MCMC diagnostics for higher dimensions using Kullback Leibler divergence

Anand Dixit and Vivekananda Roy
Department of Statistics, Iowa State University

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

In the existing literature of MCMC diagnostics, we have identified two areas for improvement. Firstly, the density based diagnostic tools currently available in the literature are not equipped to assess the joint convergence of multiple variables. Secondly, in case of multi-modal target distribution if the MCMC sampler gets stuck in one of the modes, then the current diagnostic tools may falsely detect convergence. The Tool 1 proposed in this article makes use of adaptive kernel density estimation, symmetric Kullback Leibler (KL) divergence and a testing of hypothesis framework to assess the joint convergence of multiple variables. In cases where Tool 1 detects divergence of multiple chains, started at distinct initial values, we propose a visualization tool that can help to investigate reasons behind their divergence. The Tool 2 proposed in this article makes a novel use of the target distribution (known up till the unknown normalizing constant), to detect divergence when an MCMC sampler gets stuck in one of the modes of a multi-modal target distribution. The usefulness of the tools proposed in this article is illustrated using a multi-modal distribution, a mixture of bivariate normal distribution and a Bayesian logit model example.

Keywords: Adaptive kernel density estimation; Convergence diagnostics; Kullback Leibler divergence; MCMC; Monte Carlo.

1. Introduction

The process of simulating observations from a fully specified distribution is generally carried out using a traditional technique like inversion sampling. But if the distribution is analytically intractable, then there may not be any efficient methods for direct simulations from it, and in such cases, one often relies on Markov Chain Monte Carlo (MCMC) sampling techniques. Specifically, in Bayesian analysis, we often come across situations where the posterior distribution of parameters of interest is only known up to some unknown normalizing constant. In such situations, one often uses MCMC algorithms to produce approximate observations from the analytically intractable posterior distributions to make inference about the parameters of interest.

MCMC samplers are iterative in nature and require a starting observation. If \( \pi(\theta) \) is the analytically intractable target distribution from which we wish to simulate observations, then MCMC samplers like the Gibbs sampler and the Metropolis Hasting sampler construct a Markov chain \( \{ \theta_n : n = 0, 1, 2, \cdots \} \) started at \( \theta_0 \) such that the stationary distribution of the chain is equal to \( \pi(\theta) \). Thus, as \( n \to \infty \), the distribution of \( \theta_n \) converges to \( \pi(\theta) \). In the Bayesian inference framework, the target distribution is the posterior distribution of the unknown parameters \( \theta \) given the data \( y \) and is commonly

*Corresponding author. Email: adixit@iastate.edu
denoted as $\pi(\theta | y)$. Hence if the stationary distribution exists and is unique, then at some point the MCMC sampler will start producing approximate observations from the target distribution. Generally, deep mathematical analysis is needed to establish quantitative convergence bounds for determining the sample size ($n$) required for the Markov chain to be sufficiently close to the target distribution (see e.g. Rosenthal [1], Jones and Hobert [2], Roy and Hobert [3]). In the absence of such theoretical analysis, often empirical diagnostic tools are used to check the convergence of MCMC samplers.

In the early 90’s there was an interesting debate on whether one should use multiple chains or a single long chain to diagnose convergence. Gelman and Rubin [4] and Brooks and Gelman [5] advocated the usage of multiple chains, while Raftery and Lewis [6] and Geweke [7] believed a single long chain is sufficient for assessing convergence. Gelman and Rubin [4] used the potential scale reduction factor (PSRF) to monitor convergence in the univariate case. Suppose we are working with $m$ chains and each chain has $n$ iterations. Let $\{\theta_{ij} : i = 1, 2, \ldots, m \text{ and } j = 1, 2, \ldots, n\}$ be the observations generated from the $m$ chains. Then the PSRF, denoted by $\hat{R}$ is defined as,

$$\hat{R} = \frac{\hat{V}}{W}, \quad (1)$$

where $\hat{V} = ((n−1)/n)W + (1 + (1/m))(B/n)$ is the pooled variance estimate,

$$B/n = \frac{\sum_{i=1}^{m} (\bar{\theta}_i - \bar{\theta})^2}{m−1}$$

is the between chain variance estimate,

$$W = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} (\theta_{ij} - \bar{\theta}_i)^2}{m(n−1)}$$

is the within chain variance estimate and $\bar{\theta}_i$ and $\bar{\theta}$ are the $i$th chain mean and the overall mean respectively where $i = 1, 2, \ldots, m$.

Brooks and Gelman [5] came up with the multivariate PSRF (MPSRF) to diagnose convergence in the multivariate case. It is denoted by $\hat{R}_p$ and is given by,

$$\hat{R}_p = \max_a \frac{a^T \hat{V}^* a}{a^T W^* a} = \frac{n−1}{n} + \left(1 + \frac{1}{m}\right) \lambda_1, \quad (2)$$

where $\hat{V}^*$ is the pooled covariance matrix, $W^*$ is the within chain covariance matrix, $B^*$ is the between chain covariance matrix and $\lambda_1$ is the largest eigenvalue of the matrix $(W^*^{-1}B^*)/n$. In this diagnostic tool, convergence is detected when $\hat{R} \approx 1$ in the univariate case and $\hat{R}_p \approx 1$ in the multivariate case. Raftery and Lewis [6] proposed a univariate diagnostic tool based on the specified level of accuracy desired by the user in quantile estimation. In this tool, the chain obtained from the sampler is used to find the number of initial iterations that should be discarded (burn-in period), how long the chain should be run after the burn-in and how should the chain be thinned in order to obtain the desired level of accuracy in quantile estimation. Geweke [7] also proposed a univariate diagnostic tool in which he used a test statistic, to compare the mean of a function of the samples from two non overlapping parts of the chain. Usually, the choice is the first 10% of the chain and the last 50% of the chain. Thus, Gelman and Rubin [4], Brooks and Gelman [5], Raftery and Lewis [6] and Geweke [7] are all moment based diagnostic tools.

More recently, researchers have come up with density based diagnostic tools. Boone et al. [8] and Hjorth and Vadeby [9] used divergence measures to come up with univariate
diagnostic tools. Boone et al. [8] estimated the Hellinger distance between the kernel density estimates of two chains or two parts of a single chain. If the estimated distance was close to zero (i.e. less than 0.10), then the Markov chains were said to have converged else not. Hjorth and Vadeby [9] used a measure imitating KL divergence to compare the empirical distributions of subsequences of chains to the empirical distribution of the whole chain, and in the case of multiple chains, they compared empirical distribution of individual chains to the empirical distribution of the combination of all chains.

