# Annexure-II

DECLARATION

I/We hereby declare that the work presented in the project report entitled Changepoint Analysis for MCMC contains my own ideas in my own words. At places, where ideas and words are borrowed from other sources, proper references, as applicable, have been cited. To the best of our knowledge this work does not emanate from or resemble other work created by person(s) other than mentioned herein.

Name and Signature Ayush Agarwal

Date: 9th Dec 2020

# Change-point analysis for Markov chains

## Undergraduate Project Report

---

---

*Submitted To:*
Prof. Dootika Vats

*Submitted By :*
Ayush Agarwal

# Contents

# 1 Introduction

It is often difficult to come across independent and identically distributed random samples. More often that not, we require specialised sampling techniques and obtain dependent samples, which in turn are used to estimate the true statistics. In this paper, we look at the class of Markov chain Monte Carlo (MCMC) methods. Since we are using sample statistics as an estimate, it is essential that the samples are representative of the true underlying distribution. In case of Markov chains, if we are unable to sample from stationarity, the estimators will likely suffer from convergence problems. Consequently a bias in the sample will lead to inaccurate estimates. To address this problem, it is a common practice to discard a *fraction* of the chain from the start called the "burn-in" period. However, in practice the said *fraction* is often arbitrary, wastes samples and cause inefficiency. Most methods in literature are either parametric, assume stationarity or identify the most significant change-point rather than the first significant one.

We propose a novel change-point algorithm for the Markov chain Monte Carlo setup. In particular, we show how procedures based on the cumulative sum, CUSUM, statistics can be modified to work for data exhibiting serial dependence such as a Markov Chain. This allows us to find mean structural breaks under a non-parametric framework, without any strong assumptions on the underlying distribution. We argue that this helps us identify the ideal "burn-in" period for the chain, and subsequently improves the MCMC estimates.

# 2 Literature Review

The existing literature on a non-parametric change point analysis is wide, however, a discussion on using change-point methods to identify the initial "transient" (or "burn-in") period seems to be missing. In particular, the methods based on the CUSUM max-type statistics described in [1] and [9] suffer from a few major shortcomings as described below. We look at few of these methods in the Appendix.

- The CUSUM method described is "offline" in the sense that while we're dealing with sequential data, the chain is not processed in a strictly sequential (or "online") sense. Max-type statistics that operate on the whole process generally identify the most significant change-point rather than the first significant one. Hence, we require methods that are tuned to identify change-point in the initial part of the chain.
- The power of a max-type statistic is maximum when the two fragments of data are of roughly equal size and is significantly affected when we move to either side.
- Another cause for concern in hypothesis based testing is the assumption of stationarity as the null hypothesis. These methods require the estimation the long-run variance-covariance matrix under the stationarity assumption. However, the chain is often not stationary; the estimates are biased and tend to affect the performance of the test.

- Another potential source of error are multivariate chains with high correlation among components. Normalising CUSUM by the covariance estimator tends to bias the change-point towards the faster converging components.

To summarize, a method that is robust to changes after the first significant change-point, avoids hypothesis based testing and variance estimation is of interest. The rest of the paper is organised as follows, we describe our modified CUSUM method for both univariate and multivariate chains in section 3, then we present a few examples in section 4, we make some additional considerations and recommendations in section 6.

# 3 CUSUM Local-Max method

## 3.1 Univariate sample

We start with the hypothesis that once the Markov chain reaches an area of high probability under a distribution $F$, it is likely to remain in thereabouts. To rephrase, we use the convergence in $1^{st}$ moment as a proxy for T.V. convergence. Note that this is a reasonably valid claim as for a univariate case, there is no logical reason for the chain to move away from there. It may happen that the chain has high variance around that region, however, that is of lesser interest. We then look at CUSUM process.

Assume $\{X_1, X_2, ...X_n\}$ to be a Markov chain. Define the CUSUM process $Z_n = \{Z_n(x) : x \in [0,1]\}$ of the observations by:

$$Z_n(k) = \frac{1}{\sqrt{n}}\left(\sum_{t=1}^{k} X_t - \frac{k}{n}\sum_{t=1}^{n} X_t\right), \quad k \in \{1, 2, ..., n\} \tag{1}$$

where $\lfloor . \rfloor$ denotes the greatest integer. Note that $Z_n$ is a zero-mean process.
Since the observations are correlated, we need additional conditions for an FCLT to hold for $\alpha$-mixing processes. Then under these conditions, the following weak convergence holds,

$$Z_n \Rightarrow \omega B \quad (n \to \infty), \tag{2}$$

in the Skorohod space $D[0,1]$. Here, $B = (B(t) : t \in [0,1])$, with $B(t) = W(t) - tW(1)$, is a standard Brownian bridge and $\omega > 0$ is the scaling parameter. In application, $\omega^2$ is the long-run variance and can be estimated using the Bartlett estimator[3] $\hat{\omega}_n^2$. This estimator is weakly consistent, and (2) implies that

$$Z_n \Rightarrow \hat{\omega}_n B \quad (n \to \infty), \tag{3}$$

By definition, the Brownian bridge is a classical Brownian motion defined on the interval $[0,1]$ with property $B(0) = B(1) = 0$. It is useful for modeling a system that starts at some

given level and is expected to return to that level at some specified future time. Note that $B(0) = 0$, and $E[B(t)] = 0$. Since Brownian bridge has to return to 0, we make use of the fact that $|B(t)|$ has to dip and achieve a maxima(local or global). A max-type statistic, $\max|B|$ identifies only the global max, however, we are interested in the local maxima. In particular, the first dipping point will indicate the movement of the chain around the mean, and since we are interested in the earliest change-point, the first local maxima would be the estimate of our change-point.

Define the first difference $Y_n = \{Y_n(x) : x \in [0,1]\}$ as,

$$Y_{k+1} = |Z_n(k+1)| - |Z_n(k)| \tag{4}$$

Note that for a local maxima $Y_{k+1} < 0$, since $Y_n$ is an increasing function till the first maxima. This could only occur if,

$$\begin{aligned} Z_n(k+1) &= Z_n(k) + \frac{1}{\sqrt{n}}\left(X_{k+1} - \bar{X_n}\right) \\ Z_n(k+1) - Z_n(k) &= \frac{1}{\sqrt{n}}\left(X_{k+1} - \bar{X_n}\right) \\ \frac{1}{\sqrt{n}}\left(X_{k+1} - \bar{X_n}\right) < 0 &\rightarrow X_{k+1} < \bar{X_n} \end{aligned}$$

where $\bar{X_n}$ is the sample mean. This is to say, that once chain reaches the sample mean, we expect it to dip and thus produce a local maxima.

