian hierarchical model flow downward from founder nodes to dependent nodes. Note that these conditionals result explicitly from the model rather than through algorithmic conditioning in the Gibbs sampler sense. In addition, parameters can be highly correlated without these structural relationships. The difficulty posed by all these characteristics is that they generally slow the mixing of the chain, making convergence and full exploration more difficult.

6 Empirical Examples

6.1 *Convergence Problems from Priors Specifications (Gibbs Sampling)*

A standard regression tool from the empirical modeling toolbox for dealing with censored data is the tobit model (Tobin 1958), which is described in detail in Amemiya (1985, chap. 10). Suppose an interval-measured outcome variable is censored such that all values that would have naturally been observed as negative are reported as 0 (generalizable to other values). A classic example is public support for some practice where the practice is adopted in only some cases (trade protection, death penalty, campaign contributions). So support for the death penalty in murder cases in Hawaii is recorded as 0, but is unlikely to actually be 0. If $\mathbf{z}$ is a latent outcome variable in this context with the assumed relation $\mathbf{z} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ (and $z_i \sim N(\mathbf{x}\boldsymbol{\beta}, \sigma^2)$), then the observed outcome variable is produced according to: $y_i = z_i$ if $z_i > 0$ and $y_i = 0$ if $z_i \leq 0$. The resulting likelihood function is:

$$L(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}, \mathbf{X}) = \prod_{y_i=0} \left[1 - \Phi\left(\frac{x_i\boldsymbol{\beta}}{\sigma}\right)\right] \prod_{y_i>0} (\sigma^{-1}) \exp\left[-\frac{1}{2\sigma^2}(y_i - x_i\boldsymbol{\beta})^2\right], \qquad (18)$$

where Φ denotes the normal CDF. Chib (1992) introduces a blocked Gibbs sampling estimation process for this model using data augmentation, and Albert and Chib (1993) extend this to discrete choice outcomes. This is a quite natural approach since augmentation can be done with the latent variable $\mathbf{z}$. A flexible parameterization for the priors is given by Gawande (1998):

$$\boldsymbol{\beta} \mid \sigma^2 \sim \mathcal{N}(\boldsymbol{\beta}_0, \mathbf{I}\sigma^2 B_0^{-1}) \quad \sigma^2 \sim \mathcal{IG}\left(\frac{\gamma_0}{2}, \frac{\gamma_1}{2}\right) \qquad (19)$$

with vector hyperparameter $\boldsymbol{\beta}_0$, scalar hyperparameters B_0, $\gamma_0 > 2$, $\gamma_1 > 0$ ($\mathcal{IG}$ here denotes the inverse gamma PDF), and appropriately sized identity matrix $\mathbf{I}$. Substantial prior flexibility can be achieved with varied levels of these parameters. For instance, a likelihoodist-like result for $\boldsymbol{\beta}$ can be specified with $\boldsymbol{\beta}_0 = 0$ and small B_0. A diffuse prior specification for σ^2 is achieved by stipulating small γ_0 and large γ_1 since $E_{\mathcal{IG}}[\sigma^2] = (\gamma_1/(\gamma_0 - 1))$.

The resulting full conditional distributions for Gibbs sampling are given for the $\boldsymbol{\beta}$ block, σ^2, and the individual $z_i|y_i = 0$ as:

$$\begin{aligned}
&\boldsymbol{\beta} \mid \sigma^2, \mathbf{z}, \mathbf{y}, \mathbf{X} \sim \mathcal{N}((B_0 + \mathbf{X}'\mathbf{X})^{-1})(\boldsymbol{\beta}_0 B_0 + \mathbf{X}'\mathbf{z}), (\sigma^{-2} B_0 + \sigma^{-2}\mathbf{X}'\mathbf{X})^{-1}))\\
&\sigma^2 \mid \boldsymbol{\beta}, \mathbf{z}, \mathbf{y}, \mathbf{X} \sim \mathcal{IG}\left(\frac{\gamma_0 + n}{2}, \frac{\gamma_1 + (\mathbf{z} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{z} - \mathbf{X}\boldsymbol{\beta})}{2}\right) \qquad (20)\\
&z_i \mid y_i = 0, \boldsymbol{\beta}, \sigma, \mathbf{X} \sim \mathcal{TN}(\mathbf{X}\boldsymbol{\beta}, \sigma^2) I(-\infty, 0),
\end{aligned}$$

where $\mathcal{TN}()$ denotes the truncated normal and the indicator function $I(-\infty, 0)$ provides the limits of truncation.

Norrander (2000) uses tobit models to look at social and political influences on U.S. state decisions to impose the death penalty since the Supreme Court ruled the practice constitutional in *Furman v. Georgia* 1972. The central question addressed is whether the ideology, racial and religious makeup, political culture, and urbanization are causal effects for state-level death sentences from 1993 to 1995. Norrander posits a causal model whereby public opinion centrally, influenced by past policies and demographic factors, determines death penalty rates by legitimating the practice over time. Her results are convincing, but some curious "nonfindings" also emerge, in particular the lack of support (across several model specifications) for state ideology and the current murder rate as affecting sentencing rates.

The tobit model is appropriate here because 15 states did not have capital punishment provisions on the books in the studied period and so the effect of public opinion on death penalty rates is therefore censored in the data. That is, if these states had the legal ability to impose death penalty sentences, then we would see evidence of some relationship between the explanatory variables and the count. Note that the data are also truncated at 0 since states cannot impose a negative number of death penalty sentences.

Unfortunately, the estimation process is sensitive to values of B_0, and this is why it is common to see very diffuse priors in this specification. Suppose we wanted to approach this project in a much more canonical Bayesian fashion by actually specifying informed prior information. After all, we actually know about some of these effects: a state's willingness to sentence convicted murders to death should be positively affected by past rates of executions, influenced by current state culture and opinion, increased with increasing average level of conservative ideology, and increased with murder rates. In order to avoid a lengthy policy discussion for this pedagogical example, we will simply pull prior means from *bivariate* regression slope coefficients for each of these variables against the sentencing rate, shown in the left-hand side of Table 1. These prior means are then used in the specification of normal distributions all with variance 10 and $B_0 = 0.02$, as a way to be somewhat skeptical.[5]

A Gibbs sampler is applied using the full conditional distributions given above for $\boldsymbol{\beta}$, σ^2, and the z_i using estimates from a regular non-Bayesian treatment as starting points. What we see from the results in the right-hand side of Table 1 (10,000 iterations total summarizing the last 2000) is that the introduction of prior information, albeit reasonably mild prior information, leads to a defensible finding on ideology (although still not on the murder rate). More interesting for our purposes, altering the precision value of B_0 upwards

[5]Although this makes use of the data to stipulate prior distributions, these are bivariate relationships and thus not indicative of a full multivariate relationship. The signs of these slopes are all in normally expected directions for affecting death penalty rates and the priors are all given large variance. In practice, the rationale for specifying prior distributions should be more substantively driven.

