To appear in the *Journal of Statistical Computation and Simulation*
Vol. 00, No. 00, Month 20XX, 1–16

Equivalence testing for MCMC convergence assessment

Michael Lerch[a*] Megan Higgs[a]

[a]*Montana State University, Bozeman, Montana*

(*Received 00 Month 20XX; final version received 00 Month 20XX*)

Practitioners of Bayesian data analysis often find the need to use Markov chain Monte Carlo to estimate posterior quantities. A necessary step in using MCMC is convergence assessment. The practitioner must carefully assess whether whether the draws from the chains accurately represent the target distribution, and this is often done with some convergence diagnostic. We review multiple prominent strategies for assessing convergence and propose a method designed for problems where posterior quantiles or probabilities are the quantities of interest, a common inferential objective. The framework of our diagnostic is equivalence testing which is novel in MCMC convergence assessment. We demonstrate the use of this new diagnostic in assessing convergence and contrast to some existing diagnostics for a simulated example and for a hierarchical modeling example.

Keywords: Markov chain Monte Carlo; Bayesian data analysis; convergence; equivalence testing

1. Introduction

1.1. *MCMC and convergence background*

A typical objective of Bayesian data analysis is to obtain summaries, such as credible intervals, of a posterior distribution, $p(\theta|y)$. Often, analytically calculating these summary values is intractable and the analyst turns to a Markov chain Monte Carlo (MCMC) method in order to produce a set of samples $\Theta_n = \{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(n)}\}$ to approximate $p(\theta|y)$ and compute summary values such as medians, means, intervals and probabilities. Though there may be strong autocorrelation within a chain of samples, we are assured that under the appropriate conditions [1], an estimated summary value calculated with Θ_n approaches the corresponding summary value of the true distribution, $p(\theta|y)$, as the number of samples, or the length of the chain, increases. MCMC is widely used because it is typically easier to produce a set of samples Θ_n, than to analytically calculate summary values of $p(\theta|y)$ [2–5]. In fact, today there are many software solutions that automate much of the process to obtain such a sample Θ_n [6–8, etc.].

Convergence of the estimated value (the summary of the samples Θ_n) to the true value (the summary of the true posterior) is what makes MCMC such a valuable tool, but researchers must still ask themselves the practical question of how long to sample until convergence kicks in and the estimates are appropriate to report. Posterior MCMC estimates are subject to error if the algorithm has not effectively explored the entire parameter space [9, 10]. Tail quantiles, as endpoints of credible intervals, are often summary values of interest and these tail quantile estimates may be particularly sensitive to MCMC chains that have not appropriately explored the parameter space. Therefore, the researcher should explicitly assess convergence in the context of the desired quantities.

*Corresponding author. Email: michael.lerch@msu.montana.edu

There are multiple reasons why an MCMC chain may not provide a good estimate of a posterior summary value. First, draws in MCMC chains are, by definition, autocorrelated. It has been suggested [11] that MCMC chains are initiated via draws from an overly dispersed starting distribution, with the goal of better assessing convergence, and we echo these sentiments. However, autocorrelation in a chain with an overly dispersed starting point may bias samples in the chain towards that starting value. To combat this bias, many MCMC software packages remove initial parts of chains, called 'burn-in', by default. Some early work in calculating an appropriate burn-in length was performed by Schruben and others [12–14]. Today, generating samples is computationally cheaper and software MCMC solutions like RStan [15] default to large burn-in lengths (half of the chain) by default. In the end, chains need to be sampled long enough so that the *collection* of draws mimics a sample of independent draws from the posterior distribution. If a chain is stopped too soon, then the distribution of draws across the posterior space may not adequately represent the true posterior.

A second potential source of error in MCMC estimates is the fact that the estimates are based on a finite number of samples For example, 100 independent draws might be taken from a target, say, normal distribution. Although these draws are unaffected by a starting value and are not autocorrelated, they may still be insufficient to precisely estimate, say, the 0.975 quantile depending on the desired precision. Thus, even in exact or perfect sampling [16, 17], where autocorrelation is not an issue, estimates can be produced that do not meet the precision required. The accumulating effects of a finite sample and autocorrelation in draws are sometimes referred to as Monte Carlo error [2, 18].

In this paper, we propose the Quantile Equivalence Diagnostic (QED), along with a graphical display, the Quantile Equivalence plot (QEplot). Our diagnostic and plot supplement a number of existing convergence assessment tools. As we will see, many diagnostics are used as generic tests with difficult to interpret measures of convergence without deference to specific goals of the researchers. We approach the problem by specifically addressing quantities of interest and using equivalence testing as the framework to provide readily interpretable measures of convergence.

1.2. *Brief history of MCMC convergence*

There are many existing diagnostics for assessing MCMC convergence that have been developed as the technique has become more widely used. Here, we review the Geweke diagnostic, the potential scale reduction factor ($\hat{R}$), Boone's Hellinger distance, and the Raftery-Lewis diagnostics. We have selected these diagnostics due to their popularity and relation to our Quantile Equivalence Diagnostic. Cowles and Carlin [9] can be consulted for a more in-depth review that, although published in 1996, covers most commonly used diagnostics.

1.2.1. Geweke

Markov chains can be fraught with autocorrelation. Though this is the nature of a Markov process, low autocorrelation is a property of an efficient MCMC algorithm. If autocorrelation is very high, the beginning part of a chain may not match the end part of a chain. Geweke's diagnostic calculates a two sample z-score where one sample is the first n_1 draws of the chain and the other sample is the last n_2 draws of the chain. The `geweke.diag()` function in the CODA R package [19] defaults to n_1 and n_2 being the first 10% and the last 50% of the chain, respectively. A large z-score indicates the autocorrelation is so high that the two parts of the chain are shifted apart and therefore, at minimum, one

part does not represent draws from the posterior, thus implying a lack of convergence. A small z-score indicates the beginning and end draws are similar to each other, with the assumption that if both parts are similar to each other it is because they are both similar to the posterior. However, Geweke's diagnostic could suggest a lack of convergence if the whole chain provides a good approximation to the posterior despite both the end part and beginning part individually being poor approximations to the posterior. Of course, the more serious error is to suggest convergence when it is not appropriate, but it is still undesirable to unnecessarily suggest lack of convergence particularly if running the sampler is costly. Diagnostics like the Potential Scale Reduction Factor use whole chains as the comparison unit rather than parts of chains, which can prevent high autocorrelation from falsely indicating a lack of convergence.

