ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Convergence Diagnostics for Markov Chain Monte Carlo

Vivekananda Roy

Department of Statistics, Iowa State University, Ames, Iowa 50011, USA; email: vroy@iastate.edu

Annu. Rev. Stat. Appl. 2020. 7:387–412

First published as a Review in Advance on November 20, 2019

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-031219-041300

Keywords

autocorrelation, empirical diagnostics, Gibbs sampler, Metropolis algorithm, MCMC, Markov chain Monte Carlo, stopping rules

Abstract

Markov chain Monte Carlo (MCMC) is one of the most useful approaches to scientific computing because of its flexible construction, ease of use, and generality. Indeed, MCMC is indispensable for performing Bayesian analysis. Two critical questions that MCMC practitioners need to address are where to start and when to stop the simulation. Although a great amount of research has gone into establishing convergence criteria and stopping rules with sound theoretical foundation, in practice, MCMC users often decide convergence by applying empirical diagnostic tools. This review article discusses the most widely used MCMC convergence diagnostic tools. Some recently proposed stopping rules with firm theoretical footing are also presented. The convergence diagnostics and stopping rules are illustrated using three detailed examples.

1. INTRODUCTION

Markov chain Monte Carlo (MCMC) methods are now routinely used to fit complex models in diverse disciplines. A Google search for "Markov chain Monte Carlo" returns more than 11.5 million hits. The popularity of MCMC is mainly due to its widespread usage in computational physics and Bayesian statistics, although it is also used in frequentist inference (see, e.g., Geyer & Thompson 1995, Christensen 2004).

The fundamental idea of MCMC is that if simulating from a target density π is difficult so that the ordinary Monte Carlo method based on independent and identically distributed (i.i.d.) samples cannot be used for making inference on π, it may be possible to construct a Markov chain $\{X_n\}_{n\geq 0}$ with stationary density π for forming Monte Carlo estimators. An introduction to construction of such Markov chains, including the Gibbs sampler and the Metropolis–Hastings (MH) sampler, is provided by Geyer (2011) (see also Robert & Casella 2004). General purpose MH algorithms are available in the R packages `mcmc` (Geyer & Johnson 2017) and `MCMCpack` (Martin et al. 2011). There are several R (R Core Team 2018) packages implementing specific MCMC algorithms for a number of statistical models [see, e.g., `MCMCpack` (Martin et al. 2011), `MCMCglmm` (Hadfield 2010), and `geoBayes` (Evangelou & Roy 2019)]. Here, we do not discuss development of MCMC algorithms, but rather focus on analyzing the Markov chain obtained from running such an algorithm for determining its convergence.

Two important issues that must be addressed while implementing MCMC are where to start and when to stop the algorithm. As we discuss now, these two tasks are related to determining convergence of the underlying Markov chain to stationarity and convergence of Monte Carlo estimators to population quantities, respectively. It is known that under some standard conditions on the Markov chain, for any initial value, the distribution of X_n converges to the stationary distribution as $n \to \infty$ (see, e.g., Meyn & Tweedie 1993, chapter 13; Roberts & Rosenthal 2004). Since $X_0 \not\sim \pi$ and MCMC algorithms produce (serially) correlated samples, the further the initial distribution from π, the longer it takes for X_n to approximate π. In particular, if the initial value is not in a high-density (π) region, the samples at the earlier iterations may not be close to the target distribution. In such cases, a common practice is to discard early realizations in the chain and start collecting samples only after the effect of the initial value has (practically) worn off. The main idea behind this method, known as burn-in, is to use samples only after the Markov chain gets sufficiently close to the stationary distribution, although its usefulness for Monte Carlo estimation has been questioned in the MCMC community (Geyer 2011). Thus, ideally, MCMC algorithms should be initialized at a high-density region, but if finding such areas is difficult, collection of Monte Carlo samples can be started only after a certain iteration n' when approximately $X_{n'} \sim \pi$.

Once the starting value is determined, one needs to decide when to stop the simulation. (Note that the starting value here refers to the beginning of collection of samples, as opposed to the initial value of X_0 of the Markov chain, although these two values can be the same.) Often the quantities of interest regarding the target density π can be expressed as means of certain functions, say, $E_\pi g \equiv \int_{\mathcal{X}} g(x)\pi(x)\mathrm{d}x$, where g is a real-valued function. For example, appropriate choices of g make $E_\pi g$ different measures of location, spread, and other summary features of π. Here, the support of the target density π is denoted by $\mathcal{X}$, which is generally $\mathbb{R}^d$ for some $d \geq 1$, although it can be non-Euclidean as well. We later consider vector valued functions g as well (Section 2). The MCMC estimator of the population mean $E_\pi g$ is the sample average $\bar{g}_{n',n} \equiv \sum_{i=n'+1}^{n} g(X_i)/(n-n')$. If no burn-in is used, then $n' = 0$. It is known that usually $\bar{g}_{n',n} \to E_\pi g$ as $n \to \infty$ (see Section 2 for details). In practice, however, MCMC users run the Markov chain for a finite n^* number of iterations, thus MCMC simulation should be stopped only when $\bar{g}_{n',n^*}$ has sufficiently converged to $E_\pi g$. The accuracy of the time average estimator $\bar{g}_{n',n}$ obviously depends on the quality of the

samples. Thus, when implementing MCMC methods, it is necessary to wisely conclude Markov chain convergence and subsequently determine when to stop the simulation. In particular, while premature termination of the simulation will most likely lead to inaccurate inference, unnecessarily running longer chains is not desirable either, as it eats up resources.

By performing theoretical analysis on the underlying Markov chain, an analytical upper bound on its distance to stationarity may be obtained (Rosenthal 1995), which in turn can provide a rigorous method for deciding MCMC convergence and thus finding n' (Jones & Hobert 2001). Similarly, using a sample size calculation based on an asymptotic distribution of the (appropriately scaled) Monte Carlo error $\bar{g}_{n',n^*} - E_\pi g$, an honest stopping value n^* can be found. In the absence of such theoretical analysis, often empirical diagnostic tools are used to check convergence of MCMC samplers and estimators, although, as shown through examples in Section 3, these tools cannot determine convergence with certainty. Since early 1990s, with the increasing use of MCMC, a great deal of research effort has gone into developing convergence diagnostic tools. These diagnostic methods can be classified into several categories. For example, corresponding to the two types of convergence mentioned before, some of these diagnostic tools are designed to assess convergence of the Markov chain to the stationary distribution, whereas others check for convergence of the summary statistics like sample means and sample quantiles to the corresponding population quantities. The available MCMC diagnostic methods can be categorized according to other criteria as well, for example, their level of theoretical foundation, if they are suitable for checking joint convergence of multiple variables, whether they are based on multiple (parallel) chains or a single chain or both, if they are complemented by a visualization tool or not, if they are based on moments and quantiles or the kernel density of the observed chain, and so on. Several review articles on MCMC convergence diagnostics are available in the literature (see, e.g., Cowles & Carlin 1996, Brooks & Roberts 1998, Mengersen et al. 1999). Cowles & Carlin (1996) provide a description of 13 convergence diagnostics and summarize these according to the different criteria mentioned above. While some of these methods are widely used in practice, several new approaches have been proposed since then. In this article, we review some of these tools that are commonly used by MCMC practitioners or that we find promising.

2. MCMC DIAGNOSTICS

As mentioned in the introduction, MCMC diagnostic tools are needed for deciding convergence of Markov chains to the stationarity. Although in general, the longer the chain is run, the better Monte Carlo estimates it produces, in practice it is desirable to use some stopping rules for prudent use of resources. In this section, we describe some MCMC diagnostics that may be used for deciding Markov chain convergence or stopping MCMC sampling. In the context of each method, we also report if it is designed particularly for one of these two objectives.

2.1. Honest MCMC

In this section, we describe some rigorous methods for finding n' and n^* that were mentioned in the introduction. Let f_n be the density of X_n. It is known that under some standard conditions (see, e.g., Meyn & Tweedie 1993, chapter 13), $\frac{1}{2}\int_{\mathcal{X}} |f_n(x) - \pi(x)|\mathrm{d}x \downarrow 0$ as $n \to \infty$, that is, X_n converges in the total variation (TV) norm to a random variable following π. Jones & Hobert (2001) mention that a rigorous way of deciding the convergence of the Markov chain to π is by finding an iteration number n' such that

$$\frac{1}{2}\int_{\mathcal{X}} |f_{n'}(x) - \pi(x)|\mathrm{d}x < 0.01. \qquad 1.$$

(The cutoff value 0.01 is arbitrary, and any predetermined precision level can be used.) Jones & Hobert (2001) propose to use the smallest n' for which Equation 1 holds as the honest value for burn-in.

The above-mentioned burn-in hinges on the TV norm in Equation 1, which is generally not available. Constructing a quantitative bound to the TV norm is also often difficult, although significant progress has been made in this direction (Rosenthal 1995, 2002; Baxendale 2005; Andrieu et al. 2015). In particular, a key tool for constructing a quantitative bound to the TV norm is using the drift and minorization (d&m) technique (Rosenthal 1995). The d&m technique has been successfully used to analyze a variety of MCMC algorithms (see, e.g., Fort et al. 2003, Jones & Hobert 2004, Roy & Hobert 2010, Vats 2017). The d&m conditions, as we explain later in this section, are also crucial to provide an honest way to check convergence of MCMC estimators of popular summary measures like moments and quantiles of the target distributions. Although we consider the TV norm here, over the last few years, other metrics like the Wasserstein distance have also been used to study Markov chain convergence (see, e.g., Durmus & Moulines 2015, Qin & Hobert 2019). Using Stein's method, Gorham & Mackey (2015) propose a computable discrepancy measure that seems promising as it depends on the target only through the derivative of $\log \pi$, and hence is appropriate in Bayesian settings where the target is generally known up to the intractable normalizing constant.

As in the Introduction, let a particular feature of the target density be expressed as $E_\pi g$, where g is a real valued function. By the strong law of large numbers for Markov chains, it is known that if $\{X_n\}_{n\geq 0}$ is appropriately irreducible, then $\bar{g}_{n',n} \equiv \sum_{i=n'+1}^{n} g(X_i)/(n-n')$ is a strongly consistent estimator of $E_\pi g$, that is, $\bar{g}_{n',n} \to E_\pi g$ almost surely as $n \to \infty$ for any fixed n' (Asmussen & Glynn 2011). Without loss of generality, we let $n' = 0$ when discussing stopping rules, and for the ease of notation, we simply write $\bar{g}_n$ for $\bar{g}_{0,n}$. The law of large numbers justifies estimating $E_\pi g$ by the sample (time) average estimator $\bar{g}_n$, as in the ordinary Monte Carlo. If a central limit theorem (CLT) is available for $\bar{g}_n$ (that is, for the error $\bar{g}_n - E_\pi g$) then a sample size calculation based on the width of an interval estimator for $E_\pi g$ can be performed for choosing an appropriate value for n^*. Indeed, under some regularity conditions,

$$\sqrt{n}(\bar{g}_n - E_\pi g) \xrightarrow{d} N(0, \sigma_g^2) \text{ as } n \to \infty, \tag{2.}$$

where $\sigma_g^2 \equiv \mathrm{Var}_\pi(g(X_0)) + 2\sum_{i=1}^{\infty} \mathrm{Cov}_\pi(g(X_0), g(X_i)) < \infty$; the subscript π indicates that the expectations are calculated assuming $X_0 \sim \pi$. (Note that, due to the autocorrelations present in a Markov chain, $\sigma_g^2 \neq \mathrm{Var}_\pi(g(X_0)) = \lambda_g^2$, say.) If $\widehat{\sigma}_{g,n}$ is a consistent estimator of σ_g, then an estimator of the standard error of $\bar{g}_n$, based on the sample size n, is $\widehat{\sigma}_{g,n}/\sqrt{n}$. Since the standard error $\widehat{\sigma}_{g,n}/\sqrt{n}$ allows one to judge the reliability of the MCMC estimate, it should always be reported along with the point estimate $\bar{g}_n$. The standard error also leads to a $100(1-\alpha)\%$ confidence interval for $E_\pi g$, namely $\bar{g}_n \mp z_{\alpha/2}\widehat{\sigma}_{g,n}/\sqrt{n}$. Here, $z_{\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution. The MCMC simulation can be stopped if the half-width of the $100(1-\alpha)\%$ confidence interval falls below a prespecified threshold, say ϵ. Jones & Hobert (2001) refer to this method as the honest way to stop the chain. Indeed, the fixed-width stopping rule (FWSR) (Flegal et al. 2008, Jones et al. 2006) terminates the simulation the first time after some user-specified $\tilde{n}$ iterations that

$$t_* \frac{\widehat{\sigma}_{g,n}}{\sqrt{n}} + \frac{1}{n} \leq \epsilon. \tag{3.}$$

Here, t_* is an appropriate quantile. The role of $\tilde{n}$ is to make sure that the simulation is not stopped prematurely due to a poor estimate of $\widehat{\sigma}_{g,n}$. The value of $\tilde{n}$ should depend on the complexity of the problem. Gong & Flegal (2016) suggest that using $\tilde{n} = 10^4$ works well in practice.