Table 1 Tobit model for death sentences by state

	Prior		*Posterior*	
	β_0	σ^2	$\bar{\beta}$	σ_β
Constant	1.0000	10	−14.2951	3.4128
Past rates	180.2485	10	171.2920	8.2988
Political culture	0.9596	10	0.3323	0.1415
Current opinion	5.7686	10	3.8777	1.0645
Ideology	6.4436	10	3.1074	1.0714
Murder rate	0.5019	10	0.0152	0.0761

slightly ruins the stability of the Markov chain. This is dangerous since researchers can set this parameter to reflect varying prior weight, moving the full conditional specification for $\boldsymbol{\beta}$ away from the influence of the data in equation (20) and making the Gibbs sampler less stable. We also know that precision terms are more susceptible to mixing problems than unbounded cases. For instance, if we change B_0 to 0.1, then at iteration 160, the subchain for ideology begins a march toward negative infinity. This dimension is shown to be more variable than the others anyway in Fig. 1, where traceplots for the stable Markov chain for $B_0 = 0.02$ are shown down the first column and traceplots for the problematic version for $B_0 = 0.1$ are shown down the second column. The values that the stable chain displays here endure through the 10,000 iterations.

A few iterations after reaching 160, the effect from the ideology dimension ripples through the conditional mechanism of the Gibbs sampler and all the dimensions become unstable. Although this example exhibited problems early in the life of the chain, there is no guarantee that such problems cannot emerge at any time. The important point to note is that once problems emerged in one dimension, they did not stay confined there.

We can also learn something about convergence in this example by applying some of the standard empirical diagnostic tools commonly used in practice. The BGR diagnostic is based on an ANOVA comparison of multiple parallel runs of the Markov chain started from widely dispersed positions in the sample space. First, run $m \geq 2$ chains of length $2n$ from these overdispersed starting points and dispose of the first n. For a single parameter k out of d in the model, we denote $\theta_{(j)}^{[t,k]}$ as the tth value ($n < t \leq 2n$) from the jth parallel chain ($1 \leq j \leq m$). For the kth parameter, we calculate two quantities: the within-chain variance, $W = \frac{1}{m(n-1)} \sum_{j=1}^{m} \sum_{t=1}^{n} (\theta_{(j)}^{[t,k]} - \bar{\theta}_{(j)}^{[\cdot,k]})^2$, and the between-chain variance, $B = \frac{n}{m-1} \sum_{j=1}^{m} (\bar{\theta}_{(j)}^{[\cdot,k]} - \bar{\bar{\theta}}^{[\cdot,k]})^2$. From standard ANOVA theory, the marginal posterior variance is given by $\widehat{\text{var}}(\theta) = (1 - 1/n)W + (1/n)B$, and the *scale reduction factor* is calculated as: $\hat{R} = \sqrt{\widehat{\text{var}}(\theta)/W}$. The value of $\hat{R}$ needs to be close to one to claim convergence.

The Geweke diagnostic proceeds by preselecting for the kth parameter two nonoverlapping window proportions, one early in the chain and one later in the chain: $\theta_1^{[k]}$ of length n_1 and $\theta_2^{[k]}$ of length n_2. For each window, calculate some function of interest $g()$, which is generally the mean. The diagnostic is given by: $G = (g(\theta_1^{[k]}) - g(\theta_2^{[k]}))/\sqrt{\frac{s_1(0)}{n_1} + \frac{s_2(0)}{n_2}}$, where $s_1(0)$ and $s_2(0)$ are the symmetric spectral density functions for each window. Large values show a discrepancy and therefore imply Markov chain instability.

For the Heidelberger and Welch diagnostic, define the following quantities: (1) $T =$ the length of the chain after burn-in, $s \in [0 : 1] =$ the test chain proportion,

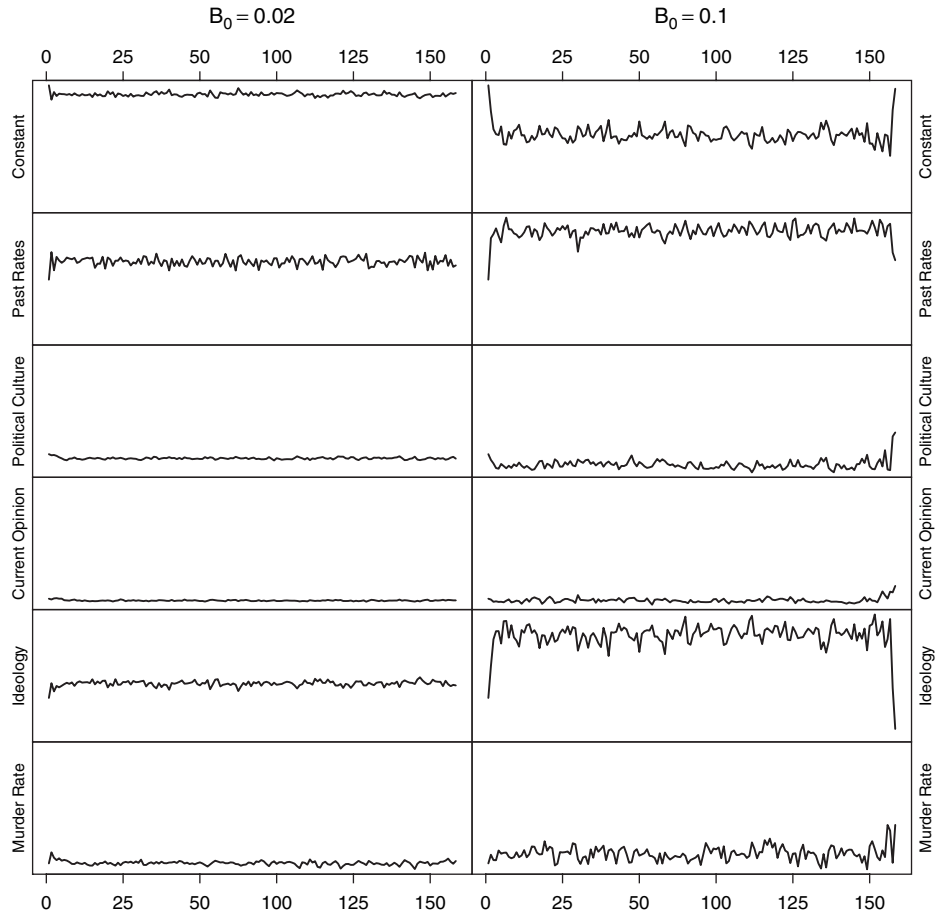

Fig. 1 Comparison of Bayesian tobit models.