1.2.2. The Potential Scale Reduction Factor

The Potential Scale Reduction Factor (PSRF), or $\hat{R}$, has been published multiple times with slight changes [5, 11, 20]. With multiple chains, the posterior variance of a scalar parameter can be estimated by the within-chain variances as well as the variance of the means of the chains (multiplied by the number of samples within a chain) assuming the draws all come from the posterior distribution. The concept is similar to a one-way ANOVA problem with the chains acting as the different groups. By starting the chains from overly dispersed locations in the posterior space, it is expected that before convergence, the within-chain variance estimate will be less than the between chain estimate. As sampling continues and convergence is approached, the two estimates of variance will both approach the true posterior variance and the diagnostic, $\hat{R}$, will approach 1. The recommendation of Gelman et al. [5] is to compute $\hat{R}$ for all scalar parameters of interest and verify that each is near 1.0. The authors suggest that values below 1.1 are acceptable for most cases and that values closer to 1.0 may be necessary for more critical problems. However, such cutoffs are arbitrary and often there is little intuition about what an acceptable value may be for specific posterior quantities of interest.

Further, the authors recommend the parameter be transformed so that the posterior distribution is approximately normal before computing the PSRF. However, this may not be realistic for some distributions like bimodal distributions, and many users of the methods are not likely to implement this additional step. Brooks and Gelman [20] note that the PSRF may be inadequate for quantities other than the posterior mean and Gelman et al. [5] state that 'when performing inference for extreme quantiles, or for parameters with multimodal posterior distributions, one should monitor also extreme quantiles of the "between" and "within" sequences.' In our experience, it is common for the inferential objectives to include extreme quantiles to construct credible intervals. While the calculation of the PSRF is reliant on only the mean and variance of each chain, a diagnostic proposed by Boone et al. aims to use the *entire* chain distributions to assess convergence.

1.2.3. Boone's Hellinger distance

In a recent paper, Boone et al. [21] suggest a test for convergence using the Hellinger distance to measure discrepancies between the approximate posterior distributions obtained from different MCMC chains, with smaller discrepancies between chains indicating convergence. As given in their paper, the Hellinger distance is a symmetric distance between two probability distributions f and g,

$$H(f,g) = \sqrt{\frac{1}{2}\int \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 dx}\,.$$

The authors calculate approximate Hellinger distances between pairs of chains via two approximations. First, the chain densities are approximated along a grid in the posterior space via a Gaussian kernel density, and then, the Hellinger distance integral is approximated with a summation over those points. In examples, they show the distance is not overly sensitive to these approximations. They propose several cutoff values for the approximation, $\hat{H}$, above which the researcher should conclude the chains have not reached convergence and continue sampling. Their proposed values are motivated by, and demonstrated in, simulation studies where samples are drawn from known distributions in order to mimic MCMC chains. Still, a practitioner may find it difficult to choose a cutoff value for this diagnostic that has some intuitive meaning for specific inferences such as posterior quantiles.

While the comparison of entire distributions is certainly an admirable and ideal situation, it may also prove excessive depending on the inferences that are of interest to the researcher. Further, the Hellinger distance does not provide an indication of *where* two chains may deviate and therefore is not explicitly connected to the ultimate use of the posterior distribution for inference. If estimating posterior quantiles or probabilities is an objective, it makes sense to use a diagnostic that explicitly assesses these quantities, possibly in addition to Boone's or other diagnostics.

1.2.4. Raftery-Lewis

One diagnostic that explicitly addresses posterior quantiles and probabilities is the Raftery-Lewis diagnostic [22]. This diagnostic differs from the preceding diagnostics in a major way. Raftery-Lewis diagnostic does not diagnose convergence *after* sampling, but rather predicts the total number of MCMC samples needed in a single chain to meet a desired precision specified by the user, based on a pilot run of the sampler. Specifically, the proposed number of samples is the number to ensure with probability s that $p(\theta < C|y)$ is found to within $\pm r$, where θ is the parameter of interest, C is a chosen value of θ and y is the data. This calculation is reliant on the pilot run providing a good indication of the 'real' run. Also, as Cowles and Carlin [9] note, diagnostics based on calculating a necessary number of samples may suggest an impractical sampling length, especially for complicated models.

One could use this diagnostic to obtain a preliminary number of samples to draw and supplement with a convergence check on the final samples, perhaps even using the Raftery-Lewis diagnostic again to check if the goals were met or using a different diagnostic. We consider the specificity of the Raftery-Lewis diagnostic to quantiles and the interpretability of the diagnostic to be advantages over the previously discussed diagnostics. With our proposed diagnostic, we aim to keep these advantages while incorporating the power of multiple chains in diagnosing convergence and the framework of equivalence testing. The QED also operates on a relatively interpretable scale, similar to the Raftery-Lewis diagnostic, which contrasts to the Hellinger distance or PSRF convergence measure. Unlike the Raftery-Lewis diagnostic, the QED uses multiple chains, thus providing a more sensitive assessment of convergence as single chain diagnostics may be led to suggest continued sampling is needed by high autocorrelation. The use of multiple chains is shared by the Hellinger distance diagnostic and PSRF.

1.3. *Equivalence testing*

The backbone of the QED is equivalence testing, which is a novel framework within the convergence testing literature. Much of statistical practice is predicated on a search for, or an assessment of, differences. Alternatively, the principle of our QED is to check the degree of consistency among chains. The difference between these two goals may seem subtle, but the key is to recognize that the objective is not to assess the degree to which the chains are different, but rather if they are similar enough to consider them agreeing on the inference in question.

Null hypothesis significance testing (NHST) is a widely implemented statistical technique in which a null hypothesis of equivalence is posed and the strength of the evidence *against* equivalence is assessed. Equivalence testing is often contrasted to NHST as a "shifting of the burden of proof." In equivalence testing, a null hypothesis of a *difference* between two quantities is posed and the strength of evidence for *similarity* is assessed. This framework more accurately depicts how we think about convergence in MCMC. We purposefully start multiple chains from an overly dispersed starting distribution and wait for them to come together. Thus, our default stance should be that the chains are different and we need evidence in order to conclude the chains are similar.

Early work in, and advocation for, equivalence testing was performed by Altman and Bland [23, 24]. Much of the original use of equivalence testing was in the biomedical field where it has been referred to as non-inferiority testing and bioequivalence [25] due to the nature of its application. Now, equivalence testing is used in a variety of applications. Our diagnostic will rely on equivalence testing as the tool for assessing whether multiple chains report values for the estimate of a posterior quantity that are similar enough to conclude convergence.