The density based diagnostic tools mentioned before are univariate tools and cannot assess the convergence of multiple variables jointly. The Tool 1 proposed in this article, computes the adaptive kernel density estimate of the joint distribution of each multivariate chain, and then compares the estimated symmetric KL divergence between them to a cut-off value, to assess convergence. Since the adaptive kernel density estimation suffers from the curse of dimensionality, for higher dimensions, Tool 1 monitors convergence marginally i.e. one variable at a time, and since we determine the cut-off values for KL divergence measure using a testing of hypothesis framework, they can be easily adjusted for multiple comparison. Thus, Tool 1 is a density based diagnostic tool that can assess convergence of multiple variables jointly. Other notable differences between Tool 1 and the density based diagnostic tools mentioned before are, Boone et al. [8] used numerical integration to compute the estimated Hellinger distance and Hjorth and Vadeby [9] compute differences between empirical distribution functions over a partition of the real line, while we provide a Monte Carlo estimate of the symmetric KL divergence to compare the adaptive kernel density estimates of multiple chains.

If the target distribution is multi-modal, then the MCMC chain might get stuck in one of the modes. In such cases, even if the MCMC sampler is run for a reasonably long time, it continues to produce observations around that mode. Many of the current diagnostic tools only make use of the iterations obtained from the MCMC samplers to diagnose convergence, and so, in cases where the chain gets stuck in one of the modes, they get fooled into thinking that the target distribution is unimodal, and hence, falsely detect convergence. Most of the times, we do not know a priori how many modes the target distribution has, and hence, even if we use multiple chains, there is a chance that all chains might be stuck at the same mode. In order to overcome this difficulty, we propose Tool 2 that makes a novel use of the target distribution known up till the unknown normalizing constant in the diagnostic tool. Yu [10] proposed a tool which also incorporates the target distribution, in which she estimated the unknown normalizing constant using the MCMC samples, and then estimated the $L^1$ distance between the kernel density estimate of the chain and the estimated target distribution over a compact set, where the difference between the two was most likely. But if the MCMC sampler is stuck in a particular mode, then the normalizing constant estimator proposed by Yu [10] is no longer reliable.

Many other diagnostic tools are available in the literature, and a very nice review of them can be found in Brooks and Roberts [11]. Our tools can be used when either a single chain or multiple chain samplers are available. This article is structured as follows. In Section 2, we provide definition and certain properties of KL divergence. In Section 3, we propose two new MCMC convergence diagnostic tools and a visualization tool. In Section 4, we provide three examples to illustrate the usefulness of the proposed diagnostic and visualization tools. Some concluding remarks are given in Section 5.
2. **Kullback Leibler Divergence**

KL divergence is a measure used to calculate the difference between two probability distributions. If $P(\theta)$ and $Q(\theta)$ are any two probability density functions on $\Theta \subseteq \mathbb{R}^d$, then the KL divergence between $P(\theta)$ and $Q(\theta)$ is defined as,

$$KL(P|Q) = \int_{\Theta} \log \left( \frac{P(\theta)}{Q(\theta)} \right) P(\theta)d\theta. \quad (3)$$

Some important properties of the KL Divergence are as follows,

- $KL(P|Q) \geq 0$,
- $KL(P|Q) = 0$ iff $P = Q$ almost everywhere wrt the Lebesgue measure, and
- KL divergence is not symmetric in $P$ and $Q$.

The KL divergence is not symmetric because $KL(P|Q)$ is the expected difference between log of densities $P$ and $Q$ with respect to $P$ while $KL(Q|P)$ is the expected difference between log of densities $Q$ and $P$ with respect to $Q$. The symmetric KL divergence between $P$ and $Q$, denoted by $KL_{sy}(P,Q)$ is given as,

$$KL_{sy}(P,Q) = \frac{KL(P|Q) + KL(Q|P)}{2}. \quad (4)$$

3. **Diagnostic tools**

3.1. **Tool 1**

Let $\pi(\theta)$ be the target distribution where $\theta \in \Theta \subseteq \mathbb{R}^d$. In order to explore the target distribution, two chains are initiated at different starting points and each chain produces $n$ observations. As prescribed in Gelman and Rubin [4], the starting points should be over dispersed with respect to the target distribution. Let $\{\theta_{ij} : i = 1, 2 \text{ and } j = 1, 2, ..., n\}$ be the $n$ observations obtained from each of the two chains where $\theta_{ij} \in \Theta \subseteq \mathbb{R}^d \forall i \text{ and } \forall j$. The adaptive kernel density estimates of observations obtained from the two chains are denoted by $P_{1n}(\theta)$ and $P_{2n}(\theta)$ and are found by substituting $i = 1$ and $i = 2$ respectively in the following equation,

$$P_{in}(\theta) = \frac{1}{n} \sum_{j=1}^{n} \prod_{k=1}^{d} \frac{1}{h_{ij}^{(k)}} K \left( \frac{\theta_{ij}^{(k)} - \theta_{ij}^{(k)}}{h_{ij}^{(k)}} \right), \quad (5)$$

where,

- $\theta_{ij}^{(k)}$ denotes the $k^{th}$ dimension in the $j^{th}$ observation of the $i^{th}$ chain, where $i = 1, 2; j = 1, 2, \cdots, n \text{ and } k = 1, 2, \cdots, d$,
- $\theta^{(k)}$ denotes the $k^{th}$ dimension of a $d$ dimensional vector at which the adaptive kernel density estimate is evaluated,
- $\{h_{ij}^{(k)} : j = 1, 2, \cdots, n \text{ and } k = 1, 2, \cdots, d\}$ are smoothing parameters and $K(\cdot)$ is a Gaussian kernel.