$T_{\lfloor sT \rfloor} = \sum_{i=1}^{\lfloor sT \rfloor} \theta_i =$ the sum of the chain values from one to the integer value just below sT, $\lfloor sT \rfloor \bar{\theta} =$ the chain mean times the integer value just below sT, and $s(0) =$ the spectral density of the chain (the notation $\lfloor X \rfloor$, here, denotes the integer value of X. This lets us construct a Cramér-von Mises test statistic for sums as cumulative values scaled by the spectral density, for any given s: $B_T(s) = (T_{\lfloor sT \rfloor} - \lfloor sT \rfloor \bar{\theta}) / \sqrt{Ts(0)}$. The mechanics of this test increment the s parameter by units of 0.10 until the resulting chain either passes (e.g., the test statistic is within $[-1.96 : 1.96]$ for the default $\alpha = 0.05$) or more than half of the available iterations have been discarded and general failure is prescribed.

The regular diagnostics are somewhat disturbing for this example. Notice from Table 2 that for the model with $B_0 = 0.02$, all three of these tests show no cause for concern. The BGR Scale Reduction Factors (using four comparison chains) all hover around one (the value for Past Rates is slightly less than one due to rounding/truncating), the Geweke z-scores are all well-within $[-1.96 : 1.96]$, and the Heidelberger and Welch Cramér-von Mises test statistics are similarly small. Unfortunately, moving B_0 to 0.1 does not allow us to collect many observations for the diagnostics, but running them on the last 200 of 300 total gives the results in the second part of Table 2 (it is nearly impossible to get accepted values after this point, so the chain remains in place). First notice that the BGR tests show

Table 2 MCMC empirical diagnostics, tobit model

	BGR Scale Reduction Factor	*Geweke (at defaults)*	*Heidelberger and Welch Cramér-von Mises*
$\mathbf{B}_0 = 0.02$			
Constant	1.000010	−0.692213	0.1317036
Past rates	0.999808	−0.875609	0.2302676
Political culture	1.000052	−0.329635	0.0667068
Current opinion	1.000909	0.821076	0.1370106
Ideology	1.000250	0.518202	0.1112524
Murder rate	1.000436	0.417410	0.0581120
$\mathbf{B}_0 = 0.1$			
Constant	1.016265	1.429219	0.317495
Past rates	1.001267	−1.312726	0.2663231
Political culture	1.015474	−0.191747	0.0734364
Current opinion	1.047117	−1.869762	0.2664384
Ideology	1.003489	−0.653879	0.1268610
Murder rate	1.007958	−2.454625	0.2949386

no problem whatsoever. This is because all the chains are simultaneously becoming unbounded and there is no notable difference between them as they go. The Geweke diagnostic shows several large values, the one for Murder Rate in particular. Although Table 2 gives the Geweke values at the default window widths, varying these parameters provided very similar statistics. Regretfully, the Heidelberger and Welch diagnostic does not catch problems here. Manipulating the two available parameters for the second half of the test (error "accuracy" of the posterior estimates for the parameters set at 0.1, and the alpha level for the confidence in the sample mean of the retained iterations set at 0.05), a less useful "halfwidth" comparison, can reveal problems. However, like Geweke, this is not common practice among practitioners. Fortunately, in this case the eventual decline of the truncated normal draw acceptance rate to 0 indicates the existence of an algorithmic problem independently of these formal tests.

What is the lesson here? The BGR diagnostic only provides measures of similarity between the parallel chains. Normally, this is enormously useful since well-chosen starting points spread widely throughout the sample space that lead to chain co-location at a later time are a reassuring phenomenon, but it does not preclude the possibility that all the parallel chains are seduced by a nonoptimal region for the time period being considered. Researchers working with suspect chains may consider repeating the BGR test with additional starting points dissimilar to the original set. The Geweke test also comes with a strong caution. The selected comparison windows are totally arbitrary, and it is possible to imply stability on instability with some chains simply by altering this selection. Unfortunately, most researchers simply leave the defaults in place. The Heidelberger and Welch diagnostic comes with similar concerns since the manner in which the *s* parameter is used is not changeable in coda or boa.

6.2 *Convergence Problems from Identification Error (Metropolis-Hastings)*

In this section, we develop a Bayesian multinomial logit model for party choice in the 1972 Italian election. This is an interesting illustration because estimation of multinomial logit models is more complex than standard generalized linear models, and multinomial

consideration of nine party choices is also difficult. The data come from a postelection national survey of 1841 Italian citizens (reduced here to 1839 due to coding issues with two cases; there is no missing data) querying their political attitudes and obtaining demographic information (ICPSR-7954). The respondents give their vote choice by party,[6] which is the studied outcome variable. The explanatory variables are education level (ordinal from 1 = never been to school to 9 = postgraduate study), marriage status (0 = unmarried, 1 = married), sex (0 = male, 1 = female), age (reported value in years), religiosity (ordinal with 1 = very, 2 = somewhat, 3 = little, 4 = not), and employment (0 = no, 1 = yes). Details and frequencies are available online at http://www.icpsr.umich.edu for study number 07954. All coefficient priors are specified as improper uninformed forms: $p(\beta_j) \propto k$ over $[-\infty : \infty]$.

As a reminder, a multinomial logit model with J choices is estimated with respect to a reference category in order to be identified so that the resulting coefficient vectors, $\boldsymbol{\beta}_j,\ j = 1 \ldots J-1$ (the reference category gets $\boldsymbol{\beta}_1 = \mathbf{0}$), provide the relative effect through the logit function of that explanatory variable on the probability that the respondent chose a specific category rather than the reference category. So given the observed data matrix $\mathbf{X}$, the probability that respondent i chooses category j over category 1 is given by: $P(y_{ij}) = \exp(\mathbf{X}_i\boldsymbol{\beta}_j)/\sum_{k=1}^{J}\exp(\mathbf{X}_i\boldsymbol{\beta}_k)$. Estimation of the Bayesian multinomial logit combines log likelihood function, $\ell(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J) = \sum_{i=1}^{n}\sum_{j=1}^{J} y_i \log P(y_{ij})$ (Amemiya 1985, p. 295), with prior distributions for all unknown parameters using Bayes law. This model is implemented with a random walk Metropolis-Hastings algorithm from the MCMCpack package in R (Martin and Quinn 2005, http://mcmcpack.wustl.edu/wiki/index.php/Main_page).

The primary issue of concern with this example is the mixing effect on nonconvergence. The single source of mixing issues is that the third party category (Partito Socialista Italiano [PSI]) has only 13 votes in the survey providing a serious identification problem in the model. As seen in Fig. 2, the coefficient for the impact of being in the not religious category on voting for the third party relative to a minor party or abstaining has poor mixing qualities (panel 1). The figure shows the traceplot for the first 10,000 iterations after a burn-in period of 1000 iterations for four coefficient dimensions. The poor mixing for RELIGIOUSNot.3 is particularly manifest here and would have provided an incorrect view of the posterior distribution if we had stopped here (a posterior mean of 163.71 versus the correct one of 1467.41). The chain is subsequently run up to 300,000 iterations with the last 100,000 shown. We can still see, however, that there is significant "snaking" in the first panel indicating a stickiness in the path of the chain (i.e., poor mixing even in the long run). Surprisingly, these last 100,000 values provide good convergence diagnostics for all dimensions including this problematic one (see below).