2. The Quantile Equivalence Diagnostic

Here, we present the theoretical motivation of the QED in a general sense. We follow the general statements with specific calculations for quantiles and probabilities which are our focus for the QED. Diagnostics based on the same theoretical underpinnings but for different posterior summaries are also possible using a direct extension of the methods described here.

2.1. *Motivation*

In much of Bayesian data analysis, posterior distributions are summarized with simple measures such as means, probabilities, and credible intervals. In many cases, the summary value can be expressed as the posterior expectation of some function, g, of a posterior parameter θ; $E(g) = \int g(\theta)p(\theta|y)d\theta$ where $p(\theta|y)$ is the posterior distribution of θ. For example,the posterior mean is calculated with $g(\theta) = \theta$. When we approximate a posterior distribution with samples via MCMC, we approximate the posterior expectation of $g(\theta)$ by calculating the average value of the function over the n MCMC samples,

$$\hat{g}_n = \frac{1}{n}\sum_{i=1}^{n} g(\theta_i) \ , \tag{1}$$

where the notation $\hat{g}_n$ indicates it is an estimate based on n samples. Under the appropriate conditions (see for example Meyn and Tweedie [1]), $\hat{g}_n$ converges to the true

posterior value $E(g)$ as the number of MCMC samples approaches infinity. Researchers may informally state that chains have converged to the posterior distribution to mean that any $\hat{g}_n$ that may be of interest has converged (to at least some tolerance) to $E(g)$; claims that the discrete chain of samples have converged to a (potentially) continuous posterior distribution are difficult to comprehend in any other sense–though Boone's diagnostic perhaps makes the best attempt at honestly assessing such claims. Our approach to convergence is based on the particular $g(\theta)$ of interest, which allows us to perform specific tests.

If a chain has not reached convergence, then given a specific starting value, length of chain (n), and the sampling algorithm of the chain, the expectation of $\hat{g}_n$ over all possible pseudo-random chains is not likely to be the true posterior value, $E(g)$. With m chains indexed by j, we calculate the estimator for each chain, $\hat{g}_n^j$. Unless the chain length is long enough that convergence has been reached, the expectation of $\hat{g}_n^j$ is likely different for each chain given the starting value, length, and sampling algorithm of each chain. The collection of these $\hat{g}_n^j$'s can provide insight to whether the chains are near convergence.

To make a heuristic argument, we consider the possible outcomes for the m observed $\hat{g}_n^j$'s. In the most general cases, the $\hat{g}_n^j$ could be approximately centered about the true quantity $E(g)$ or be biased away from $E(g)$. By using overly dispersed starting values, we hope to avoid the second scenario except potentially in the case that n is so small that all the $\hat{g}_n^j$'s are quite spread out in addition to not being centered on the true value. After an adequate number of samples, then, in addition to being centered at or near the true quantity $E(g)$, the $\hat{g}_n^j$ will occupy a small interval, which, given the overly dispersed starting values, is unlikely to occur unless convergence is reached. Thus, given overly dispersed starting values for the chains, $\hat{g}_n^j$'s that are close together indicate convergence and that the $\hat{g}_n^j$'s are near the true value.

Now, the question becomes in how small of an interval must the $\hat{g}_n^j$'s lie in order for the researcher to be comfortable reporting an estimate from the amalgamation of the chains. There is no single answer to this question as it will depend on the particular definition of $g(\cdot)$ and the precision desired by the researcher. The role of the QED for probability and quantile problems is to take a statement from the researcher about *their* desired precision on an easily interpretable scale and translate that to an interval in which all the $\hat{g}_n^j$'s must be contained in order to meet the specified tolerance level.

2.2. *Specific to probabilities and quantiles*

We now apply the general framework to our main quantities of interest, posterior probabilities and quantiles. In this case, the statement of interest is $p = Pr(\theta < C|y)$. For such a problem, $g(\theta)$ is defined as

$$g(\theta) = \begin{cases} 1 \text{ if } \theta < C \\ 0 \text{ if } \theta \geq C \end{cases} \tag{2}$$

and thus, $p = E(g)$. We estimate p with Equation (1) which is simply the proportion of the MCMC samples that are less than C. This formulation is pertinent for two types of problems. First, C is given and we are looking for p; we refer to these as *probability* problems. An example is reporting the posterior probability that a parameter is less than zero. Second, p is given and we are looking for C; we refer to these as *quantile* problems. An example is reporting the endpoints of a $(1 - p/2)$ posterior credible interval. For *probability* problems, $\hat{g}_n^j$ is calculated with Equations (1) and (2) where the specified value for C is used. Rather than the $\hat{g}_n^j$ notation, we will use $\hat{p}_n^j$. Likewise, we will simply

use p in place of $E(g)$.

For *quantile* problems, there is no specified C to plug into Equation (2). Instead, with the specified p, we calculate $\hat{C}$ as the empirical p^{th} quantile of the MCMC samples from the amalgamation of all the chains, and use this $\hat{C}$ as our C. Then we utilize the strategy described *probability* problems where the objective is to find the posterior probability p associated with this C. Raftery and Lewis [22] use a very similar procedure in the setup of their convergence diagnostic when the quantity of interest is a *quantile*. For both types of problems, the researcher reports a pair of numbers, a probability and a quantile, the only difference being which was specified first. So, *flipping* the problem for *quantile* problems makes sense when we consider the inference to be the pair of numbers.

2.3. *Equivalence test*

With C defined for either type of problem, we can calculate the empirical chain probability $\hat{p}_n^j$ for each of our $j = 1, \ldots, m$ chains. Given our overly dispersed starting positions of the chains, we should not expect the $\hat{p}_n^j$ to be unbiased estimators of p before convergence. Instead, we introduce p_n^j which we call the expected chain probability. This quantity is the unknown expected value of $\hat{p}_n^j$ for chain j given its length n, starting value, and the sampling algorithm. However, if we have sampled long enough, the dependence of p_n^j on the starting value, length of chain, and the MCMC algorithm will disappear and this value will approach the true posterior p.

As n increases, the p_n^j's will become closer to each other as they approach the true posterior probability p. We connect the unknown p_n^j's to the observed $\hat{p}_n^j$'s through a sampling distribution. By saying that $\hat{p}_n^j$ has a sampling distribution, we imply that there is some different $\hat{p}_n^j$ that *could have* (in some sense) been observed. One way to make this sampling distribution tangible is to prepare a set of chains all with the same starting value. Sampling from these chains with different random seeds will produce a distribution of $\hat{p}_n^j$ given that starting value. Another way we might think of this sampling distribution is to investigate the bootstrap distribution of $\hat{p}_n^j$ for a single chain. Here, we might resample with replacement the entire length of the chain and thus produce a bootstrap resample of $\hat{p}_n^j$. This may give an indication of the variability of the procured $\hat{p}_n^j$ and thus suggest the distribution for $\hat{p}_n^j$.