In (5), the smoothing parameters are chosen using Silverman [12] (Sec 5.3.1) wherein observations in sparse regions are assigned Gaussian kernels with high bandwidth and observations in high probability regions are assigned Gaussian kernels with low band-
width. In our examples, we use the kepdf function in the R package pdfCluster (Azzalini and Menardi [13]) to compute the adaptive kernel density estimate of the chains. The KL divergence between $P_{1n}$ and $P_{2n}$, denoted by $KL(P_{1n}|P_{2n})$ and KL divergence between $P_{2n}$ and $P_{1n}$, denoted by $KL(P_{2n}|P_{1n})$ can be obtained after substituting appropriate values of $i$ and $j$ in the equation given below,

$$KL(P_{1n}|P_{2n}) = \int_\Theta \left( \log(P_{1n}(\theta)) - \log(P_{2n}(\theta)) \right) P_{1n}(\theta) d\theta. \quad (6)$$

The symmetric KL divergence between $P_{1n}$ and $P_{2n}$, denoted by $KL_{sy}(P_{1n}, P_{2n})$ is given below,

$$KL_{sy}(P_{1n}, P_{2n}) = \frac{KL(P_{1n}|P_{2n}) + KL(P_{2n}|P_{1n})}{2}. \quad (7)$$

We can find the Monte Carlo estimate of $KL(P_{1n}|P_{2n})$ and $KL(P_{2n}|P_{1n})$ using (8) and (9) respectively,

$$\hat{KL}(P_{1n}|P_{2n}) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \log(P_{1n}(\theta^*_i)) - \log(P_{2n}(\theta^*_i)) \right\}, \quad (8)$$

$$\hat{KL}(P_{2n}|P_{1n}) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \log(P_{2n}(\theta^*_i)) - \log(P_{1n}(\theta^*_i)) \right\}, \quad (9)$$

where $\{\theta^*_i\}_{i=1}^{n}$ are the observations simulated from $P_{1n}(\theta)$ using the technique proposed by Silverman [12](Sec 6.4.1). Similarly,

$$\hat{KL}_{sy}(P_{1n}, P_{2n}) = \frac{\hat{KL}(P_{1n}|P_{2n}) + \hat{KL}(P_{2n}|P_{1n})}{2}. \quad (10)$$

Adaptive kernel density estimation suffers from the curse of dimensionality. Hence we need to increase our sample size ($n$) as the dimension ($d$) increases in order to obtain a good estimate of the symmetric KL divergence between any two distributions.

In order to find the appropriate sample size ($n$) required for achieving convergence when univariate and bivariate chains are drawn from similar distributions, we conduct a simulation study. In the univariate case, for each $n$, we generate 1000 datasets each from $f_1 \equiv N(0, 1)$ and $f_2 \equiv N(0, 1)$. Let $f_1$ and $f_2$ be the adaptive kernel density estimates of observations drawn from $f_1$ and $f_2$ respectively. The estimated symmetric KL divergence between $f_1$ and $f_2$ for each pair can be computed using (10) in which $P_{1n}$ and $P_{2n}$ are replaced by $f_1$ and $f_2$. The true symmetric KL divergence between $f_1$ and $f_2$ is known to be zero. Thus we can then find the bias, standard deviation and root mean square error (RMSE) of $\hat{KL}_{sy}(f_1, f_2)$. In the bivariate case, for each $n$, we generate 1000 datasets each from $f_3 \equiv N(0, I_2)$ and $f_4 \equiv N(0, I_2)$ and carry out a similar procedure as before to find the bias, standard deviation and RMSE of $\hat{KL}_{sy}(f_3, f_4)$. The results are tabulated in Table 1.
Table 1. Bias, Standard Deviation and RMSE of (a) $KL_{sy}(\hat{f}_1, \hat{f}_2)$ where $\hat{f}_1$ and $\hat{f}_2$ are adaptive kernel density estimates of observations drawn from $f_1 \equiv N(0, 1)$ and $f_2 \equiv N(0, 1)$ (b) $KL_{sy}(\hat{f}_3, \hat{f}_4)$ where $\hat{f}_3$ and $\hat{f}_4$ are adaptive kernel density estimates of observations drawn from $f_3 \equiv N(0, I_2)$ and $f_4 \equiv N(0, I_2)$.

<table>
<thead>
<tr>
<th>(a) Univariate Distribution</th>
<th>(b) Bivariate Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>$\text{Bias}$</td>
</tr>
<tr>
<td>1000</td>
<td>0.0066</td>
</tr>
<tr>
<td>1500</td>
<td>0.0046</td>
</tr>
<tr>
<td>2000</td>
<td>0.0039</td>
</tr>
<tr>
<td>2500</td>
<td>0.0032</td>
</tr>
<tr>
<td>3000</td>
<td>0.0027</td>
</tr>
</tbody>
</table>

In Table 1 we observe that for both univariate and bivariate chains, the bias, standard deviation and RMSE go on reducing as the sample size ($n$) increases. In order to assess the convergence of Markov chains, we intend to accurately estimate the symmetric KL divergence up till two decimal points. In Table 1, we observe that for $n = 2000$ in the univariate case and for $n = 12,000$ in the bivariate case, the bias, standard deviation and RMSE are significantly small, and hence, the symmetric KL divergence can be estimated efficiently up till two decimal points. Thus if we wish to use Tool 1 for assessing convergence, we need to run the chains for at least 2000 iterations in the univariate case and for at least 12,000 iterations in the bivariate case. We next check if the above mentioned sample sizes hold true, when instead of the Gaussian distribution the above simulation study is carried out using a heavy tailed distribution, skewed distribution or a distribution with dependent coordinates. For this purpose, the above simulation study was repeated with t distribution ($df=5$), chi-square distribution ($df=10$) and bivariate normal distribution (with correlation coefficient equal to 0.3). The sample sizes prescribed above were found to be sufficient for efficiently estimating the symmetric KL divergence up till two decimal points in these cases. The tabulated results are provided in the supplementary material. Thus we can safely use the Gaussian distribution for studying the bias, standard deviation and RMSE associated with the symmetric KL divergence estimator.

Tool 1 will detect convergence i.e. indicate that the chains have mixed adequately, when the estimated symmetric KL divergence between $P_{1n}$ and $P_{2n}$ will be close to zero. Hence we now need to identify cut-off points, so that estimated symmetric KL divergence below or equal to that point will indicate convergence. Boone et al. [8] carried out a simulation study and came up with a criteria, wherein if the estimated Hellinger distance between the kernel density estimates of univariate chains is less than 0.10, then the Markov chains have converged else not.

