Testing for a Change in Mean of a Weakly Stationary Time Series

Shreyas Kowshik (17MA20039)
Under the guidance of : Prof. Buddhananda Banerjee

Indian Institute of Technology Kharagpur

shreyaskowshik@iitkgp.ac.in

21/04/2022

Outline

Background

Time Series

Definition

Time series is a collection of random variables $\{X_t | t \in T\}$ over a time index set T, which might be a finite, countably infinite or an uncountable set.

- What we observe are the realized values of the time series i.e. the data set is $\{X_1 = x_1, \cdots, X_n = x_n\}$, where the x_is are some numeric or categorical values.

For example : **Population of India**, **Stock Prices**, **Rainfall in a city**, etc.

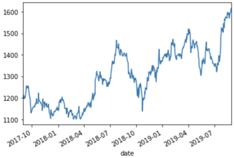


Figure: Time Series Data : Stock Prices

Time Series

Mean and Covariance

- The mean $\mu_X(t)$ of a series $\{X_t\}$ is : $\mu_X(t) = \mathbb{E}[X_t]$
- Covariance (autocovariance) function of $\{X_t\}$:
  $\gamma_X(r,s) = \mathrm{Cov}(X_r, X_s) = \mathbb{E}\left[(X_r - \mu_X(r))\,(X_s - \mu_X(s))\right]$

Weak Stationarity

A time series $\{X_t\}$ is said to be weakly stationary if :

- $\mu_X(t)$ is independent of t
- For every $h \in \mathbb{Z}$, $\gamma_X(t+h, t)$ is independent of t

Strong Stationarity

A time series $\{X_t\}$ is said to be strongly stationary if for all $k, h, t_1 \cdots t_k, x_1 \cdots x_k$, shift of the time axis does not affect the distribution i.e. $P\left(X_{t_1} \leq x_1, \cdots, X_{t_k} \leq x_k\right) = P\left(X_{t_1+h} \leq x_1, \cdots, X_{t_k+h} \leq x_k\right)$

Time Series

AR(p) Process

An $AR(p)$ (autoregressive) process of order p is defined as :

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + W_t$$

where $W_t \sim WN(0, \sigma^2)$ is white noise.

- The series $\{W_t\}$ is a white noise process.
- For the rest of this work, we will specifically consider AR processes of order 1 i.e. $AR(1)$ processes i.e. :

$$X_t = \rho X_{t-1} + W_t$$

where $|\rho| < 1$, $\rho \neq 0$

Changepoint Detection

Changepoint Detection

- Detection of the existence of an abrupt change in the distribution of a time series
- Can be change in mean, variance, parameter value, etc.
- We focus our attention on detecting a **change in mean** of a **weakly stationary** time series

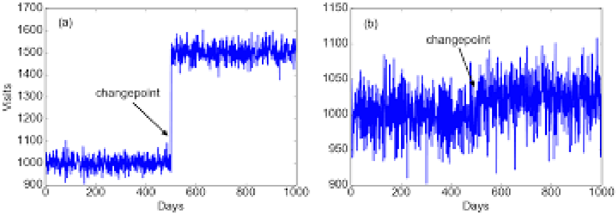


Figure: Example changepoint in a time series

Change in Mean Detection

Problem Statement

Given a sample $\mathbf{X} = \{X_1, \cdots, X_n\}$ from a time-series $\{X_t\}$, we are interested in testing the following hypothesis :

$$H_0 : \mathbb{E}[X_1] = \cdots = \mathbb{E}[X_n]$$

versus

$$H_1 : \mathbb{E}[X_1] = \cdots = \mathbb{E}[X_{k^*}] \neq \mathbb{E}[X_{k^*+1}] = \cdots = \mathbb{E}[X_n]$$

where, $1 \leq k^* < n$ is the location of the changepoint and is unknown.