We suggest a sampling distribution for $\hat{p}_n^j$ centered at p_n^j motivated by the normal approximation to the binomial distribution:

$$\hat{p}_n^j \overset{approx}{\sim} N\left(p_n^j, \frac{p_n^j(1-p_n^j)}{n}\right) . \tag{3}$$

Under this distributional assumption, we can perform an equivalence test for each chain. Our strategy to assert that the p_n^j are all close enough is to demonstrate they are each within some ϵ of $\hat{p}$, calculated with Equations (1) and (2) using the amalgamation of the chains. Thus, we test the hypotheses

$$H_0 : |p_n^j - \hat{p}| \geq \epsilon \text{ vs } H_A : |p_n^j - \hat{p}| < \epsilon$$

simultaneously for all m chains. Wellek [26, Chapter 4] specifies the uniformly most powerful test for one parameter problems with normal sampling distributions with known variance. To utilize Wellek's procedure, we must assume a known variance. We suggest approximating the variance in Equation (3) by replacing p_n^j with $\hat{p}$.

With the distributional approximation, we use Wellek's uniformly most powerful test for the equivalence test on each chain. The α level UMP equivalence test rejects the null hypothesis when $|\hat{p}_n^j - \hat{p}|\sqrt{n/(\hat{p}(1-\hat{p}))}$ is less than the square root of the α^{th} quantile from a chi-squared distribution with a single degree of freedom and a non-centrality parameter of $\frac{n}{\hat{p}(1-\hat{p})}\epsilon^2$. We combine the results of the hypothesis test for each of the chains with the intersection-union principle [25]. If, for every chain, the null hypothesis is rejected, the QED suggests the desired precision of ϵ has been obtained and no further sampling is needed. However, if, for *any* chain, the null hypothesis is not rejected, the QED suggests the researcher continue sampling in order to reach convergence at the specified precision.

2.4. *Choosing the desired precision*

Convergence as diagnosed by the QED implies that each of the p_n^j are within ϵ of $\hat{p}$. The QED only accommodates specification of tolerance on the probability scale; there is no ϵ equivalent for the QED on the quantile scale. If the researcher's desired precision can only be expressed on the quantile scale, we recommend careful evaluation of the accompanying plot we discuss below. To come up with a method of choosing a value for ϵ, we investigate the implication of the QED suggesting convergence for a given ϵ. First, consider the m starting values drawn for our chains as a realization of the many possible values that could be drawn from a starting value distribution. Thus, we can consider the m p_n^j's to be a sample of the many possible values that could arise from the many possible chain starting points. Let's assume that the distribution of the expected chain probabilities is centered at p, the true posterior probability, with some standard deviation σ, that is dependent on n. By considering our m chains as a sample of the many chains that could have arisen from the distribution of starting values, we can consider a sampling distribution of the average of the m p_n^j's, $\bar{p_n} = \frac{1}{m}\sum_{j=1}^{m} p_n^j$, that is centered on p with a standard deviation of $\sigma/\sqrt{m}$. Borrowing a multiplier of 2 from elementary frequentist statistics, it is reasonable to expect the average of m p_n^j's to be within $2\sigma/\sqrt{m}$ of p. Though we do not know the p_n^j's, we have the estimators $\hat{p}_n^j$ and it is reasonable to estimate the average of the p_n^j's with $\hat{p}$. We could estimate σ with the standard deviation of the $\hat{p}_n^j$. However, a more conservative strategy is to realize that if the QED suggests convergence, we should believe that each p_n^j falls within an interval of length 2ϵ centered at $\hat{p}$. Under this scenario, the largest possible standard deviation of the p_n^j's would occur if half of the p_n^j are ϵ above $\hat{p}$ and the other half are ϵ below, giving a sample standard deviation of $\sqrt{m\epsilon^2/(m-1)}$. Plugging this worst case scenario in for σ, we should expect that $\hat{p}$ is within $b = 2\epsilon/\sqrt{m-1}$ of p. As a readily interpretable quantity, b can be specified by the researcher and the corresponding ϵ value can be used for the QED. For example, if we are willing for $\hat{p}$ to deviate from p by as much as $b = 0.02$, we can solve for ϵ: $\epsilon = 0.02\sqrt{m-1}/2$. For $m = 5$ chains, we have $\epsilon = 0.02 \cdot 2/2 = 0.02$.

2.5. *Graphical Display*

We recommend using graphical displays to assess convergence in conjunction with diagnostics. If the conditions to use the QED are satisfied (running multiple chains and interest in probabilities or quantiles), the Quantile Equivalence plot (QEplot) may also be employed. This plot does not aim to replace other MCMC diagnostic plots, like a sample path plot, but rather to supplement.

The idea behind the QEplot is to display the variability in the calculated quantile and probability estimates among chains (Figure 1). To display the variability in the m $\hat{p}_n^j$'s,

we plot a point for each chain (x-value of $\hat{p}_n^j$ and a y value of C) creating a set of m pairs $(\hat{p}_n^j,\ C)$. It is this horizontal band of points that goes into the QED. To gain intuition about the variability among the chains on the quantile scale, we also plot a vertical band of points where $\hat{p}$ is the x value for all chains and the y-values are the empirical p^{th} chain quantiles. We label these values $\hat{C}^j$ to give the m pairs $(\hat{p},\ \hat{C}^j)$, plotted as points. The two points from chain j, $(\hat{p}_n^j,\ C)$ and $(\hat{p},\ \hat{C}^j)$, are connected with a line to indicate these points come from the same chain. Finally, the pair $(\hat{p},\ C)$, which is the intersection of the horizontal and vertical bands, is plotted for reference. Further, we have demarcated the approximate range on the x-axis in which the $\hat{p}_n^j$'s must lie in order for the QED to suggest convergence.

One of the objectives of the QED is to provide more intuition behind the decision criteria compared to some other common diagnostics. The QEplot adds to this by clearly displaying the variability of the estimates across the chains. The plot can be used on its own, or simply to understand the conclusion from the QED. If the points on the QEplot are variable enough that the researcher is uncomfortable reporting the estimates, then the researcher should continue sampling the chains regardless of the conclusion of any convergence diagnostic.