For validity of the honest stopping rule, a CLT (Equation 2) for $\bar{g}_n$ needs to exist, and one would need a consistent estimator $\widehat{\sigma}_{g,n}$ of σ_g. For the CLT to hold, the TV norm in Equation 1 needs to converge to zero at a certain rate (see Jones 2004 for different conditions guaranteeing a Markov chain CLT). The most common method of establishing a CLT (Equation 2), as well as providing a consistent estimator of σ_g, has been by showing the Markov chain $\{X_n\}_{n\geq 0}$ is geometrically ergodic, that is, the TV norm (Equation 1) converges at an exponential rate (Jones & Hobert 2001, Roberts & Rosenthal 2004). Generally, geometric ergodicity of a Markov chain is proven by constructing an appropriate d&m condition (Rosenthal 1995, Roy & Hobert 2010). For estimation of σ_g^2, while Mykland et al. (1995) and Hobert et al. (2002) discuss the regenerative simulation method, Jones et al. (2006) and Flegal & Jones (2010) provide consistent batch means and spectral variance methods. Availability of a Markov chain CLT has been demonstrated for myriad MCMC algorithms for common statistical models. Here, we provide an incomplete list: linear models (Román & Hobert 2012, 2015), generalized linear models including the probit model (Roy & Hobert 2007, Chakraborty & Khare 2017), the popular logistic model (Choi & Hobert 2013, Wang & Roy 2018c) and the robit model (Roy 2012), generalized linear mixed models including the probit mixed model (Wang & Roy 2018b) and the logistic mixed model (Wang & Roy 2018a), quantile regression models (Khare & Hobert 2012), multivariate regression models (Roy & Hobert 2010, Hobert et al. 2018), and penalized regression and variable selection models (Khare & Hobert 2013, Roy & Chakraborty 2017, Vats 2017).

So far we have described the honest MCMC in the context of estimating means of univariate functions. The method is applicable to estimation of vector valued functions as well. In particular, if g is a $\mathbb{R}^p$ valued function, and if a CLT holds for $\bar{g}_n$, that is, if $\sqrt{n}(\bar{g}_n - E_\pi g) \xrightarrow{d} N(0, \Sigma_g)$ as $n \to \infty$, for some $p \times p$ covariance matrix Σ_g, then using a consistent estimator $\widehat{\Sigma}_{g,n}$ of Σ_g, a $100(1-\alpha)\%$ asymptotic confidence region $C_\alpha(n)$ for $E_\pi g$ can be formed (for details, see Vats et al. 2019). Vats et al. (2019) propose a fixed volume stopping rule, which terminates the simulation the first time after $\tilde{n}$ iterations that

$$(\text{Vol}\{C_\alpha(n)\})^{1/p} + \frac{1}{n} \leq \varepsilon,$$

where, as in Equation 3, ε is the user's desired level of accuracy. Note that when $p = 1$, except the $1/n$ terms, the expression above is the same as Equation 3 with $\varepsilon = 2\epsilon$. Honest MCMC can also be implemented for estimation of the quantiles (Doss et al. 2014). In order to reduce computational burden, the sequential stopping rules should be checked only at every l iterations, where l is appropriately chosen. Finally, even if theoretical d&m analysis is not carried out establishing a Markov chain CLT, in practice, FWSR can be implemented using the batch means and spectral variance estimators of $\sigma_g(\Sigma_g)$ available in the R package `mcmcse` (Flegal et al. 2012).

2.2. Relative Fixed-Width Stopping Rules

FWSR (described in Section 2.1) explicitly addresses how well the estimator $\bar{g}_n$ approximates $E_\pi g$. Flegal & Gong (2015) and Gong & Flegal (2016) discuss relative FWSR in the MCMC setting. Flegal & Gong (2015) consider a relative magnitude rule that terminates the simulation when, after $\tilde{n}$ iterations, $t_* \widehat{\sigma}_{g,n} n^{-1/2} + n^{-1} \leq \epsilon \bar{g}_n$. Flegal & Gong (2015) also consider a relative standard deviation FWSR (SDFWSR) that terminates the simulation when, after $\tilde{n}$ iterations, $t_* \widehat{\sigma}_{g,n} n^{-1/2} + n^{-1} \leq \epsilon \widehat{\lambda}_{g,n}$, where $\widehat{\lambda}_{g,n}$ is a strongly consistent estimator of the population standard deviation λ_g. Asymptotic validity of the relative magnitude and relative standard deviation stopping rules are

established by Glynn & Whitt (1992) and Flegal & Gong (2015), respectively. This ensures that the simulation will terminate in a finite time with probability 1.

In Bayesian statistics applications, Flegal & Gong (2015) advocate the use of relative SDFWSR. In the high-dimensional settings, that is, where g is an $\mathbb{R}^p$ valued function and p is large, without a priori knowledge of the magnitude of $E_\pi g$, Gong & Flegal (2016) prefer relative SDFWSR over FWSR based on the marginal chains. In the multivariate settings, Vats et al. (2019) argue that stopping rules based on p marginal chains may not be appropriate, as these ignore cross-correlations between components and may be dictated by the slowest mixing marginal chain. Vats et al. (2019) propose a multivariate relative standard deviation stopping rule involving the volume of the $100(1-\alpha)\%$ asymptotic confidence region, that is, $\text{Vol}\{C_\alpha(n)\}$. Let $\widehat{\Lambda}_{g,n}$ be the sample covariance matrix. Vats et al. (2019) propose to stop the simulation the first time after $\tilde{n}$ iterations that

$$(\text{Vol}\{C_\alpha(n)\})^{1/p} + \frac{1}{n} \leq \varepsilon(|\widehat{\Lambda}_{g,n}|)^{1/2p}, \tag{4.}$$

where $|\cdot|$ denotes the determinant.

2.3. Effective Sample Size

For an MCMC-based estimator, effective sample size (ESS) is the number of independent samples equivalent to (that is, having the same standard error as) a set of correlated Markov chain samples. Although ESS (based on n correlated samples) is not uniquely defined, the most common definition (Robert & Casella 2004) is

$$\text{ESS} = \frac{n}{1 + 2\sum_{i=1}^{\infty} \text{Corr}_\pi(g(X_0), g(X_i))}.$$

Gong & Flegal (2016) rewrite the above definition as $\text{ESS} = n\lambda_g^2/\sigma_g^2$. In the multivariate setting, that is, when g is $\mathbb{R}^p$ valued for some $p \geq 1$, Vats et al. (2019) define multivariate ESS (mESS) as

$$\text{mESS} = n\left(\frac{|\Lambda_g|}{|\Sigma_g|}\right)^{1/p}, \tag{5.}$$

where Λ_g is the population covariance matrix. An approach to terminate MCMC simulation is when $\widehat{\text{ESS}}$ ($\widehat{\text{mESS}}$) takes a value larger than a prespecified number, where $\widehat{\text{ESS}}$ ($\widehat{\text{mESS}}$) is a consistent estimator of ESS (mESS). Indeed, Vats et al. (2019) mention that simulation can be terminated the first time that

$$\widehat{\text{mESS}} = n\left(\frac{|\widehat{\Lambda_{g,n}}|}{|\widehat{\Sigma_{g,n}}|}\right)^{1/p} \geq \frac{2^{2/p}\pi}{(p\Gamma(p/2))^{2/p}} \frac{\chi^2_{1-\alpha,p}}{\varepsilon^2}, \tag{6.}$$

where ε is the desired level of precision for the volume of the $100(1-\alpha)\%$ asymptotic confidence region, and $\chi^2_{1-\alpha,p}$ is the $(1-\alpha)$ quantile of χ^2_p. This ESS stopping rule is (approximately) equivalent to the multivariate relative standard deviation stopping rule given in Equation 4 (for details, see Vats et al. 2019). Note that $\widehat{\text{ESS}}$ ($\widehat{\text{mESS}}$) per unit time can be used to compare different MCMC algorithms (with the same stationary distribution) in terms of both computational and statistical efficiency. ESS is implemented in several R packages, including `CODA` (Plummer et al. 2006) and `mcmcse` (Flegal et al. 2012). In the `mcmcse` package, estimates of ESS in both univariate and multivariate settings are available. While Gong & Flegal (2016) and Vats et al. (2019) provide

a connection between ESS and relative SDFWSR stopping rules, Vats & Knudson (2018) draw correspondence between ESS and a version of the widely used Gelman–Rubin (GR) diagnostic presented in the next section.

2.4. Gelman–Rubin Diagnostic

The GR diagnostic appears to be the most popular method for assessing samples obtained from running MCMC algorithms. The GR diagnostic relies on multiple chains $\{X_{i0}, X_{i1}, \ldots, X_{in-1}\}, i = 1, \ldots, m$ starting at initial points that are drawn from a density that is overdispersed with respect to the target density π. Gelman & Rubin (1992) describe methods of creating an initial distribution, although in practice, these initial points are usually chosen in some ad hoc way. Using parallel chains, Gelman & Rubin (1992) construct two estimators of the variance of X where $X \sim \pi$, namely, the within-chain variance estimate, $W = \sum_{i=1}^{m} \sum_{j=0}^{n-1} (X_{ij} - \bar{X}_{i\cdot})^2/(m(n-1))$, and the pooled variance estimate $\hat{V} = ((n-1)/n)W + B/n$, where $B/n = \sum_{i=1}^{m} (\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2/(m-1)$ is the between-chain variance estimate, and $\bar{X}_{i\cdot}$ and $\bar{X}_{\cdot\cdot}$ are the ith chain mean and the overall mean, respectively, $i = 1, 2, \ldots, m$. Finally, Gelman & Rubin (1992) compare the ratio of these two estimators to one. In particular, they calculate the potential scale reduction factor (PSRF) defined by

$$\hat{R} = \frac{\hat{V}}{W} \qquad 7.$$

and compare it to one.

Gelman & Rubin (1992) argue that since the chains are started from an overdispersed initial distribution, in finite samples, the numerator in Equation 7 overestimates the target variance, whereas the denominator underestimates it, making $\hat{R}$ larger than one. Simulation is stopped when $\hat{R}$ is sufficiently close to one. The cutoff value 1.1 is generally used by MCMC practitioners, as recommended by Gelman et al. (2014). Vats & Knudson (2018) propose a modified GR statistic where the between-chain variance (B/n) is replaced with a particular batch means estimator of the asymptotic variance for the Monte Carlo averages $\bar{X}_n$. This modified definition allows for a connection with ESS and, more importantly, computation of the GR diagnostic based on a single chain. We would like to point out that the expression of $\hat{R}$ given in Equation 7, although widely used in practice, differs slightly from the original definition given by Gelman & Rubin (1992).

Brooks & Gelman (1998) propose the multivariate PSRF (MPSRF) to diagnose convergence in the multivariate case. It is denoted by $\hat{R}_p$ and is given by

$$\hat{R}_p = \max_a \frac{a^T \widehat{V^*} a}{a^T W^* a} = \frac{n-1}{n} + \left(1 + \frac{1}{m}\right)\lambda_1, \qquad 8.$$

where $\widehat{V^*}$ is the pooled covariance matrix, W^* is the within-chain covariance matrix, B^* is the between-chain covariance matrix, and λ_1 is the largest eigenvalue of the matrix $(W^{*-1}B^*)/n$. As in the univariate case, simulation is stopped when $\hat{R}_p \approx 1$. Peltonen et al. (2009) have proposed a visualization tool based on linear discriminant analysis and discriminant component analysis, which can be used to complement the diagnostic tools proposed by Gelman & Rubin (1992) and Brooks & Gelman (1998). The GR diagnostic can be easily calculated and is available in different statistical packages including the `CODA` package (Plummer et al. 2006) in R. To conclude our discussion on the GR diagnostic, note that originally Gelman & Rubin (1992) suggested running m

parallel chains, each of length $2n$. Then, discarding the first n simulations, $\hat{R}$ is computed based on the last n iterations. This leads to the waste of too many samples, and we do not recommend it.