Interestingly, the anticipated effect on other dimensions is visible for three example paths in the remaining panels of Fig. 2. The horizontal bars in all four panels show 90% highest posterior density interval thresholds. What is clear, to varying degrees, is that while the early period of the three chains has close to the same posterior mean (-1.25 versus -1.30, 1.12 versus 1.23, -0.21 versus 0.17, accordingly), but noticeably different ranges of travel. So while the long-term effect on the other parameters dissipates to 0, stopping the chain earlier would have been detrimental even to dimensions seemingly unconcerned

[6]The listed parties are Partito Comunista Italiano (PCI), Socialist Party of Proletarian Unity (PSIUP), Partito Socialista Italiano (PSI), Partito Socialista Democratico Italiano (PSDI), Partito Republicano Italiano (PRI), Democrazia Cristiana (DC), Partito Liberale Italiano (PLI), and Movimento Sociale Italiano (MSI). The "Other" category is a collapsing of indicated votes for minor parties where only 0–3 of the survey respondents indicated a positive vote choice.

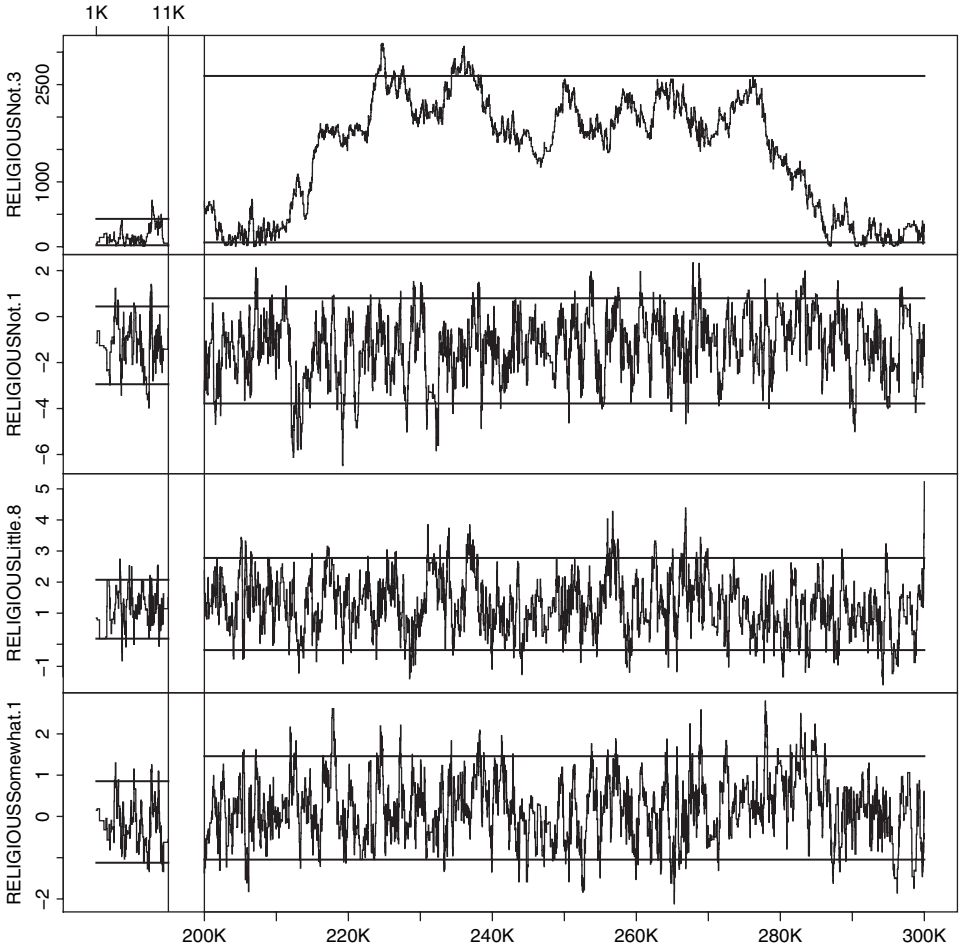

Fig. 2 Traceplots and 90% HPD lines, Italy election model.

with the problems from the PSI party choice. Furthermore, rerunning the model but moving this party into the "other" category alleviates the mixing problem.

The formal diagnostic tests are difficult and incomplete for this example. Restricting ourselves to the four parameters of interest above, the BGR test fails in all cases using boa or coda. Worse yet, boa and coda have difficulty even reporting values for Geweke and Heidelberger and Welch. For Geweke, coda was able to report two parameter values but crashed on the other two, whereas boa crashed on all four. Note that the two Geweke values reported in Table 3 are extremely large. For Heidelberger and Welch, boa crashed on all four, and coda was able to report only one parameter test, which surprisingly passes. Obviously, it is better here to fail to report test values than to erroneously imply convergence. Still, it is disappointing to not have more information about the reasons for the software behavior, although we know that the available time period here is insufficient for these two diagnostics based on time-series principles. Clearly, though, the contents of Table 3 point towards nonconvergence.

Table 3 MCMC empirical diagnostics, Italy election model

	BGR Scale Reduction Factor	*Geweke (at defaults)*	*Heidelberger and Welch Cramér-von Mises*
RELIGIOUSNot.3	2.552787	—	—
RELIGIOUSNot.1	3.601752	11.45	—
RELIGIOUSLittle.8	2.792330	—	0.07870
RELIGIOUSSomewhat.1	4.524198	30.94	—

7 Recommendations

The paradox here is that we really care about *joint* posterior convergence but end up analytically looking strictly at *marginal* posterior convergence. That is, almost every diagnostic, formal or graphic, evaluates convergence one dimension at a time (the BGR diagnostic does have an omnibus statistic, the *Multivariate Potential Scale Reduction Factor*, but this is rarely reported by authors). The key problem is that marginal convergence, partial or complete, is not the same thing as joint convergence. So the question remains what should a conscientious practitioner do to maximize their confidence about convergence? Can we define some sense of the "best practices" here that will increase reader confidence in MCMC results?

One strategy is to be thoughtful about starting values for the Markov chain. If the maximum likelihood estimation is used as a starting point for the Markov chain, then one can be more certain beginning at an area of high posterior density. Robert and Casella (2004, p. 461) note that in cases where the target distribution is approximately known, it is possible to start the chain as if it were already in stationary regime since this point belongs to a region of high probability. Gelman and Rubin (1992, p. 459) suggest using the EM algorithm to find posterior modes noting that EM should be started from several different points if multimodality is suspected. Another, more commonly ignored, value is the random seed. Generally, it is easy to vary this value for different runs, and WinBUGS gives a simple pull-down menu for this purpose. One interesting feature of WinBUGS is that when the same model is rerun with the same seed, the exact same chain values are produced.