3. Examples

Boone et al. [21] provide a number of examples using artificial MCMC chains constructed via independent draws from known distributions to demonstrate weaknesses of the PSRF, as well as a data analysis demonstrating the use of their diagnostic. We provide an additional example using independent draws from known distributions and compare results of the QED to the PSRF and Boone's Hellinger distance diagnostic and also provide an example for a familiar hierarchical Bayesian analysis used in examples for Gelman et al. [27] and Gelman and Hill [28].

3.1. *Clipped Tails*

Our first example aims to simulate a scenario where the tails of a posterior distribution have not been adequately explored by one or more chains. We will use a normal distribution as our target distribution and simulate three chains by simply drawing three sets of independent draws from the normal distribution and then manipulating them. Each chain contains 10500 draws, but we remove the smallest 500 from the first chain, the largest 500 from the second chain, and the smallest and largest 250 each from the third chain for a total of 10000 draws in each chain. Our objective for this example is to create a scenario where the upper tail of one chain has not been explored, the lower tail of another, and both tails of a third (Figure 2 and Table 1).

A glance at the summary statistics in Table 1 shows that if a 95% credible interval was of interest, the intervals reported by each chain differ. For example, the 0.025 quantile varies from about -1.5 to -2.0, which is also clearly seen in the QEplot (Figure 1). Depending on the aims of the researcher, this difference may exceed the tolerance desired and immediately suggest continued sampling. Though the third chain (with both ends clipped) is consistent with the estimates from the amalgamation of the chains, the other two chains are not and the QED will fail to suggest convergence for $\epsilon = 0.01$ and $\alpha = 0.05$ (Table 1). Depending on the inferences of interest and the desired precision, the summary statistics in Table 1 and an investigation of Figures 1 and 2 may convince the researcher that further sampling is required before concluding convergence.

We have assessed convergence for this example with several diagnostics (Table 1). We

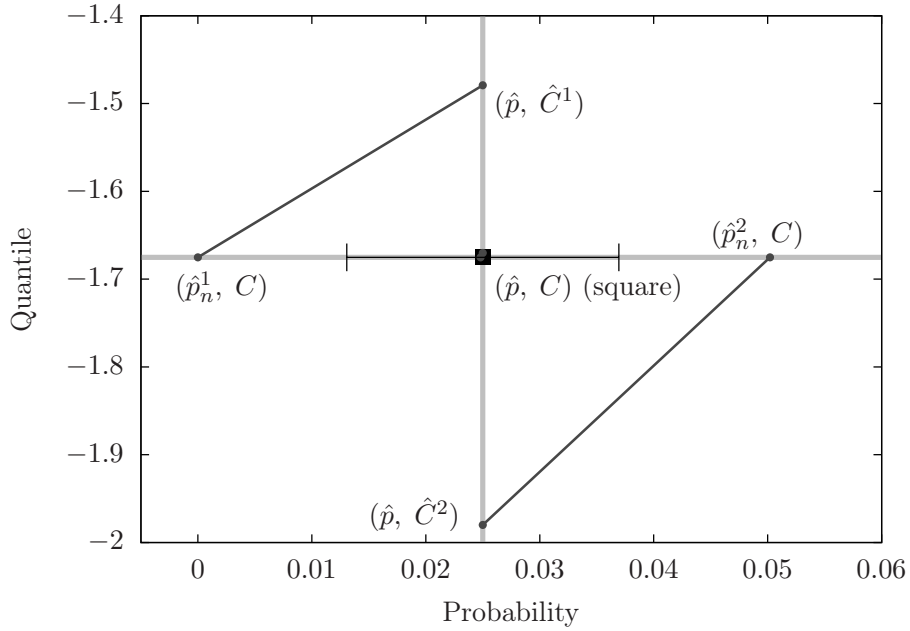

Figure 1. QEplot for the 0.025 quantile of the three simulated clipped chains. Points for chain 3 graphically overlap the central overall value and so are not explicitly labeled. The bars indicate the approximate range in which the estimates must lie in order for the QED to indicate convergence at $\epsilon = 0.01$ and $\alpha = 0.05$.

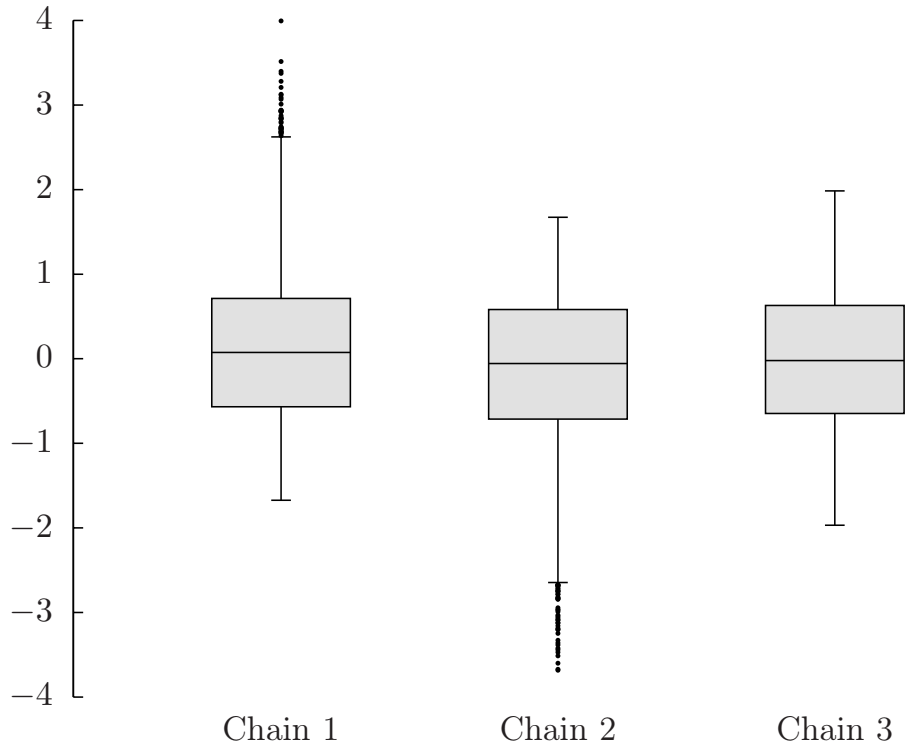

Figure 2. Boxplots of the samples from three simulated chains after removing low, high, and both low and high samples from chains 1, 2, and 3, respectively

used our QED to assess convergence for both the 0.025 and 0.975 quantiles, testing at $\epsilon = 0.01$ (corresponds to $b \approx 0.014$) as well as $\epsilon = 0.05$ (corresponds to $b \approx 0.07$) both with $\alpha = 0.05$. At $\epsilon = 0.01$, both quantiles' QED indicate further sampling is needed. However, if precision to only $b = 0.07$ is required, then the QED does suggest the chains have been run long enough. Of course, $b = 0.07$ may be an exceptionally

Table 1. Summary statistics for each of the three clipped normal distribution chains and also the amalgamation of the chains. Convergence diagnostics are shown in the All Chains row. For the QED, a 0 indicates a failure to suggest convergence and that further sampling is required; a 1 indicates that convergence to the specified tolerance has been obtained at significance level $\alpha = 0.05$.