- This framework is usually considered in retrospective changepoint study
- The other paradigm is online changepoint analysis

Change in Mean Detection : Contribution Overview

- We survey different statistics for the given problem statement
- Propose a new **self-normalizing statistic** that has a **sharper rise in power** upon deviation from the null hypothesis
- We show theoretical analysis and simulation studies on the same

Approaches to Change in Mean Detection

KS Statistic

Construction

Given a sample $\mathbf{X} = \{X_1, \cdots, X_n\}$ from a weakly stationary time-series $\{X_t\}$ and defining $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, construct :

$$T_n(\lfloor nt \rfloor) = \frac{1}{\sqrt{n}} \sum_{t=1}^{\lfloor nt \rfloor} (X_t - \overline{X}_n)$$

Kolmogorov-Smirnov Statistic

KS_n statistic is defined as : $KS_n = \sup_{t \in [0,1]} \left| \frac{T_n(\lfloor nt \rfloor)}{\hat{\sigma}_n} \right|$

KS Statistic

Kolmogorov-Smirnov Statistic

KS_n statistic is defined as : $KS_n = \sup_{t \in [0,1]} \left| \frac{T_n(\lfloor nt \rfloor)}{\hat{\sigma}_n} \right|$

- $\hat{\sigma}_n$ is a consistent estimator of σ where $\sigma^2 = lim_{n \to \infty} n\text{Var}(\overline{X}_n) = \Sigma_{k \in \mathbb{Z}} \gamma(k)$ is the long run variance
- Estimating $\hat{\sigma}_n$ generally requires using a kernel-based estimate : $\hat{\sigma}_n^2 = \sum_{k=-l_n}^{l_n} \hat{\gamma}(k) \mathbb{K}\left(\frac{k}{l_n}\right)$
- Here l_n is a bandwidth parameter which can be a function of sample size n or be chosen from the data
- l_n as function of sample size : **not adaptive to presence of changepoint**
- l_n that is data-dependent : **can introduce bias in estimation of σ^2 under alternative hypothesis**

Self-Normalizing Statistic

- Construct a statistic which is pointwise scaled with its estimated pointwise standard deviation
- This construction can help avoid direct estimation of σ^2.

$\widetilde{KS}_n$ Statistic

$\widetilde{KS}_n$ statistic is defined as (Shao'10) :

$$\widetilde{KS}_n = \sup_{t \in [0,1]} \left| \frac{T_n(\lfloor nt \rfloor)}{D_n} \right|$$

where $D_n^2 = n^{-2} \sum_{t=1}^{n} (\sum_{j=1}^{t} (X_j - \bar{X}_n)^2)$.

- No need to estimate σ^2 : Avoids bandwidth selection

$\widetilde{KS}_n$ Statistic

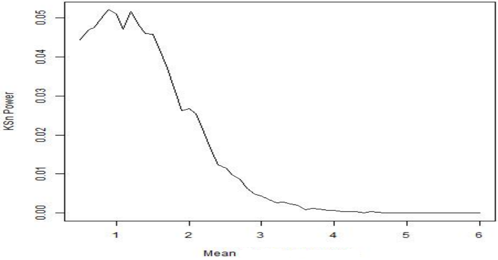


Figure: Power of $\widetilde{KS}_n$

- The mean of data before changepoint is fixed at 1 and the mean after the changepoint is varired above.
- As one moves away from the null hypothesis, the power decreases
- Reason : D_n does not take the alternative into account i.e. the presence and location of a changepoint

G_n Statistic

G_n statistic is defined as :

$$T_n(\lfloor nk \rfloor) = \frac{1}{\sqrt{n}} \sum_{t=1}^{[nk]} (X_t - \overline{X}_n)$$

$$S_{t_1,t_2} = \sum_{j=t_1}^{t_2} X_j \text{ if } t_1 \leq t_2, 0 \text{ otherwise}$$