2.5. Two Spectral Density–Based Methods

In this section, we discuss two diagnostic tools based on asymptotic variance estimates of certain statistics to check for convergence to stationarity. Geweke (1992) proposes a diagnostic tool based on the assumption of existence of the spectral density of a related time series. Indeed, for the estimation of $E_\pi g$, the asymptotic variance of $\bar{g}_n$ is $S_g(0)$, the spectral density of $\{g(X_n), n \geq 0\}$ (treated as a time series) evaluated at zero. After n iterations of the Markov chain, let $\bar{g}_{n_A}$ and $\bar{g}_{n_B}$ be the time averages based on the first n_A and the last n_B observations. Geweke's (1992) statistic is the difference $\bar{g}_{nA} - \bar{g}_{nB}$, normalized by its standard error calculated using a nonparametric estimate of $S_g(0)$ for the two parts of the Markov chain. Thus, Geweke's statistic is

$$Z_n = \left(\bar{g}_{n_A} - \bar{g}_{n_B}\right) \Big/ \sqrt{\widehat{S_g(0)}/n_A + \widehat{S_g(0)}/n_B}.$$

Geweke (1992) suggests using $n_A = 0.1n$ and $n_B = 0.5n$. The Z score is calculated under the assumption of independence of the two parts of the chain. Thus Geweke's (1992) convergence diagnostic is a Z test of equality of means where autocorrelation in the samples is taken into account while calculating the standard error.

Heidelberger & Welch (1983) propose another method based on spectral density estimates. Heidelberger & Welch's (1983) diagnostic is based on

$$B_n(t) = \left(\sum_{i=0}^{\lfloor nt \rfloor} g(X_i) - \lfloor nt \rfloor \bar{g}_n\right) \Big/ \sqrt{n\widehat{S_g(0)}}.$$

Assuming that $\{B_n(t), 0 \leq t \leq 1\}$ is distributed asymptotically as a Brownian bridge, the Cramer–von Mises statistic $\int_0^1 B_n(t)^2 \mathrm{d}t$ may be used to test the stationarity of the Markov chain. The stationarity test is successively applied, first on the whole chain, and then rejecting the first 10%, 20%, and so forth of the samples until the test is passed or 50% of the samples have been rejected. Both of these two spectral density–based tools presented here are implemented in the `CODA` package (Plummer et al. 2006). These are univariate diagnostics, although Cowles & Carlin (1996) mention that for Geweke's (1992) statistic, g may be taken to be -2 times the log of the target density when $\mathcal{X} = \mathbb{R}^d$ for some $d > 1$. Finally, we would like to mention that the two spectral density–based methods mentioned here, just like the ESS and the GR diagnostic, assume the existence of a Markov chain CLT (Equation 2), emphasizing the importance of the theoretical analysis discussed in Section 2.1.

2.6. Raftery–Lewis Diagnostic

Suppose the goal is to estimate a quantile of $g(X)$, that is, to estimate u such that $P_\pi(g(X) \leq u) = q$ for some prespecified q. Raftery & Lewis (1992) propose a method for calculating an appropriate burn-in. They also discuss choosing a run length so that the resulting probability estimate lies in $[q - \epsilon, q + \epsilon]$ with probability $(1 - \alpha)$. Thus, the required accuracy ϵ is achieved with probability $(1 - \alpha)$. Raftery & Lewis (1992) consider the binary process $W_n \equiv I(g(X_n) \leq u), n \geq 0$. Although, in general, $\{W_n\}_{n\geq 0}$ itself is not a Markov chain, Raftery & Lewis (1992) assume that for sufficiently large k, the subsequence $\{W_{nk}\}_{n\geq 0}$ is approximately a Markov chain. They discuss a method for

choosing k using model selection techniques. The transition probability $P(W_{nk} = j|W_{(n-1)k} = i)$ is estimated by the usual estimator

$$\frac{\sum_{l=1}^{n} I(W_{lk} = j, W_{(l-1)k} = i)}{\sum_{l=1}^{n} I(W_{lk} = i)},$$

for $i, j = 0, 1$. Here, $I(\cdot)$ is the indicator function. Using a straightforward eigenvalue analysis of the two-state empirical transition matrix of $\{W_{nk}\}_{n\geq 0}$, Raftery & Lewis (1992) provide an estimate of the burn-in. Using a CLT for $\sum_{j=0}^{n-1} W_{jk}/n$, they also give a stopping rule to achieve the desired level of accuracy.

To implement this univariate method, an initial number $n_{\min}$ of iterations is used, and then it is determined if any additional runs are required using the above techniques. The value $n_{\min} = \{\Phi^{-1}(1 - \alpha/2)\}^2 q(1 - q)/\epsilon^2$ is based on the standard asymptotic sample size calculation for Bernoulli (q) population. Since the diagnostic depends on the q values, the method should be repeated for different quantiles, and the largest among these burn-in estimates can be used. The diagnostic of Raftery & Lewis (1992) is available in the `CODA` package (Plummer et al. 2006).

2.7. Kernel Density–Based Methods

There are MCMC diagnostics that compute the distance between the kernel density estimates of two chains or two parts of a single chain and conclude convergence when the distance is close to zero. Unlike the widely used GR diagnostic (Gelman & Rubin 1992), which is based on comparison of some summary moments of MCMC chains, these tools are intended to assess the convergence of the whole distributions. Yu (1994) and Boone et al. (2014) estimate the L^1 distance and Hellinger distance between the kernel density estimates respectively. More recently, Dixit & Roy (2017) use the symmetric Kullback–Leibler (KL) divergence to produce two diagnostic tools based on kernel density estimates of the chains. Below, we briefly describe the method of Dixit & Roy (2017).

Let $\{X_{ij} : i = 1, 2; j = 1, 2, \ldots, n\}$ be the n observations obtained from each of the two Markov chains initialized from two points well-separated with respect to the target density π. The adaptive kernel density estimates of observations obtained from the two chains are denoted by p_{1n} and p_{2n}, respectively. The KL divergence between p_{in} and p_{jn} is denoted by $\mathrm{KL}(p_{in}|p_{jn}), i \neq j,\ i, j = 1, 2$, that is,

$$\mathrm{KL}(p_{in}|p_{jn}) = \int_{\mathcal{X}} p_{in}(x) \log \frac{p_{in}(x)}{p_{jn}(x)} \mathrm{d}x.$$

Dixit & Roy (2017) find the Monte Carlo estimates of $\mathrm{KL}(p_{in}|p_{jn})$ using samples simulated from p_{in} using the technique proposed by Silverman (1986, section 6.4.1). They use the estimated symmetric KL divergence $([\mathrm{KL}(p_{1n}|p_{2n}) + \mathrm{KL}(p_{2n}|p_{1n})]/2)$ between p_{1n} and p_{2n} to assess convergence where a testing of hypothesis framework is used to determine the cutoff points. The hypotheses are chosen such that the type 1 error is concluding that the Markov chains have converged when in fact they have not. The cutoff points for the symmetric KL divergence are selected to ensure that the probability of type 1 error is below some level, say, 0.05. In case of multiple $(m > 2)$ chains, the maximum among $\binom{m}{2}$ estimated symmetric KL divergences (referred to as Tool 1) is used to diagnose MCMC convergence. Finally, for multivariate examples—that is, when $\mathcal{X} = \mathbb{R}^d$ for some $d > 1$—although multivariate Tool 1 can be used, in higher dimensions when kernel density estimation is not reliable, Dixit & Roy (2017) recommend assessing convergence

marginally, i.e., one variable at a time, where appropriate cutoff points are found by adjusting the level of significance using Bonferroni's correction for multiple comparison.

For multimodal target distributions, if all chains are stuck at the same mode, then empirical convergence diagnostics based solely on MCMC samples may falsely treat the target density as unimodal and are prone to failure. In such situations, Dixit & Roy (2017) propose another tool (Tool 2) that makes use of the KL divergence between the kernel density estimate of MCMC samples and the target density (generally known up to the unknown normalizing constant) to detect divergence. In particular, let $\pi(x) = f(x)/c$, where $c = \int_{\mathcal{X}} f(x)\mathrm{d}x$ is the unknown normalizing constant. Dixit & Roy's Tool 2 is given by

$$T_2^* = \frac{|\,\hat{c} - c^*\,|}{c^*}, \tag{9.}$$

where $\hat{c}$ is a Monte Carlo estimate, as described in section 3.3 of Dixit & Roy (2017), of the unknown normalizing constant (c), based on the KL divergence between the adaptive kernel density estimate of the chain and π, and c^* is an estimate of c obtained by numerical integration. Dixit & Roy (2017) discuss that T_2^* can be interpreted as the percentage of the target distribution not yet captured by the Markov chain. Using this interpretation, they advocate that if $T_2^* > 0.05$, then the Markov chain has not yet captured the target distribution adequately. Since Equation 9 involves numerical integration, it cannot be used in high-dimensional examples.

2.7.1. A visualization tool. Dixit & Roy (2017) propose a simple visualization tool to complement their KL divergence diagnostic tool. This tool can be used for any diagnostic method (including the GR diagnostic) based on multiple chains started at distinct initial values, to investigate reasons behind their divergence. Suppose $m(\geq 3)$ chains are run, and a diagnostic tool has revealed that the m chains have not mixed adequately and thus the chains have not yet converged. This indication of divergence could be due to a variety of reasons. A common reason for divergence is formation of clusters among multiple chains. Dixit & Roy's (2017) visualization tool utilizes the tile plot to display these clusters. As mentioned in Section 2.7, for m chains, the KL divergence tool finds the estimated symmetric KL divergence between each of the $\binom{m}{2}$ combinations of chains and reports the maximum among them. In the visualization tool, if the estimated symmetric KL divergence for a particular combination is less than or equal to the cutoff value, then the tool utilizes a gray tile to represent that the two chains belong to the same cluster, or else it uses a black tile to represent that the two chains belong to different clusters.

This visualization tool can also be used for multivariate chains. In cases where the diagnostic tool for d variate chains indicates divergence, for further investigation, the user can choose a chain from each cluster and implement the visualization tool marginally, i.e., one variable at a time. This will help the user identify which among the d variables are responsible for inadequate mixing among the m multivariate chains.

2.8. Graphical Methods

In addition to the visualization tool mentioned in Section 2.7.1, we now discuss some of the widely used graphical methods for MCMC convergence diagnosis. The most common graphical convergence diagnostic method is the trace plot. The trace plot is a time series plot that shows the realizations of the Markov chain at each iteration against the iteration numbers. This graphical method is used to visualize how the Markov chain is moving around the state space, that is, how well it is mixing. If the MCMC chain is stuck in some part of the state space, the trace plot shows flat bits indicating slow convergence. Such a trace plot is observed for an MH chain if

too many proposals are rejected consecutively. In contrast, if too many proposals are accepted consecutively, then trace plots may move slowly, not exploring the rest of the state space. Visible trends or changes in spread of the trace plot imply that the stationarity has not been reached yet. It is often said that a good trace plot should look like a hairy caterpillar. For an efficient MCMC algorithm, if the initial value is not in the high-density region, the beginning of the trace plot shows back-to-back steps in one direction. In contrast, if the trace plot shows similar pattern throughout, then there is no use in throwing burn-in samples.

Unlike i.i.d. sampling, MCMC algorithms result in correlated samples. The lag-k (sample) autocorrelation is defined to be the correlation between the samples k steps apart. The autocorrelation plot shows values of the lag-k autocorrelation function (ACF) against increasing k values. For fast-mixing Markov chains, lag-k autocorrelation values drop down to (practically) zero quickly as k increases. In contrast, high lag-k autocorrelation values for larger k indicate the presence of a high degree of correlation and slow mixing of the Markov chain. Generally, in order to get precise Monte Carlo estimates, Markov chains need to be run a large multiple of the amount of time it takes the ACF to be practically zero.