	2.5%	97.5%	PSRF	BMK	QED 2.5% $\epsilon = 0.01$	QED 97.5% $\epsilon = 0.01$	QED 2.5% $\epsilon = 0.05$	QED 97.5% $\epsilon = 0.05$
Chain 1	-1.48	2.00						
Chain 2	-1.98	1.49						
Chain 3	-1.67	1.67						
All Chains	-1.68	1.68	1.04	0.18	0	0	1	1

generous tolerance when the quantity of interest is a 0.025 or 0.975 quantile. Meanwhile, the PSRF is smaller than the 1.1 often used as a cutoff point in evaluating convergence. Boone's Hellinger diagnostic (labeled BMK in Table 1) was calculated using the BMK R package [29]. The smallest Hellinger distance discrepancy among the pairs of chains is 0.10 and the largest, the value we have reported in the table, is 0.18. Boone provided a set of potential cutoff values above which indicate a lack of convergence; the largest proposed value was 0.1. The reported BMK diagnostic of 0.18 is well above this cutoff as expected with the chain distributions differing considerably as seen in Figure 2.

3.2. *Eight schools example*

In eight different schools, SAT coaching was employed as a means to help students improve scores on the SAT-verbal test. In each of the schools, the estimated coaching effect on exam scores and its associated standard error are reported. The reader may be familiar with this dataset which was used as an example in Gelman and Hill [28] and Gelman et al. [27] and earlier published in Rubin [30] and Alderman and Powers [31], where the details of estimating the coaching effect can be found. This dataset is also used as an example for the Stan [6] computer package for Bayesian inference. Example code is given in Gelman et al. 2014 [27] which fits the following hierarchical model:

$$\begin{aligned} y_j &\sim N(\theta_j, \sigma_j) \\ \theta_j &= \mu + \tau\eta_j \\ \eta_j &\sim N(0,1) \\ p(\tau) &\propto 1 I_{0,\infty}(\tau) \\ p(\mu) &\propto 1 \end{aligned}$$

where $j = 1 \ldots 8$ indexes the schools, y_j is the average observed difference at the j^{th} school, and σ_j is the assumed known standard error of y_j (see Rubin 1981 [30] for details). The school effect was decomposed into τ and η_j to improve convergence. For our example, we narrow our focus to the parameter μ which captures the mean change in SAT scores over all eight schools and consider two potential quantities of interest: a 95% credible interval for μ and the posterior probability that $\mu > 0$ which in context is the posterior probability that coaching is effective across these schools.

We used RStan [15] to sample eight chains of length 20000. We took snapshots of the chains after 500, 1000, 2000, 4000, 10000 and 20000 samples. For each snapshot, we discard the first half of each chain as burn-in, the default in RStan, and then calculate the quantities of interest (Table 2) and calculate the PSRF, BMK, and QED convergence

diagnostics (Table 3). For the PSRF and BMK, a single value is reported corresponding to the μ parameter, and for the QED, a value is reported for each of the posterior summaries of interest (each endpoint of the interval and the posterior probability that $\mu > 0$).

Table 2. Summary values of interest at each snapshot. The chain length is the number of post burin-in samples.

Chain length	0.025	0.975	$Pr(\mu > 0\|y)$
250	-1.40	17.09	0.951
500	-0.87	17.39	0.957
1000	-1.48	17.58	0.953
2000	-1.80	17.58	0.947
5000	-1.83	17.71	0.947
10000	-2.22	18.17	0.943

Table 3. Convergence diagnostics at six snapshots of the chains. The chain length is the number of post burn-in samples and the QED are calculated using $\alpha = 0.05$ and $\epsilon = 0.015$. For the QED, a 0 indicates additional sampling is required, and a 1 indicates that convergence to the specified tolerance has been obtained at significance level $\alpha = 0.05$.

Chain length	PSRF	BMK	QED 0.025	QED 0.975	QED >0
250	1.02	0.10	0	0	0
500	1.00	0.08	0	0	0
1000	1.00	0.06	0	0	0
2000	1.00	0.07	0	0	0
5000	1.01	0.08	1	1	0
10000	1.00	0.04	1	1	1

After just 250 post burn-in samples, the PSRF indicates adequate convergence according to the cutoff of 1.1, the BMK value is at the cusp of the 0.10 cutoff, and the QED does not suggest convergence. By inspecting a QEplot, we can judge whether we are comfortable reporting the single summary quantity based on the variability in the quantity across our chains. In Figure 3, we see how the QEplot changes as the chain length increases. If all chains are in close enough agreement with the single summary value, we may have no hesitation in reporting the value, however if the chain values vary greatly we make the decision to continue sampling before reporting results. In this example, the discrepancy of the interval endpoint values in Table 2 could be explained by long tails where a small amount of posterior probability is spread across a long range of parameter values. If this were the case, we might see in a QEplot where quantile values at a single snapshot take on a large range of values while the associated probabilities are all similar. How large these ranges could be and remain acceptable depends on the context of the research problem. In Figure 4, we plot the evolution of the $\hat{C}^j$'s (the y-values of the QEplots) as the chain lengths increases.

If the researcher does not find the QEplot sufficient for diagnosing convergence, the QED may be calculated for a more automatic decision similar to how the BMK or PSRF are often used in practice. Here, we have used $\epsilon = 0.015$ ($b \approx 0.01$) and $\alpha = 0.05$ for each QED in Table 3. We implore the researcher to consider their own requirements in choosing b or ϵ which need not be the same for all quantities of interest.

