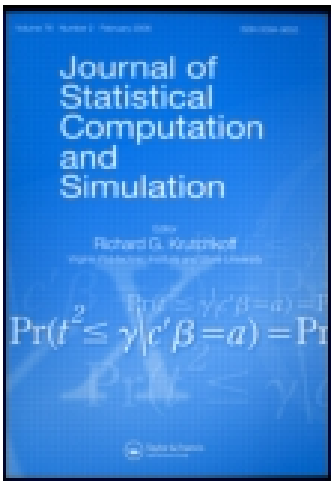

Journal of Statistical Computation and Simulation

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gscs20

A Hellinger distance approach to MCMC diagnostics

Edward L. Boone[a], Jason R.W. Merrick[a] & Matthew J. Krachey[b]

[a] Department of Statistical Sciences and Operations Research, Virginia Commonwealth University, Richmond, VAm USA

[b] Department of Biology, North Carolina State University, Raleigh, NC, USA

Published online: 09 Oct 2012.

To cite this article: Edward L. Boone, Jason R.W. Merrick & Matthew J. Krachey (2014) A Hellinger distance approach to MCMC diagnostics, Journal of Statistical Computation and Simulation, 84:4, 833-849, DOI: 10.1080/00949655.2012.729588

To link to this article: http://dx.doi.org/10.1080/00949655.2012.729588

Journal of Statistical Computation and Simulation, 2014
Vol. 84, No. 4, 833–849, http://dx.doi.org/10.1080/00949655.2012.729588

Taylor & Francis
Taylor & Francis Group

A Hellinger distance approach to MCMC diagnostics

Edward L. Boone[a]*, Jason R.W. Merrick[a] and Matthew J. Krachey[b]

[a]*Department of Statistical Sciences and Operations Research, Virginia Commonwealth University, Richmond, VA, USA;* [b]*Department of Biology, North Carolina State University, Raleigh, NC, USA*

(*Received 9 April 2012; final version received 10 September 2012*)

Bayesian analysis often requires the researcher to employ Markov Chain Monte Carlo (MCMC) techniques to draw samples from a posterior distribution which in turn is used to make inferences. Currently, several approaches to determine convergence of the chain as well as sensitivities of the resulting inferences have been developed. This work develops a Hellinger distance approach to MCMC diagnostics. An approximation to the Hellinger distance between two distributions f and g based on sampling is introduced. This approximation is studied via simulation to determine the accuracy. A criterion for using this Hellinger distance for determining chain convergence is proposed as well as a criterion for sensitivity studies. These criteria are illustrated using a dataset concerning the *Anguilla australis*, an eel native to New Zealand.

Keywords: Hellinger distance; kernel density estimation; Markov chain Monte Carlo; Bayesian robustness

1. Introduction

Markov Chain Monte Carlo (MCMC) techniques are now a permanent part of the Bayesian analysis toolbox. As with any other approximation method, MCMC-based estimates require diagnostics to ensure their validity. In this paper, we examine three such diagnostics, single-chain convergence diagnostics, parallel-chain convergence diagnostics, and robustness diagnostics. Let $\theta_0^i, \theta_1^i, \theta_2^i \ldots$ represent a sample obtained from the ith chain, where θ_0^i is an arbitrary starting value for the chain. In single-chain convergence diagnostics, we seek to determine for a single chain whether the distribution represented by $\theta_w^i, \theta_{w+1}^i, \ldots \theta_{w+b-1}^i$ and the distribution represented by $\theta_{w+b}^i, \theta_{w+b+1}^i, \ldots \theta_{w+2b-1}^i$ are sufficiently similar for given values of w and b to conclude that the ith chain is drawing samples from its stationary distribution. In parallel-chain convergence diagnostics, we seek to determine for two (or more) identical chains with $\theta_0^i \neq \theta_0^j$ whether the distribution represented by $\theta_w^i, \theta_{w+1}^i, \ldots \theta_{w+b-1}^i$ and the distribution represented by $\theta_w^j, \theta_{w+1}^j, \ldots \theta_{w+b-1}^j$ are sufficiently similar to conclude that the ith chain and the jth chain are drawing samples from the same stationary distribution. In robustness diagnostics, we seek to determine for two (or more) chains with different prior distributions whether the distribution represented by $\theta_w^i, \theta_{w+1}^i, \ldots \theta_{w+b-1}^i$ and the distribution represented by $\theta_w^j, \theta_{w+1}^j, \ldots \theta_{w+b-1}^j$ are sufficiently similar to conclude that the different prior distributions have not affected the stationary distribution. In essence, each of these diagnostics deals with the similarity or dissimilarity of distributions represented by batches of samples from chains.

*Corresponding author. Email: elboone@vcu.edu

Early work on convergence of iterative simulations came from the discrete-event simulation literature. Schruben [1,2] and Schruben *et al.* [3] determine appropriate burn-in periods to remove initialization bias in estimates of the mean of simulation output. Diagnostics have also been created specifically for MCMC diagnostics. Gelman and Rubin [4], Brooks and Roberts [5], Brooks and Gelman [5,6], Gelman *et al.* [7] and Gelman and Shirley [8] suggest using the potential scale reduction across parallel chains, estimated by $\hat{R}$. This is given by

$$\hat{R} = \sqrt{\frac{\hat{\text{var}}^{+}(\theta|D)}{W}}, \tag{1}$$

where

$$\hat{\text{var}}^{+}(\theta|D) = \frac{n-1}{n}W + \frac{1}{n}B. \tag{2}$$

with $B = (n/(m-1))\sum_{j=1}^{m}(\bar{\theta}_{.j} - \bar{\theta}_{..})^2$ and $W = (1/m)\sum_{j=1}^{m} s_j^2$, where $s_j^2 = (1/(n-1))\sum_{i=1}^{n}(\theta_{ij} - \bar{\theta}_{.j})^2$ and $\bar{\theta}_{.j} = (1/n)\sum_{i=1}^{n}\theta_{ij}$ and $\bar{\theta}_{..} = (1/m)\sum_{j=1}^{n}\bar{\theta}_{.j}$. Values of $\hat{R}$ near 1 indicate convergence and that further sampling may not improve inferences on θ. Essentially, $\hat{R}$ is monitoring the within- and between-chain variance, and is thus a method based on moments. Geweke [9] proposes a statistic to test convergence when the mean of a function of the samples from a single chain is to be estimated. The statistic is a test on differences between the sample means of successive batches using spectral density estimates of the variance. Jones *et al.* [10], Flegal *et al.* [11] and Flegal and Jones [12] explore the batch means, noting the Geweke batch means are not consistent, and provide the necessary theory for strong consistency batch means as a function of the chain length.

To look beyond the moments of the sample(s), one can also examine a set of quantiles to ensure similarity across several points in the distributions' support. Heidelberger and Welch [13] develop confidence intervals on simulation output quantiles using three methods, one using the spectral density of the process and two using batched means approximations. They then determine the number of iterations necessary to achieve a given level of accuracy. Raftery and Lewis [14] thin a single chain and transform to a binary chain using an indicator function with a given cut-off value. The thinned, binary chain then approximates a two-state Markov chain and the Markov chain theory is used to determine an appropriate burn-in period for a specified level of accuracy in the quantile estimate.