Hence we estimate the change point as the first local maxima (or dipping point):

$$\hat{k} = \min_k \ \{k \in [1,n] \ \mid Y_k < 0\} \tag{5}$$

Note that this method is tuned to find the first significant change point, and is therefore robust to changes occurring past the estimation point. Further, this avoids a hypothesis based test which requires the estimation of long-run variance altogether.

## 3.2 Multivariate sample

The next natural step is to extend to a multivariate setup. We want to avoid joint covariance estimation as discussed above, therefore it is reasonable to process each component individually. Note that the marginal samples are not neccessarily Markovian. Even though we lose the Markov structure, we retain the $\alpha$-mixing properties and the method is valid. We argue that just like in the univariate case, once the Markov chain reaches a "good" area of high density, it is likely to remain in thereabouts. However, it is likely that some components have a slower rate of convergence and cause the chain to move as a whole. Therefore, we define our "good" area as the point when *all* the components are in a high density region marginally. This is a reasonable assumption as the only logical reason for one of the components to move

away from a high density region would be due to its correlation with a component that is still in a low density area.

Formally, assume $\{X_1, X_2, ...X_n\}$ to be a $d$-dimensional Markov Chain, where each observation $X_i = (X_{i1}, X_{i2}, ..., X_{id})'$. Define the $d$-dimensional CUSUM process by:

$$\begin{gathered} Z_n = \left(Z_n^1, Z_n^2, ..., Z_n^d\right)' \quad \text{where,} \\ Z_n^i(k) = \frac{1}{\sqrt{n}} \sum_{t=1}^{k} (X_{it} - \bar{X_{in}}), \quad k \in \{1, 2, ..., n\}; \quad i = 1, 2..., d \end{gathered} \tag{6}$$

where $\lfloor . \rfloor$ denotes the greatest integer.

Since the observations are correlated, we need additional conditions for an FCLT to hold for $\alpha$-mixing processes. Then under these conditions, the following weak convergence holds,

$$Z_n \Rightarrow W_\Omega \quad (n \to \infty) \tag{7}$$

in the d-dimensional Skorohod space $D^d[0, 1]$ holds. The limit $W_\Omega = (W_\Omega(x) : x \in [0, 1])$ is a d-dimensional Brownian bridge with covariance matrix $\Omega$, i.e. $W_\Omega$ is Gaussian with $E[W_\Omega(x)] = 0$ and $E[W_\Omega(x)W_\Omega(y)'] = \Omega \min\{x, y\}$. One can construct the weak estimator for the long run covariance matrix as

$$\hat{\Omega} \xrightarrow{p} \Omega \quad (n \to \infty). \tag{8}$$

If (7) and (8) are satisfied, one gets

$$Z_n \Rightarrow \hat{\Omega}^{1/2} W = (B_1, B_2, ..., B_d)' \quad (n \to \infty) \tag{9}$$

where $B_1, ..., B_d$ denote independent standard Brownian bridges.

Now we apply the univariate process as in section 3.1 to each component. The change-point is estimated by picking the maximum of all estimated change-points as follows:

$$\begin{gathered} \hat{k}^i = \min_k \ \left\{k \in [1, n] \ \mid \ Y_k^i < 0\right\}, \quad i = 1, 2, .., d \\ \hat{k} = \max_k \ \left\{\hat{k}^i\right\} \end{gathered} \tag{10}$$

**Note:** Here, we use the principle that a Markov chain converges to its stationary distribution implies all the individual marginal processes will converge to their marginal target distributions. Of course, a challenge is that convergence of the marginal distributions does not imply convergence of the joint, however, this is the best we can do. Further, this method avoids a hypothesis based test which requires the estimation of long-run variance altogether. The method although is applied component wise, it takes into consideration the overall structure and picks the first change-point for the Markov chain as whole.

To summarize our discussion we present the following algorithm(for univariate $d = 1$),

**Algorithm 1** CUSUM Local-Max algorithm

1: Given N samples, $\{X_1, X_2, ...X_n\}$ from a $d$-dimensional Markov Chain.
2: Define CUSUM process, $Z_n^i(k) \quad k = 1, 2.., N; i = 1, 2.., d.$
3: For each $i$, calculate the first local maxima, and identify the change point $\hat{k}^i$ as $\hat{k^i} = \min_k \{k \in [1, n] \mid Y_k^i < 0\}$
4: Find $\hat{k}$ as $\max_k \left\{\hat{k}^i\right\}$ and identify the component $i = \arg\max_i \hat{k}^i$.

### 3.3 Modification to the Local-Max method

Note that in the CUSUM process we estimate our sample mean using the whole chain. Since the initial transient period will cause a bias in the mean estimate, we propose a running mean estimate which estimates the sample mean at a point as the mean of data points after that point. Repeat the procedure in section 3.2 by replacing $Z_n$ by $A_n$.

Formally, assume $\{X_1, X_2, ...X_n\}$ to be a $d$-dimensional Markov Chain, where each observation $X_i = (X_{i1}, X_{i2}, ..., X_{id})'$. Define the $d$-dimensional running CUSUM process by:

$$A_n = \left(A_n^1, A_n^2, ..., A_n^d\right)' \quad \text{where,}$$
$$A_n^i(k) = \frac{1}{\sqrt{n}}\left(\sum_{t=1}^{k} X_{it} - \frac{k}{n-k}\sum_{t=k+1}^{n} X_{it}\right), \quad k \in \{1, 2, ..., n\}; \;\; i = 1, 2, .., d \tag{11}$$

To summarize our discussion we present the following algorithm(for univariate $d = 1$),

**Algorithm 2** CUSUM Modified Local-Max algorithm

1: Given N samples, $\{X_1, X_2, ...X_n\}$ from a $d$-dimensional Markov Chain.
2: Define CUSUM process, $A_n^i(k) \quad k = 1, 2.., N; i = 1, 2.., d.$
3: For each $i$, calculate the first local maxima, and identify the change point $\hat{k}^i$ as $\hat{k^i} = \min_k \{k \in [1, n] \mid Y_k^i < 0\}$
4: Find $\hat{k}$ as $\max_k \left\{\hat{k}^i\right\}$ and identify the component $i = \arg\max_i \hat{k}^i$.