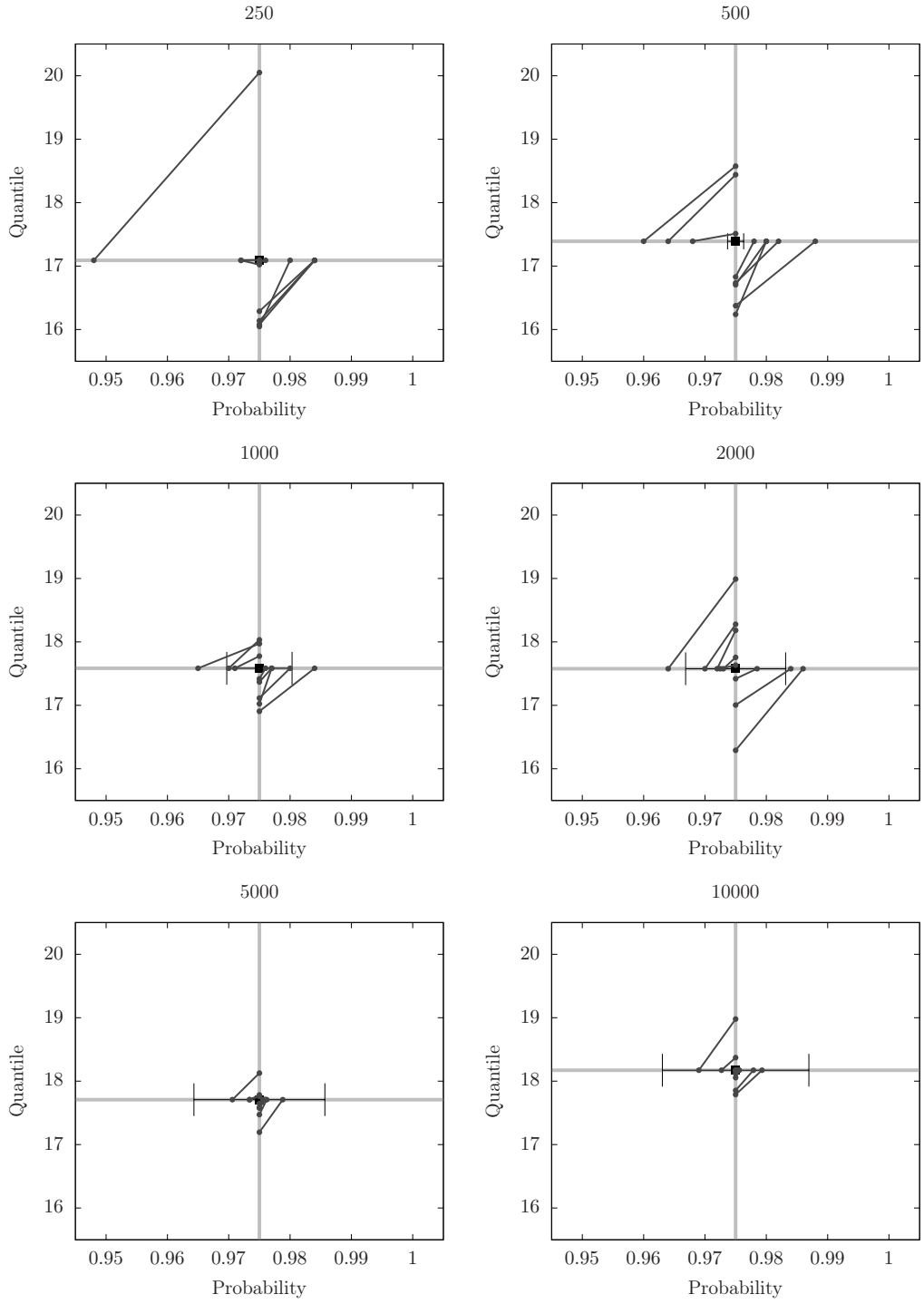

Figure 3. QEplots for the 0.975 quantile for each snapshots. Plot titles indicate the number of post burn-in samples. Plot x- and y-ranges are held fixed across the sequence. Bars indicate the approximate range of values within which each chain estimate must lie in order for the QED to suggest convergence for $\epsilon = 0.015$ and $\alpha = 0.05$.

4. Discussion

In this paper, we motivated and specified a new MCMC convergence diagnostic, the QED and the accompanying QEplot. We provided examples of the use and interpretation of the QED, contrasted the conclusions of the QED to the PSRF and BMK, and identified scenarios where the QED is an improvement over existing diagnostics. Catch

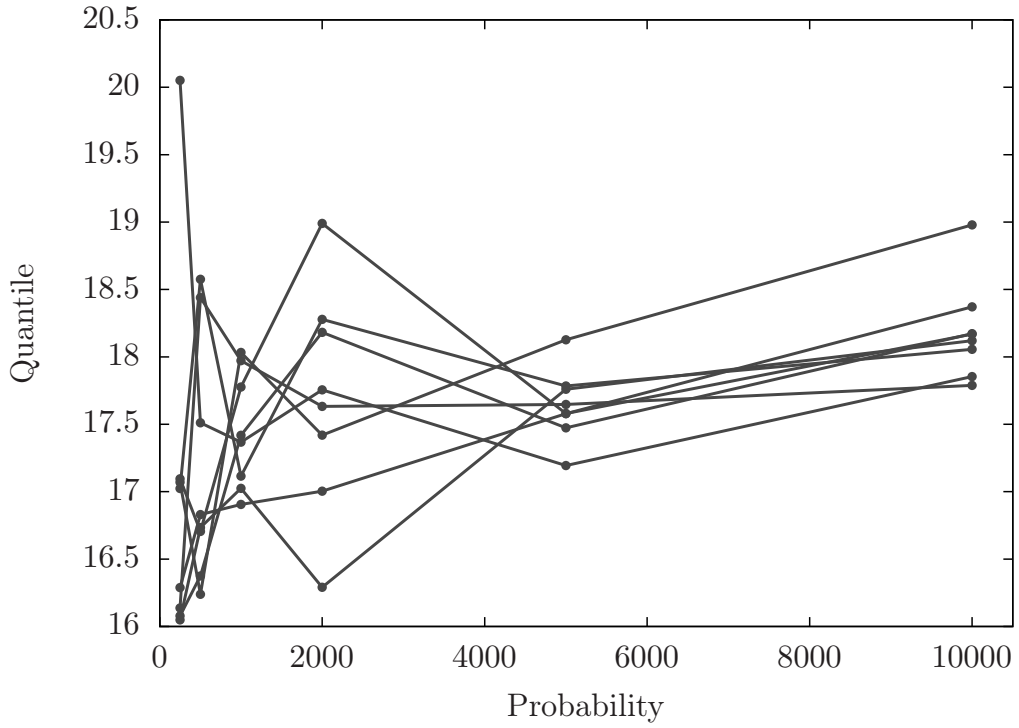

Figure 4. For each chain, the 0.975 quantile is plotted on the y-axis. The x-axis is the number of post burn-in samples for each chain. Quantiles from a single chain are connected with lines.

all convergence tests like the PSRF and BMK may fail to diagnose issues that are of primary importance to the researcher or, similarly, may diagnose issues that are not important to the researcher. The QED specifically addresses issues related to quantiles and probabilities. If obtaining quantiles or probabilities is part of the inferential objective, we advocate the use of the QED and QEplot as important additions to the suite of tools used by researchers to diagnose convergence of MCMC.