Another graphical method used in practice is the running mean plot, although its use has faced criticism (Geyer 2011). The running mean plot shows the Monte Carlo (time average) estimates against the iterations. This line plot should stabilize to a fixed number as iteration increases, but nonconvergence of the plot indicates that the simulation cannot be stopped yet. While the trace plot is used to diagnose a Markov chain's convergence to stationarity, the running mean plot is used to decide stopping times.

In the multivariate case, individual trace, autocorrelation, and running mean plots are generally made based on realizations of each marginal chain. Thus, the correlations that may be present among different components are not visualized through these plots. In multivariate settings, investigating correlation across different variables is required to check for the presence of high cross-correlation (Cowles & Carlin 1996).

3. EXAMPLES

In this section, we use three detailed examples to illustrate the convergence diagnostics presented in Section 2. Using these examples, we also demonstrate that empirical convergence diagnostic tools may give false indication of convergence to stationarity as well as convergence of Monte Carlo estimates.

3.1. An Exponential Target Distribution

Let the target distribution be Exp(1), that is, $\pi(x) = \exp(-x)$, $x > 0$. We consider an independence Metropolis sampler with Exp(θ) proposal, that is, the proposal density is $q(x) = \theta \exp(-\theta x)$, $x > 0$. We study the independence chain corresponding to two values of θ, namely, $\theta = 0.5$ and $\theta = 5$. Using this example, we illustrate the honest choices of burn-in and stopping time described in Section 2.1, as well as several other diagnostic tools. It turns out that, even in this unimodal example, some empirical diagnostics may lead to premature termination of the simulation. We first consider some graphical diagnostics for Markov chain convergence. Since the target density is a strictly decreasing function on $(0, \infty)$, a small value may serve as a reasonable starting value. We run the Markov chains for 1,000 iterations initialized at $X_0 = 0.1$. **Figure 1** shows the trace plots and autocorrelation plots of the Markov chain samples. From the trace plots (*left panels*) we see that while the first chain ($\theta = 0.5$) mixes well, the second chain exhibits several flat bits and suffers from slow mixing. Thus, from the trace plots, we see that there is no need for burn-in for

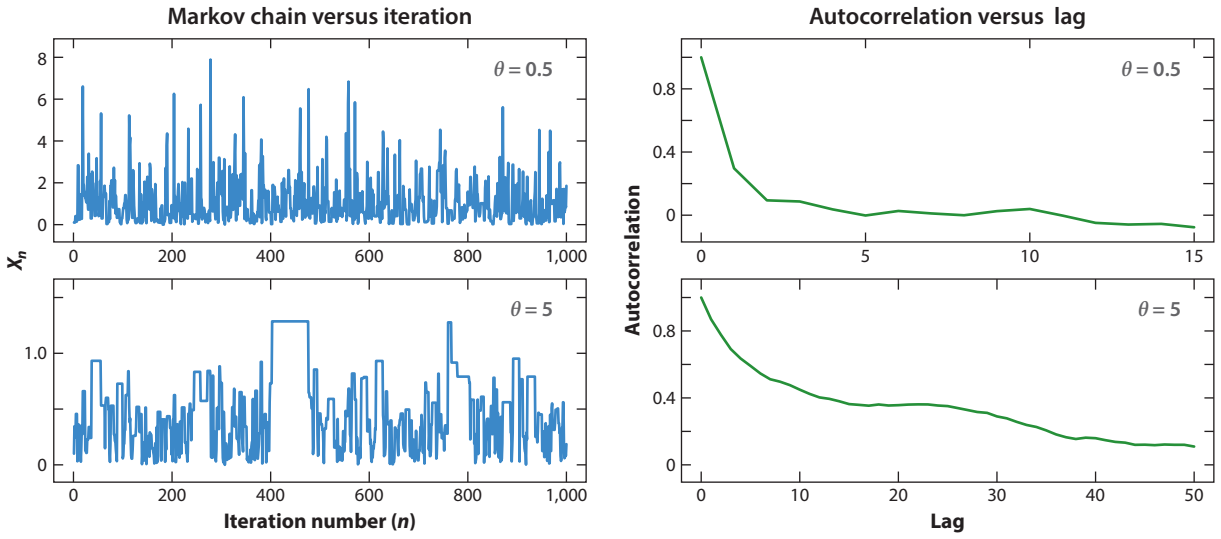

Figure 1

Trace (*left panels*) and autocorrelation function (*right panels*) plots of the independence Metropolis chains (*top row*, $\theta = 0.5$; *bottom row*, $\theta = 5$) for the exponential target example. The presence of frequent flat bits in the trace plot and high autocorrelation values indicate slow mixing of the Markov chain with $\theta = 5$. $1/\theta$ represents the mean of the proposal exponential distribution.

$\theta = 0.5$, that is, $X_0 = 0.1$ seems to be a reasonable starting value. In contrast, for $\theta = 5$, the chain can be run longer to find an appropriate burn-in. This is also corroborated by the autocorrelation plots (*right panels*). When $\theta = 0.5$, autocorrelation is almost negligible after lag 4. For $\theta = 5$, there is significant autocorrelation even after lag 50. Next, using the `CODA` package (Plummer et al. 2006), we compute Geweke's (1992) and Heidelberger & Welch's (1983) convergence diagnostics for the identity function $g(x) = x$. Using the default $n_A = 0.1n$ and $n_B = 0.5n$, Geweke's Z scores for the $\theta = 0.5$ and $\theta = 5$ chains are 0.733 and 0.605, respectively, failing to reject the hypothesis of the equality of means from the beginning and end parts of the chains. Similarly, both the chains pass the Heidelberger & Welch (1983) test for stationarity. Next, we consider the Raftery & Lewis (1992) diagnostic. When the two samplers are run for 38,415 ($\lceil n_{\min} \rceil$ corresponding to $\epsilon = 0.005, \alpha = 0.05$, and $q = 0.5$) iterations, and the Raftery–Lewis diagnostic is applied for different q values $(0.1, \ldots, 0.9)$, the burn-in estimates for the $\theta = 5$ chain are larger than those for the $\theta = 0.5$ chain, although the overall maximum burn-in (981) is less than 1,000. Finally, we consider the choice of honest burn-in. Since for $\theta < 1$, $\pi(x)/q(x) = \theta^{-1} \exp(x(\theta - 1)) \leq \theta^{-1}$ for all $x > 0$, according to Mengersen & Tweedie (1996), we know that

$$\frac{1}{2} \int_{\mathcal{X}} |f_n(x) - \pi(x)| \mathrm{d}x \leq (1 - \theta)^n,$$

that is, an analytical upper bound to the TV norm can be obtained. Thus for $\theta = 0.5$, if $n' = \lceil \log(0.01)/\log(0.5) \rceil = 7$, then Equation 1 holds. Thus, $n' = 7$ can be an honest burn-in for the independence Metropolis chain with $\theta = 0.5$. Note that, for $\theta < 1$, the independence chain is geometrically ergodic; for $\theta = 1$, the chain produces i.i.d. draws from the target; and for $\theta > 1$, by Mengersen & Tweedie (1996), the independence chain is subgeometric. As mentioned by Jones & Hobert (2001), when $\theta > 1$, the tail of the proposal density is much lighter than that of the target,

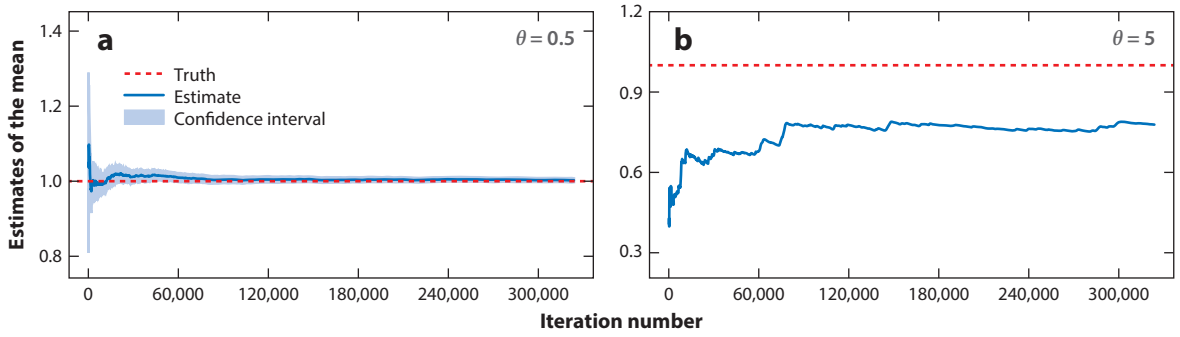

Figure 2

(*a*) Running estimates of the mean with confidence interval for $\theta = 0.5$. (*b*) Running mean plot for $\theta = 5$. Panel *b* reveals that even after 300,000 iterations, the Monte Carlo estimate for the chain with $\theta = 5$ is far off from the truth. $1/\theta$ is the mean of the proposal exponential distribution.

making it difficult for the chain to move to larger values, and when it does move there, it tends to get stuck.

Next, we consider stopping rules for estimation of the mean of the stationary distribution, that is, $E_\pi X = 1$. Based on a single chain, we apply the FWSR (Equation 3) to determine the sample size for $\epsilon = 0.005$ and $\alpha = 0.05$ (that is, $t_* = 1.96$). For the independence Metropolis chain with $\theta = 0.5$ starting at $X_8 = 0.1545$, it takes $n^* = 323,693$ iterations to achieve the cutoff 0.005. The running estimates of the mean, along with confidence intervals, are given in **Figure 2*a***. We next run the independence Metropolis chain with $\theta = 5$ for 323,700 iterations starting at $X_0 = 0.1$. The corresponding running estimates are given in **Figure 2*b***. Since a Markov chain CLT is not available for the independence chain with $\theta > 1$, we cannot compute asymptotic confidence intervals in this case. From the plot, we see that the final estimate (0.778) is far off from the truth ($E_\pi X = 1$). Next, we consider ESS. The cutoff value for ESS mentioned in Equation 6, with $\varepsilon = 2 * 0.005 = 0.01$, is 153,658. The ESSs for the two chains are 163,955 and 1,166, respectively, which again shows the presence of large correlation among the MCMC samples for $\theta = 5$. We use the R package `mcmcse` (Flegal et al. 2012) for computing ESS. Finally, we consider the GR diagnostic. We run four parallel chains for 2,000 iterations starting at 0.1, 1, 2, and 3, respectively, each with both $\theta = 0.5$ and $\theta = 5$. We calculate the Gelman & Rubin (1992) PSRF (Equation 7) based on these chains. The plots of iterative $\hat{R}$ at the increment of every 100 iterations are given in **Figure 3**. We see that $\hat{R}$ for the chain with $\theta = 0.5$ reaches below 1.1 in 100 iterations (**Figure 3*a***). In contrast, the Monte Carlo estimate for $E_\pi X$ and its standard error based on the first 100 iterations for the chain started at 0.1 are 1.109 and 0.111, respectively. Thus, the GR diagnostic leads to premature termination of simulation, and the inference drawn from the resulting samples can be unreliable. Finally, we note that $\hat{R}$ for the chains with $\theta = 5$ takes large (>16) values even after 2,000 iterations (**Figure 3*b***), showing that simulation cannot be stopped yet in this case.