## 4 `cptmcmc`: R package

An `R` package on GitHub has been developed for the methods developed in the previous section. The package contains the following function for change-point analysis:

```
library(cptmcmc)
##returns a dataframe with the truncated_chain and estimated change-points
truncated_chain <- cptmcmc(chain, method = 2, logpost = FALSE,
                           plot = TRUE, thresh = 0.5)
```

- `chain:` matrix or vector of the Markov chain to be analysed. For matrix each column is treated as a component.
- `method:` '1' for local-max method in section 3.1 and 3.2, '2' as in section 3.3
- `logpost:` F by default, T if using log-posterior vector. Discussed in section 6.1
- `plot:` T by default, plots the change-point on the trace-plot of the component with the maximum change-point
- `thresh:` in (0,1). Threshold of maximum viable break-pt fraction.

# 5 Examples

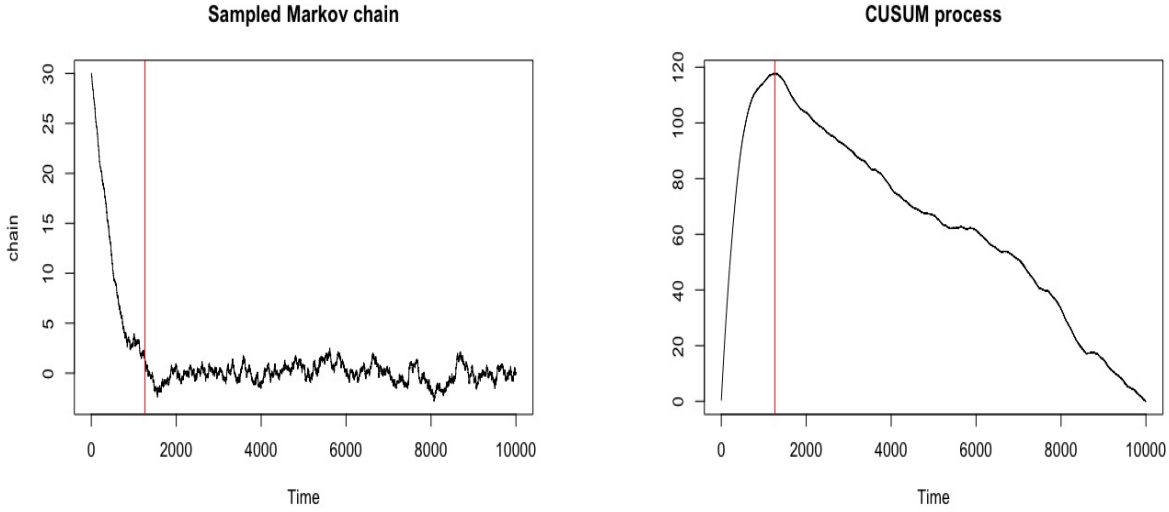
## 5.1 Local-Max method: MH for Normal distribution

Let us generate a Markov chain using the Metropolis-Hastings (MH) algorithm. Consider a univariate standard normal distribution, $f(x) \equiv N(0,1)$ target, and at $n^{th}$ step, the proposal is $g(x) \equiv N(x_{n-1}, h^2)$. $h$ is the step size standard deviation to be user optimised.

Assume that $\{X_1, X_2, ...X_n\}$ be a Markov chain sampled from the above MH algorithm. Define the univariate CUSUM process $Z_n$ as in equation (1). Now, we look at few test runs using different values of $h$ and $X_1$. This gives us a different correlation structure, and/or convergence period each time.

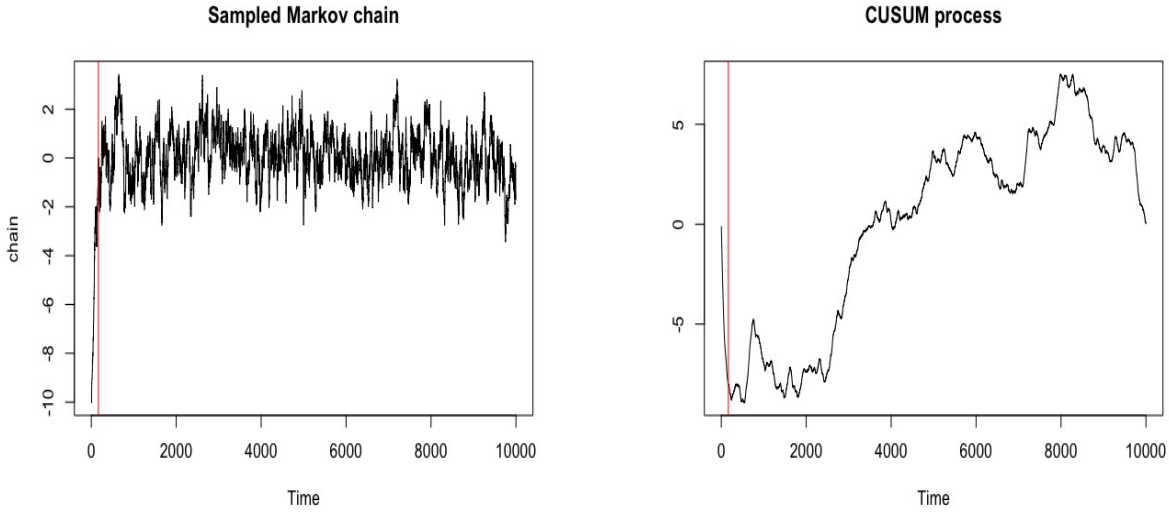
Figure 1,2 and 3 show the Markov chain (with different starting values and step sizes) on the left, and the CUSUM process $Z_n$ on the right. The red line indicates our change-point estimates. Figure 2 shows the minima, however, notice that for the absolute proccess, the indicated value is the first local maxima.