The three most practical and convenient convergence diagnostics are Brooks, Gelman and Rubin; Geweke; and Heidelberger and Welch. These are easily understood tools that are offered in menu-drive form in the R packages coda and boa. Other tools either require coding by the practitioner or have been shown to possess nonoptimal properties. The general advice here is to run all these diagnostics with long chains and close inspections.

Of these three, BGR requires the most experience and sophistication to use correctly. This is because the diagnostic requires overdispersed starting points for running multiple chain in order to make within- and between-variance comparisons. Here "overdispersed" means spread out through the sample space sufficiently widely apart that their relative distances exceed the major range of the density. In small dimensions this is not a difficult or time-consuming enterprise, although it may require some analytical knowledge of the marginal posterior forms or the application of the EM algorithm or a grid search to find areas of high density. However, with high dimensional posterior forms, this can be a daunting task, and in practice, it is often eased by either guessing the range or randomly assigning starting points. Simply guessing what counts as overdispersion can be deleterious and may dramatically alter the results. Using randomly selected starting points for the

separate chain values is better if done with caution and repeated several times. An advantage, though, is that normally BGR does not require long runs of the Markov chain to see useful convergence information, so repeated analyses with different sets of starting points is not overly taxing.

Consider again a Metropolis-Hastings algorithm stopped at time t where the BGR test provides Scale Reduction Factors that differ by dimension: dimensions $1:e$ have values near 1, but $e+1:d$ are considerably larger. From this we would surmise that $\| f(\boldsymbol{\theta}_t^*) - \pi(\boldsymbol{\theta}^*) \| \approx \delta$, $\forall \boldsymbol{\theta}^* \in \Omega^e$, and $\| f(\boldsymbol{\theta}_t^\dagger) - \pi(\boldsymbol{\theta}^\dagger) \| \gg \delta$, $\forall \boldsymbol{\theta}^\dagger \in \Omega^{d-e}$. This situation should trigger concern about mixing, which can easily be checked by adding code to accumulate acceptance decisions (MCMCpack reports this to the user by default). Very small acceptance rates, say below 1%, are an indication that this problem may be seriously affecting the algorithm. Since Gibbs sampling moves on every iteration, such partial results from the BGR diagnostic have a different interpretation. For dimensions where the posterior form is relatively flat and featureless, there is much less of an inclination for the multiple chains to converge to the same area. So with Gibbs sampling, this behavior may indicate a poorly specified model.

As seen in the examples, both Geweke and Heidelberger and Welch rely on running the Markov chain for a relatively long period of time in order to view possibly different eras. Thus, by comparing early periods to late periods, a distributional test can be performed to make assertions about changing chain behavior with simple summary statistics. So this is an intuitive approach, even though it requires relatively longer runs than BGR. The biggest concern with the Geweke diagnostics is that the results can vary with differing window selections and bin sizes, producing different views of overall convergence. Cautious researchers should change these values (again, simply a menu-driven process in coda and boa) several times and watch for any evidence of nonconvergence in any of the dimensions.

How can the time-series nature of the Geweke and Heidelberger and Welch diagnostics help when they indicate partial convergence. With both tools we are comparing different eras of the chain, and with Metropolis-Hastings it may be the case that the acceptance rates differ dramatically by period. With minor coding we can turn on accumulation of acceptance decisions over these periods. This is not directly available in WinBUGS, JAGS, or MCMCpack, but implementations coded by the research in C, R, or some other programming language can be modified to record such information. For example, at the Geweke defaults, if the [0.4 : 0.5] era has significantly lower values of $a(\boldsymbol{\theta}, \boldsymbol{\theta}')$ or acceptance decisions than the [0.5 : 1.0] era, it would explain large test statistics for some dimensions. In the case of Gibbs sampling, this provides less prescriptive information but may indicate the existence of wide, flat posterior areas in some dimensions as noted above with BGR.

While discussing these diagnostics, it is critical to make one important point about the required strength of evidence that is provided. It would be highly unrealistic to require that every diagnostic pass every dimension, particularly with a large number of dimensions. Such reasoning would ignore the fact that these diagnostics make use of formal hypothesis tests and thus will incorrectly reject the null of stationarity for some dimensions at the prechosen level. Suppose we had hundreds or thousands of parameters under consideration, even in complete convergence 5% of the Geweke z-statistics will be greater than 1.96 in absolute value. This is obviously less critical in model specifications with many fewer explanatory variables, but the intention here is to recommend cautious inspection by multiple means rather than dogmatic application of the single-dimension formal tests with a 100% pass criteria.

Graphical diagnostics require special caution. Perhaps the worst case is the traceplot in WinBUGS because it dynamically resizes the *y* axis of the plot as it runs thus giving the illusion of chain stability under all but the most extreme circumstances. Users of this software should consider the dynamic traceplot (labeled as just "traceplot" in the analysis window) as merely an amusement. The "history" button provides a full graph of the chain path, and more usefully coda and boa provide flexible implementations. One more caveat applies to the use of traceplots. Occasionally, one sees traceplots in published or unpublished work where the starting points are included and these starting points are far from the main area of the stationary distribution for that marginal posterior. The net effect is to make the chain appear very stable due to scale effects in the graph, even when there is substantial snaking or alternating between modes. This effect can also be seen on running mean/cumsum graphs.

Another graphical summary to worry about, although much more benignly, is the smoothed density graph provided by all relevant packages and easily programmed on one's own. These are not formal tests of convergence in any sense but can reveal some sense of marginal instability. Since the smoothing parameter is set by the user, or more likely left at a default, a marginal posterior can appear more smooth than it should be. Furthermore, overly smooth specifications can hide multimodal features that may be substantively important.

An unappreciated approach that adds to one's confidence is the idea of performing an "insurance run." The notion here is to add an additional step after the analysis and therefore after one has done everything possible to provide convincing evidence of convergence. While the researcher moves on to writing up the results and discussing the findings, run one extremely long chain (perhaps in the background, depending on computing platform). This is where thinning the chain (only saving every 10th iteration or so) is useful for storage purposes. After a few days or longer, the chain can be checked, and if the results do not comport with the previous findings, then this is cause for alarm. Experience shows that really is just a means of providing extra confidence for those who were careful in the ways described above.

The mixing rate is another concern that bears on convergence. Poor mixing does not necessarily mean that there exist fundamental model or algorithmic problems but it does slow progress towards convergence and then full exploration of the posterior. Since all Markov chains in applied settings are run in finite time, slow mixing can lead to inference problems. Tests for mixing can be quite complex and time consuming, so most wary practitioners compare the range of chain visits in each dimension to the known distributional support for that dimension. This is certainly an ad hoc approach but one that provides useful information. For the Metropolis-Hastings algorithm, one easy diagnostic is to look at the acceptance rate of new destinations. Generally, something between 20% and 50% is a sign of healthy mixing properties (more specific guidance is given in Gelman, Roberts, and Gilks 1996). A creative strategy for checking mixing in general is to manipulate the Geweke diagnostic to check small intervals rather than the large sweeping percentages normally used. If chain averages are not different across small ranges of hundreds or more, then there is evidence of stickiness. More elaborate procedures for improving Markov chain mixing (generally called "acceleration algorithms") include simulated annealing (Kirkpatrick, Gelatt, and Vecchi 1983), simulated tempering (Marinari and Parisi 1992; Geyer and Thompson 1995), and related dynamic variants (Gill and Casella 2004), blocking where parameters are updated in groups (Liu 1994, Roberts and Sahu 1997), and collapsing where parameters values are generated partial marginalizations (Gill 2007, chap. 12). Chen, Shao, and Ibrahim (2000, p. 43–52) give detailed

descriptions of grouped move, multigrid Monte Carlo sampling, and covariance-adjusted MCMC, all to speed mixing.