3.2. A Sixmodal Target Distribution

This example is proposed by Leman et al. (2009), where the target density is as follows:

$$\pi(x,y) \propto \exp\left(\frac{-x^2}{2}\right) \exp\left(\frac{((\csc y)^5 - x)^2}{2}\right), \quad -10 \le x, y \le 10. \qquad 10.$$

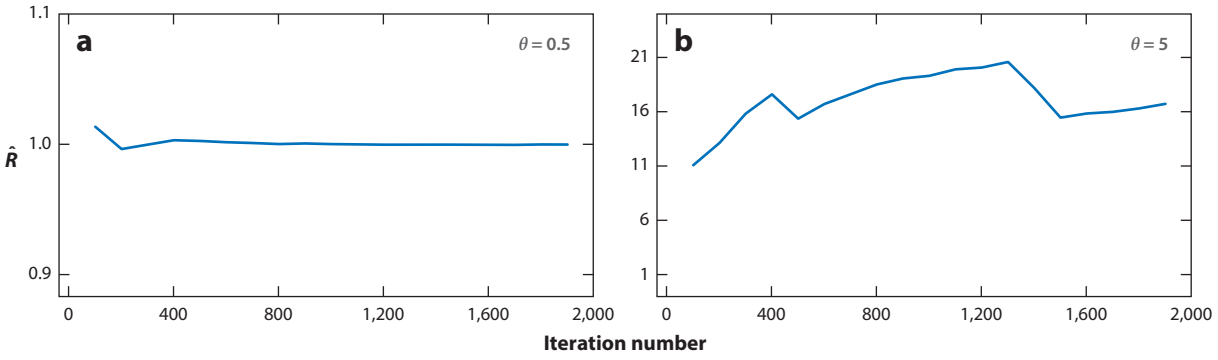

Figure 3

Iterative $\hat{R}$ plot (from four parallel chains) for the independence chains. (*a*) $\theta = 0.5$, (*b*) $\theta = 5$. In panel *a*, the PSRF reaches below the cutoff (1.1) before 100 iterations, leading to premature termination of the chain. $1/\theta$ is the mean of the proposal exponential distribution, and $\hat{R}$ is the potential scale reduction factor (PSRF).

The contour plot of the target distribution (known up to the normalizing constant) is given in **Figure 4**, and marginal densities are plotted in **Figure 5**. The plots of the joint and marginal distributions clearly show that the target distribution is multimodal in nature.

To draw MCMC samples from the target density (Equation 10), we use a Metropolis within Gibbs sampler in which X is drawn first, and then Y. In this example, we consider only convergence to stationarity, that is, we do not discuss stopping rules here. Through this example, we illustrate that when an MCMC sampler is stuck in a local mode, the empirical convergence diagnostic tools may give false indication of convergence. [Empirical diagnostics may fail even when modes are not well defined (Geyer & Thompson 1995).] In order to illustrate the diagnostic tools, as in Dixit & Roy (2017), we consider two cases.

In case 1, we run four chains wherein two chains (chains 1 and 2) are started at a particular mode, while the remaining two chains (chains 3 and 4) are started at some other mode. Each of the four chains is run for 30,000 iterations. Trace plots of the last 1,000 iterations of the four parallel X and Y marginal chains are given in panel *a* of **Figures 6** and **7**, respectively. Trace plots show the divergence of the Markov chains. High ACF values can also be seen from the autocorrelation plots of the marginal chains in **Figures 6** and **7**.

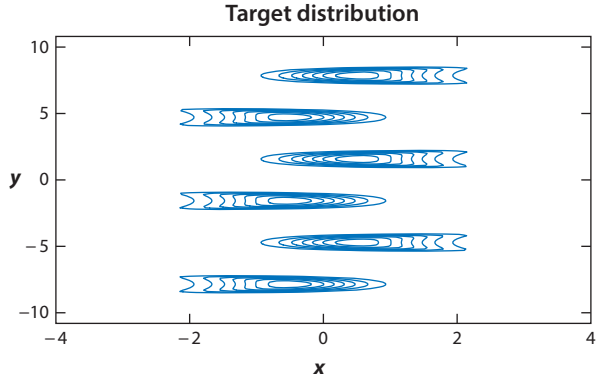

Figure 4

Contour plot of the target distribution in the sixmodal example.

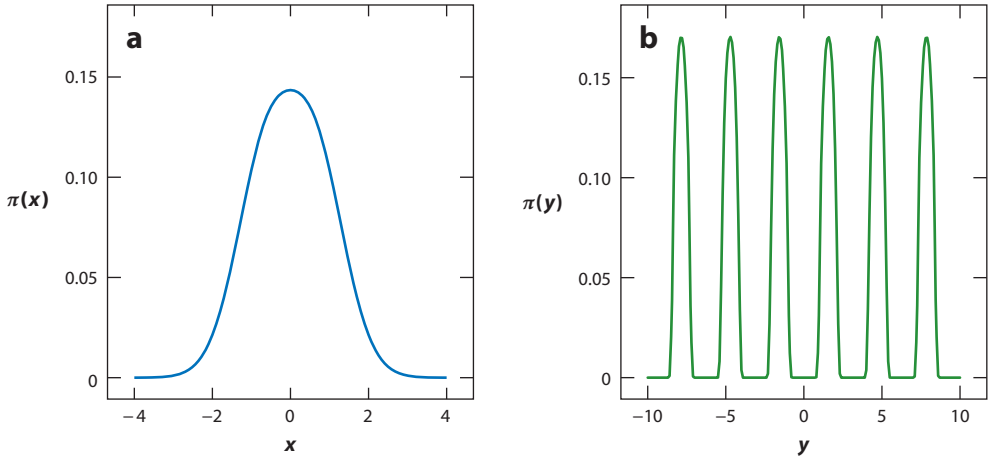

Figure 5

Marginal densities of X and Y in the sixmodal example.

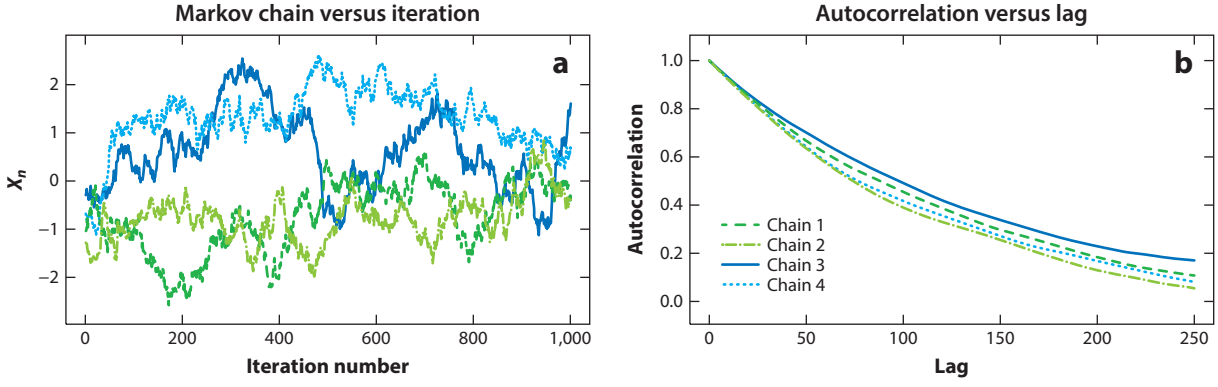

Figure 6

(*a*) Trace and (*b*) autocorrelation function plots of the X marginal of the four chains for the sixmodal example in case 1.

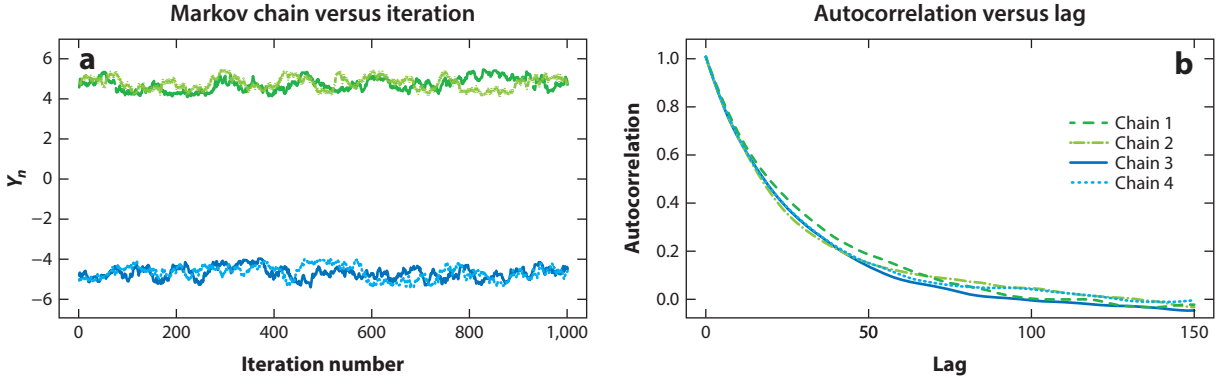

Figure 7

(*a*) Trace and (*b*) autocorrelation function plots of the Y marginal of the four chains for the sixmodal example in case 1. Unlike for the X marginal chains, trace plots of some of the Y marginal chains do not have any overlap demonstrating divergence of the Markov chains.

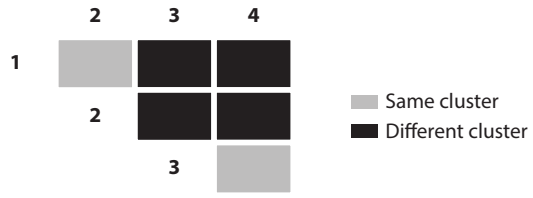

Figure 8

Dixit & Roy's (2017) tile plot in case 1 of the sixmodal example. The plot shows formation of two distinct clusters by the four chains.

Next, we apply Dixit & Roy's (2017) bivariate KL divergence Tool 1 on the joint chain. The maximum symmetric KL divergence among the six pairs is 104.89, significantly larger than the cutoff value, 0.06. Finally, we use Dixit & Roy's (2017) visualization tool to identify clusters among the four chains. The result is given in **Figure 8**, which shows that there are two clusters among the four chains, wherein chain 1 and chain 2 form one cluster, while chain 3 and chain 4 form another cluster.

In case 2, we also run four chains, but all the chains are started at the same local mode. As in case 1, all four chains are run for 30,000 iterations. The trace and autocorrelation plots of the marginal chains are given in **Figures 9** and **10**. From these plots, one may conclude mixing of the Markov chains, although the large autocorrelations result in low ESS for the chains. The minimum and maximum mESS (Equation 5) computed using the R package `mcmcse` for the four chains are 412 and 469, respectively.

The adaptive kernel density estimates of the four chains are visualized in **Figure 11**. This bivariate density plot does not reveal nonconvergence of the chains to the stationary distribution. Next, we compute the Geweke (1992) and Heidelberger & Welch (1983) convergence diagnostics for the identity function $g(x) = x$ for all four individual chains. At level 0.05, the Geweke (1992) diagnostic fails to reject the hypothesis of the equality of means from the beginning and end parts of each chain. Similarly, all chains pass the Heidelberger & Welch (1983) test for stationarity. Thus, both Geweke's and Heidelberger & Welch's diagnostics fail to detect the nonconvergence of the chains to the target distribution. Also, the Raftery–Lewis diagnostic fails to distinguish between the chains in case 1 and case 2, as it results in similar burn-in estimates in both cases.

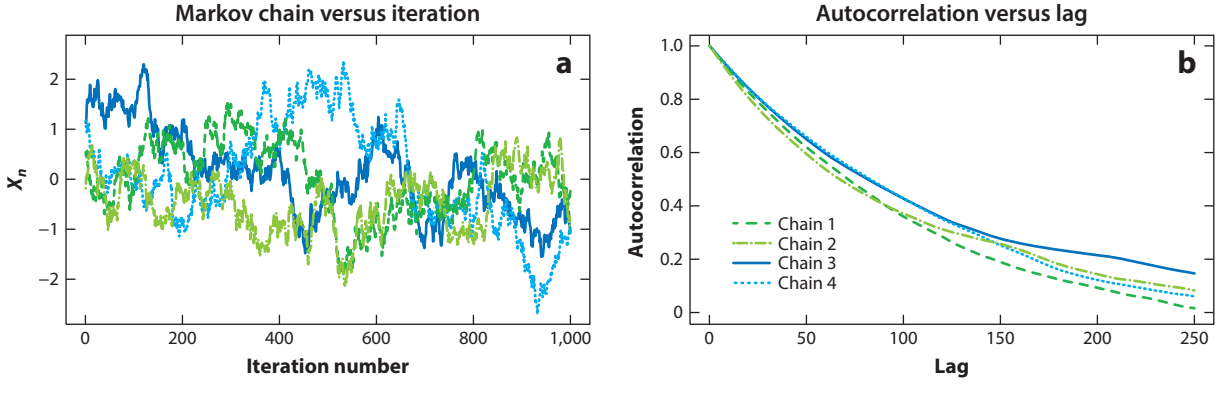

Figure 9

(*a*) Trace and (*b*) autocorrelation function plots of the X marginal of the four chains for the sixmodal example in case 2.

