1 Abstract

The R package **SNSeg** provides functions to perform change-point segmentation and estimation for univariate and multivariate time series through Self-Normalization (SN). While many existing nonparametric approaches support change in only one type of parameters (e.g. mean or variance) for time series and are tuning-parameter-dependent, SN-based algorithms are offered with more convinence in any user-chosen number or smooth parameter, robust to temporal dependence, and can implement change-point estimation using change in more than one type of parameters (e.g. mean, variance and acf) for the given time series. The nested local-window segmentation algorithm (SNCP) is applied to conduct multiple change-point estimation, and an extension of SNCP, named SNHD, is designed for change-point estimation in high-dimensional time series. Graphics for time series and SN test statistics segmentation plots are available with certain options. Detailed examples are given in simulations of different MAR processes.

2 Introduction

In recent years, change-point analysis has become an increasingly research area in statistics and other related fields, including finance, economics, signal processing, and medical research to study diseases such as COVID-19. (Here needs a large number of citations and some other complement stuff) There are a number of R packages avaliable for change-point analysis, see **bf** (Erdman and Emerson 2007), **cpm** (Ross 2013), **changepoint** (Killick and Eckley 2014), and ecp (James et al. 2020) for some reviews.

In this paper, we concern the R package **SNSeg** to implement SN-based framework for time series change-point estimation and segmentation. SN techniques were first introduced by Shao (2010) for confidence interval construction in time series, and furtherly developed in Shao (2015) for univariate time series and Zhao et al. (2021) for multivariate and high-dimensional time series. (blablabla)

The rest of the paper is organized as follows: (blablabla)

3 SN Change-Point Estimation Framework

3.1 Single Change-Point Estimation

We start with a single change-point estimation in a general parameter $\theta = \theta(F_t)$ for a univariate time series $Y_{tt=1}^{n}$, where F_t denotes the CDF of Y_t and $\theta(\cdot)$ denotes a functional. The SN-based testing method is defined as

$$SN_n = \max_{k=1,\cdots,n-1} T_n(k), T_n(k) = D_n(k)^2/V_n(k),$$

where

$$D_n(k) = \frac{k(n-k)}{n^{3/2}}(\hat{\theta}_{1,k} - \hat{\theta}_{k+1,n}), \tag{1}$$

$$V_n(k) = \sum_{i=1}^{k} \frac{i^2(k-i)^2}{n^2k^2}(\hat{\theta}_{1,i} - \hat{\theta}_{i+1,k})^2 + \sum_{i=k+1}^{n} \frac{(n-i+1)^2(i-k-1)^2}{n^2(n-k)^2}(\hat{\theta}_{i,n} - \hat{\theta}_{k+1,i-1}). \tag{2}$$

If $\theta(\cdot)$ is a mean functional, SN_n is denoted as the CUSUM-based SN test statistic for all observations in Shao and Zhang (2010). If it is not a mean functional (for instance, variance, acf, etc.), SN_n is not the CUSUM-based test statistic and can be referred to to Zhang and Lavitas (2018) for more details. $T_n(k)$ is referred as the SN test statistic for the k^{th} observation. For a pre-specified threshold K_n, we declare no change point when $SN_n < K_n$. Given that SN_n exceeds the threshold, we estimate the single change-point location via

$$\hat{k} = \arg \max_{k=1,\cdots,n-1} T_n(k). \tag{3}$$

Intuitively, the change point takes place at the location of the maximum test statistic among all observations if it exceeds the SN threshold. Note that this SN-based procedure is a general framework since it can be implemented with any functional $\theta(\cdot)$ with a nonparametric estimator based on the empirical distribution. Assumptions and theoretical justifications can be found in Zhao et al. (2021) for more discussion.

3.2 Multiple Change-Point Estimation

We furtherly extend the SN-based test to multiple change-point estimation. To proceed, we assume $m_0 \geq 0$ unknown number of change points $0 < k_1 < \cdots < k_{m_0} < n$ that partition Y_t into $m_0 + 1$ stationary segments. Define $k_0 = 0$ and $k_{m_0+1} = n$, and then the i_{th} segment, denoted as

$$Y_t = Y_t^{(i)}, k_{i-1} + 1 \leq t \leq k_i, fori = 1, \cdots, m_0 + 1, \tag{4}$$

contains observations between two change points $t = k_{i-1} + 1$ and k_i, which share common feature characterized by the functional θ_i, for $i = 1, \cdots, m_0 + 1$.