Our introduction of equivalence testing to convergence assessment in MCMC is a new contribution to convergence checking. Equivalence testing naturally fits with convergence assessment. For many convergence diagnostics, the strategy is to make comparisons between chains to check for consistency. However, the mechanics of most diagnostics are to quantify the degree of difference between chains. With our diagnostic, we make the step to equivalence testing where we can explicitly check similarity to a desired precision. Further, our diagnostic works on specific quantities, quantiles and probabilities, that are often of interest to researchers.

A valid concern that may be raised against the QED is that as more chains are run, *passing* the intersection-union hypothesis test underlying the QED becomes more difficult. The diagnostic consists of a test for each chain and *fails* as soon as one chain does not pass. Hence, increasing the number of chains increases the number of opportunities the QED has to *fail* the test. This issue is not unique to the QED as decisions made using other diagnostics are also dependent on the number of chains. For the reader concerned with this behavior, we argue that the researcher's objective is to create accurate and precise inferences rather than to pass a convergence diagnostic. Also, in Section 2.4 and the examples, we provided a method of selecting ϵ given the overall tolerance b that scales with the number of chains. Further, we believe the reader should be reassured of this response to more chains rather than concerned. Some practitioners favor a single chain approach to MCMC arguing arguing that a single chain of length nm is more likely to have reached convergence than m chains of length n. Though this is true, we argue for multiple chains, not for improved convergence, but rather improved testing of conver-

gence. Running more chains *should* lead to a more sensitive test. Further, with parallel processing available on most modern personal computers, chains can often be sampled in parallel and researchers may be able to run *many* long chains rather than choosing between a single long chain and multiple short chains.

Finally, the most important message we wish to convey in this paper is that when reporting quantiles or probabilities based on MCMC techniques, the values reported are estimates of the true posterior quantiles. Though this may be an obvious observation, it may be under appreciated. With this understanding, we built our Quantile Equivalence Diagnostic to check for consistency of estimates of quantiles or probabilities across chains which we argue indicates the estimates are near the true value. Between computing the diagnostic and generating the QEplots, we hope researchers will carefully consider the variability of quantile estimates and the magnitude of errors that are acceptable for a problem when reporting values obtained with MCMC techniques. The explicit focus of the QED on the quantities of interest makes this easier.

Acknowledgements

The authors would like to thank Dr. Jim Robison-Cox for reading an early version of this manuscript and providing valuable feedback.

References

[1] Meyn S, Tweedie R. Markov chains and stochastic stability. London: Springer-Verlag; 1993.
[2] Flegal JM, Haran M, Jones GL. Markov Chain Monte Carlo: Can We Trust the Third Significant Figure? Statistical Science. 2008 May;23(2):250–260.
[3] Robert CP, Casella G. Monte Carlo statistical methods. Springer; 1999.
[4] Liu JS. Monte Carlo Strategies in Scientific Computing. Springer; 2001.
[5] Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. CRC press; 2003.
[6] Stan Development Team. Stan: A C++ Library for Probability and Sampling, Version 2.5.0. 2014.
[7] Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS A Bayesian modelling framework: Concepts, structure, and extensibility. Statistics and Computing. 2000;10:325–337.
[8] Plummer M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. 2003.
[9] Cowles M, Carlin B. Markov chain Monte Carlo convergence diagnostics: a comparative review. Journal of the American Statistical Association. 1996;91(434):883–904.
[10] Brooks SP, Roberts GO. Convergence assessment techniques for Markov chain Monte Carlo. Statistics and Computing. 1998;8(4):319–335.
[11] Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. Statistical science. 1992;7(4):457–472.
[12] Schruben LW. Detecting Initialization Bias in Simulation Output. Operations Research. 1982;30(3):569–590.
[13] Schruben L. Confidence Interval Estimation Using Standardized Time Series. Operations Research. 1983;31(6):1090–1108.
[14] Schruben L, Singh H, Tierney L. Optimal Tests for Initialization Bias in Simulation Output. Operations Research. 1983;31(6):1167–1178.
[15] Stan Development Team. RStan: the R interface to Stan, Version 2.5.0. 2014.
[16] Casella G, Lavine M, Robert CP. Explaining the Perfect Sampler. The American Statistician. 2001 Nov;55(4):299–305.
[17] Propp JG, Wilson DB. Exact sampling with coupled Markov chains and applications to statistical mechanics. Random Structures & Algorithms. 1996;9:223–252.
[18] Geyer C. Practical Markov Chain Monte Markov. Statistical Science. 1992;7(4):473–483.

[19] Plummer M, Best N, Cowles MK, Vines K. CODA: Convergence Diagnosis and Output Analysis for MCMC. R News. 2006;6(1):7–11.
[20] Brooks SP, Gelman A. General methods for monitoring convergence of iterative simulations. Journal of computational and graphical statistics. 1998;7(4).
[21] Boone EL, Merrick JRW, Krachey MJ. A Hellinger distance approach to MCMC diagnostics. Journal of Statistical Computation and Simulation. 2014 Apr;84(4):833–849.
[22] Raftery AE, Lewis SM. How many iterations in the Gibbs sampler. Bayesian statistics. 1992; 4(2):763–773.
[23] Altman D, Bland J. Measurement in medicine: the analysis of method comparison studies. The statistician. 1983;32(July 1981):307–317.
[24] Altman DG, Bland JM. Statistics notes: Absence of evidence is not evidence of absence. BMJ. 1995;311(7003):485.
[25] Berger R, Hsu J. Bioequivalence trials, intersection-union tests and equivalence confidence sets. Statistical Science. 1996;11(4):283–302.
[26] Wellek S. Testing statistical hypotheses of equivalence and noninferiority. CRC Press; 2010.
[27] Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. Vol. 3. Taylor & Francis; 2014.
[28] Gelman A, Hill J. Data analysis using regression and multilevel/hierarchical models. Vol. 1. Cambridge University Press New York; 2007.
[29] Krachey M, Boone EL. bmk: MCMC diagnostics package. 2012.
[30] Rubin DB. Estimation in Parallel Randomized Experiments. Journal of Educational and Behavioral Statistics. 1981;6(4):377–401.
[31] Alderman DL, Powers DE. The Effects of Special Preparation on SAT-Verbal Scores. American Educational Research Journal. 1980;17(2):239–251.