These methods are commonly applied as they can be used with any MCMC procedure and do not require analysis specific to the posterior distribution being simulated. Yu and Mykland [15] also propose such a method using cumulative sum plots used in quality control to test convergence of mean estimates. Roberts [16,17], Ritter and Tanner [18], Zellner and Min [19], Liu *et al.* [20], Garren and Smith [21], Johnson [22], Mykland *et al.* [23] and Yu [24] offer methods that are applicable to certain MCMC methods. Each of these methods involves convergence of either moment or quantile estimates, not the whole distribution. Heidelberger and Welch [25] standardize a single simulation output process using test statistics from time-series analysis that converges to a Brownian bridge process under stationarity assumptions. They then create hypothesis tests on the standardized series using the properties of the Brownian bridge process. Heidelberger and Welch's tests are based on Cramer–von Mises, Anderson–Darling (AD), Kolmogorov–Smirnov (KS) statistics for differences between the distributions of successive batches. Brooks *et al.* [26] also propose a method applying the chi-square and KS tests for differences between the distributions of successive batches, but the method is specific to reversible

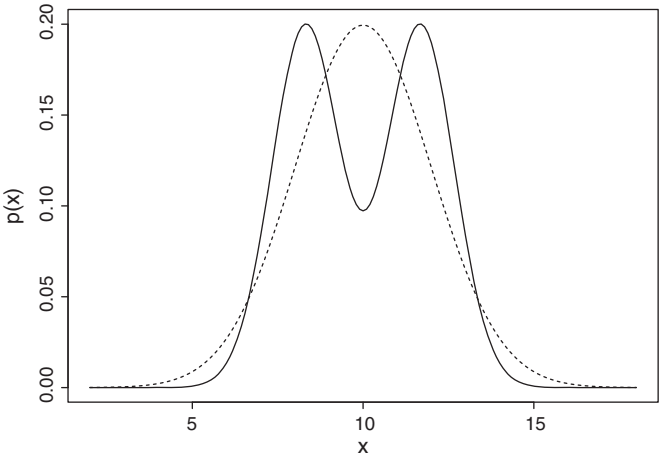

Figure 1. Density plots for $N(10, 2)$ (dashed) and $1/2N(8.32, 1) + 1/2N(11.68, 1)$ (solid).

jump MCMC applications. For a comparison of various MCMC diagnostic methods, see Cowles and Carlin [27].

While we certainly do not propose to replace the existing diagnostic tools, there are situations that they would have trouble diagnosing. Brooks and Roberts [5,6] found that the moment methods from Gelman *et al.* [7] and Geweke [9] and the quantile method from Raftery and Lewis [14] are most widely applied, but that no one tool should be relied upon. Consider the following two distributions $f(x) = N(10, 2)$ and $g(x) = 1/2N(8.32, 1) + 1/2N(11.68, 1)$. Both f and g have mean of 10 and standard deviation of 2. Furthermore, both f and g have the same 2.5%, 50% and 97.5% quantiles. However, simply looking at the plotted densities one can see the discrepancy. Figure 1 shows the plot of f and g. Notice the discrepancy when x is near 10. Such mixed posterior distributions are common in modern Bayesian analysis. Hence, the need for a diagnostic to detect this type of discrepancy where moments and quantiles are equivalent, but the distributions are not. The method must also be applicable to both single-chain and parallel-chain situations and not be limited to specific MCMC algorithms.

A few obvious approaches to considering a non-moment based approach to MCMC diagnostics would be the KS, AD and Kullback–Leibler (KL) based tests. The two sample KS test statistic is given by

$$D_{n_1,n_2} = \sup_x |F_{1,n_1}(x) - F_{2,n_2}(x)|, \tag{3}$$

where F_{1,n_1} and F_{2,n_2} are empirical distribution functions of their respective samples [28]. Brooks *et al.* [29] consider using this for MCMC diagnostics for reversible jump MCMC. The k-sample AD test statistic is given by

$$\text{ADK} = \frac{n-1}{n^2(k-1)} \sum_{i=1}^{k} \left[\frac{1}{n_i} \sum_{j=1}^{L} h_j \frac{(nF_{ij} - n_j H_j)^2}{H_j(n - H_j) - nh_j/4} \right], \tag{4}$$

where h_j is the number of values in the combined samples equal to z_j, H_j is the number of values in the combined samples that are less than z_j plus half of the number of values equal to z_j, F_{ij} is the number of values in group i that are less than z_j plus half of the number of values in group i equal to z_j and $z_1, z_2, \ldots, z_L$ are the unique values in the combined dataset [30]. Hjorth and Vadeby [31]

develop a KL distance approach to MCMC diagnostics:

$$\mathrm{KL}(f, g) = \int \ln\left(\frac{g(x)}{f(x)}\right) g(x)\,\mathrm{d}x. \tag{5}$$

Two major drawbacks to the KL distance are: KL is not symmetric with respect to g and f; magnitude of the KL value has no clear interpretation. Another drawback to all of the above methods is that they are based on the empirical distributions and ties are a complication. In MCMC computations, ties are frequent in cases where there is a low acceptance probability. Using a kernel density estimate instead of the empirical density estimate naturally handles the issues of ties.

We propose a sampling-based estimate of the Hellinger distance between two distributions. For two probability distributions f and g, the Hellinger distance is defined to be

$$H(f, g) = \sqrt{\frac{1}{2}\int (\sqrt{f(x)} - \sqrt{g(x)})^2\,\mathrm{d}x}. \tag{6}$$

This distance or divergence measure is bounded by $0 \leq H(f, g) \leq 1$, where 0 corresponds to no divergence and 1 corresponds to the probability distributions that share no common support. This interpretability makes the measure easily interpretable, as compared with the KL distance, for instance, that is just non-negative, leading to problems interpreting how large a distance is too large. Figure 2 shows the Hellinger distance between two normal distributions as the difference in the mean increases (left) and the difference in the variance increases (right). The Hellinger distance can be used to determine convergence of the whole distribution. Its applicability is not limited to specific MCMC algorithms and our approach does not involve problem- specific analysis. Furthermore, it can be applied to either single-chain or parallel-chain convergence diagnostics, and robustness diagnostics, as we will demonstrate.

The paper is organized as follows. Section 2 states and proves a general theorem and two corollaries that show that the Hellinger distance for our three applications will converge to zero in the limit and further studies the proposed sample-based approximation of the Hellinger distance using a simulation study for accuracy. Section 3 demonstrates how to use the Hellinger distance for MCMC diagnostics. Section 4 provides an example using logistic regression, and Section 5 gives a general discussion of the method and issues that may arise when using it.

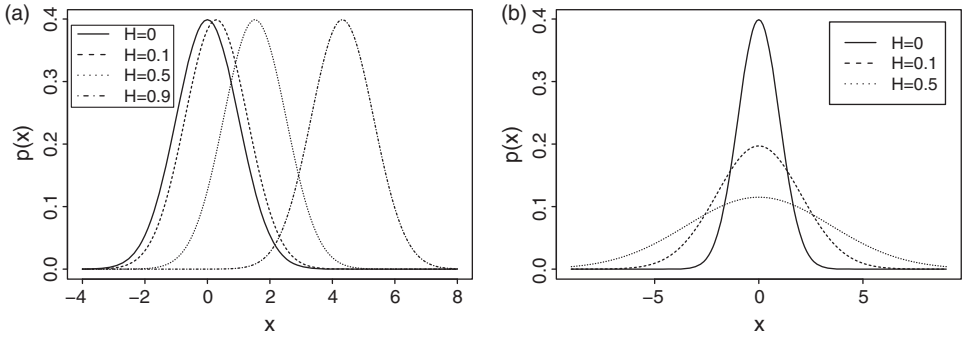

Figure 2. Examples of different normal distributions for various Hellinger distances. (a) Shows the effect of μ on the Hellinger distance. (b) Shows the effect of σ on the Hellinger distance.

2. A Hellinger distance estimate using kernel density estimatess for MCMC diagnostics

Let $x_1, \ldots, x_n$ be a sample from some probability density f. The kernel density estimate of f is given by

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \tag{7}$$

where K is a kernel.

For this work, the Gaussian kernel K is used and is given by

$$K\left(\frac{x - x_i}{h}\right) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-(1/2h^2)(x-x_i)^2}. \tag{8}$$

The bandwidth h is the standard deviation used by the kernel. To select the bandwidth this work uses 'Silverman's Rule of Thumb' [32]

$$h = \frac{0.9 \min(S, \mathrm{IQR})}{1.34} n^{1/5}. \tag{9}$$

For more on kernel density estimation, see Silverman [32].

To estimate the Hellinger distance between densities f and g we can use

$$\begin{aligned} \hat{H}(f, g) &= \left[\frac{1}{2} \int (\sqrt{\hat{f}(x)} - \sqrt{\hat{g}(x)})^2 \, \mathrm{d}x\right]^{1/2} \\ &\approx \left[\frac{1}{2} \sum_{l=1}^{k} (\sqrt{\hat{f}(x_l)} - \sqrt{\hat{g}(x_l)})^2 (x_l - x_{l-1})\right]^{1/2}. \end{aligned} \tag{10}$$

There are two questions that must be answered to use this estimate in MCMC diagnostics. First, if the chain or chains have converged will the Hellinger distance diagnostic tend to zero in each of our applications. Second, is the kernel density estimate of the Hellinger distance $\hat{H}(f, g)$ accurate?