To recover the unknown locations of the change points, we start by introduce some notations. Similar to the single change-point estimation framework, for $1 \leq t1 < k < t2 \leq n$, we define

$$T_n(t_1, k, t_2) = D_n(t_1, k, t_2)^2 / V_n(t_1, k, t_2), \tag{5}$$

where $D_n(t_1, k, t_2) = \frac{(k-t_1+1)(t_2-k)}{t_2-t_1+1}^{3/2}(\hat{\theta}_{t_1,k} - \hat{\theta}_{k+1,t_2})$, $V_n(t_1, k, t_2) = L_n(t_1, k, t_2) +$

$R_n(t_1, k, t_2)$ and

$$L_n(t_1, k, t_2) = \sum_{i=t_1}^{k} \frac{(i - t_1 + 1)^2(k - i)^2}{(t_2 - t_1 + 1)^2(k - t_1 + 1)^2}(\hat{\theta}_{t_1,i} - \hat{\theta}_{i+1,k}), \tag{6}$$

$$R_n(t_1, k, t_2) = \sum_{i=k+1}^{t_2} \frac{(t_2 - i + 1)^2(i - 1 - k)^2}{(t_2 - t_1 + 1)^2(t_2 - k)^2}(\hat{\theta}_{i,t_2} - \hat{\theta}_{k+1,i-1}). \tag{7}$$

Here $T_n(t_1, k, t_2)$ represents the SN test defined on the subsample $Y_{t t=t_1}^{t_2}$. In other words, we find the SN test statistic for each observation within the interval $[t_1, t_2]$. Set $t_1 = 1$ and $t_2 = n$, $T_n(t_1, k, t_2) = T_n(1, k, n)$ will be reduced to the single global test change-point estimation framework in ().

Then we combine the SN framework with a nested local-window segmentation algorithm proposed in Zhao et al. (2021). For each k, instead of one global test for $T_n(1, k, n)$, we compute a maximal SN test based on a collection of nested windows covering k. Specifically, we fix a small trimming papameter $\epsilon \in (0, 1/2)$ and define the window size $h = \lfloor n\epsilon \rfloor$. For each $k = h, \cdots, n - h$, we define the nexted local-window set $H_{1:n}(k)$ where

$$H_{1:n}(k) = (t_1, t_2)|t_1 = k - j_1 h + 1, j_1 = 1, \cdots, \lfloor k/h \rfloor; t_2 = k + j_2 h, j_2 = 1, \cdots, \lfloor (n - k)/h \rfloor. \tag{8}$$

Note that for $k < h$ and $k > n - h$, we have $H_{1:n}(k) = \emptyset$.

For each $k = 1, \cdots, n$, based on its nested local-window set $H_{1:n}(k)$, we define a maximal SN test statistic such that

$$T_{1,n}(k) = \max_{(t1,t2) \in H_{1:n}(k)} T_n(t_1, k, t_2), \tag{9}$$

where we set $\max_{(t1,t2)\in\emptyset} T_n(t_1, k, t_2) = 0$. Intuitively, the nested local-window framework requires that there exist some local window sets $(t1, t2)$ regards each k as the only change-point. This means each k will have several SN test statistics, and the maximal test statistic $T_n(t_1, k, t_2)$ will be equivalent to the original global test statistic $T_{1,n}(k)$.