# Changepoint for MCMC



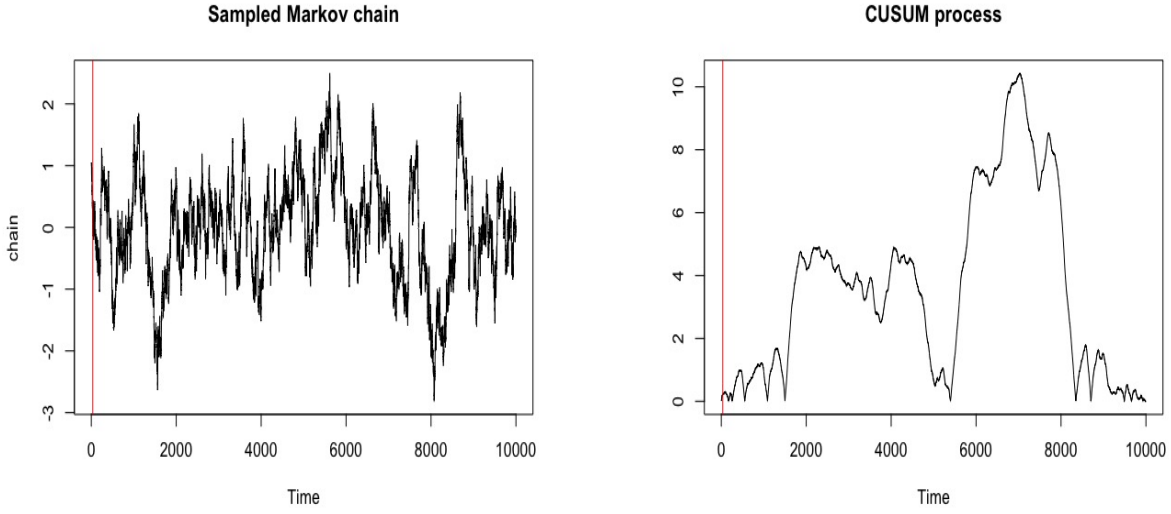

(a) Chain (b) CUSUM

Figure 1: $h = 0.10$, start $= 30$, estimate $= 1264$




(a) Chain (b) CUSUM

Figure 2: $h = 0.25$, start $= -10$, estimate $= 166$




(a) Chain (b) CUSUM

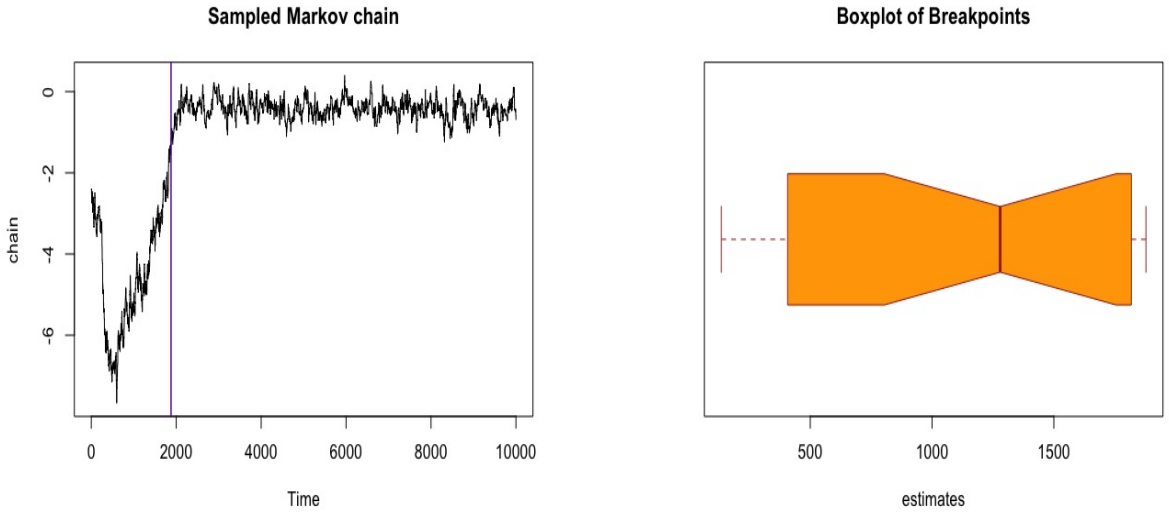
Figure 3: $h = 0.10$, start $= 1$, estimate $= 35$

## 5.2 Local-Max method: Bayesian Multinomial Regression

Assume that $\{X_1, X_2, ...X_N\}$ be a Markov chain sampled from a Bayesian multinomial regression model as described.

Consider the 1989 Dutch parliamentary election study (Anker and Oppenhuis, 1993) which contains 1754 self-reported voting choices of survey respondents among four contesting parties and 11 explanatory variables. The dataset is available in `Nethvote` in the `R` package `MCMCpack` (Martin et al.,2011). We consider a Bayesian multinomial regression with four levels for the response variable, each for a contesting political party. To generate our Markov chain sample, we use the `MCMCmnl` function in the `R` package `MCMCpack` which uses a random walk Metropolis-Hastings sampler with a Gaussian proposal distribution.

Define the $d$-dimensional CUSUM process $Z_n$ as in equation (6). Then we run the change-point algorithm component wise and get the following results. Figure 4 shows the component with the maximum estimated change point on the left and a boxplot of all change-points on the right.