2.1. *Convergence of a sampling-based Hellinger distance for MCMC diagnostics*

If the Hellinger distance is to be a suitable diagnostic tool to test convergence, we must first prove that the Hellinger distance between successive draws from two Markov chains with the same stationary distribution tending to zero. Using the result of Campos and Dorea [33], provided that each Markov Chain has a continuous stationary distribution, the following results hold.

THEOREM *If $f_n \to g$ and $f'_n \to g$ as $n \to \infty$ then $H(f_n, f'_n) \to 0$ as $n \to \infty$.*

Proof By Campos and Dorea [33], as $f_n \to g$ then $H(f_n, g) \to 0$ and as $f'_n \to g$ then $H(f'_n, g) \to 0$. Thus, for all $\epsilon/2 > 0$ there exists a k such that $H(f_n, g) < \epsilon/2$ for all $n > k$ and k' such that $H(f'_n, g) < \epsilon/2$ for all $n > k'$. Hence for all $n > \max\{k, k'\}$ then $H(f_n, g) < \epsilon/2$ and $H(f'_n, g) <$

$\epsilon/2$. This implies $\{\int (f_n^{1/2} - g^{1/2})\}^2 < \epsilon/2$ and $\{\int (f_n'^{1/2} - g^{1/2})\}^2 < \epsilon/2$. Therefore,

$$\left(\int (f_n^{1/2} - g^{1/2})\right)^2 + \left(\int (f_n'^{1/2} - g^{1/2})\right)^2 < \epsilon$$
$$\Rightarrow \left(\int (f_n^{1/2} - g^{1/2}) - \int (f_n'^{1/2} - g^{1/2})\right)^2 < \epsilon$$
$$\Rightarrow \left(\int (f_n^{1/2} - f_n'^{1/2})\right)^2 < \epsilon.$$

This gives $H(f_n, f_n') \to 0$ as $n \to \infty$.

This theorem provides the basic result. More specifically for single-chain diagnostics, if we form batches of size b from a single chain, then the Hellinger distance between two successive batches converges to zero. ■

COROLLARY *If f_n and f_{n-1} are two density estimates from successive batches of size b then $H(f_n, f_n') \to 0$ as $n \to \infty$ where*

$$f_n(x) = \frac{1}{bh} \sum_{k=1}^{b} K\left(\frac{x - x_k}{h}\right).$$

Proof By Campos and Dorea [33] $f_n \to g$ as $n \to \infty$ and $f_{n-1} \to g$ as $n \to \infty$. Then from the above Theorem $H(f_n, f_n') \to 0$ as $n \to \infty$.

Furthermore, if we are drawing samples from two chains that converge to the same stationary distribution then the Hellinger distance between them tends to zero. This result applies to the parallel-chain mixing case as each case converges to the same stationary distribution. It also applies to the robustness case if the perturbations of the prior distribution do not affect the posterior (stationary) distribution. ■

THEOREM *If $f_n \to g$ and $g_n \to g$ as $n \to \infty$, then $H(f_n, g_n) \to 0$ as $n \to \infty$ where*

$$f_n(x) = \frac{1}{nh} \sum_{k=1}^{n} K\left(\frac{x - x_k}{h}\right) \tag{11}$$

and

$$g_n(y) = \frac{1}{nh} \sum_{k=1}^{n} K\left(\frac{y - y_k}{h}\right). \tag{12}$$

Proof By Campos and Dorea [33] $f_n, g_n \to g$ as $n \to \infty$ and from the above Theorem $H(f_n, g_n) \to 0$ as $n \to \infty$.

Given these results, we can expect the Hellinger distance to converge to zero in each diagnostic scenario. Now we must show that $\hat{H}(f, g)$ is a suitable estimate of $H(f, g)$. ■

2.2. *Accuracy of the Hellinger distance approximation using kernel density estimates*

When applying a method based on asymptotic results, it is often wise to determine the performance of the method under practical conditions. The accuracy of the estimated Hellinger distance $\hat{H}(f, g)$ may be affected by the size of the set of posterior samples, n_s, from f and g as well as the number

of partitions k used to evaluate the integral. A simulation study varying the sample size and the partition width was performed. Sample sizes of $n_s = 100$, 1000, 10,000, and 25,000 were taken from each distribution f and g. Furthermore, the number of partitions $k = 512$ and 1000 were used. For each sample size, n_s, and the number of partitions, k, 1000 datasets each from $f(x) = N(\mu, 1)$ and $g(x) = N(0, 1)$ were generated and the estimated Hellinger distance $\hat{H}(f, g)$ was computed as given in Equation (10). The Hellinger distance between two normal distributions $f(x) = N_1(\mu_1, \sigma_1^2)$ and $g(x) = N_2(\mu_2, \sigma_2^2)$ is given by

$$H(f,g) = \left(1 - \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}}\, e^{-(1/4)((\mu_1-\mu_2)^2/(\sigma_1^2+\sigma_2^2))}\right)^{1/2}. \tag{13}$$

The estimate and true Hellinger distances were computed for $\mu = 0, 0.5, 1, 2$, and 4. The mean and standard deviation were computed for each n_s, k, and μ combination and are presented in Table 1. One can note from Table 1 that the estimated Hellinger distance is quite accurate for sample sizes n_s of 1000 or greater. Furthermore, it does not appear that the number of partitions k has much effect on the accuracy since the values for $k = 512$ and $k = 1000$ are almost identical. In addition, notice that as n_s gets large for all values μ that $\hat{H}(f, g)$ is close to $H(f, g)$ and $H(f, g)$ the $\hat{H}(f, g)$ values are quite close to 0.

To further study the accuracy of the estimated Hellinger distance, a similar simulation study was conducted. In this study the estimated Hellinger distance $\hat{H}(f, g)$ of $f(x) = N(\mu, 1)$ from $g(x) = N(0, 1)$ used 1000 simulated datasets where μ varied from 0 to 10. For each μ value, the true Hellinger distance $H(f, g)$ and the estimated Hellinger distance $\hat{H}(f, g)$ were computed. The results of this study are given in Figure 3. Notice that the mean estimated Hellinger distance is almost indistinguishable from the true Hellinger distance. The only exception is when μ is near 0. In this case, the estimated Hellinger distance is larger than the true Hellinger distance. This represents the error due to estimating the density and error induced by sampling from f. Notice that this error is small.

Many decisions will be based on determining if two distributions f and g are similar based on samples from their respective distribution. Hence the interest is sampling variability when the distributions are the same. One can note from Table 1 that the accuracy of the estimated Hellinger distance $\hat{H}(f, g)$ is better for larger sample sizes n_s when $\mu = 0$ which is when the distributions are the same. To further study the sampling error in $\hat{H}(f, g)$ when $\mu = 0$, a simulation study was conducted using 1000 simulated datasets across values of n_s and the standard deviation of $\hat{H}(f, g)$ was computed. Since the number of partition points k did not seem to have an effect on the results, we chose $k = 512$. Figure 4 shows the standard deviation of the estimated Hellinger

Table 1. The mean (standard deviation) of the estimated Hellinger distance $\hat{H}(f, g)$ of $f(x) = N(\mu, 1)$ to $g(x) = N(0, 1)$ using 1000 simulated data sets across sample size n_s and number of partition points k for $\mu = 0$, 0.5, 1, 2, and 4.

k	n_s	$\mu = 0$	$\mu = 0.5$	$\mu = 1$	$\mu = 2$	$\mu = 4$
512	100	0.094(0.029)	0.186(0.045)	0.333(0.046)	0.611(0.043)	0.922(0.025)
	1000	0.043(0.009)	0.176(0.015)	0.339(0.014)	0.620(0.014)	0.929(0.009)
	10,000	0.019(0.003)	0.175(0.005)	0.340(0.005)	0.624(0.004)	0.929(0.003)
	25,000	0.013(0.002)	0.175(0.003)	0.341(0.003)	0.624(0.003)	0.929(0.002)
1000	100	0.094(0.029)	0.186(0.043)	0.334(0.046)	0.611(0.043)	0.922(0.024)
	1000	0.043(0.009)	0.176(0.015)	0.339(0.015)	0.620(0.014)	0.929(0.009)
	10,000	0.019(0.003)	0.175(0.005)	0.340(0.005)	0.624(0.004)	0.929(0.003)
	25,000	0.013(0.002)	0.175(0.003)	0.341(0.003)	0.625(0.003)	0.929(0.002)
True $H(f, g)$		**0**	**0.175**	**0.343**	**0.627**	**0.929**

Note: True value of $H(f, g)$ is given in the bottom row.