Perhaps the most common and practical way to alleviate mixing problems is to reparameterize model variables. This works because poor mixing is often caused by high correlations between these terms and can be detected through graphical diagnostics and checking parameter correlation matrix (Hobert, Robert, and Goutis 1997), and there is evidence that the Gibbs sampler is especially sensitive to parameter correlation (Gilks and Roberts 1996). Although such approaches depend heavily on the particular model and data at hand and thus very general recommendations do not make sense, two strategies dominate. First, one should try reparameterizations that make each model component as independent of the others as possible. Common reparameterization tricks here are specific parametric forms, for example, $\gamma_1 = \theta_1 + \theta_2$, $\gamma_2 = \theta_1 - \theta_2$ and possibly quadratics, centering the covariates $\mathbf{x}'_j = \mathbf{x}_j - \bar{x}$ and centering by "sweeping" all the parameter means into a single term (Vines and Gilks 1994). Also, Robert and Casella (2004, p. 398–399) show two particular examples where reparameterization and subsequent imposition of identity constraints improves the mixing of the Gibbs sampler. Second, since specific boundings of the parameters tend to slow down mixing, consider transformations onto the entire real line (for continuous-measured parameters). The workhorse here is log transformation, but other forms are often reasonable (Gilks and Roberts 1996).

Finally, data augmentation (Tanner and Wong 1987) can help with mixing properties even though it most often posited as a means of dealing with missing data in Bayesian models (Liu 2001, p. 135–8). In particular, see Carlin and Polson (1991); Albert and Chib (1993); Rosenthal (1993); Lui, Wong, and Kong (1994); and Swendsen and Wang (1987) and more recently Imai and van Dyk (2005) and Gelman et al. (2006). The basic idea behind data augmentation in the context here is that if an MCMC algorithm to marginalize $\pi(\boldsymbol{\theta} \mid \mathbf{X})$ is mixing slowly, then it is sometimes possible to find another convenient random variable $\boldsymbol{\Upsilon}$ such that the algorithm operates more efficiently through sampling from $\pi(\boldsymbol{\theta}, \boldsymbol{\Upsilon} \mid \mathbf{X})$ by alternating between $\pi(\boldsymbol{\theta} \mid \boldsymbol{\Upsilon}, \mathbf{X})$ and $\pi(\boldsymbol{\Upsilon} \mid \boldsymbol{\theta}, \mathbf{X})$. Chib (1992) explains how this can be quite simple and elegant when $\boldsymbol{\Upsilon}$ is a latent feature in the model specification. Meng and van Dyk (1999) and van Dyk and Meng (2001) extend this idea with conditional and marginal versions of data augmentation and take advantage of unidentifiable parameters to improve the rate of convergence.

8 Conclusion

In an era when just about anyone can run MCMC in WinBUGS, MCMCpack, or other convenient software, it is important to think seriously about how we assess convergence. It is not hard to imagine a number of well-intentioned researchers who become lulled into a false sense of security by the mechanical nature of standard empirical diagnostics. In an experiment, Cowles, Roberts, and Rosenthal (1999) ran 300 parallel chains for 1000 iterations starting already in stationarity with a simple statistical model, and still obtained conflicting guidance from a common diagnostic (Geweke with different options tried) about how long the burn-in period should be (4–49). Gill (2007, chap. 12) shows further that the different commonly used diagnostics can provide conflicting results relative to each other, even about the same parameters.

We need to think more about what the standard empirical diagnostics *are* or *are not* telling us when there is slow mixing and only partial evidence. We have shown here that the performance across multiple dimensions is far from independent, and therefore we

ignore nonconvergence in some dimensions at our own peril (although it differs in effect by algorithm and model). Furthermore, the subchains that appear to have converged could be simply exploring a false attraction imposed by nonstationarity. This, of course, is related to mixing which was discussed here as well. Another linkage between mixing and convergence was noted by Brooks and Roberts (1998, p. 320). "In particular, for slow mixing Markov chains, convergence diagnostics are likely to be unreliable since their conclusions will be based upon output from only a small region of the state space."

It is surprising that no serious attention has been given to considering the implication for partial convergence, although a related problem in statistical physics has been considered (Manita [1997] looks at multiple particles moving Markovian, but assumes complete independence making the results not applicable to the statistical problems encountered routinely used with MCMC in applied statistics). Therefore, it is hoped that this discussion points practitioners of MCMC, a growing group in political science, towards more caution and therefore better statistical inferences.

Finally, this paper is neither intended as a polemic against extensive use of MCMC in political science nor to imply that things often go awry. In fact, most chains readily converge with typical models and provide strong evidence that they have done so. We know a lot about the traversal of mode-finding algorithms in maximum likelihood estimation settings, including the things that could wrong. The purpose here is merely to provide a similar level of knowledge and comfort about this newer estimation process.

References

Albert, James H., and Chib, Siddhartha. 1993. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* 88:669–79.

Amemiya, Takeshi. 1985. *Advanced econometrics.* Cambridge, MA: Harvard University Press.

Amit, Y. 1991. On rates of convergence of stochastic relaxation for Gaussian and non-Gaussian distributions. *Journal of Multivariate Analysis* 38:82–99.

Amit, Y., and Grenander, U. 1991. Comparing sweep strategies for stochastic relaxation. *Journal of Multivariate Analysis* 37:197–222.

Asmussen, S. P., Glynn, P., and Thorisson, H. 1992. Stationarity detection in the initial transient problem. *ACM Transactions on Modeling and Computer Simulation* 2:130–57.

Athreya, K. B., Doss, Hani, and Sethuraman, Jayaram. 1996. On the convergence of the Markov chain simulation method. *Annals of Statistics* 24:69–100.

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

Barker, A. A. 1965. Monte Carlo calculation of the radical distribution functions for a proton-electron plasma. *Australian Journal of Physics* 18:119–33.

Brooks, S. P., Dellaportas, P., and Roberts, G. O. 1997. An approach to diagnosing total variation convergence of MCMC algorithms. *Journal of Computational and Graphical Statistics* 6:251–65.

Brooks, S. P., and Roberts, G. O. 1998. Convergence assessment techniques for Markov chain Monte Carlo. *Statistics and Computing* 8:319–35.