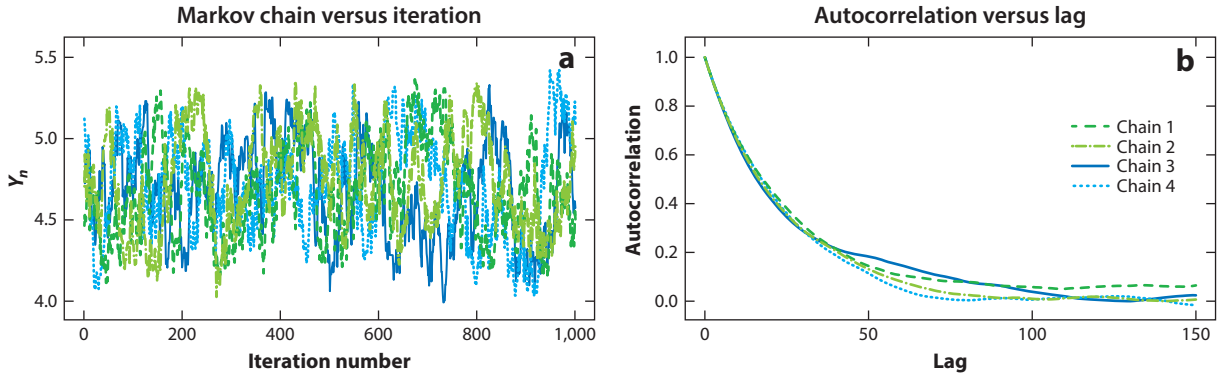

Figure 10

(*a*) Trace and (*b*) autocorrelation function plots of the Y marginal of the four chains for the sixmodal example in case 2. The large amount of overlap between the trace plots of the four marginal chains fails to indicate nonconvergence of the Markov chains to stationarity.

We also calculate the PSRF for the marginal chains, as well as the MPSRF for the joint chain based on the four parallel chains, as the GR diagnostic is often used by practitioners for determining burn-in (Flegal et al. 2008, p. 256). The plots of iterative $\hat{R}$ at increments of 200 iterations are given in **Figure 12**. PSRFs for the marginal chains reach below 1.1 before 3,000 iterations. The MPSRF (not shown in the plot) also reaches below 1.1 before 6,000 iterations. Both the PSRF and MPSRF values are close to one, which is often used as a sign of convergence to stationarity.

Since all four chains are stuck at the same local mode, that is, these are not run long enough to move between the modes, the convergence diagnostics, including PSRF and MPSRF, get fooled into thinking that the target distribution is unimodal and hence falsely detect convergence.

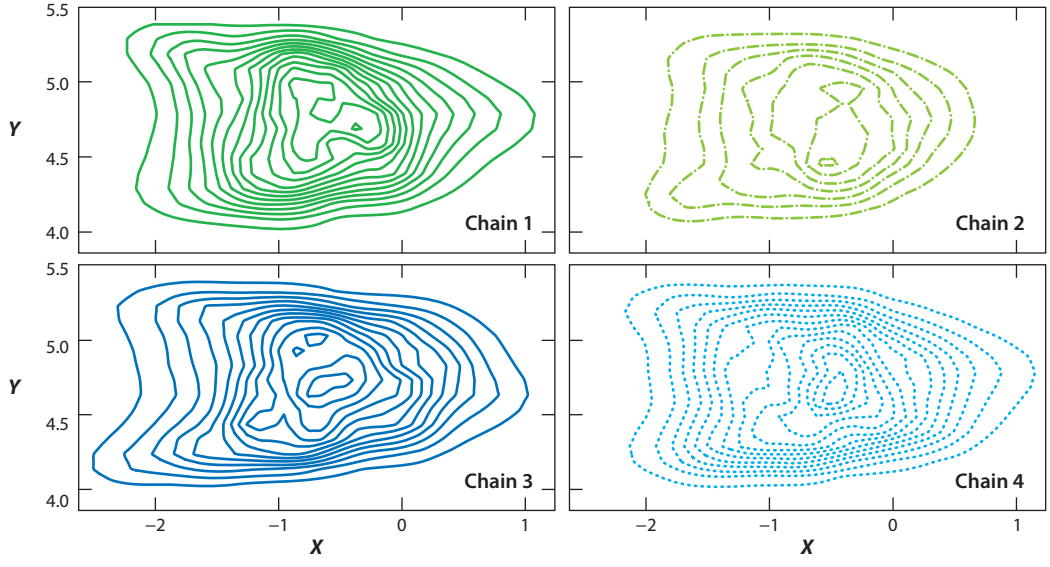

Figure 11

Visualizations of the adaptive kernel density estimates of the four chains in case 2 of the sixmodal example. Since the bivariate density plots look similar, they fail to provide indication of nonconvergence of the chains to the stationary distribution.

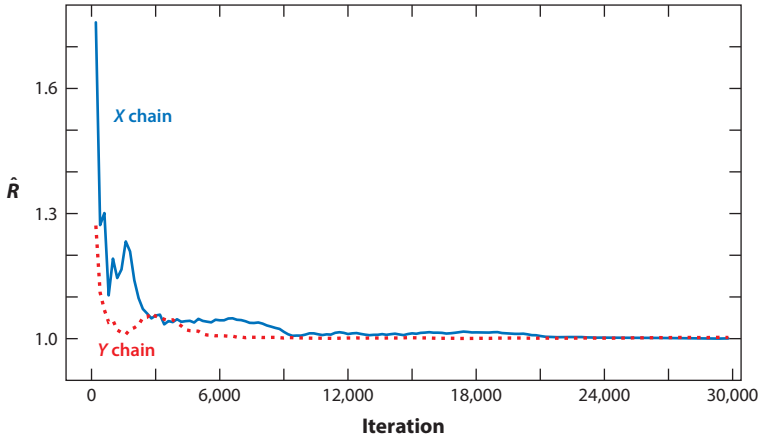

Figure 12

Iterative $\hat{R}$ plot from four parallel chains for the sixmodal example in case 2.

Laha et al. (2016) demonstrate failures of trace plots, autocorrelation plots, and PSRF in diagnosing nonconvergence of MCMC samplers in the context of a statistical model used for analyzing rank data. [See Hobert et al. (2011) for examples of multimodal targets arising from the popular Bayesian finite mixture models where empirical convergence diagnostic tools face similar issues.] Since these diagnostic tools make use of (only) the samples obtained from the MCMC algorithm, and all observations lie around the same mode, they fail to diagnose nonconvergence. In contrast, Dixit & Roy's (2017) Tool 2 (Equation 9) uses both MCMC samples and the target density. Since Tool 2 requires only one chain, and since the PSRF suggests that the four chains are similar, we simply choose one of the four chains. Now, $T_2^* = 0.88$ is significantly greater than zero and thus indicates that the chain is stuck at a local mode. Furthermore, it also indicates that 88% of the target distribution is not yet captured by the Markov chain. Thus, Dixit & Roy's (2017) Tool 2 is successful in detecting the divergence of the chains.

3.3. A Bayesian Logistic Model

In this section, we illustrate MCMC convergence diagnostics in the context of a real data analysis using a popular statistical model. In particular, we fit a Bayesian logistic model on the *Anguilla australis* distribution data set provided in the R package `dismo` (Hijmans et al. 2016). Data are available on a number of sites with the presence or absence of the short-finned eel (*Anguilla australis*) in New Zealand, and some environmental variables at these sites. In particular, we fit the Anguilla_train data available in the `dismo` package. Here, the response variable is the presence or absence of short-finned eel, and six other variables are included as covariates. The six covariates are: summer air temperature (SeqSumT), distance to coast (DSDist), area with indigenous forest (USNative), average slope in the upstream catchment (USSlope), maximum downstream slope (DSMaxSlope), and fishing method (categorical variable with five classes: electric, mixture, net, spot and trap). Thus, the data set consists of $(y_i, x_i), i = 1, \ldots, 1{,}000$, where y_i is the ith observation of the response variable taking value 1 (presence) or 0 (absence), and $x_i = (1, \tilde{x}_i)$ is the ten-dimensional covariate vector, 1 for the intercept, and $\tilde{x}_i$ for the other nine covariates (with four components for the categorical variable fishing method). This example was also used by Dixit & Roy (2017) and Boone et al. (2014) to illustrate their MCMC convergence diagnostic tools.

Denote $\beta = (\beta_0, \beta_1, \ldots, \beta_9)$ where β_0 is the intercept and $(\beta_1, \ldots, \beta_9)$ is the 9×1 vector of unknown regression coefficients. We consider the logistic regression model

$$Y_i|\beta \overset{\text{ind}}{\sim} \text{Bernoulli}(F(x_i^T\beta)), i = 1, \ldots, 1{,}000,$$

where $F(\cdot)$ is the cumulative distribution function of the logistic distribution, that is,

$$F(x_i^T\beta) = \frac{\exp(x_i^T\beta)}{1 + \exp(x_i^T\beta)}, i = 1, \ldots, 1{,}000.$$

We consider a Bayesian analysis with a diffuse normal prior on β. Thus, the posterior density is

$$\pi(\beta|y) \propto \ell(\beta|y)\phi_{10}(\beta) = \prod_{i=1}^{n} F(x_i^T\beta)^{y_i}\{1 - F(x_i^T\beta)\}^{1-y_i}\phi_{10}(\beta), \qquad 11.$$

where $\ell(\beta|y)$ is the likelihood function and $\phi_{10}(\beta)$ is the density of $N(\mathbf{0}, 100\, I_{10})$. The posterior density (Equation 11) is intractable in the sense that means with respect to this density, which are required for Bayesian inference, are not available in closed form.

As in Dixit & Roy (2017) and Boone et al. (2014), we use the `MCMClogit` function in the R package `MCMCpack` (Martin et al. 2011) to draw MCMC samples from the target density $\pi(\beta|y)$. The maximum likelihood estimate (MLE) of β is the value of the parameter where the likelihood function $\ell(\beta|y)$ is maximized. Exact MLE is not available for the logistic likelihood function, and neither is the mode of the posterior density (Equation 11). But numerical optimization methods can be used to find an approximate MLE or posterior mode, which may then be used as starting values. In order to assess convergence to stationarity, we run three parallel chains with the default tuning values for 5,000 iterations, one initialized at the MLE and the other two initialized at points away from the MLE. Trace plots of the three chains for the last 1,000 iterations for the regression coefficients of summer air temperature (panel *a*) and distance to coast (panel *b*) are given in **Figure 13**. Trace plots of the other variables look similar. From these plots, we see that there is not much overlap between the three parallel chains. From the frequent flat bits, it follows that the Markov chains move tardily and suffer from slow mixing. Indeed, the default tuning parameters in the `MCMClogit` function result in a low (0.11) acceptance rate. We next set the tuning

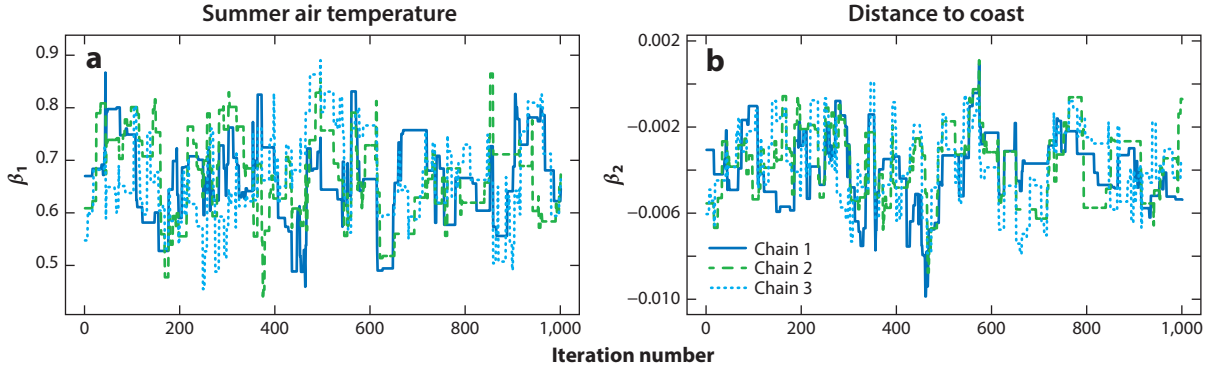

Figure 13

Trace plots of the three chains with default tuning for the regression coefficients of (*a*) summer air temperature and (*b*) distance to coast for the Bayesian logistic model example. The presence of frequent flat bits indicates slow mixing of the Markov chains.