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

$$G_n = \sup_{k=1,\cdots,n-1} T_n(k)V_n^{-1}(k)T_n(k)$$

G_n Statistic

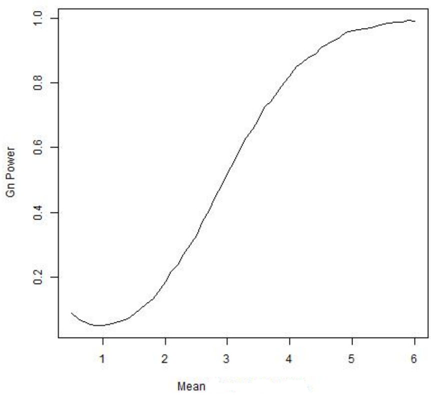


Figure: Power of G_n, $0 \leq \mu \leq 6$

G_n Statistic

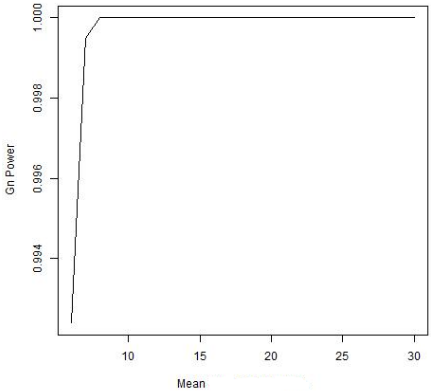


Figure: Power of G_n, $6 \leq \mu \leq 30$

Proposed Statistic

Proposed Statistic

H_n Statistic

Given a sample $\mathbf{X} = \{X_1, \cdots, X_n\}$, the H_n statistic is defined as :

$$T_r(\mathbf{X}) = \frac{1}{n}\left[\left(1 - \frac{r}{n}\right)\sum_{i=1}^{r} X_i + \left(-\frac{r}{n}\right)\sum_{i=r+1}^{n} X_i\right]$$

$$H_n = \sup_{r=1,\cdots,n-1} \frac{\sqrt{n}T_r(\mathbf{X})}{\sqrt{\sum_{|h|<n} w(r,h,n)\gamma(h)}}$$

where $w(r,h,n)$ is a weighting function the details of which we will soon derive and $r \in \{1, \cdots, n-1\}$.

- To obtain a normalized form of H_n, we need to compute $\mathrm{Var}(\sqrt{n}T_r(\mathbf{X}))$

Proposed Statistic

Variance Computation

$$\text{Var}\left[\sqrt{n}T_r(\mathbf{X})\right] = \frac{1}{n}Var\left[\mathbf{a}^T\mathbf{X}\right] = \frac{1}{n}\mathbf{a}^T\Sigma\mathbf{a}$$

$$\text{where : } \mathbf{a} = \left[\underbrace{\left(1-\frac{r}{n}\right),\cdots,\left(1-\frac{r}{n}\right)}_{\text{r times}}\underbrace{\left(,-\frac{r}{n}\right),\cdots,\left(-\frac{r}{n}\right)}_{\text{n-r times}}\right]^T$$

$$\Sigma = \text{Cov}(\mathbf{X})$$

Definining $\alpha = \frac{r}{n}, \beta = \frac{l}{n}$ for lag l, it can be shown that :
$\mathbf{a}^T\Sigma\mathbf{a} = \sum_{l=-(n-1)}^{n-1}\gamma(l)w(r,l,n)$, where :

$$w(r,l,n) = \begin{cases} n\left[\alpha\left(1-\alpha\right)-\beta\left(1-\alpha+\alpha^2\right)\right], & \text{if } \alpha \geq \beta \\ -n\beta\alpha^2, & \text{if } \alpha < \beta \end{cases}$$