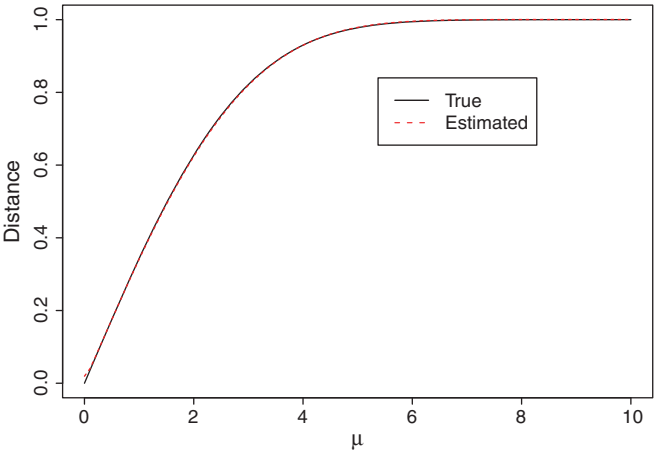

Figure 3. Simulation of mean estimated Hellinger distance $\hat{H}(f,g)$ (dashed) to true Hellinger distance $H(f,g)$ (solid) for $f(x) = N(\mu, 1)$ to $g(x) = N(0,1)$. Mean estimated Hellinger distance is based on 1000 simulations.

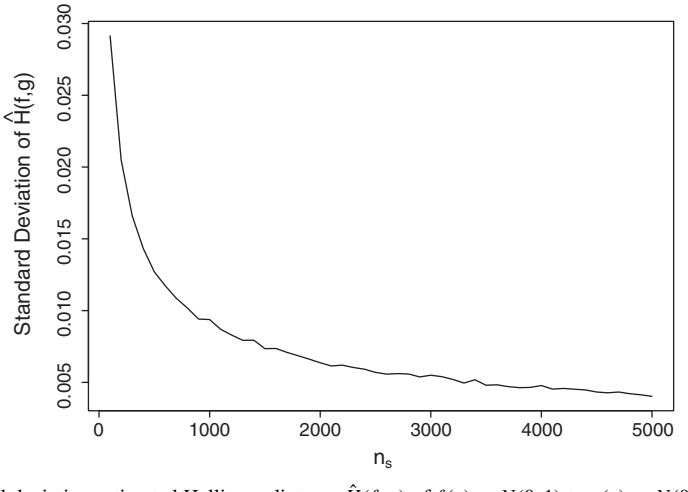

Figure 4. Standard deviation estimated Hellinger distance $\hat{H}(f,g)$ of $f(x) = N(0,1)$ to $g(x) = N(0,1)$ versus number of samples n_s. Standard deviation of estimated Hellinger distance based on 1000 simulations.

distance $\hat{H}(f,g)$ of $f(x) = N(0,1)$ to $g(x) = N(0,1)$ versus number of samples n_s. Notice that the standard deviation reduces dramatically to near 0 around 2500 samples.

Another question of interest is the sample size n_s needed to determine if two distributions f and g are dissimilar. To determine if the distributions are dissimilar, a cut-off value that leads to a the decision of dissimilarity is needed. Based on the simulations above, candidate cut-off values could be 0.05, 0.075 , and 0.1. Hence if the estimated Hellinger distance $\hat{H}(f,g)$ is less than C, then we conclude that the two distributions are similar, otherwise they are dissimilar. A simulation study was performed to determine the estimated probability of misclassifying the two distributions as dissimilar when they are actually similar. This study utilized 1000 simulated data sets from $f(x) = N(0,1)$ and $g(x) = N(0,1)$ and for each dataset pair $\hat{H}(f,g)$ was calculated and for the proportion of datasets where the conclusion was dissimilar. This study was performed across various sample sizes n_s. Figure 5 shows that the estimated probability of deciding $f(x) = N(0,1)$ to $g(x) = N(0,1)$ is dissimilar using cut-off values $C = 0.05, 0.075$ and 0.1 versus n_s. Notice that when $n_s > 2500$ the probability of concluding dissimilarity is near 0 for all cut-off values. For the

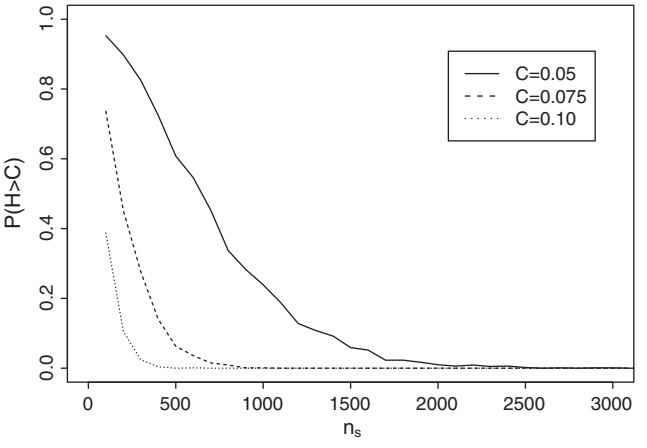

Figure 5. Estimated probability of deciding $f(x) = N(0,1)$ to $g(x) = N(0,1)$ is dissimilar using cut-off values $C = 0.05$, 0.075 and 0.1 versus n_s. Each estimated probability is based on 1000 simulations.

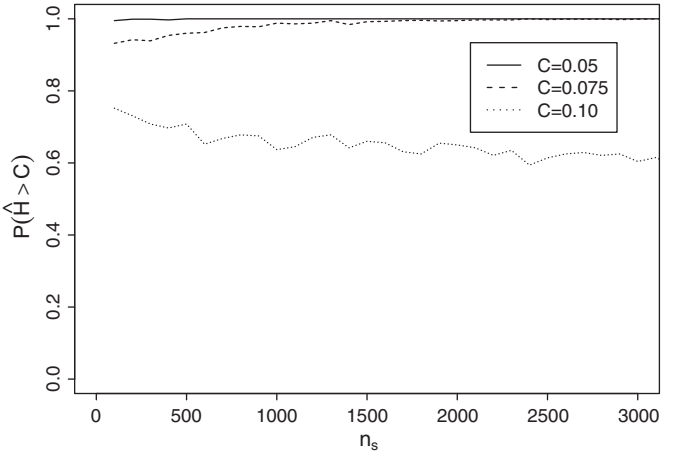

Figure 6. Estimated probability of deciding $f(x) = N(0.2835,1)$ to $g(x) = N(0,1)$ are dissimilar using cut-off values $C = 0.05$, 0.075 and 0.1 versus n_s. Each estimated probability is based on 10,000 simulations.

$C = 0.1$, $n_s > 500$ appears to be large enough so that the probability of concluding dissimilarity is near 0. For the $C = 0.075$, $n_s > 500$ appears to be large enough so that the probability of concluding dissimilarity is near 0.

Researchers also may be interested in the ability of the decision rule to determine a dissimilarity when one exists. A $H(f,g)$ value was chosen to be 0.1 which corresponds to $\mu = 0.2835$. To study this, a simulation study was conducted comparing $\hat{H}(f,g)$ for $f(x) = N(0.2835,1)$ to $g(x) = N(0,1)$ across values of n_s. For each value of n_s, 1000 datasets were simulated and $\hat{H}(f,g)$ was calculated and the proportion of datasets where dissimilarity was concluded was determined. This was performed for $C = 0.05$, 0.075, and 0.1. Figure 6 shows the plot of the estimated probability of detecting dissimilarity against n_s. Notice for $C = 0.05, 0.075$, $n_s > 1000$ seemed to be enough to detect the difference. However, due to the true underlying Hellinger distance being 0.1, using a cut-off of $C = 0.1$ does not have a good ability to determine the dissimilarity.