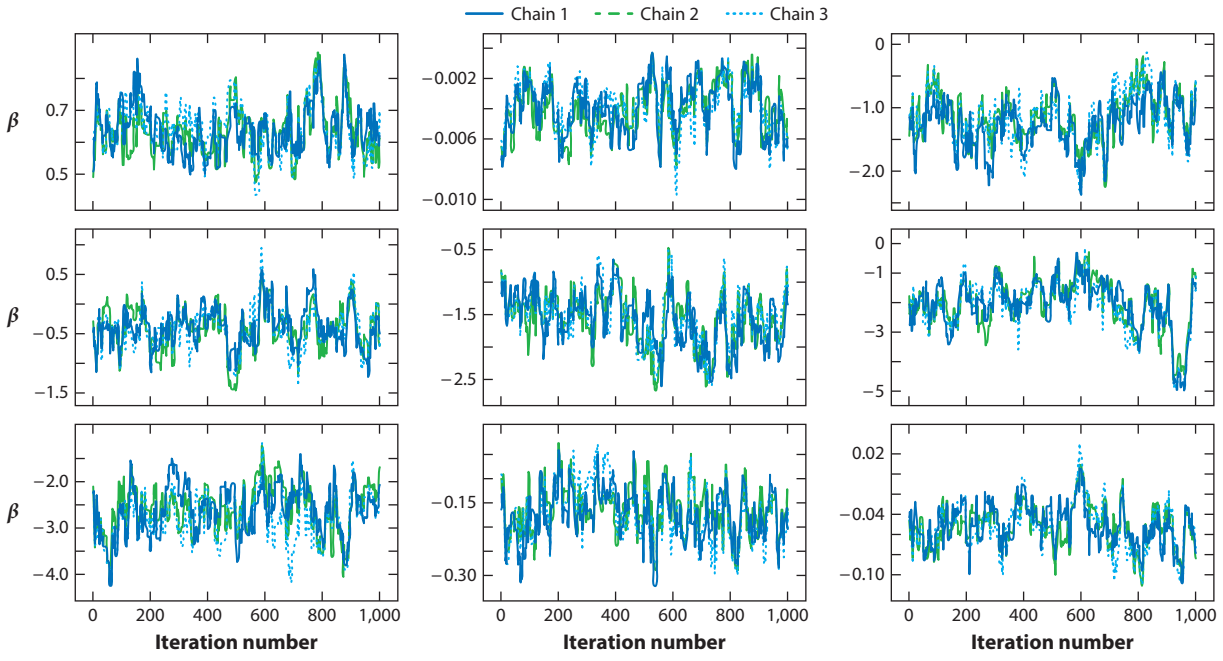

Figure 14

Trace plots of the three chains for the nine regression coefficient variables for the Bayesian logistic model example. The plots show improved mixing from tuning the acceptance rate of the Markov chains.

parameters to achieve around a 40% acceptance rate, and all analysis in the remaining section is based on these new tuning values. We run the three chains longer (30,000 iterations) to obtain reliable ACF plots. Trace plots of the last 1,000 iterations for each of the three chains for the nine regression coefficient variables are given in **Figure 14**. From the trace plots, we see that convergence of the chains can be further improved. Autocorrelations for all ten variables for one of the chains based on all 30,000 draws are given in **Figure 15**. Autocorrelations for the other two chains look similar (not included here). Like the trace plots, the autocorrelation plots also reveal that the Markov chains suffer from high autocorrelations. It is further corroborated by the mESS values, which are less than 1,000 for all the three chains. To sample from Equation 11, one may use an alternative MCMC sampler, e.g., the Pólya–Gamma Gibbs sampler (Polson et al. 2013), which is known to be geometrically ergodic (Choi & Hobert 2013, Wang & Roy 2018c). Here we do not use the Pólya–Gamma Gibbs sampler, as our goal is to illustrate the convergence diagnostic methods. The MPSRF reaches close to one before 30,000 iterations. Since the Markov chains are 10-dimensional, to maintain an overall type 1 error rate of $\alpha = 0.05$, using Bonferroni's correction, Dixit & Roy (2017) advocate the cutoff point 0.01 for the KL Tool 1 for marginal chains. For each of the ten variables, the maximum symmetric KL divergence among the three pairs of chains is computed. It turns out that the marginal chains do not pass the KL Tool 1 test, as the maximum symmetric KL divergence takes the value 7.26 for the variable USSlope. After 30,000 iterations, all marginal chains pass the Heidelberger & Welch (1983) stationarity test. In contrast, for each of the three parallel chains, for some of the variables, the Geweke (1992) Z test turns out to be significant at the 0.05 level. Next, we run the chains for another 40,000 iterations. For the last 40,000 iterations, all marginal chains pass the Geweke (1992) Z test, as well as the KL Tool 1 test. Also, based on these 40,000 iterations, the maximum burn-in estimate from the Raftery–Lewis

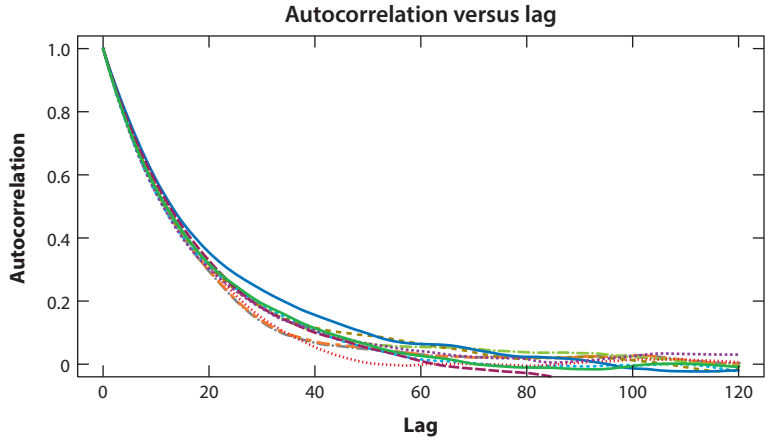

Figure 15

Autocorrelation plots of the ten marginal chains for the Bayesian logistic model example. For all variables, correlations between samples with more than 100 steps apart are small.

diagnostic (with $\epsilon = 0.005, \alpha = 0.05$) over different quantiles ($q = 0.1, \ldots, 0.9$) is less than 100 for all 10 variables. We thus use $n' = 70{,}000$ as the burn-in value.

After removing the first 70,000 iterations as initial burn-in, each of the three chains is run for an additional 15,000 iterations. **Table 1** presents the PSRF and the maximum symmetric KL divergence [Dixit & Roy's (2017) KL Tool 1] values based on three parallel chains for all 10 variables. The half-widths of the 95% confidence intervals based on the first chain (started at the MLE) are also tabulated in **Table 1**. All values are given up to three decimal places. MPSRF takes the value 1.004. For the three chains, mESS takes the values 515, 520, and 502, respectively. High cross-correlation between the Intercept and SeqSumT regression coefficient parameters (−0.984) and between USNative and USSlope (−0.558) suggests that mixing of the Markov chain can improve if it is run on an appropriate lower dimensional space (that is, after dropping some variables) or a reparameterization is used. From **Table 1**, we see that all marginal chains pass the KL Tool 1 diagnostic. Also, all PSRF values, as well as the MPSRF value, reach

Table 1 Application of various MCMC convergence diagnostic tools to the Bayesian logistic model

Variable	$\hat{R}$	Half-width	Tool 1
Intercept	1.000	0.112	0.008
SeqSumT	1.000	0.006	0.007
DSDist	1.001	0.000	0.005
USNative	1.001	0.025	0.004
M - mix	1.000	0.031	0.005
M - net	1.001	0.031	0.004
M - spot	1.000	0.048	0.004
M - trap	1.002	0.051	0.004
DSMaxSlope	1.000	0.005	0.007
USSlope	1.001	0.002	0.004

Abbreviation: MCMC, Markov chain Monte Carlo.

Table 2 Estimates of posterior means and standard errors of regression coefficients for the Bayesian logistic model

Variable	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9
Estimate	−10.46	0.66	−0.00	−1.17	−0.47	−1.53	−1.83	−2.59	−0.17	−0.05
SE $\times 10^3$	5.73	0.32	0.00	1.52	1.46	1.65	2.76	2.35	0.24	0.08

Abbreviation: SE, Monte Carlo standard error.

below the cutoff 1.1. In contrast, the maximum half-width among the 10 regression parameters is 0.112, much larger than the cutoff 0.01. Doing a simple sample size calculation, based on the pilot sample size 15,000, we find that we need $15{,}000 \times (0.112/0.01)^2 = 1{,}881{,}600$ samples for obtaining confidence intervals with half-widths below 0.01.

Finally, we run one of the chains (the chain started at the MLE) for 1,881,600 iterations after a burn-in of $n' = 70{,}000$ iterations. Thus, the chain is stopped after $n^* = 1{,}951{,}600$ iterations. In this case, as expected, the maximum half-width of the 95% confidence interval is below 0.01. An estimate of mESS calculated using the `mcmcse` package is 55,775, which is larger than the cutoff value 55,191 given in Equation 6 for $p = 10, \alpha = 0.05$, and $\varepsilon = 0.02$. In contrast, the chain needs to be run longer to achieve the cutoff value 220,766 (Equation 6) corresponding to $\varepsilon = 0.01$. **Table 2** gives the estimates of posterior means of all regression coefficients and their corresponding Monte Carlo standard errors.

4. CONCLUSIONS AND DISCUSSION

In this article, we discuss several measures for diagnosing convergence of Monte Carlo Markov chains to stationarity as well as convergence of the sample averages based on these chains. Detection of the first is often used to decide a suitable burn-in period, while the second leads to termination of the MCMC simulation. Analytical upper bounds to the TV norm required to obtain an honest burn-in may be difficult to find in practice or may lead to very conservative burn-in values. In contrast, empirical diagnostics can falsely detect convergence when the chains are not run long enough to move between the modes. For the chains initialized at high-density density regions, there is no need for burn-in. If the global mode of the target density can be (approximately) found by optimization, then it can be used as the starting value.

Some of the empirical diagnostics for convergence of sample averages may prematurely terminate the simulation, and the resulting inference can be far from the truth. Thus, use of fixed-width– and ESS-based stopping rules is recommended. Most of the quantitative convergence diagnostics assume a Markov chain CLT. While demonstrating the existence of a Markov chain CLT requires some rigorous theoretical analysis of the Markov chain, given the great amount of work done in this direction, validating honest stopping rules does not present as much of an obstacle as in the past.

None of the three examples discussed here use thinning. Thinning, that is, discarding all but every kth observation, is often used by MCMC practitioners to reduce high autocorrelations present in the Markov chain samples. Since it wastes too many samples, it should be used only if computer storage of the samples is an issue or evaluating the functions of interest (g) is more expensive than sampling the Markov chain. If thinning is used, convergence diagnostics can be used on the thinned samples.

Some convergence diagnostic tools use parallel chains initialized at different points, or two parts of a single chain. In the presence of multiple modes, if the initial points of the parallel chains are not in distinct high-density regions, or the chain is not run long enough to move between the modes, the diagnostics fail to detect the nonconvergence. Thus, single long runs should be used

to make a final inference. Running the chain longer may also result in discovering new parts of the support of the target distribution. In contrast, recently, Jacob et al. (2017) propose a method for parallelizing MCMC computations using couplings of Markov chains.

Practitioners should be careful while depending purely on empirical convergence diagnostic tools, especially if the presence of multiple modes is suspected. Empirical diagnostics cannot detect convergence with certainty. Also, if the target is incorrectly assumed to be a proper density, the empirical diagnostic tools may not provide a red flag indicating its impropriety (Athreya & Roy 2014, Hobert & Casella 1996). Over the past two decades, much research has been done to provide honest Monte Carlo sample size calculation for myriad MCMC algorithms for common statistical models. However, theoretical analysis of MCMC algorithms is an ongoing area of research, and further important work needs to be done. A potential future study involves theoretically verifying the convergence (to zero) of Dixit & Roy's (2017) statistics based on the KL divergence. Another possible research problem is to construct theoretically valid and computationally efficient MCMC convergence diagnostics in ultrahigh-dimensional settings.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The author thanks one anonymous editor for careful and detailed comments on an earlier version of the manuscript. The author thanks Evangelos Evangelou, Mark Kaiser, and Dootika Vats for helpful comments. These valuable suggestions have substantially improved the article. The author also thanks Chris Oats for suggesting two references and Anand Dixit for providing some R codes used in the second and third examples.