Variance Computation in H_n

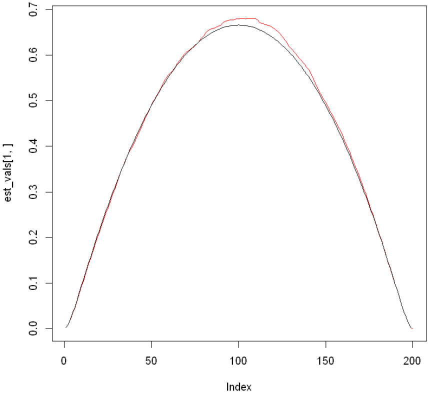
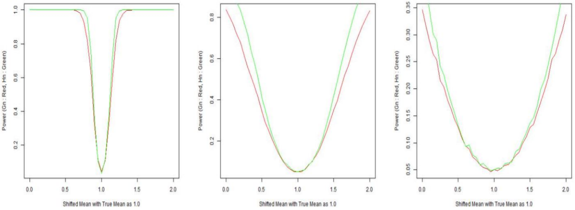


Figure: Predicted Variance (Black) v/s Sample variance of $T_r(\mathbf{X})$ (Red)

Variance Computation in H_n

- Black curve denotes variance at a particular index r predicted from the above formula for an $AR(1)$ process
- Red curve denotes sample variance at a particular index r obtained by simulating multiple $AR(1)$ processes and computing $T_r(\mathbf{X})$
- Very significant overlap!



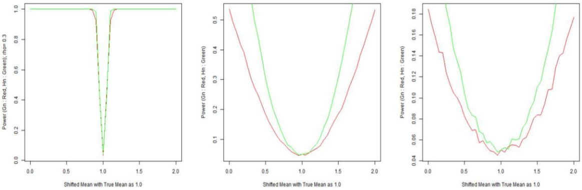
(a) $\rho = 0.7$, $\sigma_{WN} = 0.1, k = 100$

(b) $\rho = 0.7$, $\sigma_{WN} = 0.5, k = 100$

(c) $\rho = 0.7$, $\sigma_{WN} = 1, k = 100$

Figure: Power Curves comparing G_n v/s H_n statistics. Green : H_n, Red : G_n. On X-axis is plotted the new mean after the changepoint with mean before as $\mu = 1$. On Y-axis is the power of the corresponding statistic's test.
ρ:$AR(1)$ coefficient, k:changepoint location, $n = 200$, σ_{WN}:White-Noise Std Dev

Power of H_n



(d) $\rho = 0.3$, $\sigma_{WN} = 0.1, k = 100$ (e) $\rho = 0.7$, $\sigma_{WN} = 0.5, k = 50$ (f) $\rho = 0.7$, $\sigma_{WN} = 1, k = 50$

Figure: Power Curves comparing G_n v/s H_n statistics. Green : H_n, Red : G_n. On X-axis is plotted the new mean after the changepoint with mean before as $\mu = 1$. On Y-axis is the power of the corresponding statistic's test.
ρ:$AR(1)$ coefficient, k:changepoint location, $n = 200$, σ_{WN}:White-Noise Std Dev

- Power of H_n is **better (sharper)** if not the same as G_n across different values of model parameters ρ, σ_{WN}, k
- H_n shows promise to investigate it further

Normalizing Factor Estimation for Proposed Statistic

Variance Estimation

- Want to estimate : $\mathbf{a}^T \Sigma \mathbf{a} = \sum_{l=-(n-1)}^{n-1} \gamma(l) w(r, l, n)$
- Use a kernel based estimate as they are found to be consistent in the literature :

$$\hat{\sigma}_n^2 = \sum_{k=-l_n}^{l_n} w(r, k, n) \hat{\gamma}(k) \mathbb{K}\left(\frac{k}{l_n}\right)$$

  where l_n is a bandwidth parameter
- The above estimate does not account for the presence of a changepoint i.e. it does nothing special for that
- Consequently, using such an estimate can do good under the null hypothesis, but need not be that good under the alternative hypothesis
- We introduce a variable transformation to address this

Variance Estimation

Transformation of Series

Given a sample $\mathbf{X} = \{X_1, \cdots, X_n\}$, define the following : $\overline{X}_r = \frac{1}{r}\sum_{i=1}^{r} X_i$ and $\overline{\overline{X}}_r = \frac{1}{n-r}\sum_{i=r+1}^{n} X_i$. The series is then transformed as follows :

$$Z_1 = X_1 - \overline{X}_r$$
$$\vdots$$
$$Z_r = X_r - \overline{X}_r$$
$$Z_{r+1} = X_{r+1} - \overline{\overline{X}}_r$$
$$\vdots$$
$$Z_n = X_n - \overline{\overline{X}}_r$$

The transformed series $\mathbf{Z} = \{Z_1, \cdots, Z_n\}$ is used for computing the autocovariance estimates $\hat{\gamma}(h)$.