To study how $\hat{H}(f,g)$ compares with the KS and AD tests, a simulation study was performed. This study compared the $\hat{H}(f,g)$, KS and AD for $f(x) = N(\mu,1)$ to $g(x) = N(0,1)$ for $\mu = 0, 0.01, 0.05, 0.1, 0.5, 1, 5, 10$. Table 2 shows the mean and standard deviation of $\hat{H}$, KS power,

Table 2. Comparison of mean(standard deviation) of $\hat{H}$, KS power, and AD power for $f(x) = N(\mu, 1)$ to $g(x) = N(0, 1)$. Results are based on 1000 simulated data sets $n_s = 10,000$ and $k = 512$. KS and AD powers are based on a significance level of 0.05.

μ	$\hat{H}$ Mean (SD)	KS	AD
0	0.019(0.003)	0.049	0.052
0.01	0.020(0.003)	0.087	0.099
0.05	0.025(0.004)	0.867	0.942
0.1	0.040(0.005)	1	1
0.5	0.186(0.045)	1	1
1	0.333(0.046)	1	1
5	0.979(0.002)	1	1
10	1.000(0.001)	1	1

and AD power for $f(x) = N(\mu, 1)$ to $g(x) = N(0, 1)$ for the various values of μ. Results are based on 1000 simulated data sets $n_s = 10,000$ and $k = 512$. KS and AD power is based on a significance level of 0.05. Notice that for $\mu = 0.05$ gives $\hat{H}(f, g) = 0.0259$ and the power for the KS test is 0.867 and the power for the AD test is 0.942. This suggests that both the KS and AD tests have high power, which may be too powerful for MCMC diagnostics considering the small shift of $\mu = 0.05$. Notice that for large shifts of $\mu > 0.5$ all of the proposed methods agree.

3. Hellinger distance-based MCMC diagnostics

3.1. *Single-chain similarity diagnostics*

Using the estimated Hellinger distance, one can create various diagnostics for MCMC samples. One cannot determine if MCMC chains have truly converged. However, one can determine if the chain is internally similar. This can be done by splitting the set of samples into an early set of samples $\theta_E^{(i)}$ where $i \in I_E$ and a later set of samples $\theta_L^{(j)}$ where $j \in I_L$ where I_E and I_L are index sets such that $\max(I_E) < \min(I_L)$. Then the estimated Hellinger distance can be computed between the early and later set of samples. If the estimated Hellinger distance is near 0 then the chain can be deemed internally similar, otherwise the chain is not internally similar. If the chain is not internally similar then the chain is not suitable for posterior inferences. This can also be used to determine how many samples need to be discarded as 'burn-in' samples.

Consider two MCMC sample chains given in Figure 7, the chain depicted in (a) is converged and the chain depicted in (b) converges at some point in the sample. Notice the chain in (b) needs to have more samples discarded as 'burn-in' samples. If you divide the samples in half and compare the first 5000 samples to the second 5000 samples, for the chain in (a) $\hat{H}(f, g) = 0.019$ which is small for the number of samples used. Both the KS and AD tests agree with the internal similarity with the p-value for the KS test of 0.627 and p-value for the AD test of 0.423. For the chain in Figure 7 (b) $\hat{H}(f, g) = 0.228$, which is sufficient to indicate that the chain is not internally similar. The KS and AD tests agree that the chain is not internally similar with both tests producing p-values less than 0.0001. To determine how many samples should be discarded, the chain was divided into 10 batches of 1000 samples each. For each successive batch, $\hat{H}(f, g)$ was calculated. This results in batches after 3000 samples had $\hat{H}(f, g) < 0.05$, indicating that the first 3000 samples should be discarded as 'burn-in' samples. The KS and AD tests agree with this result by producing p-values > 0.05 for tests on sequential batches of 1000 after the first 3000 samples.

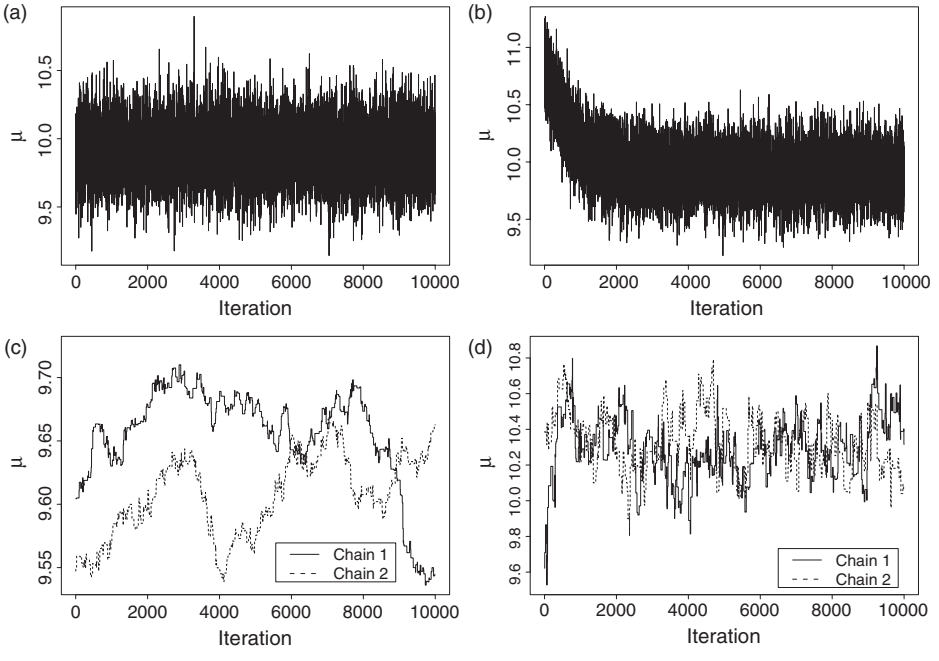

Figure 7. A converged MCMC sample chain (a) a nonconverged MCMC sample chain (b) a well-mixed MCMC sample chain (c), and a poor mixing MCMC sample chain (d).

3.2. *Parallel-chain similarity diagnostics*

Another question of interest is whether multiple chains 'mix well'. This is the concept that the separate chains are covering the same space and hence are not converged to separate localized regions. Figure 7 shows two plots: one where there is poor mixing (d) and one with good mixing (c). The estimated Hellinger distance can be used to determine the discrepancy between the two chains. If the estimated Hellinger distance is near 0 then the chains are mixing well, otherwise the chains are not mixing well.

Consider again the simple Bayesian analysis on a mean μ where $x_1, x_2, \ldots, x_{100} \sim N(\mu, \sigma^2)$ with prior distribution $\mu \sim N(0, 100)$ and $\sigma^2 \sim \text{Inv} - \chi^2(1, 1)$. Three chains of 10,000 were pulled from the Gibbs sampler and $\hat{R}$ and $\max(\hat{H}(f, g))$ was computed. This resulted in $\hat{R} = 1.001$ and $\max(\hat{H}(f, g)) = 0.022$, each indicating that chains are mixing well and that the chains are similar to each other.

Now consider the two distributions $f(x) = N(10, 2)$ and $g(x) = 1/2N(8.32, 1) + 1/2N(11.68, 1)$. If two chains of 10,000 are taken where chain 1 is taken from f and chain 2 is taken from g and $\hat{R}$ and $\hat{H}(f, g)$ are computed. This results in $\hat{R} = 0.999$ and $\hat{H}(f, g) = 0.156$. Since $\hat{R}$ depends on moments of the data, it is unable to discriminate between the two distributions. Whereas $\hat{H}(f, g)$ does not depend on moments and can determine a discrepancy in the two chains. Similarly, the KS and AD tests give p-values less than 0.0001, indicating that the two distributions are different.