We will utilize a testing of hypothesis framework to come up with cut-off points. In our framework, the null hypothesis states that the Markov chains have diverged, i.e., the chains have not yet mixed adequately and our alternative hypothesis states that the Markov chains have converged, i.e., the chains have mixed adequately. In this scenario, the probability of Type 1 error will be the probability of concluding that the Markov chains have converged when in fact they have not. We would like to limit this probability to some level $\alpha$ which is typically chosen to be 0.05. As mentioned earlier, we will be estimating the symmetric KL divergence up till two decimal points, hence the cut-off value should also be reported up till the second decimal point. In order to find our cut-off values, we will first generate 1000 datasets each from two dissimilar distributions.
and the maximum value of $C$ for which $P(\hat{KL}_{sy}(P_{1n}, P_{2n}) \leq C) \leq \alpha = 0.05$, i.e. the probability of Type 1 error is less than 0.05, will be our cut-off value.

In the univariate case, we will generate 1000 datasets with suitable sample size ($n$) each from $f_5 \equiv N(0, 1)$ and $f_6 \equiv N(\mu, 1)$ where $\mu \neq 0$. The adaptive kernel density estimate of observations drawn from $f_5$ and $f_6$ are denoted by $\hat{f}_5$ and $\hat{f}_6$ respectively. The maximum value of $C$ for which $P(\hat{KL}_{sy}(\hat{f}_5, \hat{f}_6) \leq C) \leq \alpha = 0.05$ will be our cut-off value in the univariate case. In the bivariate case, we will generate 1000 datasets with a suitable sample size ($n$) each from $f_7 \equiv N(0, I_2)$ and $f_8 \equiv N(\mu I_2, I_2)$ where $\mu \neq 0$ and then carry out a similar procedure as before to identify our cut-off value in the bivariate case. By limiting the probability of Type 1 error, we are exercising control over variability of estimated symmetric KL divergence, but we need to make sure that its bias does not affect the cut-off value adversely, and hence for this purpose, the suitable sample size ($n$) should be chosen in such a way that the bias is significantly small and does not affect the first two decimal points of the estimated value. The $\mu$ can be chosen by the users as per their requirement. If the users are dealing with a very sensitive experiment, then the choice of $\mu$ should be small otherwise $\mu$ can be chosen to be slightly bigger. For simulation studies corresponding to identifying cut-off values using dissimilar distributions, Boone et al. [8] utilized $N(0, 1)$ and $N(0.2835, 1)$ since the true Hellinger distance between them is known to be 0.10. Hence a possible choice for $\mu$ can be 0.2835. For this choice of $\mu$, as mentioned before, the sample size ($n$) should be chosen based on the bias of $\hat{KL}_{sy}(\hat{f}_5, \hat{f}_6)$ and $\hat{KL}_{sy}(\hat{f}_7, \hat{f}_8)$ for univariate and bivariate cases respectively.

The symmetric KL divergence between any two univariate Gaussian distributions with the same variance parameter, and any two multivariate Gaussian distributions with the same covariance matrix can be computed analytically using (11) and (12),

$$KL_{sy}(g_1, g_2) = \frac{(\mu_1 - \mu_2)^2}{2\sigma^2},$$  \hspace{1cm} (11)

where $g_1 \equiv N(\mu_1, \sigma^2)$ and $g_2 \equiv N(\mu_2, \sigma^2)$, $\mu_1 \in R$, $\mu_2 \in R$, and

$$KL_{sy}(g_3, g_4) = \frac{1}{2}\left\{ (\mu_4 - \mu_3)^T \Sigma^{-1} (\mu_4 - \mu_3) \right\},$$  \hspace{1cm} (12)

where $g_3 \equiv N(\mu_3, \Sigma)$ and $g_4 \equiv N(\mu_4, \Sigma)$, $\mu_3 \in R^d$, $\mu_4 \in R^d$.

Thus using (11) and (12), for $\mu = 0.2835$, true symmetric KL divergence between $f_5$ and $f_6$ was found to be 0.04 and the true symmetric KL divergence between $f_7$ and $f_8$ was found to be 0.08. We then compute the bias, standard deviation and RMSE of $\hat{KL}_{sy}(\hat{f}_5, \hat{f}_6)$ and $\hat{KL}_{sy}(\hat{f}_7, \hat{f}_8)$. The results are provided in Table 2.

In Table 2 we observe that, for $n = 2000$ in the univariate case and $n = 12,000$ in the bivariate case, the bias associated with $\hat{KL}_{sy}(\hat{f}_5, \hat{f}_6)$ and $\hat{KL}_{sy}(\hat{f}_7, \hat{f}_8)$ is significantly small and does not affect the first two decimal points of the estimate. In Table 2 we also observe that, the standard deviation and RMSE are also reasonably small for $n = 2000$ and $n = 12,000$ in the univariate and bivariate cases respectively. Hence we can carry out our cut-off procedure for $\mu = 0.2835$ with $n = 2000$ in the univariate case and $n = 12,000$ in the bivariate case. The Type 1 error associated with different cut-off values is given in Table 3.

In Table 3 we observe that $C = 0.02$ and $C = 0.06$ are ideal cut-off points for univariate and bivariate distributions respectively. The users should also be aware that if they choose a larger sample size than the one prescribed before, then the Type 1 error associated
Table 2. Bias, Standard Deviation and RMSE of (a) $\overline{KL}_{sy}(\hat{f}_5, \hat{f}_6)$ where $\hat{f}_5$ and $\hat{f}_6$ are adaptive kernel density estimates of observations drawn from $f_5 \equiv N(0, 1)$ and $f_6 \equiv N(0.2835, 1)$ (b) $\overline{KL}_{sy}(\hat{f}_7, \hat{f}_8)$ where $\hat{f}_7$ and $\hat{f}_8$ are adaptive kernel density estimates of observations drawn from $f_7 \equiv N(0.2835, I_2)$ and $f_8 \equiv N(0.2835, I_2)$.

<table>
<thead>
<tr>
<th>(a) Univariate Distribution</th>
<th>(b) Bivariate Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>Bias</td>
</tr>
<tr>
<td>1000</td>
<td>0.0054</td>
</tr>
<tr>
<td>1500</td>
<td>0.0035</td>
</tr>
<tr>
<td>2000</td>
<td>0.0024</td>
</tr>
<tr>
<td>2500</td>
<td>0.0019</td>
</tr>
<tr>
<td>3000</td>
<td>0.0018</td>
</tr>
</tbody>
</table>

Table 3. Probability of Type 1 error associated with different cut-off values in the univariate and bivariate case.