Variance Estimation

Transformation of Series

This transformation can also be seen as a matrix multiplication

$$\mathbf{Z} = Br\mathbf{X}$$

$$Br = \left(\begin{array}{c|c} I_r - \frac{1}{r}11^T & \mathbf{0} \\ \dots & \\ \hline & \\ \mathbf{0} & I_{n-r} - \frac{1}{n-r}11^T \\ & \dots \end{array}\right)$$

Thus we have $Cov(\mathbf{Z}) = Br\Sigma Br^T$

Any consistent estimator of $Cov(\mathbf{Z})$ will converge to the above result

Variance Estimation

Optimization Problem

- Under the null hypothesis, we want the variance after transformation to be same as variance of $\sqrt{n}T_r(\mathbf{X})$ i.e. $a^T\Sigma a$
- Given a sample $\mathbf{X} = \{X_1, \cdots, X_n\}$, and a $r \in \{1, \cdots, n-1\}$, define M_x to be the sample covariance matrix using only $\mathbf{X}$ and M_z the sample covariance matrix after transforming to $\mathbf{Z}$. Using $\widetilde{Br} = Br + \lambda I$ and fixing a constraint threshold ϵ, with $x \in \mathbb{R}^n$ such that $\widetilde{Br}x = a$, we have

$$\begin{aligned} \min_{\lambda} \quad & \left| a^T M_x a - (\widetilde{Br}x)^T M_z (\widetilde{Br}x) \right| \\ \text{s.t.} \quad & \left\| \widetilde{Br}x - a \right\|_2^2 < \epsilon, \lambda > 0 \end{aligned}$$

- Absolute of quadratic in scalar, thus $\exists \lambda^* \in \mathbb{R}$ such that it is minimizer

Variance Estimation

- The final variance estimate is obtained as $x^T M_z x$
- Different kernels exist for the smooth estimation of variance
- We manually select the bandwidth for the kernels by tuning parameters under the null hypothesis

Truncated: $$k_{TR}(x) = \begin{cases} 1 & \text{for } |x| \leqslant 1, \\ 0 & \text{otherwise,} \end{cases}$$

Bartlett: $$k_{BT}(x) = \begin{cases} 1 - |x| & \text{for } |x| \leqslant 1, \\ 0 & \text{otherwise,} \end{cases}$$

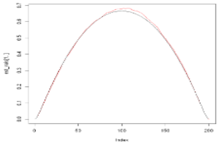
Parzen: $$k_{PR}(x) = \begin{cases} 1 - 6x^2 + 6|x|^3 & \text{for } 0 \leqslant |x| \leqslant 1/2, \\ 2(1 - |x|)^3 & \text{for } 1/2 \leqslant |x| \leqslant 1, \\ 0 & \text{otherwise} \end{cases}$$

Tukey-Hanning: $$k_{TH}(x) = \begin{cases} (1 + \cos(\pi x))/2 & \text{for } |x| \leqslant 1, \\ 0 & \text{otherwise,} \end{cases}$$

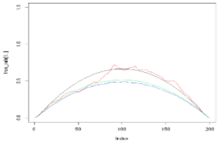
Quadratic Spectral: $$k_{QS}(x) = \frac{25}{12\pi^2 x^2} \left(\frac{\sin(6\pi x/5)}{6\pi x/5} - \cos(6\pi x/5) \right)$$