LITERATURE CITED

Andrieu C, Fort G, Vihola M. 2015. Quantitative convergence rates for subgeometric Markov chains. *J. Appl. Probab.* 52:391–404

Asmussen S, Glynn PW. 2011. A new proof of convergence of MCMC via the ergodic theorem. *Stat. Probab. Lett.* 81:1482–85

Athreya KB, Roy V. 2014. Monte Carlo methods for improper target distributions. *Electron. J. Stat.* 8:2664–92

Baxendale PH. 2005. Renewal theory and computable convergence rates for geometrically ergodic Markov chains. *Ann. Appl. Probab.* 15:700–38

Boone E, Merrick J, Krachey M. 2014. A Hellinger distance approach to MCMC diagnostics. *J. Stat. Comput. Simul.* 84:833–49

Brooks SP, Gelman A. 1998. General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Stat.* 7:434–55

Brooks SP, Roberts GO. 1998. Assessing convergence of Markov chain Monte Carlo algorithms. *Stat. Comput.* 8:319–35

Chakraborty S, Khare K. 2017. Convergence properties of Gibbs samplers for Bayesian probit regression with proper priors. *Electron. J. Stat.* 11:177–210

Choi HM, Hobert JP. 2013. The Polya-Gamma Gibbs sampler for Bayesian logistic regression is uniformly ergodic. *Electron. J. Stat.* 7:2054–64

Christensen OF. 2004. Monte Carlo maximum likelihood in model based geostatistics. *J. Comput. Graph. Stat.* 13:702–18

Cowles MK, Carlin BP. 1996. Markov chain Monte Carlo convergence diagnostics: a comparative review. *J. Am. Stat. Assoc.* 91:883–904

Dixit A, Roy V. 2017. MCMC diagnostics for higher dimensions using Kullback Leibler divergence. *J. Stat. Comput. Simul.* 87:2622–38

Doss CR, Flegal JM, Jones GL, Neath RC. 2014. Markov chain Monte Carlo estimation of quantiles. *Electron. J. Stat.* 8:2448–78

Durmus A, Moulines E. 2015. Quantitative bounds of convergence for geometrically ergodic Markov chain in the Wasserstein distance with application to the Metropolis adjusted Langevin algorithm. *Stat. Comput.* 25:5–19

Evangelou E, Roy V. 2019. `geoBayes`: analysis of geostatistical data using Bayes and empirical Bayes methods. *R package*, version 0.6.2. **https://cran.r-project.org/web/packages/geoBayes/index.html**

Flegal JM, Gong L. 2015. Relative fixed-width stopping rules for Markov chain Monte Carlo simulations. *Stat. Sinica* 25:655–75

Flegal JM, Haran M, Jones GL. 2008. Markov chain Monte Carlo: Can we trust the third significant figure? *Stat. Sci.* 23:250–60

Flegal JM, Hughes J, Vats D, Dai N. 2012. `mcmcse`: Monte Carlo standard errors for MCMC. *R package*, version 0.1. **https://cran.r-project.org/web/packages/mcmcse/index.html**

Flegal JM, Jones GL. 2010. Batch means and spectral variance estimators in Markov chain Monte Carlo. *Ann. Stat.* 38:1034–70

Fort G, Moulines E, Roberts G, Rosenthal J. 2003. On the geometric ergodicity of hybrid samplers. *J. Appl. Probab.* 40:123–46

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. 2014. *Bayesian Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC

Gelman A, Rubin DB. 1992. Inference from iterative simulation using multiple sequences. *Stat. Sci.* 7:457–72

Geweke J. 1992. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In *Bayesian Statistics 4*, ed. JM Bernardo, JO Berger, AP Dawid, AFM Smith, pp. 169–93. Oxford, UK: Clarendon

Geyer CJ. 2011. Introduction to Markov chain Monte Carlo. In *Handbook of Markov Chain Monte Carlo*, ed. S Brooks, A Gelman, GL Jones, XL Meng, pp. 3–48. Boca Raton, FL: Chapman & Hall/CRC

Geyer CJ, Johnson LT. 2017. `mcmc`: Markov chain Monte Carlo. *R package*, version 0.9-5. **https://cran.r-project.org/web/packages/mcmc/index.html**

Geyer CJ, Thompson EA. 1995. Annealing Markov chain Monte Carlo with applications to ancestral inference. *J. Am. Stat. Assoc.* 90:909–20

Glynn PW, Whitt W. 1992. The asymptotic validity of sequential stopping rules for stochastic simulations. *Ann. Appl. Probab.* 2:180–98

Gong L, Flegal JM. 2016. A practical sequential stopping rule for high-dimensional Markov chain Monte Carlo. *J. Comput. Graph. Stat.* 25:684–700

Gorham J, Mackey L. 2015. Measuring sample quality with Stein's method. In *Advances in Neural Information Processing Systems 28*, ed. C Cortes, ND Lawrence, DD Lee, M Sugiyama, R Garnett, pp. 226–34. San Diego, CA: NeurIPS

Hadfield JD. 2010. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *J. Stat. Softw.* 33:1–22

Heidelberger P, Welch PD. 1983. Simulation run length control in the presence of an initial transient. *Oper. Res.* 31:1109–44

Hijmans RJ, Phillips S, Leathwick J, Elith J. 2016. `dismo`: species distribution modeling. *R package*, version 1.0-15. **https://cran.r-project.org/web/packages/dismo/index.html**

Hobert JP, Casella G. 1996. The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *J. Am. Stat. Assoc.* 91:1461–73

Hobert JP, Jones GL, Presnell B, Rosenthal JS. 2002. On the applicability of regenerative simulation in Markov chain Monte Carlo. *Biometrika* 89:731–43

Hobert JP, Jung YJ, Khare K, Qin Q. 2018. Convergence analysis of MCMC algorithms for Bayesian multivariate linear regression with non-Gaussian errors. *Scand. J. Stat.* 45:513–33

Hobert JP, Roy V, Robert CP. 2011. Improving the convergence properties of the data augmentation algorithm with an application to Bayesian mixture modelling. *Stat. Sci.* 26:332–51
Jacob PE, O'Leary J, Atchadé YF. 2017. Unbiased Markov chain Monte Carlo with couplings. arXiv:1708.03625 [stat.ME]
Jones GL. 2004. On the Markov chain central limit theorem. *Probab. Surv.* 1:299–320
Jones GL, Haran M, Caffo BS, Neath R. 2006. Fixed-width output analysis for Markov chain Monte Carlo. *J. Am. Stat. Assoc.* 101:1537–47
Jones GL, Hobert JP. 2001. Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Stat. Sci.* 16:312–34
Jones GL, Hobert JP. 2004. Sufficient burn-in for Gibbs samplers for a hierarchical random effects model. *Ann. Stat.* 32:784–817
Khare K, Hobert JP. 2012. Geometric ergodicity of the Gibbs sampler for Bayesian quantile regression. *J. Multivar. Anal.* 112:108–16
Khare K, Hobert JP. 2013. Geometric ergodicity of Bayesian lasso. *Electron. J. Stat.* 7:2150–63
Laha A, Dutta S, Roy V. 2016. A novel sandwich algorithm for empirical Bayes analysis of rank data. *Stat. Interface* 10:543–56
Leman SC, Chen Y, Lavine M. 2009. The multiset sampler. *J. Am. Stat. Assoc.* 104:1029–41
Martin AD, Quinn KM, Park JH. 2011. MCMCpack: Markov chain Monte Carlo in R. *J. Stat. Softw.* 42:22
Mengersen KL, Robert CP, Guihenneuc-Jouyaux C. 1999. MCMC convergence diagnostics: a reviewww. *Bayesian Stat.* 6:415–40
Mengersen KL, Tweedie RL. 1996. Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Stat.* 24:101–21
Meyn SP, Tweedie RL. 1993. *Markov Chains and Stochastic Stability*. London: Springer
Mykland P, Tierney L, Yu B. 1995. Regeneration in Markov chain samplers. *J. Am. Stat. Assoc.* 90:233–41
Peltonen J, Venna J, Kaski S. 2009. Visualizations for assessing convergence and mixing of Markov chain Monte Carlo simulations. *Comput. Stat. Data Anal.* 53:4453–70
Plummer M, Best N, Cowles K, Vines K. 2006. Coda: convergence diagnosis and output analysis for MCMC. *R News* 6:7–11
Polson NG, Scott JG, Windle J. 2013. Bayesian inference for logistic models using Pólya-Gamma latent variables. *J. Am. Stat. Assoc.* 108:1339–49
Qin Q, Hobert JP. 2019. Geometric convergence bounds for Markov chains in Wasserstein distance based on generalized drift and contraction conditions. arXiv:1902.02964 [math.PR]
R Core Team. 2018. R: A language and environment for statistical computing. *Statistical Software*, R Found. Stat. Comput., Vienna
Raftery AE, Lewis SM. 1992. How many iterations in the Gibbs sampler? In *Bayesian Statistics 4*, ed. JM Bernardo, JO Berger, AP Dawid, AFM Smith, pp. 763–73. Oxford, UK: Clarendon
Robert C, Casella G. 2004. *Monte Carlo Statistical Methods*. New York: Springer. 2nd ed.
Roberts GO, Rosenthal JS. 2004. General state space Markov chains and MCMC algorithms. *Probab. Surv.* 1:20–71
Román JC, Hobert JP. 2012. Convergence analysis of the Gibbs sampler for Bayesian general linear mixed models with improper priors. *Ann. Stat.* 40:2823–49
Román JC, Hobert JP. 2015. Geometric ergodicity of Gibbs samplers for Bayesian general linear mixed models with proper priors. *Linear Algebra Appl.* 473:54–77
Rosenthal JS. 1995. Minorization conditions and convergence rates for Markov chain Monte Carlo. *J. Am. Stat. Assoc.* 90:558–66
Rosenthal JS. 2002. Quantitative convergence rates of Markov chains: a simple account. *Electron. Commun. Probab.* 7:123–28
Roy V. 2012. Convergence rates for MCMC algorithms for a robust Bayesian binary regression model. *Electron. J. Stat.* 6:2463–85
Roy V, Chakraborty S. 2017. Selection of tuning parameters, solution paths and standard errors for Bayesian lassos. *Bayesian Anal.* 12:753–78

Roy V, Hobert JP. 2007. Convergence rates and asymptotic standard errors for MCMC algorithms for Bayesian probit regression. *J. R. Stat. Soc. B* 69:607–23

Roy V, Hobert JP. 2010. On Monte Carlo methods for Bayesian regression models with heavy-tailed errors. *J. Multivar. Anal.* 101:1190–202

Silverman BW. 1986. *Density Estimation for Statistics and Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC

Vats D. 2017. Geometric ergodicity of Gibbs samplers in Bayesian penalized regression models. *Electron. J. Stat.* 11:4033–64

Vats D, Flegal JM, Jones GL. 2019. Multivariate output analysis for Markov chain Monte Carlo. *Biometrika* 106:321–37

Vats D, Knudson C. 2018. Revisiting the Gelman-Rubin diagnostic. arXiv:1812.09384 [stat.CO]

Wang X, Roy V. 2018a. Analysis of the Pólya-Gamma block Gibbs sampler for Bayesian logistic linear mixed models. *Stat. Probab. Lett.* 137:251–56

Wang X, Roy V. 2018b. Convergence analysis of the block Gibbs sampler for Bayesian probit linear mixed models with improper priors. *Electron. J. Stat.* 12:4412–39

Wang X, Roy V. 2018c. Geometric ergodicity of Pólya-Gamma Gibbs sampler for Bayesian logistic regression with a flat prior. *Electron. J. Stat.* 12:3295–311

Yu B. 1994. *Estimating* L^1 *error of kernel estimator: monitoring convergence of Markov samplers*. Tech. Rep., Univ. Calif., Berkeley. **https://pdfs.semanticscholar.org/78ea/f3290a6f612ac5b8f43c9cc8a5a03d267084.pdf**