In addition to being able to detect odd features of posterior distributions, the estimated Hellinger distance can also help to diagnose 'sticky' MCMC chains. A 'sticky' chain is one that infrequently moves to a new state, i.e. becomes 'stuck' in a current state. A 'sticky' chain often results from a Metropolis–Hastings sampler in which the acceptance probability to move to a proposed sample is very low. Figure 8 shows the histograms of three MCMC chains and their combined trace plot. Notice that the histograms look different and that the trace plot is 'sticky'. In this case $\hat{R} = 1.001$,

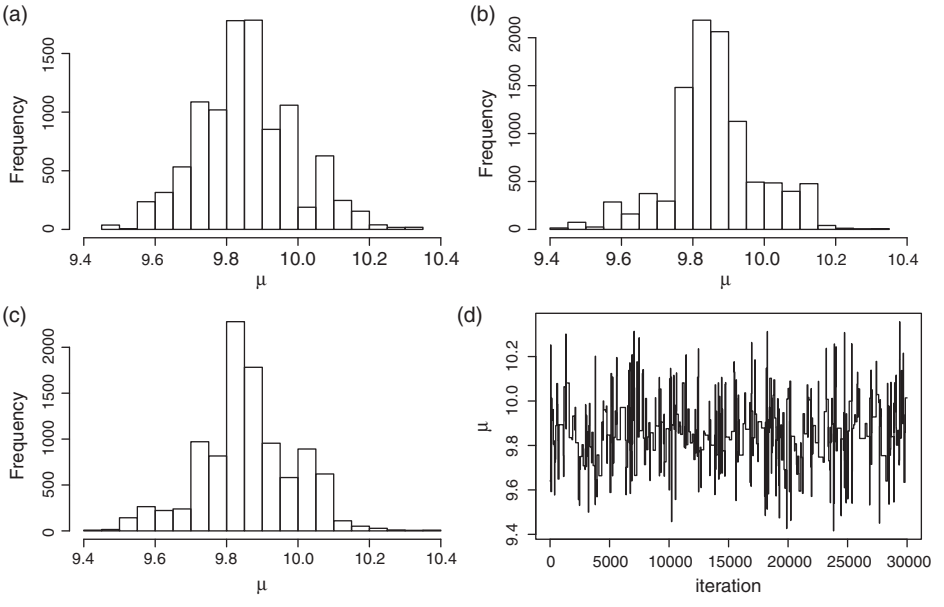

Figure 8. Histograms of 3 MCMC chains (a), (b), (c) and traceplot of the MCMC chains combined (d).

indicating no further sampling is necessary. However, the $\max(\hat{H}(f, g)) = 0.18$ indicating that the three chains are not similar. Furthermore, both the AD test and KS test give p-values smaller than 0.0001 indicating that the three chains are not similar.

3.3. *Robustness diagnostics*

To determine how sensitive inferences are to prior distribution specifications, often researcher simply compare the moments or the 2.5%, 50% and 97.5% quantiles under the two different prior distributions. While this can capture some of the possible discrepancies in posterior inferences induced by prior distribution specification. Recall the two distributions $f(x) = N(10, 2)$ and $g(x) = 1/2N(8.32, 1) + 1/2N(11.68, 1)$. Both f and g have mean of 10 and standard deviation of 2. Furthermore, both f and g have the same 2.5%, 50% and 97.5% quantiles. However, simply looking at the plotted densities one can see the discrepancy. Figure 1 shows the plot of f and g. In this case, $\hat{H}(f, g) = 0.159$ which indicates that the distributions are not similar even though any analysis of the moments or 2.5%, 50% and 97.5% quantiles would suggest that the distributions are similar.

Suppose that one wants to determine the sensitivity of posterior inferences based on two different prior distribution specifications $\pi_1(\theta)$ and $\pi_2(\theta)$. Let $p_1(\theta|D)$ be the posterior distribution for θ using prior $\pi_1(\theta)$ and $p_2(\theta|D)$ be the posterior distribution using prior $\pi_2(\theta)$. If one takes samples $\theta_1^{(i)} \sim p_1(\theta|D)$ and $\theta_2^{(i)} \sim p_2(\theta|D)$ then the estimated Hellinger distance between $p_1(\theta|D)$ and $p_2(\theta|D)$, $\hat{H}(p_1(\theta|D), p_2(\theta|D))$ captures the discrepancy due to $\pi_1(\theta)$ and $\pi_2(\theta)$.

Consider a simple Bayesian analysis on a mean μ where $x_1, x_2, \ldots, x_{100} \sim N(\mu, \sigma^2)$ with prior distribution $\mu \sim N(0, \tau^2)$ and $\sigma^2 \sim \text{Inv} - \chi^2(1, 1)$ where τ^2 is the prior variance. Using 10,000 samples from the Gibbs sampler the mean standard deviation, 2.5%, 50% and 97.5% quantiles were computed for μ when $\tau^2 = 1$, 10, 20, 30, 40, 50, and 100. Table 2 shows the results from this sensitivity study as well as $\hat{H}(f, g)$, AD and KS tests calculated for each posterior distribution against the case when $\tau^2 = 100$. Notice the mean is sensitive as well as the quantiles. However, using only the mean and quantiles it is difficult to determine how close the results actually are.

Table 3. Example sensitivity study for posterior inferences on a mean μ for prior variances $\tau^2 = 1, 10, 20, 30, 40, 50,$ and 100. The mean, standard deviation, quantiles, $\hat{H}$, and p-values for the KS and AD tests.

τ^2	Mean	Standard deviation	(2.5%, 50%, 97.5%)	$\hat{H}$	AD	KS
1	9.406	0.198	(8.985, 9.410, 9.802)	0.573	<0.0001	<0.0001
10	9.748	0.206	(9.360, 9.748, 10.134)	0.072	<0.0001	<0.0001
20	9.772	0.196	(9.383, 9.771, 10.156)	0.037	<0.0001	<0.0001
30	9.775	0.196	(9.392, 9.774, 10.166)	0.033	<0.0001	0.0001
40	9.779	0.196	(9.390, 9.778, 10.165)	0.026	0.003	0.031
50	9.783	0.197	(9.401, 9.784, 10.168)	0.023	0.193	0.281
100	**9.787**	**0.199**	(**9.392**, **9.788**, **10.175**)			

Note: Here $\hat{H}$, AD and KS tests compare each prior variance specification with the model with prior variance $\tau^2 = 100$.

The estimated Hellinger distance $\hat{H}(f, g)$ shows that when $\tau^2 = 10$ and $\tau^2 = 100$, the results are quite close with a value of 0.072 with both the KS and AD tests producing p-values less than 0.0001. In contrast, $\hat{H}(f, g)$ shows that when $\tau^2 = 1$ and $\tau^2 = 100$, the results are quite different with a value of 0.573 with both KS and AD tests producing p-values less than 0.0001. One interesting scenario is when $\tau^2 = 30$, which gives $\hat{H}(f, g) = 0.033$ and where both KS and AD produce p-values less than 0.001. Hence, based on the KS and AD tests one would conclude that the posterior distributions based on $\tau^2 = 30$ and $\tau^2 = 100$ are different. However, looking at the estimated Hellinger distance of 0.03 one may conclude that the posterior distributions are similar. Furthermore, looking at the moments and the quantiles one could conclude that the posterior distributions are similar. This simple example shows how $\hat{H}(f, g)$ compares with traditional sensitivity analysis (Table 3).

4. Example

In order to demonstrate the use of Hellinger distance for MCMC diagnostics, we take an example of a Bayesian logistic regression. We used the testing data for the presence or absence of the freshwater eel *Anguilla australis* provided in the 'dismo' R package (see [34,35]). This dataset is a subset of 1000 observations from a New Zealand survey of site-level presence-absence of *A. australis* [36]. We selected six covariates, five continuous variables: the average daily summer air temperature (SegSumT), distance in km to the coast (DSDist), proportion of area with native vegetation (USNative), the maximum downstream slope (DSSlope), and average catchment slope (DSSlope), and one categorical variable corresponding to the type of fishing method used at the survey location: electricrofishing (electric), spotlighting (Spo), trapping (Trap), fyke nets (Net), or a combination of fishing methods (Mixture) [36]

Let x_i be the regression vector of covariates for the ith observation of length k. For the presence–absence of *A. australis* let $Y_i = 1$ if present and $Y_i = 0$ if absent for the ith observation. The Bayesian logistic regression model is given by

$$Y_i \sim \text{Bernoulli}(p_i),$$
$$p_i \sim \frac{\exp(x_i'\boldsymbol{\beta})}{1 + \exp(x_i'\boldsymbol{\beta})},$$
$$\boldsymbol{\beta} \sim N(\mathbf{0}, \sigma_\beta^2 \mathbf{I}_k),$$

where $\mathbf{I}_k$ is the $k \times k$ identity matrix. For the analyses, $\sigma_\beta^2 = 100$ was chosen to represent a diffuse prior distribution on $\boldsymbol{\beta}$. A sensitivity study is conducted to determine the impact of this specification on the resulting posterior distributions.