Carlin, B. P., and Polson, N. G. 1991. Inference for nonconjugate Bayesian models using the Gibbs sampler. *Canadian Journal of Statistics* 19:399–405.

Casella, G., and Robert, C. P. 1996. Rao-Blackwellization of sampling schemes. *Biometrika* 93:81–94.

Chan, K. S. 1993. Asymptotic behavior of the Gibbs sampler. *Journal of the American Statistical Association* 88:320–6.

Chan, K. S., and Geyer, C. J. 1994. Discussion of Markov chains for exploring posterior distributions. (L. Tierney). *Annals of Statistics* 22:1747–58.

Chen, M.-H., Shao, Q.-M., and Ibrahim, J. G. 2000. *Monte Carlo methods in Bayesian computation.* New York: Springer-Verlag.

Chib, S. 1992. Bayes inference in the tobit censored regression model. *Journal of Econometrics* 51:79–99.

Chib, S., and Greenberg, E. 1995. Understanding the Metropolis-Hastings algorithm. *The American Statistician* 49:327–35.

Cowles, Mary Kathryn, and Carlin, Bradley P. 1996. Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association* 91:883–904.

Cowles, Mary Kathryn, Roberts, Gareth O., and Rosenthal, Jeffrey S. 1999. Possible biases induced by MCMC convergence diagnostics. *Journal of Statistical Computation and Simulation* 64:87–104.

Cowles, Mary Kathryn, and Rosenthal, Jeffrey S. 1998. A simulation approach to convergence rates for Markov chain Monte Carlo algorithms. *Statistics and Computing* 8:115–24.

Creutz, M. 1979. Confinement and the critical dimensionality of space-time. *Physical Review Letters* 43:553–56.

Creutz, M., Jacobs, L., and Rebbi, C. 1983. Monte Carlo computations in lattice gauge theories. *Physical Review* 95:201.

Diaconis, Persi, and Saloff-Coste, Laurent. 1993. Comparison theorems for reversible Markov chains. *Annals of Applied Probability* 3:696–730.

———. 1996. Logarithmic Sobolev inequalities for finite Markov chains. *Annals of Applied Probability* 6:695–750.

Diaconis, Persi, and Stroock, Daniel. 1991. Geometric bounds for eigenvalues of Markov chains. *Annals of Applied Probability* 1:36–61.

Doeblin, W. 1940. Élements d'Une Thérie Générale des Chaînes Simple Constantes de Markoff. *Annales Scientifiques de l'Ecole Normale Supérieure* 57:61–111.

Fill, J. 1991. Eigenvalue bounds on convergence to stationarity for nonreversible Markov chains with an application to the exclusion process. *Annals of Applied Probability* 1:62–7.

Frieze, A., Kannan, R., and Polson, N. G. 1993. Sampling from log-concave distributions. *Annals of Applied Probability* 4:812–37.

Frigessi, Arnoldo, Hwang, Chii-Ruey, Di Steffano, Patrizia, and Sheu, Shuenn-Jyi. 1993. Convergence rates of the Gibbs sampler, the Metropolis algorithm, and other single-site updating dynamics. *Journal of the Royal Statistical Society, Series B* 55:205–20.

Fulman, Jason, and Wilmer, Elizabeth L. 1999. Comparing eigenvalue bounds for Markov chains: When does Poincare beat Cheeger? *Annals of Applied Probability* 9:1–13.

Gawande, Kishore. 1998. Comparing theories of endogenous protection: Bayesian comparison of tobit models using Gibbs sampling output. *Review of Economics and Statistics* 80:128–40.

Gelman, Andrew, Huang, Zaiying, van Dyk, David A., and Boscardin, John W. 2006. Transformed and parameter-expanded Gibbs samplers for multilevel linear and generalized linear models. Working paper: http://www.stat.columbia.edu/∼gelman/research/unpublished/alpha4.pdf.

Gelman, A., Roberts, G. O., and Gilks, W. R. 1996. Efficient Metropolis jumping rules. In *Bayesian Statistics 5*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, 599–608. Oxford: Oxford University.

Gelman, A., and Rubin, D. B. 1992. Inference from iterative simulation using multiple sequences. *Statistical Science* 7:457–511.

Geman, S., and Geman, D. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6:721–41.

Geyer, C., and Thompson, E. 1995. Annealing Markov chain Monte Carlo with applications to ancestral inference. *Journal of the American Statistical Association* 90:909–20.

Gilks, W., and Roberts G. 1996. Strategies for improving MCMC. In *Markov Chain Monte Carlo in practice*, eds. W. Gilks, S. Richardson, and D. Spiegelhalter, 89–114. New York: Chapman & Hall.

Gill, Jeff. 2007. *Bayesian methods for the social and behavioral sciences.* 2nd ed. New York: Chapman & Hall.

Gill, Jeff, and Casella, George. 2004. Dynamic tempered transitions for exploring multimodal posterior distributions. *Political Analysis* 12:425–43.

Gill, Jeff, and Walker, Lee. 2005. Elicited priors for Bayesian model specifications in political science research. *Journal of Politics* 67:841–87.

Goodman, J., and Sokal, A. D. 1989. Multigrid Monte Carlo method. Conceptual foundations. *Physics Review D* 40:2035–71.

Grenander, Ulf. 1983. *Tutorial in pattern theory.* Technical Report, Division of Applied Mathematics. Providence, RI: Brown University.

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

Hill, J., and Kriesi, H. 2001. Classification by opinion changing behavior: A mixture model approach. *Political Analysis* 4:301–24.

Hobert, J., Robert, C., and Goutis, C. 1997. Connectedness conditions for the convergence of the Gibbs sampler. *Statistics and Probability Letters* 33:235–40.

Imai, Kosuke, and van Dyk, David A. 2005. A Bayesian analysis of the multinomial probit model using marginal data augmentation. *Journal of Econometrics* 124:311–34.

Ingrassia, Salvatore. 1994. On the rate of convergence of the Metropolis algorithm and Gibbs sampler by geometric bounds. *Annals of Applied Probability* 4:347–89.

Jackman, Simon. 2000a. Estimation and inference are missing data problems: Unifying social science statistics via Bayesian simulation. *Political Analysis* 8:307–32.

———. 2000b. Estimation and inference via Bayesian simulation: An introduction to Markov chain Monte Carlo. *American Journal of Political Science* 44:375–404.

———. 2001. Multidimensional analysis of Roll Call Data via Bayesian simulation: Identification, estimation, inference, and model checking. *Political Analysis* 9:227–41.

Kass, R. E., Carlin, B. P., Gelman, A., and Neal, R. M. 1998. Markov chain Monte Carlo in practice: A roundtable discussion. *American Statistician* 52:93–100.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. 1983. Optimization by simulated annealing. *Science* 220: 671–80.

Krishnan, Neeraja M., Seligman, Hervé, Stewart, Caro-Beth Stewart, de Konig, A. P. Jason, and Pollock, David D. 2004. Ancestral sequence reconstruction in primate mitochondrial DNA: Compositional bias and effect on functional inference. *Molecular Biology and Evolution* 21:1871–83.