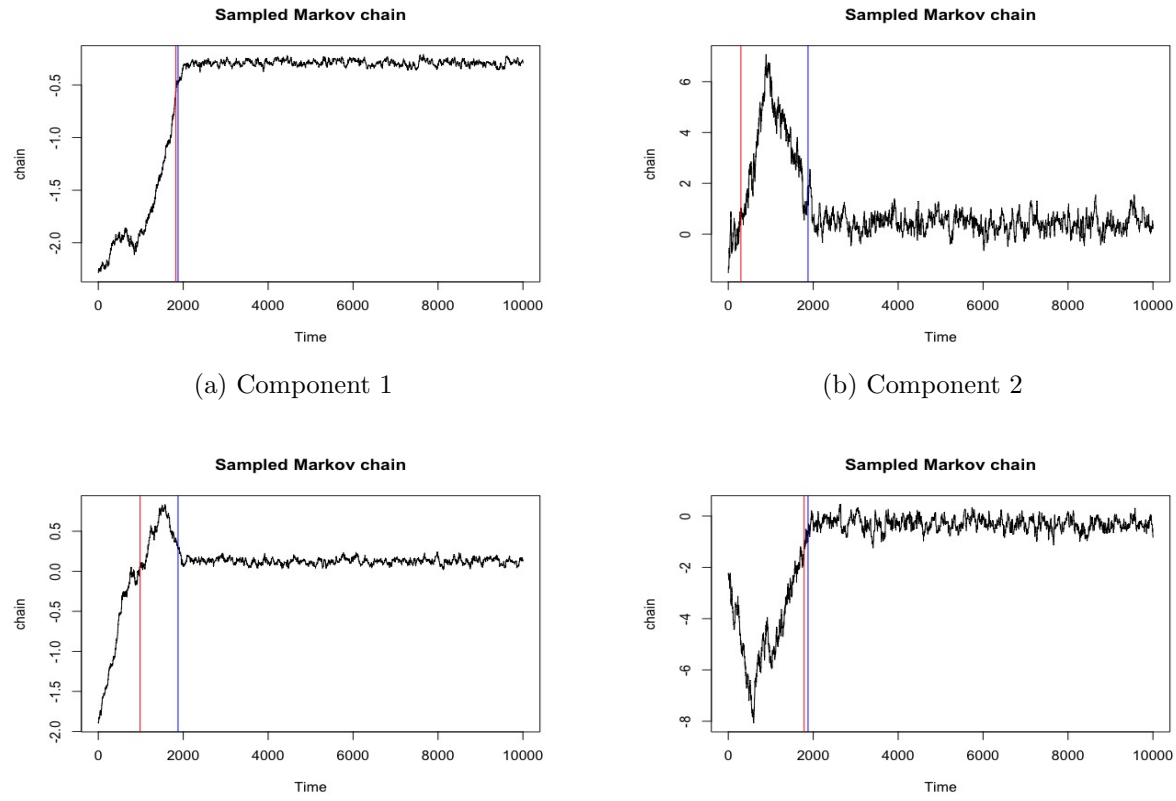



(a) Component

(b) Boxplot

Figure 4: estimate = 1878, component = $20^{th}$

Figure 5 shows 4 different kind of component behaviour in this Markov Chain, the red line on each one represent estimated change-point for the respective component, while the blue-line is the final change-point estimate. It can be clearly seen that the method takes the correlation among components into account for the final estimate.




(a) Component 1

(b) Component 2

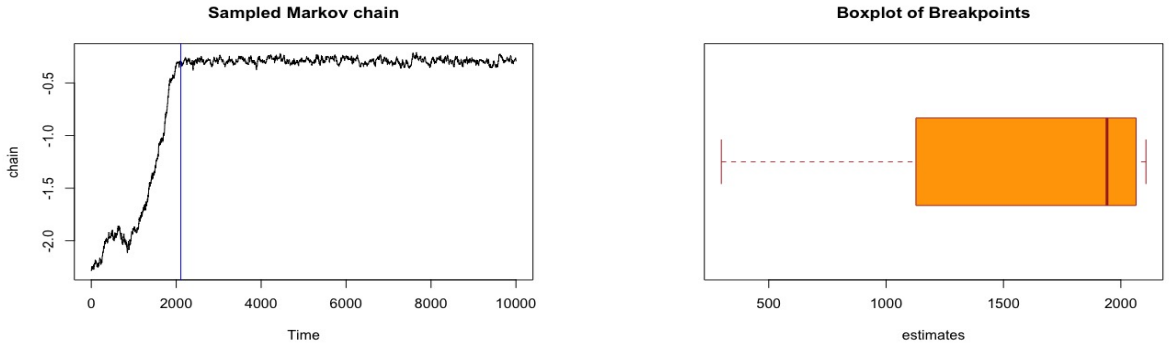
(c) Component 3

(d) Component 4

Figure 5: Final estimate = 1878

## 5.3 Modified Local-Max method: Multinomial Regression

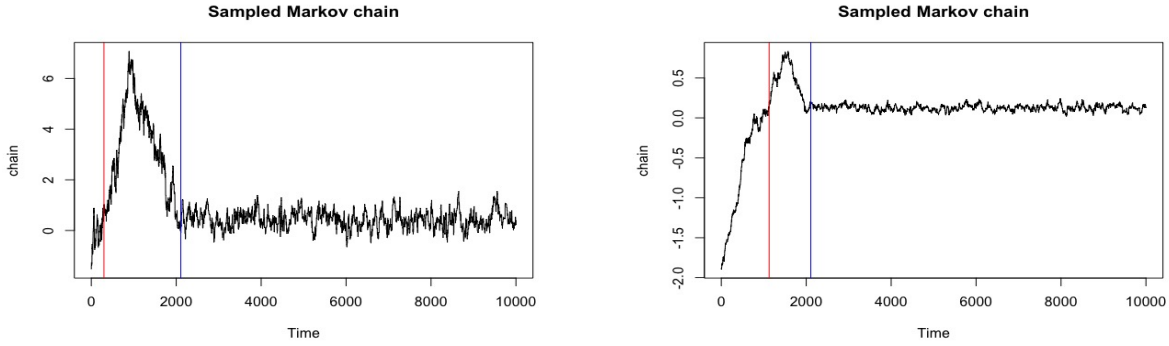
Figure 6 shows the component with the maximum estimated change point on the left and a boxplot of all change-points on the right.




(a) Component (b) Boxplot

Figure 6: estimate = 2107, component = $1^{st}$

Figure 7 is the corresponding analog to figure 5 in section 4.2




(a) Component 1 (b) Component 2

Figure 7: Final estimate = 2107

# 6 Additional considerations

## 6.1 Log-Posterior Method

In this section, we discuss a parametric change-point algorithm for when the sample distribution is known or the log-posterior values are available. Let $\{X_1, X_2, ...X_n\}$ be a $d$-dimensional Markov Chain, where each observation $X_i = (X_{i1}, X_{i2}, ..., X_{id})'$ and let $f_X(x_1, x_2, .., x_d)$ be the pdf of $X$.

Define $g : \{X_1, X_2, ..., X_n\} \to \mathbb{R}$ such that:

$$g(X_i) \mapsto \log\left(f_X(X_{i1}, X_{i2}, ..., X_{id})\right) \tag{12}$$

Note that we retain the Markov structure as $g$ is a continuous mapping on $\mathbb{R}$. We now apply the Local-Max CUSUM procedure in section 3.1 by replacing $X_t$ with $g(X_t)$.

$$G_n(k) = \frac{1}{\sqrt{n}}\left(\sum_{t=1}^{\lfloor nk \rfloor} g(X_t) - \frac{\lfloor nk \rfloor}{n}\sum_{t=1}^{n} g(X_t)\right), \quad k \in [0, 1] \tag{13}$$

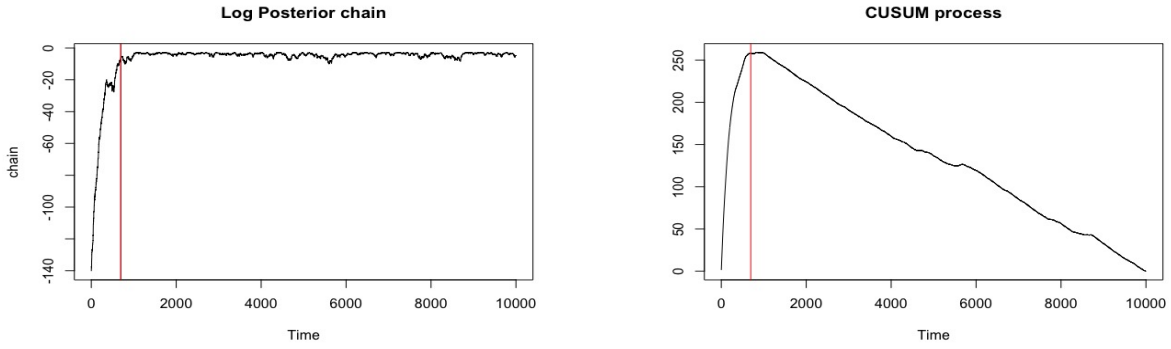
The basic argument for this method is that the trace of a log-posterior density is "directional" in the sense that the points in higher density area will have a higher log-posterior value and thus, mimics the trace plot of the chain as a whole. This eliminates the need to assume a marginal convergence in the components and takes the covariance structure of the distribution into account explicitly. Note that this method reduces the time and space complexity by projecting the dimensions onto a real line.