Table 4. Sample summaries and chain diagnostics for each of the parameters in the *A. australis* dataset.

Variable	(0.025%, 0.5%, 0.975%)	$\max(\hat{H})_W$	$\max(\hat{H})_B$	$\hat{R}$	n_{eff}	AD_B
Intercept	(−13.511, −10.674, −7.995)	0.057	0.034	1.000	14,563	<0.0001
SegSumT	(−12.720, 0.618, 0.815)	0.053	0.031	1.000	31,487	<0.0001
DSDist	(−0.007, −0.004, −0.0008)	0.053	0.041	1.000	2,644	<0.0001
USNative	(−1.885, −1.178, −0.480)	0.053	0.033	1.000	10,102	<0.0001
MethodMixture	(−1.164, −0.479, 0.213)	0.048	0.042	1.000	67,529	<0.0001
MethodNet	(−2.315, −1.528, −0.793)	0.052	0.039	1.000	10,307	<0.0001
MethodSpo	(−3.169, −1.826, −0.737)	0.048	0.046	1.000	7,673	<0.0001
MethodTrap	(−3.752, −2.600, −1.642)	0.054	0.037	1.000	22,203	<0.0001
DSMaxSlope	(−0.286, −0.167, −0.062)	0.052	0.036	1.000	42,901	<0.0001
USSlope	(−0.089, −0.052, −0.015)	0.051	0.036	1.000	6,274	<0.0001

Note: Sample summaries consist of the 2.5%, 50% and 97.5% quantiles of the posterior samples. Chain diagnostics consist of maximum within-chain Hellinger distance, $\max(\hat{H})_W$, the maximum between-chain Hellinger distance, $\max(\hat{H})_B$, the Gelman–Rubin statistic, $\hat{R}$, number of effective samples, n_{eff}, and the between chain p-values from the AD test, AD_B. Results based on 210,000 samples comprising three chains of 100,000 with 70,000 retained after a burn-in of 30,000 samples.

Table 5. Sensitivity of logistic regression parameters for the *A. australis* dataset to variations in the specification of prior variance σ_β^2.

Variable	$\sigma_\beta^2 = 50$	$\sigma_\beta^2 = 25$	$\sigma_\beta^2 = 10$	$\sigma_\beta^2 = 5$
Intercept	0.027	0.022	0.055	0.202
SegSumT	0.027	0.025	0.054	0.199
DSDist	0.022	0.024	0.024	0.043
USNative	0.022	0.022	0.019	0.029
MethodMixture	0.022	0.027	0.026	0.051
MethodNet	0.027	0.025	0.031	0.042
MethodSpo	0.021	0.024	0.022	0.031
MethodTrap	0.026	0.027	0.024	0.047
DSMaxSlope	0.023	0.024	0.021	0.023
USSlope	0.025	0.026	0.022	0.025

Note: All values represent the Hellinger distance of the posterior samples associated with $\sigma_\beta^2 = 5$, 10, 25, and 50 compared with the samples associated with $\sigma_\beta^2 = 100$. Results based on 210,000 samples comprised three chains of 100,000 with 70,000 retained after a burn-in of 30,000 samples.

The model was run in R using the MCMCpack package. Three chains of 100,000 samples were obtained. Each parameter of the three chains was divided into batches of 10,000 samples and the $\hat{H}(f, g)$ was computed for each batch sequentially with the maximum $\hat{H}(f, g)$ across the chains reported as $\hat{H}_W$. Table 4 shows the sample summaries as well as various diagnostics for the chains. The $\hat{H}_W$ shows that the first 30,000 samples should be discarded as burn-in samples since $\hat{H}(f, g)$ was larger than 0.05. Table 4 shows the sample summaries as well as various diagnostics for the chains. This example illustrates the need to have a measure of discrepancy among chains. Notice that if one were to use $\hat{R}$ then all the chains would be deemed to have converged and no further sampling is needed. However, if one were to compare the posterior distributions using the AD test between the chains, AD_B, one would conclude that the chains have not converged. While not shown in the table, the KS test was also conducted and produced results similar to the AD_B test. Using the maximum $\hat{H}(f, g)$ metric between chains, $\max(\hat{H})_B$, one can see that all of the chains have minor discrepancies with the largest $\max(\hat{H})_B$ distance being 0.046.

To determine the sensitivity of the prior variances to the posterior results, the model above was fit using prior variances of $\sigma_\beta^2 = 5$, 10, 25, 50, and 100. Three chains of 100,000 were run with 30,000 discarded as burn-in leaving 70,000 samples per chain. All inferences are made using the 210,000 samples retained. The Hellinger distance $\hat{H}(f, g)$ was computed for each set of

posterior samples. Each of $\sigma_\beta^2 = 5$, 10, 25, and 50 were compared with the samples obtained for $\sigma_\beta^2 = 100$. Table 5 shows the results of these comparisons for each of the parameters. Notice that the Intercept and SegSumT become sensitive to the prior variance at $\sigma_\beta^2 = 10$. DSDist, Mixture, Net, and Trap become sensitive to the specification of prior variance at $\sigma_\beta^2 = 5$. And Slope, DSMaxSlope, and USNative appear to be insensitive the prior variance at all of the above prior parameter specifications. Since none of the parameters are sensitive near $\sigma_\beta^2 = 100$, our choice of prior distribution is robust to these specifications. The p-values from AD and KS tests (not shown) were all less than 0.0001. This would indicate that the posterior distributions are different across all the values of σ_β^2 considered.

5. Discussion

We have introduced an MCMC diagnostic tool that measures the Hellinger distance between kernel density estimates from two samples. This diagnostic has been demonstrated to test single-chain convergence, parallel-chain convergence, and robustness. Existing tools based on moments or quantiles are effective for many situations, but we have demonstrated examples where traditional tools do not reveal problems and the Hellinger distance approximation does. The approximation is not computationally expensive and, thus, can be used alongside traditional tools. For our empirical testing of Hellinger distance approximations using kernel density estimates, we have chosen the familiar Gaussian kernel estimate. This sampling-based approach to Hellinger distance between two distributions gives researchers a measure by which different types of questions can be addressed. In addition this work considers MCMC chain diagnostics, these could easily be transferred to other sampling-based techniques such as importance sampling, resampling, or acceptance sampling.

One big issue with kernel density estimation is minimizing the 'edge effect' of density estimates near a boundary in the parameter space. For example, when estimating a variance σ^2 the value must always be above zero, which is the bound on σ^2. If the distribution for σ^2 is concentrated near zero then the density estimate for σ^2, depending on the kernel used, may give positive probability to negative values. To study the effects of 'edge effects' on our proposed method, a simulation study was conducted. Sample sizes of $n_s = 100$, 1000, 10,000, and 25,000 were taken from each distribution f and g. Furthermore, the number of partitions $k = 512$ and 1000 were used. For each sample size, n_s, and number of partitions, k, 1000 datasets each from $f(x) = \text{Exp}(\alpha)$ and $g(x) = \text{Exp}(1)$ were generated and the estimated Hellinger distance $\hat{H}(f, g)$ was computed as given in Equation (10). The true Hellinger distance between two exponential distributions

Table 6. The mean(standard deviation) of the estimated Hellinger distance $\hat{H}(f, g)$ of $f(x) = \text{Exp}(\alpha)$ to $g(x) = \text{Exp}(1)$ using 1000 simulated data sets across sample size n_s and number of partition points k for $\alpha = 1$, 2, and 10.

k	n_s	$\alpha = 1$	$\alpha = 2$	$\alpha = 10$
512	100	0.125(0.033)	0.282(0.051)	0.697(0.028)
	1000	0.054(0.009)	0.261(0.016)	0.685(0.010)
	10,000	0.025(0.003)	0.249(0.005)	0.676(0.004)
	25,000	0.017(0.002)	0.246(0.003)	0.679(0.007)
1000	100	0.116(0.033)	0.284(0.052)	0.697(0.029)
	1000	0.054(0.009)	0.262(0.015)	0.688(0.010)
	10,000	0.025(0.003)	0.249(0.005)	0.676(0.003)
	25,000	0.017(0.002)	0.246(0.003)	0.673(0.002)
True $H(f, g)$		**0**	**0.057**	**0.425**

Note: True value of $H(f, g)$ is given on the bottom row.