<table>
<thead>
<tr>
<th>(a) Univariate Distribution</th>
<th>(b) Bivariate Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C$</td>
<td>$P(\overline{KL}_{sy}(\hat{f}_5, \hat{f}_6) \leq C)$</td>
</tr>
<tr>
<td>0.01</td>
<td>0.001</td>
</tr>
<tr>
<td>0.02</td>
<td>0.040</td>
</tr>
<tr>
<td>0.03</td>
<td>0.242</td>
</tr>
</tbody>
</table>

with the cut-off value will further reduce. Thus, if the user chooses $\mu$ to be 0.2835 and at $n = n^*$, estimated symmetric KL divergence (rounded to two decimal points) is found to be less than or equal to the cut-off value, then Tool 1 indicates that the chains have mixed adequately at that point, hence the user should then discard the first $n^*$ observations and iterations obtained thereafter will be considered as approximate observations from the target distribution. On the other hand, if the estimated symmetric KL divergence (rounded to two decimal points) is found to be strictly greater than the cut-off value, then Tool 1 indicates that chains have not yet mixed adequately and the user should run the chain longer. If found necessary, the user can thin the chain and consider an initial burn-in before applying Tool 1.

Tool 1 can be used for more than two dimensions as well. For a multivariate chain, we recommend assessing convergence marginally, i.e. one variable at a time. We found our cut-off values using a testing of hypothesis approach, wherein our level of significance was chosen to be $\alpha = 0.05$. In the case of multiple comparison, we will adjust our level of significance using Bonferroni’s correction, so that the overall type 1 error does not go beyond $\alpha = 0.05$. Hence, if the Markov chain is $d$ dimensional (where $d > 2$), then our level of significance for each comparison will be $\alpha/d$. Using this adjusted level of significance in our cut-off procedure, we obtain an appropriate cut-off point for multiple comparison. For example, if our Markov chain is 10 dimensional, then in Table 3 we observe that $C = 0.01$ is an ideal cut-off point, since it maintains a type 1 error of less than $\alpha/10 = 0.005$ for each comparison. This ability to maintain the overall type 1 error at $\alpha$, by adjusting the cut-off value in the case of multiple comparison, is another advantage of our tool over the tools proposed by Boone et al. [8]. Applying Tool 1 in the univariate case is similar to the tool proposed by Boone et al. [8]. Hence, in case of multi-modal target distribution, if both chains are stuck at the same mode, then this tool is also prone to failure.

8
In case of multiple chains, say \( m \) chains, we will find the estimated symmetric KL divergence between each of the \( \binom{m}{2} \) combination of chains and find the maximum among them. If the maximum estimated symmetric KL divergence is less than or equal to the cut-off value, then Tool 1 indicates that the chains have converged else not. If the user wishes to use a single chain, then one can estimate the symmetric KL divergence between the adaptive kernel density estimate of any two non-overlapping parts of the chain.

In cases where the state space is bounded, adaptive kernel density estimation might suffer from boundary bias if high probability regions are closer to the boundary. But the objective of Tool 1 is to check if the two chains have mixed adequately or not, hence, even if it is found that the adaptive kernel density estimate of the two chains suffers from boundary bias, the final objective is not affected as long as the same density estimation procedure is used for both chains. The user must also note that, if the sample simulated from the adaptive kernel density estimate of the chain contain several observations from outside the state space, and if the target distribution is not expected to have a lot of mass close to the boundary, then it is an indication that chains have not captured the target distribution adequately and thus Tool 1 indicates divergence in such a situation.

To implement Tool 1 for two univariate chains with \( n = 2,000 \) and two bivariate chains with \( n = 12,000 \), it takes approximately 3.81s and 109.86s respectively on an Intel (R) Core (TM) i5-6300U 2.40GHz machine running Windows 10. For multivariate chains, we implement Tool 1 marginally i.e. one variable at a time, which can be done in parallel.

### 3.2. Visualization Tool

Suppose a user is using multiple chains, say \( m \) chains (where \( m \geq 3 \)). Further suppose that application of Tool 1 revealed that the \( m \) chains have not mixed adequately and thus chains have not yet converged. This indication of divergence could be due to a variety of reasons. A common reason for divergence is formation of clusters among multiple chains. A visualization tool can be helpful for identifying these clusters.

Peltonen et al. [14] had proposed a visualization tool based on Linear Discriminant Analysis and Discriminant Component Analysis which can be used to complement the diagnostic tools proposed by Gelman and Rubin [4] and Brooks and Gelman [5]. Similarly the visualization tool described in this section will complement Tool 1 proposed in Section 3.1.

In this tool we utilize the tile plot. As mentioned before, in case of multiple chains, say \( m \) chains, Tool 1 will find the estimated symmetric KL divergence between each of the \( \binom{m}{2} \) combinations and report the maximum among them. In the visualization tool, we will utilize the individual values provided by estimated symmetric KL divergence between each of the \( \binom{m}{2} \) distinct combinations. If the estimated symmetric KL divergence for a particular combination is less than or equal to the cut-off value, then we will utilize a “Grey” tile to represent that the two chains belong to the same cluster else we will use a “black” tile to represent that the two chains belong to different clusters.

In the case of a multivariate chain, we monitor convergence marginally i.e. one variable at a time. Hence two multivariate chains will be considered to be in the same cluster, only if the estimated symmetric KL divergence for each variable is less than or equal to the cut-off value, which has been adjusted for multiple comparison. For further investigation of a multivariate Markov chain, the user can consider the following steps.

Consider a \( d \) dimensional Markov chain initialized at \( m \) different points. Suppose these \( m \) chains (where \( m \geq 3 \)), were grouped into \( q \) clusters. The visualization tool is utilized when Tool 1 indicates divergence i.e. \( 2 \leq q \leq m \). In cases where Tool 1 indicates di-
vergence, for further investigation, the user can choose a chain from each cluster and implement the visualization tool marginally i.e. one variable at a time. This will help the user identify, which among the $d$ variables are responsible for inadequate mixing among the $m$ multivariate chains.

In order to provide an illustration of the visualization tool, suppose we run 5 chains for 5,000 iterations each wherein the $1^{st}$ and the $3^{rd}$ chain are drawn from $N(0, 1)$ while the $2^{nd}$, $4^{th}$ and $5^{th}$ chain are drawn from $N(10, 1)$. The application of the visualization tool for these five chains is provided in Figure 1. As expected, Figure 1 indicates presence of two clusters wherein chain 1 and chain 3 form a cluster while chain 2, chain 4 and chain 5 form another cluster.