Consider a Markov chain generated using the Metropolis-Hastings (MH) algorithm. Consider a $d$-variate standard normal distribution, $f(x) \equiv N_d(\underset{\sim}{0}, I_d)$ as the target, and at $n^{th}$ step, the proposal is $\pi(x) \equiv N_d(x_{n-1}, \Sigma)$.

Let $\{X_1, X_2, ...X_N\}$ be a Markov Chain sampled from the above MH algorithm. Then we have $g(X_t)$ and we define $G_n$ as in equation (13):

$$g(X_t) = -\frac{d}{2}\log(2\pi) - \frac{1}{2}X_t' X_t \tag{14}$$

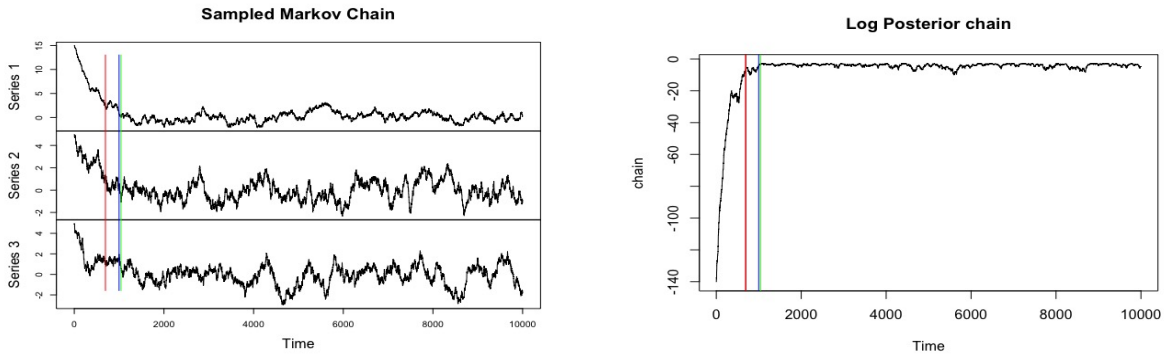
Figure 8 shows the log-posterior on the left and CUSUM process on the right with estimated change-point being the vertical red line.




(a) Log chain (b) CUSUM

Figure 8: $\Sigma = diag(0.10, 0.10, 0.10)$, start $= c(15, 5, 5)$, estimate $= 694$

Figure 9 shows the sampled chain components on the left and the log chain on the right with log-posterior estimate (current section) in red, local-max estimate (section 3.1, without using the posterior information) in blue, and modified local-max (section 5.1) in green.




(a) Chain

(b) Log-Chain

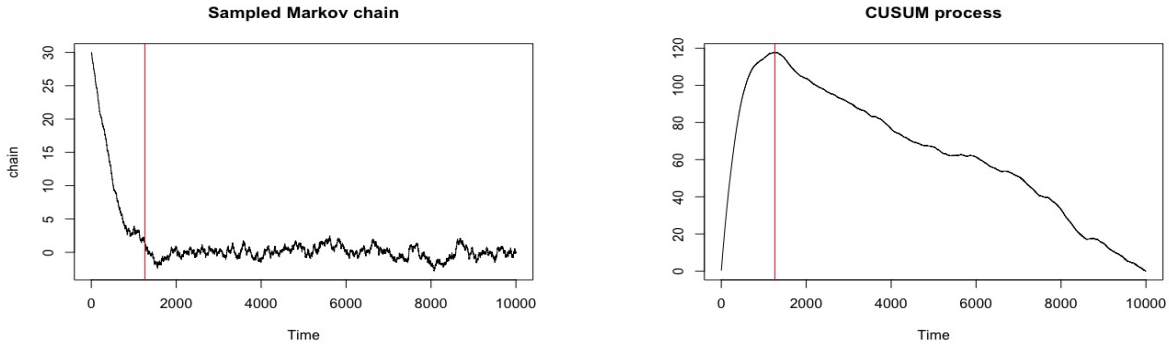
Figure 9: estimate:red = 694, estimate:blue = 999, estimate:green = 1042

## 6.2 Application to empirical distribution

In this section, we look at the equivalent of a Kolmogorov-Smirnov test for Markov chains. Note that we will compare the empirical distribution of the sample to the true distribution and demonstrate that burn-in is an essential tool for better estimates.

Let us generate a Markov chain using the Metropolis-Hastings (MH) algorithm. Consider a univariate standard normal distribution, $f(x) \equiv N(0,1)$ target, and at $n^{th}$ step, the proposal is $g(x) \equiv N(x_{n-1}, h^2)$. $h$ is the step size standard deviation to be user optimised. Let $h = 0.1$, and starting point, $start = 30$. Let us also generate a random sample $Y$ from $N(0,1)$, which will serve as our true sample.

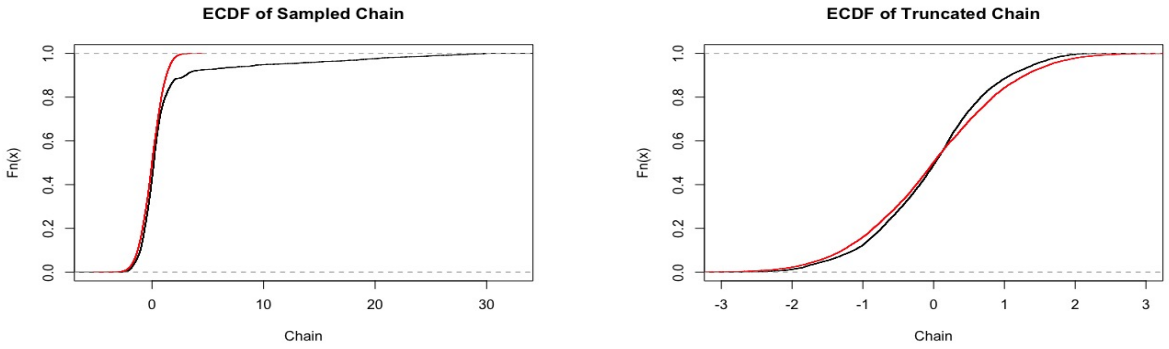
Figure 10 shows the Markov chain on the left and CUSUM process on the right. The red line indicates our change-point estimate.