The reason to apply the nested local-window algorithms is the self-normalizer $V_n(t_1, k, t_2)$ under the multiple change-point estimation framework. The intuition is as follows. For a non-change-point k, if the nested local-window $H_{1:n}(k)$ contains no change-point, the subsample test statistic $T_n(t_1, k, t_2)$ is expected to be small. If $H_{1:n}(k)$ contains at least one change-point, its self-normalizer $V_n(t_1, k, t_2)$ may be inflated since $L_n(t_1, k, t_2)$ and $R_n(t_1, k, t_2)$ are based on contrast statistics and might significantly inflate due to other change-points except for k. This means a large value of $V_n(t_1, k, t_2)$ can cause deflation to $T_n(t_1, k, t_2)$. By applying the nested local-window algorithms, with a relatively small value of the trimming parameter ϵ, there exists at least one nested window $(\tilde{t_1}, \tilde{t_2}) \in H_{1:n}(k)$ that includes k as the only change-point for any true change-point location k. Thus, the maximal statistic $T_n(t_1, k, t_2)$ will remain effective thanks to $T_n(\tilde{t_1}, k, \tilde{t_2})$.

Based on the maximal $T_{1,n}(k)$ and the pre-specified threshold K_n, the SN-based multiple change-point estimation (SNCP) proceeds as follows. Using the full sample $Y_{tt=1}^{n}$, we apply the nested local-window framework to calculate $T_{1,n}(k)$ for $k = 1, \cdots, n$. Given that $\max_{k=1,\cdots,n} T_{1,n}(k) \leq K_n$, SNCP declares no change-point. Otherwise, SNCP declares $\hat{k} = \arg\max_{k=1,\cdots,n} T_{1,n}(k)$ as the first change-point. Then we repeat the procedures of SNCP to the two segments $Y_{tt=1}^{\hat{k}}$ and $Y_{tt=\hat{k}+1}^{n}$ until all $T_{1,n}(k) \leq K_n$. In other words, no more change-point is detected and the remaining partitioned segments are homogeneous piecewise segments.

3.3 Theoretical Results of SN-based Test Threshold

In this case, it is essestial to study theoretical properties of the threshold K_n. For multiple change-point estimation, we first assume a univariate time series case whose dimension is $d = 1$. For any $u \in (\epsilon, 1-\epsilon)$, we define the scaled limit of $H_{1:n}(k)$ by $H_\epsilon(u) = (u_1, u_2)|u_1 = u - j\epsilon, j = 1, \cdots, \lfloor u/\epsilon \rfloor;\ u_2 = u + j\epsilon, j = 1, \cdots, \lfloor (1-u)/\epsilon \rfloor$. We also define $\triangle(u_1, u, u_2) = B(u) - B(u_1) - \frac{u-u_1}{u_2-u_1}[B(u_2) - B(u_1)]$ where $B(\cdot)$ is a standard brwonian motion. Then under the no-change-point case, according to *Assumptions 3.1(i), 3.2* from Zhao et al. (2021), we have **Theorem 3.1***(i)* from Zhao et al. (2021) that

$$max_{k=1,\cdots,n} T_{1,n}(k) \xrightarrow{D} G_\epsilon = sup_{u\in(\epsilon,1-\epsilon)} max_{(u_1,u_2)\in H_\epsilon(u)} D(u_1, u, u_2)^2/V(u_1, u, u_2), \tag{10}$$

where $D(u_1, u, u_2) = \frac{1}{\sqrt{u_2-u_1}} \triangle(u_1, u, u_2)$ and $V(u_1, u, u_2) = \frac{1}{(u_2-u_1)^2}\big(\int_{u_1}^{u} \triangle(u_1, s, u)^2 ds + \int_{u}^{u_2} \triangle(u, s, u_2)^2 ds\big)$.

This theorem characterizes the asymptotic behavior of SNCP under no change-point and thus provides a natural choice of threshold K_n. For a given window size ϵ, G_ϵ is a pivotal distribution and its critical values (for a *univariate* time series based on change in a single parameter) can be obtained via simulation.

We extend the theorem for time series with dimension $1 < d \leq 10$ (univariate time series with multi-parameters and multivariate time series). By **Proposition 3.1** For a given dimension d and window size ϵ, the limiting distribution $G^*_{\epsilon,d}$ is a pivotal distribution and its critical values can be obtained via simulation. Similarly, for high-dimensional time series ($d > 10$), the simulation of its critical values is included in the supplement text (**add citation**).

4 Appendix