$f(x) = \text{Exp}(\alpha)$ and $g(x) = \text{Exp}(\beta)$ is given by

$$H(f,g) = 1 - \frac{2\sqrt{\alpha\beta}}{\alpha+\beta}. \tag{14}$$

The estimated and true Hellinger distances were computed for $\alpha = 1$, 2, and 10. The mean and standard deviation were computed for each n_s, k, and α combination and are presented in Table 6. Notice from Table 6 that the estimated Hellinger distance is quite accurate for sample sizes n_s of 1000 or greater. Furthermore, it does not appear that the number of partitions k has much effect on the accuracy since the values for $k = 512$ and $k = 1000$ are almost identical. Also notice that as n_s gets large for all values μ that $\hat{H}(f,g)$ is close to $H(f,g)$ and $H(f,g)$ the $\hat{H}(f,g)$ values are quite close to 0.

One should note that poor choice of kernel K and bandwidth h can lead to oversmoothed or undersmoothed density estimates and hence affect the accuracy of the estimated Hellinger distance. When available optimal bandwidth or near optimal bandwidth should be chosen to reduce the effects of over or undersmoothing. For more on kernel density estimation, see Silverman [32].

References

[1] L.W. Schruben, *Detecting initialization Bias in simulation output*, Oper. Res. 30(3) (1982), pp. 569–590.
[2] L.W. Schruben, *Confidence interval estimate from standardized simulation output*, Oper. Res. 31 (1983), pp. 1090–1108.
[3] L.W. Schruben, H. Singh, and L. Tierney, *Optimal tests for initialization bias in simulation output*, Oper. Res. 31(6) (1983), pp. 1167–1178.
[4] A. Gelman and D.B. Rubin, *Inference from iterative simulation using multiple sequences*, Statist. Sci. 7 (1992), pp. 457–472.
[5] S.P. Brooks and G.O. Roberts, *Convergence assessment techniques for Markov chain Monte Carlo*, Statist. Comput. 8 (1998), pp. 319–335.
[6] S.P. Brooks and A. Gelman, *Alternative methods for monitoring convergence of iterative simulations*, J. Comput. Graphical Statist. 7 (1998), pp. 434–455.
[7] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin, *Bayesian Data Analysis*, 2nd ed., Chapman & Hall/CRC, Boca Raton, FL, 2004.
[8] A. Gelman and K. Shirley, *Inference from simulations and monitoring convergence*, in *Handbook of Markov Chain Monte Carlo*, S. Brooks, A. Gelman, G. Jones, and X.L. Meng, eds., CRC Press, Boca Raton, FL, 2011, pp. 163–174.
[9] J. Geweke, *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*, in *Bayesian Statistics 4*, J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, eds., Oxford University Press, Oxford, 1992, pp. 169–193.
[10] G.L. Jones, M. Haran, B.S. Caffo, and R. Neath, *Fixed-width output analysis for Markov Chain Monte Carlo*, J. Amer. Statist. Assoc. 101(476) (2006), pp. 1537–1547.
[11] J.M. Flegal, M. Haran, and G.L. Jones, *Markov Chain Monte Carlo: Can we trust the third significant figure?*, Statist. Sci. 23 (2008), pp. 250–260.
[12] J.M. Flegal and G.L. Jones, *Batch means and spectral variance estimators in Markov chain Monte Carlo*, Ann. Statist. 38(2) (2010), pp. 1034–1070.
[13] P. Heidelberger and P.D. Welch, *Quantile estimation in dependent sequences*, Oper. Res. 32(1) (1984), pp. 185–209.
[14] A.E. Raftery and S.M. Lewis, *How many iterations in the Gibbs sampler?*, in *Bayesian Statistics 4*, J.M. Bernardo, A.F. M. Smith, A.P. Dawid, and J.O. Berger, eds., Oxford University Press, Oxford, 1992, 763–773.
[15] B. Yu and P.A. Mykland, *Looking at Markov samplers through CUSUM path plots: A simple diagnostic idea*, Statist. Comput. 8 (1998), pp. 275–286.
[16] G.O. Roberts, *Convergence diagnostics of the Gibbs sampler*, in *Bayesian Statistics 4*, J.M. Bernardo, A.F.M. Smith, A.P. Dawid, and J.O. Berger, eds., Oxford University Press, Oxford, 1992, pp. 775–782.
[17] G.O. Roberts, *Methods for estimating L2 convergence of Markov chain Monte Carlo*, in *Bayesian Statistics and Econometrics: Essays in Honor of Arnold Zellner*, D.A. Berry, K.M. Chaloner, and J.K. Geweke, eds., North Holland, Amsterdam, 1994, 373–383.
[18] C. Ritter and M.A. Tanner, *Facilitating the Gibbs sampler: The Gibbs stopper and the griddy Gibbs sampler*, J. Amer. Statist. Assoc. 87 (1992), pp. 861–868.
[19] A. Zellner and C. Min, *Gibbs sampler convergence criteria*, J. Amer. Statist. Assoc. 90 (1995), pp. 921–927.
[20] J.S. Liu, W.H. Wong, and A. Kong, *Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes*, Biometrika 81 (1994), pp. 27–40.
[21] S.T. Garren and R.L. Smith, *Convergence diagnostics for Markov chain samplers*, Tech. Rep., Department of Statistics, University of North Carolina, 1993.

[22] V.E. Johnson, *Studying convergence of Markov chain Monte Carlo Algorithms using coupled sample paths*, J. Amer. Statist. Assoc. 91 (1996), pp. 154–166.
[23] P. Mykland, L. Tierney, and B. Yu, *Regeneration in Markov chain samplers*, J. Amer. Statist. Assoc. 90 (1995), pp. 233–241.
[24] B. Yu, *Estimating the L1 error of kernel estimators based on Markov samplers*, Tech. Rep. 409, Statistics Department, University of California, Berkeley, 1994.
[25] P. Heidelberger and P.D. Welch, *Simulation run length control in the presence of an initial transient*, Oper. Res. 31(6) (1983), pp. 1109–1144.
[26] S.P. Brooks, P. Giudici, and G.O. Roberts, *Efficient construction of reversible Jump Markov Chain Monte Carlo proposal distributions*, J. Roy. Statist. Soc. Ser. B 65 (2003), pp. 3–55.
[27] M.K. Cowles and B.P. Carlin, *Markov Chain Monte Carlo convergence diagnostics: A comparative review*, J. Amer. Statist. Assoc. 91 (1996), pp. 883–904.
[28] W.J. Conover, *Practical Nonparametric Statistics*, John Wiley & Sons, New York, 1971.
[29] S.P. Brooks, P. Giudici, and A. Philippe, *Nonparametric convergence assessment for MCMC model selection*, J. Comput. Graph. Statist. 12 (2003), pp. 1–22.
[30] F.W. Scholz and M.A. Stephens, *K-sample Anderson–Darling tests*, J. Amer. Statist. Assoc. 82 (1987), pp. 918–924.
[31] U. Hjorth and A. Vadeby, *Subsample distribution distance and MCMC convergence*, Scand. J. Statist. 32 (2005), pp. 313–326.
[32] B.W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London, 1986.
[33] V.S.M. Campos and C.C.Y. Dorea, *Kernel estimation for stationary density of Markov chains with general state space*, Ann. Inst. Statist. Math. 57 (2005), pp. 443–453.
[34] J. Elith, J.R. Leathwick, and T. Hastie, *A working guide to boosted regression trees*, J. Appl. Ecol. 77(4) (2008), pp. 802–813.
[35] R.J. Hijmans, S. Phillips, J. Leathwick, and J. Elith, dismo: Species distribution modeling. R package version 0.7-17, 2012. Available at http://CRAN.R-project.org/package=dismo
[36] J.R. Leathwick, J. Elith, W.L. Chadderton, D. Rowe, and T. Hastie, *Dispersal, disturbance and the contrasting biogeographies of New Zealand's diadromous and non-diadromous fish species*, J. Biogeogr. 35 (2008), pp. 1481–1497.