(a) Chain (b) CUSUM

Figure 10: $h = 0.10$, start $= 30$, estimate $= 1264$

Figure 11 shows the empirical distribution of the sampled Markov chain vs the true sample $Y$ (red) on the left, and the empirical distribution of the truncated Markov chain (after removing burn-in as suggested by the algorithm) vs the true sample $Y$ (red) on the right.




(a) Sampled Chain (b) Truncated Chain

Figure 11: Truncated chain estimates are much closer to the truth

# A Appendix

In this section, we review the methods based on the CUSUM max-type statistics described in [1] and [9]. Note that these methods suffer from the shortcomings as described in section 2.

## A.1 Structural breaks in mean for univariate processes

Let $\mathbb{Z}$ be the set of integers. Let us consider the model

$$Y_t = \mu_t + \epsilon_t, \quad t \in \mathbb{Z} \tag{15}$$

where, $\epsilon_t$ is the noise component with $E[\epsilon_t] = 0$ and $E[\epsilon_t^2] = \sigma^2$.
Assume that n observations $Y_1, ..., Y_n$ have been taken from the real-valued stochastic process in (15), and that we are interested in testing the null hypothesis of constant means:

$$H_0 : \mu_1 = ... = \mu_n = \mu$$

against the alternative that $H_0$ is not true and that mean has changed at least once. The rescaled CUSUM process $Z_n = (Z_n(x) : x \in [0, 1])$ of the observations is given by:

$$Z_n(x) = \frac{1}{\sqrt{n}} \left( \sum_{t=1}^{\lfloor nx \rfloor} Y_t - \frac{\lfloor nx \rfloor}{n} \sum_{t=1}^{n} Y_t \right), \quad x \in [0, 1] \tag{16}$$

where $\lfloor . \rfloor$ denotes the greatest integer. Note that, under $H_0$, $Z_n(x)$ is independent of the unknown mean $\mu$. Since the observations are correlated, we need additional conditions for an FCLT to hold. Define the partial sums process, $S_n = (S_n(x) : x \in [0, 1])$ by:

$$S_n(x) = \frac{1}{\sqrt{n}} \sum_{t=1}^{\lfloor nx \rfloor} \epsilon_t, \quad x \in [0, 1]$$

Then under these conditions, the following weak convergence holds

$$S_n \Rightarrow \omega W \quad (n \to \infty) \tag{17}$$

in the Skorohod space $D[0, 1]$. Here, $W = (W(t) : t \in [0, 1])$ denotes a standard Brownian motion and $\omega > 0$ is the scaling parameter. If (17) holds, one also gets

$$Z_n \Rightarrow \omega B \quad (n \to \infty), \tag{18}$$

where $B = (B(t) : t \in [0, 1])$, with $B(t) = W(t) - tW(1)$, is a standard Brownian bridge. In applications, $\omega^2$ is the long-run variance and will be estimated using the Bartlett estimator $\hat{\omega_n^2}$. This estimator is weakly consistent, and (18) implies that

$$\frac{1}{\hat{\omega_n}} Z_n \Rightarrow B \quad (n \to \infty),$$

.

With this, we can now construct the test for $H_0$. Evaluated at $x = k/n$, the CUSUM process $Z_n(x)$ basically compares the sample mean of observations upto lag $k$ with global mean of all observations.

Since the location of the break is unknown, we check for all choices of $k \in \{1, 2, ..., n\}$. This leads to a max-type statistics whose values should be 'small' if $H_0$ holds, and 'larger' otherwise. Using Continuous mapping theorem, we have

$$M_n = \frac{1}{\hat{\omega_n}} \max_{1 \le k \le n} Z_n\left(\frac{k}{n}\right),$$
$$M_n \xrightarrow{\mathbb{D}} M = \sup_{0 \le x \le 1} B(x) \quad (n \to \infty)$$

where $\xrightarrow{\mathbb{D}}$ signifies convergence in distribution. The distribution for $M$ is tabulated in Shorack and Wellner(1986). The testing procedures based on $M_n$ work best if the structural breaks occur close to the sample center, so that pre-break and post-break sample are of similar size. Power naturally decays if the breaks appear early or late.

To improve detection ability for this latter case, we introduce the weighted versions of the CUSUM procedure. This leads to test statistics based on functionals of the type $W_n = \sup_x Z_n(x)/(\hat{\omega_n} q(x))$, where supremum is taken over the interval $[\epsilon_n, 1 - \epsilon_n] \subset [0, 1]$ and $q$ is the weight function. We choose $q(x) = \sqrt{x(1-x)}$ as it is the standard deviation of the Brownian bridge $B(x)$. If $\epsilon \to \epsilon \in (0, 1/2)$, then by continuous mapping theorem,

$$W_n \xrightarrow{\mathbb{D}} W = \sup_{x \in [\epsilon, 1-\epsilon]} \frac{B(x)}{q(x)} \quad (n \to \infty)$$

Critical values for an application of the truncated CUSUM test can now be obtained via simulation of the limit process, whose distribution depends on the amount of truncation $\epsilon$.

Once a structural break has been detected, the time of its occurrence has to be estimated. The CUSUM can also be used to locate the break points. Under $H_a$, there is exactly one break at time $k^* = \lfloor n\kappa \rfloor$ with some $\kappa \in (0, 1)$. Notice that $k^*$ itself cannot be estimated consistently, but that the break point fraction $\kappa$ can be consistently estimated by

$$\hat{\kappa}_n = \frac{1}{n} \arg\max_k Z_n\left(\frac{k}{n}\right).$$

## A.2 Structural breaks in mean for multivariate processes

Let us look at the model in (15), but this time $Y_t$ is a process taking values in $\mathbb{R}^d$. $\epsilon_t$ are stationary with $E[\epsilon_t] = 0$ and $E[\epsilon_t \epsilon_t'] = \Sigma$. We wish to test the hypothesis $H_0$ as in (2) against the one-change alternative. Define the analogous d-dimensional CUSUM process $Z_n$ using (16), and d-dimensional partial sums $S_n$.