3.3. Tool 2

Suppose the target density is as follows,

\[ \pi(\theta) = \frac{g(\theta)}{k}, \theta \in \Theta, \]  

(13)

where $k$ is the unknown normalizing constant.

Suppose a single Markov chain is run for $n$ iterations and the observations obtained are $\{\theta_{1j}\}_{j=1}^{n}$. Let $P_{1n}(\theta)$ denote the adaptive kernel density estimate of the observations as mentioned in (5). The KL divergence between $P_{1n}(\theta)$ and $\pi(\theta)$ is given below,

\[ KL(P_{1n}|\pi) = G_n + \log k, \]  

(14)

where

\[ G_n = \int_{\Theta} \log(P_{1n}(\theta)) P_{1n}(\theta) d\theta - \int_{\Theta} \log(g(\theta)) P_{1n}(\theta) d\theta. \]

In the implementation of Tool 2, we will assume that $KL(P_{1n}|\pi) \to 0$ as $n \to \infty$. Under this assumption, $G_n \to -\log(k)$ as $n \to \infty$. Hence $\exp(-G_n) \to k$ as $n \to \infty$. Thus, an estimator of the normalizing constant based on KL divergence between $P_{1n}(\theta)$ and $\pi(\theta)$, denoted by $\hat{k}$ is given below,

\[ \hat{k} = \exp \left( -\frac{1}{n} \sum_{i=1}^{n} \left\{ \log(P_{1n}(\theta_{1i})) - \log(g(\theta_{1i})) \right\} \right), \]  

(15)

10
Table 4. Finding appropriate sample size \((n)\) required for implementation of Tool 2 when the samples are indeed from the target distribution i.e. standard normal distribution.

<table>
<thead>
<tr>
<th>(a) Univariate Distribution</th>
<th>(b) Bivariate Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n)</td>
<td>(P(T^*_2 \leq 0.05))</td>
</tr>
<tr>
<td>100</td>
<td>0.631</td>
</tr>
<tr>
<td>500</td>
<td>0.871</td>
</tr>
<tr>
<td>1000</td>
<td>0.950</td>
</tr>
<tr>
<td>2000</td>
<td>0.981</td>
</tr>
<tr>
<td>3000</td>
<td>0.997</td>
</tr>
</tbody>
</table>

where \(\{\theta^*_i\}_i^n\) are the observations simulated from \(P_{1n}(\theta)\) using the technique proposed by Silverman [12](Sec 6.4.1).

If the chain has converged, then the quantity \(T_2\) given below will be close to 0,

\[
T_2 = \frac{\hat{k} - k}{k}.
\]  

(16)

But \(T_2\) contains the normalizing constant \((k)\), which is unknown. Now, the unknown normalizing constant \((k)\) is nothing but the integral given below,

\[
k = \int_\Theta g(\theta) \, d\theta.
\]  

(17)

Hahn [15] came up with Cuba library which provides implementation of general purpose multidimensional integration algorithm. The R package R2Cuba (Bouvier and Kiu [16]) provides the implementation of Cuba library in R. Using the “divonne” function in R2Cuba (Bouvier and Kiu [16]) we will evaluate the integral given in (17) and thus obtain an estimate of the normalizing constant based on numerical integration which we will denote as \(k^*\). The user can also use the “adaptIntegrate” function in the R package cubature (Johnson and Narasimhan [17]) to produce an estimate \(k^*\) of the unknown normalizing constant \((k)\). Replacing \(k\) by \(k^*\) in (16) we obtain \(T^*_2\) given below,

\[
T^*_2 = \frac{|\hat{k} - k^*|}{k^*}.
\]  

(18)

In the case of multi-modal target distribution, if the Markov chain gets stuck in one of the modes, then the adaptive kernel density estimate of the chain will estimate the density around that particular mode really well, while completely ignoring the rest of the density. The estimate \(\hat{k}\) is a Monte Carlo estimate of the unknown normalizing constant \((k)\), based on the KL divergence between the adaptive kernel density estimate of the chain and the target distribution. Thus, the evaluation of \(\hat{k}\) involves drawing \(n\) observations from the adaptive kernel density of the chain. But, since the adaptive kernel density estimate captures only one mode, the estimate \(\hat{k}\) only records the KL divergence between the adaptive kernel density estimate of the chain and the target distribution for the state space around that particular mode. Hence \(\hat{k}\) will likely underestimate the true quantity \(k\), if the Markov chain gets stuck in one of the modes. On the other hand, \(k^*\) will provide a good estimate of the unknown normalizing constant \((k)\) since it uses numerical
integration to integrate over the entire state space. Thus $T_2^*$ can be interpreted as the percentage of the target distribution not yet captured by the Markov chain. A Markov chain that captures at least 95% of the target distribution can be considered to be producing approximate observations from the target distribution. Using this interpretation of $T_2^*$, we came up with a cut-off value of 0.05 wherein if $T_2^* > 0.05$, then Tool 2 indicates that the Markov chain has not yet captured the target distribution adequately.

As seen earlier, Tool 2 assumes (i) $\exp(-G_n) \to k$ as $n \to \infty$, (ii) $k$ is a consistent Monte Carlo estimate of $\exp(-G_n)$ and (iii) $k^*$ is an estimate of $k$ based on multidimensional numerical integration. Since (i) and (ii) depend on the sample size ($n$), it is important to know their convergence rates as a function of $n$. But currently in the literature, a theoretical proof of the convergence of the adaptive kernel density estimate based on a Markov chain to its stationary distribution, with respect to the KL divergence is not available. In the absence of such a theoretical result, it is difficult to find the convergence rates of (i) and (ii) as a function of $n$. Hence we conduct a simulation study to choose a suitable sample size ($n$). Since the cut-off value is based on the interpretation of $T_2^*$, we will conduct a slightly different simulation study to get an intuition of the sample size ($n$) required to implement Tool 2. In the univariate case we will consider $f_1 \equiv N(0, 1)$ to be our target distribution, generate 1000 datasets of sample size ($n$) each from $f_1 \equiv N(0, 1)$ and find $\hat{k}$ using (15) while $k^*$ will be obtained by numerically integrating the kernel of $f_1$, thus we can then find $T_2^*$ for each dataset using (18). Then we estimate $P(T_2^* \leq 0.05)$ i.e. probability of concluding convergence when the chain is indeed from the target distribution. The sample size ($n$) for which the estimate of $P(T_2^* \leq 0.05)$ is significantly high, will be our prescribed sample size ($n$) for Tool 2. In the bivariate case, similar procedure is carried out for $f_3 \equiv N(0, I_2)$. The results are given in Table 4.