Lawler, G. F., and Sokal, A. D. 1988. Bounds on the L^2 spectrum for Markov chains and their applications. *Transactions of the American Mathematical Society* 309:557–80.

Liu, J. S. 1994. The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association* 89:958–66.

———. 1996. Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statistics and Computing* 6:113–9.

———. 2001. *Monte Carlo strategies in scientific computing.* New York: Springer-Verlag.

Liu, J. S., Wong, W. H., and Kong, 1994. Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* 81:27–40.

Macintosh, Duncan. 1994. Partial convergence and approximate truth. *British Journal for the Philosophy of Science* 45:153–70.

Manita, A. D. 1997. Convergence time to equilibrium for multi-particle Markov chains. Preprint of French-Russian Institute, No. 3. Moscow University, September 1997. http://mech.math.msu.su/~manita/.

Marinari, E., and Parisi, G. 1992. Simulated tempering: A new Monte Carlo scheme. *Europhysics Letters* 19:451–8.

Martin, Andrew D., and Quinn, Kevin M. 2002. Dynamic ideal point estimation via Markov chain Monte Carlo for the U.S. Supreme Court, 1953–1999. *Political Analysis* 10:134–53.

Meng, X.-L., and van Dyk, D. A. 1999. Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika* 86:301Ð320.

Mengersen, K. L., and Tweedie, R. L. 1996. Rates of convergence of the Hastings and Metropolis algorithms. *Annals of Statistics* 24:101–21.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller E. 1953. Equations of state calculations by fast computing machine. *Journal of Chemical Physics* 21:1087–91.

Meyn, S. P., and Tweedie, R. L. 1993. *Markov chains and stochastic stability.* New York: Springer-Verlag.

———. 1994. Computable bounds for geometric convergence rates of Markov chains. *Annals of Probability* 4:981–1011.

Mira, A., and Tierney, L. 2001. Efficiency and convergence properties of slice samplers. *Scandanavian Journal of Statistics* 29:1035–53.

Norrander, Barbara. 2000. The multi-layered impact of public opinion on capital punishment implementation in the American states. *Political Research Quarterly* 53:771–93.

Norris, J. R. 1997. *Markov chains.* Cambridge: Cambridge University Press.

Nummelin, E. 1984. *General irreducible Markov chains and non-negative operators.* Cambridge: Cambridge University Press.

Peskun, P. H. 1973. Optimum Monte Carlo sampling using Markov chains. *Biometrika* 60:607–12.

Polson, N. G. 1996. Convergence of Markov chain Monte Carlo algorithm. In *Bayesian statistics 5*, eds. J. M Bernardo, A. F. M. Smith, A. P. Dawid, and J. O. Berger, 297–323. Oxford: Oxford University Press.

Quinn, Kevin M., Martin, Andrew, and Whitford, Andrew B. 1999. Voter choice in multi-party democracies: A test of competing theories and models. *American Journal of Political Science* 43:1231–47.

Ripley, B. D. 1979. Algorithm AS 137: Simulating spatial patterns: Dependent samples from a multivariate density. *Applied Statistics* 28:109–12.

Robert, C. P. 1995. Convergence control methods for Markov chain Monte Carlo algorithms. *Statistical Science* 10:231–53.

Robert, C. P., and Casella, G. 2004. *Monte Carlo statistical methods.* 2nd ed. New York: Springer-Verlag.

Robert, Christian P., and Richardson, Sylvia. 1998. Markov chain Monte Carlo methods. In *Discretization and MCMC Convergence Assessment*, ed. C. P. Robert, 1–25. New York: Springer.

Roberts, G. O., and Polson, N. G. 1994. On the geometric convergence of the Gibbs sampler. *Journal of the Royal Statistical Society, Series B* 56:377–84.

Roberts, G. O., and Rosenthal, J. S. 1998a. Markov chain Monte Carlo: Some practical implications of theoretical results. *Canadian Journal of Statistics* 26:5–32.

———. 1998b. Two convergence properties of hybrid samplers. *Annals of Applied Probability* 8:397–407.

———. 1999. Convergence of the slice sampler Markov chains. *Journal of the Royal Statistical Society, Series B* 61:643–60.

Roberts, G. O., and Sahu, S. K. 1997. Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society, Series B* 59:291–307.

Roberts, G. O., and Smith, A. F. M. 1994. Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stochastic Processes and their Applications* 44:207–16.

Roberts, G. O., and Tweedie, R. L. 1996. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika* 83:95–110.

Rosenblatt, M. 1971. *Markov processes. Structure and asymptotic behavior.* New York: Springer-Verlag.

Rosenthal, Jeffrey S. 1993. Rates of convergence for data augmentation on finite sample spaces. *Annals of Applied Probability* 3:819–39.

———. 1995a. Convergence rates for Markov chains. *SIAM Review* 37:387–405.

———. 1995b. Minorization conditions and convergence rates for Markov chain Monte Carlo. *Journal of the American Statistical Association* 90:558–66.

———. 1996. Analysis of the Gibbs sampler for a model related to James-Stein estimators. *Statistics and Computing* 6:269–75.

Salzano, Marcia, and Schonmann, Roberto H. 1997. The second lowest extremal invariant measure of the contact process. *Annals of Probability* 25:1846–71.

———. 1999. The second lowest extremal invariant measure of the contact process II. *Annals of Probability* 27:845–75.

Sinclair, A. J., and Jerrum, M. R. 1988. Conductance and the rapid mixing property for Markov chains: The approximation of the permanent resolved. *Proceedings of the 20th Annual ACM Symposium on the Theory of Computing* 235–44.

———. 1989. Approximate counting, uniform generation and rapidly mixing Markov chains. *Information and Computation* 82:93–133.

Smith, Alastair. 1999. Testing theories of strategic choice: The example of crisis escalation. *American Journal of Political Science* 43:1254–83.

Swendsen, R. H., and Wang, J. S. 1987. Nonuniversal critical dynamics in Monte Carlo simulations. *Physical Review Letters* 58:86–8.

Tanner, M. A., and Wong, W. H. 1987. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Society* 82:528–50.

Tierney, L. 1994. Markov chains for exploring posterior distributions. *Annals of Statistics* 22:1701–28.

Tobin, James. 1958. Estimation of relationships for limited dependent variables. *Econometrica* 26:24–36.

Tweedie, R. L. 1975. Sufficient conditions for ergodicity and recurrence of Markov chains on a general state space. *Stochastic Processes Applications* 3:385–403.

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

Vines, S., and Gilks, W. 1994. Technical Report, MRC Biostatistics Unit, Cambridge University.

Western, Bruce. 1998. Causal heterogeneity in comparative research: A Bayesian hierarchical modeling approach. *American Journal of Political Science* 42(4):1233–59.

Zellner, A., and Min, C.-K. 1995. Gibbs sampler convergence criteria. *Journal of the American Statistical Association* 90:921–7.