Under conditions described in Aue et al.(2009b) and Bradley (2007), the weak convergence

$$S_n \Rightarrow W_\Omega \quad (n \to \infty) \tag{19}$$

in the d-dimensional Skorohod space $D^d[0,1]$ holds. The limit $W_\Omega = (W_\Omega(x) : x \in [0,1])$ is a d-dimensional Brownian motion with covariance matrix $\Omega$, i.e. $W_\Omega$ is Gaussian with $E[W_\Omega(x)] = 0$ and $E[W_\Omega(x)W_\Omega(y)'] = \Omega \min\{x, y\}$. One can construct the weak estimator for the long run covariance matrix as

$$\hat{\Omega} \xrightarrow{p} \Omega \quad (n \to \infty). \tag{20}$$

If (20) and (21) are satisfied, one gets

$$Q_n = Z_n' \hat{\Omega_n}^{-1} Z_n \Rightarrow Q = \sum_{l=1}^{d} B_l^2 \quad (n \to \infty)$$

where $B_1, ..., B_d$ denote independent standard Brownian bridges. Using (14) and continuous mapping theorem we have,

$$M_n = \max_{1 \le k \le n} Q_n\left(\frac{k}{n}\right) \xrightarrow{\mathbb{D}} M = \sup_{0 \le x \le 1} Q(x) \quad (n \to \infty)$$

The distribution for $M$ is tabulated in Shorack and Wellner(1986). Weighted versions of the multivariate CUSUM process, such as

$$W_n = \sup_{I_n \le t \le u_n} \frac{Q_n(t)}{t(1-t)},$$

with lower bound $I_n = 1/n + 1$ and upper bound $u_n = n/n + 1$, which are more sensitive to find structural breaks in the beginning and the end of the sample, can also be used.

## A.3 An alternate statistic using Self-Normalization

The basic idea of this section is to extend the self-normalization (SN) idea (Lobato 2001; Shao 2010) into the change-point detection problem. We look at a new test statistic that bypasses the direct estimation of $\omega_n^2$ in equation (19). The SN method is a special case of the fixed-b paradigm, as advocated by Kiefer and Vogelsang (2005).

Under the null hypothesis of no change, we have equation (18). The well-known Kolmogorov–Smirnov test statistic is defined as:

$$KS_n = \sup |Z_n(\lfloor nx \rfloor)/\hat{\omega_n}| = \sup_{k=1,...,n} |Z_n(k)/\hat{\omega_n}|$$

where $\hat{\omega_n^2}$ is a consistent estimator of $\omega_n^2$ and $Z_n$ is as in equation (16). A difficult issue in practice with this is the selection of bandwidth $l$ for the estimator. A data-dependent $l$ may yield non-monotonic power.

Let us look at a modified KS test based on SN with a self-normalizer that accounts for the one change point alternative. Let $S_{t_1,t_2} = \sum_{j=t_1}^{t_2} X_j$ if $t_1 \leq t_2$ and 0 otherwise. Define the normalization process $V_n(.)$ as follows. For $k = 1, ..., n-1$ let

$$V_n(k) = n^{-2}\left[\sum_{t=1}^{k}\{S_{1,t} - (t/k)S_{1,k}\}^2 + \sum_{t=k+1}^{n}\{S_{t,n} - (n-t+1)/(n-k)S_{k+1,n}\}^2\right]$$

The test statistic is defined as,

$$G_n = \sup_{k=1,...,n-1} Z_n(k)' V_n^{-1}(k) Z_n(k),$$
$$\hat{k} = \arg\max_k Z_n(k)' V_n^{-1}(k) Z_n(k)$$

Assuming (18), limiting null distribution on $G_n$ as

$$G_n \to \sup_{r\in[0,1]} \{B(r) - rB(1)\}' V_n^{-1}(r)\{B(r) - rB(1)\},$$
$$V(r) = \int_0^r [B(s) - (s/r)B(r)]^2\, ds + \int_r^1 [B(1) - B(s) - (1-s)/(1-r)\{B(1) - B(r)\}]^2\, ds.$$

Note that the normalization factor $V_n(k)$ in our test depends on $k$, whereas the normalization for $KS$, is the same for all $k$. The SN-based test does not involve any user-chosen number or smoothing parameter. Its asymptotic null distribution is pivotal and the (approximate) critical values are tabulated through simulations.

## A.4 Testing for Brownian Bridges

The proposed methods in section A.1, A.2 require testing for stationarity and therefore a mechanism to test whether the CUSUM process looks like a Brownian Bridge. We present one of the ways to test for the same.

The Heidelberger-Welch diagnostic is based on the assumption that we have got a weakly stationary process when the chain has reached convergence. A weakly stationary process has the properties that, if $X_j$ is defined as the $j^{th}$ iteration in the sequence, the mean function $E[X^j]$ is constant in time and $Cov(\theta^j, \theta^{j+s})$ does not depend on $j$ but only on the size of $s$. This is a sensible assumption since our sequence is generated by a Markov chain and therefore satisfies full stationarity.

Let CUSUM process be defined as in (16). Then, under the null hypothesis of stationarity for large $n$, (18) holds and will be distributed approximately as a Brownian bridge. We can

use the Cramer-von Mises statistic $T = \int_0^1 B_n(t)^2 dt$, $T >$ tabulated value rejects that $B_n$ shows similar shape to a Brownian bridge, thus rejecting the hypothesis of stationarity of the chain.

# B Bibliography

1. Aue, A. and Horváth, L. (2013), Structural breaks in time series. Journal of Time Series Analysis, 34: 1-16. https://doi.org/10.1111/j.1467-9892.2012.00819.x

2. Aue, A., Hormann, S., Horva'th, L. and Reimherr, M. (2009b) Break detection in the covariance structure of multivariate nonlinear time series models. The Annals of Statistics 37, 4046–87.

3. Bartlett, P.L., Boucheron, S. & Lugosi, G. Model Selection and Error Estimation. Machine Learning 48, 85–113 (2002) https://doi.org/10.1023/A:1013999503812

4. Bradley, R. C. (2007) Introduction to Strong Mixing Conditions Vol.1–3.. Heber City, UT: Kendrick Press.

5. Gupta, K. and Vats, D. (2020). Estimating Monte Carlo variance from multiple Markov chains. arXiv preprint arXiv:2007.04229.

6. Lobato, I. N. (2001), "Testing That a Dependent Process Is Uncorrelated," Journal of the American Statistical Association, 96, 1066–1076.

7. Shao, X. (2010), "A Self-Normalized Approach to Confidence Interval Con- struction in Time Series," Journal of the Royal Statistical Society, Ser. B, 72, 343–366.

8. Shorack, G. R., & Wellner, J. A. (2009). Empirical Processes with Applications to Statistics. Society for Industrial and Applied Mathematics.

9. Xiaofeng Shao & Xianyang Zhang (2010) Testing for Change Points in Time Series, Journal of the American Statistical Association, 105:491, 1228-1240