In Table 4 we observe that for $n = 2000$ in the univariate case and $n = 12,000$ in the bivariate case, the estimated probability of concluding that the Markov chain has converged, given the fact that it is indeed drawn from the target distribution is really high. This provides us an indication that $n = 2000$ and $n = 12,000$ are sufficient for implementing Tool 2 in univariate and bivariate distributions respectively. We also check if the prescribed sample sizes are sufficient when the target distribution is heavy tailed, skewed or is a distribution with dependent coordinates and hence we replicated the above study with $t$ distribution (df=5), chi-square distribution (df=10) and bivariate normal distribution (with correlation coefficient equal to 0.3). We observed that the sample sizes mentioned above are sufficient. The detailed results are provided in the supplementary material.

Tool 2 is specifically designed for detecting divergence when the target distribution is multi-modal and the chain gets stuck in one of the modes. Thus, if the user observes that even after running the chain for a reasonably long time (sample size prescribed above), the value of $T_2^*$ was found to be greater than 0.05, then it is highly likely that the chain is stuck in one of the modes and has not yet traveled through the whole state space. The value of $T_2^*$ will also tell the user the percentage of the target distribution not yet traveled by the Markov chain.

Tool 2 involves the target distribution, hence if the target distribution is bounded and has a lot of mass close to the boundary, then Tool 2 will be affected by boundary bias. Due to boundary bias, the sample generated from the adaptive kernel density estimator might contain observations from outside the state space and $\log$ of the target distribution without the unknown normalizing constant (i.e. $\log(q(\theta))$) is not defined at these points. A possible solution to this situation is to consider a bootstrap sample instead of sampling from the adaptive kernel density estimate. Also, in the case where the bounded target distribution has very little mass close to the boundary and the sample generated from
the adaptive kernel density estimate contains none or negligible number of observations from outside the state space, we can simply ignore those and implement Tool 2.

Users must be aware that Tool 2 is vulnerable to poor adaptive kernel density estimation. Further, estimating the unknown normalizing constant using numerical integration is challenging, hence we do not claim that Tool 2 will solve all problems related to diagnosing convergence of Markov chains in the case of multi-modal target distributions. But, we hope that the tool will help the users understand the challenges associated with it, and thus further boost research in this direction.

To implement Tool 2 for a univariate chain with $n = 2,000$ and a bivariate chain with $n = 12,000$, it takes approximately 1.31s and 35.68s respectively on an Intel (R) Core (TM) i5-6300U 2.40GHz machine running Windows 10.

4. Examples

4.1. Six Mode Example

This example was proposed by Leman et al. [18]. Suppose the target density is as follows,

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

(19)

where $-10 \leq x, y \leq 10$.

The contour plot of the target distribution known up to the normalizing constant is given in Figure 2 and the marginal distribution of X and Y known up to the normalizing constant is given in Figure 3. The visualization of joint and marginal distribution clearly show that the target distribution is multi-modal in nature.

![Figure 2. Contour plot of the target distribution in the Six Mode Example.](image)

![Figure 3. Marginal Distribution of X and Y in the six mode example.](image)
We will consider two different cases to illustrate the application of the diagnostic tools and the visualization tool proposed in section 3. In order to draw observations from the target distribution, we will use a Metropolis within Gibbs sampler in which X is drawn first and then Y.

**Case 1**
In this case, we will run four chains wherein two chains (chain 1 and chain 2) will be started at a particular mode while the remaining two chains (chain 3 and chain 4) will be started at some other mode. Each of the four chains were run for \( n = 30,000 \) iterations. The adaptive kernel density estimates of the four chains are visualized in Figure 4.

In order to assess the convergence of the above Markov chains, several diagnostic tools were implemented and the results are presented in Table 5. In Table 5 we observe that the PSRF proposed by Gelman and Rubin [4], MPSRF proposed by Brooks and Gelman [5], Hellinger distance approach proposed by Boone et al. [8] and Tool 1 proposed in Section 3 correctly indicate that the chains have not yet converged.

**Case 2**
In this case as well we will run four chains but all the chains will be started at the same mode. All the four chains were run for \( n = 30,000 \) iterations and the adaptive kernel density estimates of the four chains are visualized in Figure 6.

Convergence diagnostic tools used in case 1 were applied in this case as well and the results obtained are tabulated in Table 6. In Table 6, we observe that since all chains are stuck at the same mode, the PSRF, the MPSRF and the Hellinger distance approach
false detect convergence. Now, Tool 2 requires only one chain and since the PSRF, MPSRF and the Hellinger distance suggest that the four chains are similar, we can simply choose any one among them. Now, $T^*_2 = 0.88$ is significantly greater than zero and thus indicates that the chains are stuck at the same mode, further it also indicates that 88% of the target distribution is not yet captured by the Markov chain. Thus Tool 2 is both computationally cheap as well as efficient in detecting divergence in the case of multi-modal target distributions.

Table 6. Application of various MCMC convergence diagnostic tools for Case 2 in the Six Mode Example.

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\hat{R}$</th>
<th>H Dist</th>
<th>$\hat{R}_p$</th>
<th>$T^*_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>1.00</td>
<td>0.09</td>
<td>1.00</td>
<td>0.88</td>
</tr>
<tr>
<td>Y</td>
<td>1.00</td>
<td>0.03</td>
<td>1.00</td>
<td>0.88</td>
</tr>
</tbody>
</table>

4.2. Mixture of Bivariate Normal

Suppose the target density is as given below,

$$
\pi\left( (x, y)^T \right) = \frac{1}{2} \phi_2\left( (x, y)^T, 0, \Sigma_1 \right) + \frac{1}{2} \phi_2\left( (x, y)^T, 0, \Sigma_2 \right)
$$

(20)
where \((x, y)^T \in \mathbb{R}^2\), \(\phi_2((x, y)^T, 0, \Sigma_i)\) is the density of bivariate normal distribution with mean 0, covariance matrix \(\Sigma_i\), evaluated at \((x, y)^T\) and \(\Sigma_i = \begin{bmatrix} 1 & \rho_i \\ \rho_i & 1 \end{bmatrix}\) for \(i = 1, 2\).