Figure: Kernel Formulations

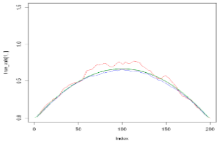
Variance Estimation



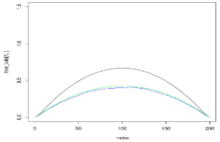
(a) True v/s Sample variance



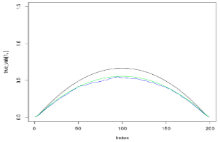
(b) Bartlett Kernel; Bandwidth=5.0 100 iterations



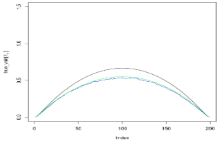
(c) Bartlett Kernel; Bandwidth=9.0 100 iterations



(d) QS Kernel; Bandwidth=5.0 100 iterations



(e) QS Kernel; Bandwidth=15.0 100 iterations



(f) QS Kernel; Bandwidth=30.0 100 iterations

Figure: Comparision of variances (X-axis : Index r, Y-axis : Variance value). Red : Sample variance of $\sqrt{n}T_r(\mathbf{X})$, Black : Variance predicted, Blue : Variance estimated after using transformation, Green : Variance estimated without using transformation, $n = 200$, AR(1) process $\rho = 0.7$

Variance Estimation

- It can be observed that over a wide range of kernels, the estimated variance with and without the transformation overlap significantly
- The estimated variance also has significant overlap with the true theoretical variance
- This shows promise in terms of using this estimate in the H_n statistic further

Variance Estimation : Power of H_n

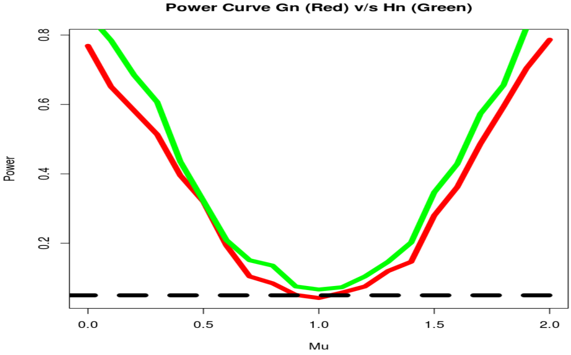


Figure: Power Curves (X-axis : Mean Value, Y-axis : Power), Red : Power of G_n, Green : Power of H_n $n = 200$, 450 iterations, AR(1) process $\rho = 0.7$, Mean before change $\mu = 1$

Variance Estimation : Power of H_n

- Power of H_n with using variance estimation has a sharper rise on deviation from the null hypothesis as compared to G_n
- This establishes our statistic's performance improvement in the given setting

Conclusion and Future Work

Conclusion and Future Work

- $\widetilde{KS}_n$ statistic suffers from **non-monotonic power** problem due to not incorporating information from alternative hypothesis
- G_n statistic takes alternative hypothesis into account and provides **monotonic power**
- Proposed self normalizing statistic H_n **is found to outperform** G_n on a wide range of model parameters under exact simulation
- A **variable transformation was introduced** to estimate the normalizer of H_n. Its **variance estimation** was conducted by framing an optimization problem.
- **Power rise was sharper** for H_n with variance estimation establishing the improvement with our proposed statistic

Future Work

- **Extensively evaluate** on different processes and parameters
- Study the **theoretical properties** and **convergence** of H_n
- Aim to **publish** the work done

Thank You

References I

Donald W. K. Andrews, *Heteroskedasticity and autocorrelation consistent covariance matrix estimation*, Econometrica **59** (1991), no. 3, 817–858.

P. C. B. Phillips, *Time series regression with a unit root*, Econometrica **55** (1987), no. 2, 277–301.

Xiaofeng Shao and Xianyang Zhang, *Testing for change points in time series*, Journal of the American Statistical Association **105** (2010), no. 491, 1228–1240.

[Phi87] [SZ10] [And91]