The target distribution with \(\rho_1 = 0.99\) and \(\rho_2 = -0.99\) is plotted in Figure 7. To simulate observations from the above target distribution, we will utilize the MCMC function in the R package adaptMCMC (Scheidegger [19]) which uses an adaptive Metropolis algorithm. We will use four chains wherein two chains each are started in neighboring corners. All four chains were run for \(n\) = 1000 iterations. The chains are visualized in Figure 8. In Figure 8 we observe that since each component is so highly correlated, chains get stuck in the component in which they were started. Hence the chains have clearly not mixed adequately.

In order to detect divergence in this example, one needs to assess the convergence of both \(X\) and \(Y\) jointly. Thus, marginal empirical convergence diagnostic tools like PSRF proposed by Gelman and Rubin [4] and Hellinger distance approach proposed by Boone et al. [8] are not applicable in such cases. The MPSRF has the ability to assess the joint convergence of \(X\) and \(Y\), but it falsely detects convergence (\(\hat{R}_p = 1.03\)) since it averages the within chain covariances and thus the positive covariances and the nega-
tive covariances cancel each other out. Using the bivariate KL Tool 1, the maximum estimated symmetric KL divergence between chains was found to be 1.98 which is significantly greater than the cut-off of $C = 0.06$. The bivariate KL Tool 1 requires at least 12,000 observations to provide a good estimate hence simply comparing the estimated symmetric KL divergence to the cut-off value is not enough. Thus, we must also look at the probability of observing an estimated symmetric KL divergence of 1.98 or less when the chains with $n = 1000$ are drawn from different distributions i.e. $N(0, I_2)$ and $N(\mu I_2, I_2)$ where $\mu = 0.2835$. This probability which can also be looked upon as the p-value in terms of the hypothesis framework given in Section 3 was found to be very large (i.e. greater than 0.999). Thus we do not reject our null hypothesis and concluded that the chains have not yet mixed adequately. Thus, in this example we observe that, even if the target distribution is unimodal, MPSRF proposed by Brooks and Gelman [5] is vulnerable to false indication of convergence. If the above chains are run for a longer period, then all the chains travel through both the components and mix adequately.

4.3. Logit Model

We use the “Anguilla_train” dataset provided in the R package dismo (Hijmans et al. [20]) to fit a Bayesian logit model with presence or absence of short finned eel considered as the response variable and the six other variables were included as covariates. The six covariates are: summer air temperature (SeqSumT), distance to coast (DSDist), area with indigenous forest (USNative), average slope in the upstream catchment (USSlope), maximum downstream slope (DSMaxSlope) and Fishing Method (categorical variable with five classes namely electric, mixture, net, spot and trap). This example was also used by Boone et al. [8] to illustrate the usefulness of the Hellinger distance approach. The model is as follows,

$$Y_i \sim Bernoulli(\mu_i),$$

$$\mu_i = \frac{\exp(X_i^T \beta)}{1 + \exp(X_i^T \beta)},$$

$$\beta \sim N(0, 100 I_{10}).$$

In order to estimate the parameters in the above model, three chains were run using the MCMClogit function in the R package MCMCpack (Martin et al. [21]). We considered an initial burn-in of 30,000 as suggested by Boone et al. [8]. Since the dimension of the MCMC chain is large, we will apply Tool 1 marginally to diagnose the convergence of each of the parameter. After the initial burn-in, each of the three chains were run for $n = 3000$, $n = 15,000$ and $n = 30,000$ iterations and for each $n$, the convergence was diagnosed using PSRF, Hellinger distance and univariate KL Tool 1. The results are tabulated in Table 7.

The Markov chain is 10 dimensional, hence in order to maintain an overall type 1 error rate of $\alpha = 0.05$, we will use Bonferroni’s correction to adjust our cut-off point. For $C = 0.01$, the type 1 error for each comparison is less than $\alpha/10 = 0.005$ while the overall type 1 error is less than $\alpha = 0.05$. The PSRF ($\hat{R}$) indicates that the chains have mixed adequately at $n = 3000$ and hence iterations obtained from 3000 onwards can be used to compute parameters estimates. But Hellinger distance approach and multivariate KL Tool 1 indicate that chains have not yet converged and hence the user must run the sampler longer. The Hellinger distance approach indicates convergence of all parameters at $n = 15,000$ while the multivariate KL Tool 1 indicates convergence of all parameters
at \( n = 30,000 \). Since the multivariate KL Tool 1 adjusts its cut-off point for multiple comparison, it is advisable for the user to use iterations from 30,000 onwards for making inference about the parameters of interest.

5. Conclusion

In this article, we have provided two new MCMC convergence diagnostic tools based on KL divergence and smoothing methods. The advantage of the first tool over existing MCMC convergence diagnostic tools is that, it has the ability to assess the joint convergence of multiple variables. For multivariate chains, we assess convergence marginally and recalibrate the cut-off point using Bonferroni’s correction to maintain the overall type 1 error at \( \alpha \). Due to the use of Bonferroni’s correction, Tool 1 can be conservative in the case of large number of variables. But in the case of MCMC diagnostics, a conservative tool is preferable as it provides the user greater assurance, that the chain is producing approximate observations from the target distribution, when it indicates convergence. In the case where the first tool indicates divergence of multiple MCMC chains, the user can use the visualization tool to further investigate reasons behind the divergence of multiple MCMC chains. The advantage of the second tool over existing MCMC convergence diagnostic tools is that, it is equipped to detect divergence when MCMC chains get stuck in a particular mode of a multi-modal target distribution. Tool 2 is vulnerable if multidimensional numerical integration does not provide a good estimate of the unknown normalizing constant. Thus the proposed methods provide a useful addition to the set of available MCMC diagnostic tools and are equipped to detect non convergence of chains when other methods might fail to do so. A possible future study involves deriving a theoretical proof of convergence of the adaptive kernel density estimate based on Markov chain samples to its stationary distribution with respect to KL divergence measure.

Supplementary Materials

Tabulated results corresponding to Tool 1 and Tool 2 simulation studies carried out using \( t \) distribution (df=5), chi-square distribution (df=10) and bivariate normal distribution (with correlation coefficient equal to 0.3) are provided in the supplementary material. The R code required to implement the tools in this article is also available online.
Acknowledgment

We would like to thank two reviewers, the AE and the Editor for their helpful comments that have improved the manuscript.

References