Accepted Manuscript

A tail adaptive approach for change point detection

Bin Liu, Cheng Zhou, Xinsheng Zhang

PII: S0047-259X(18)30210-0
DOI: https://doi.org/10.1016/j.jmva.2018.08.010
Reference: YJMVA 4405

To appear in: *Journal of Multivariate Analysis*

Received date : 22 April 2018

Please cite this article as: B. Liu, C. Zhou, X. Zhang, A tail adaptive approach for change point detection, *Journal of Multivariate Analysis* (2018), https://doi.org/10.1016/j.jmva.2018.08.010

A tail adaptive approach for change point detection

Bin Liu[b], Cheng Zhou[b,*], Xinsheng Zhang[b]

[a]*School of Management, Fudan University, Shanghai, 200433, China*

Abstract

For change point problems with Gaussian distributions, the CUSUM method is most efficient for detecting mean shifts. In contrast, it is not so efficient for heavy-tailed or contaminated data because of its sensitivity to outliers. To address this issue, Csörgő and Horváth [15] introduced the Wilcoxon–Mann–Whitney test based on two-sample U-statistics. In practice, however, the tail structure of distributions is typically unknown. For example, Barndorff-Nielsen and Shephard [4] showed that with higher frequency, stock returns' tails become heavier. To our knowledge, there are no uniformly most powerful testing methods for both heavy and light-tailed distributions. To deal with this issue, we construct a new family of test statistics and combine them to adapt to different tails. As the final test statistic is complex, we design a low-cost bootstrap procedure to approximate its limiting distribution. To capture temporal data dependence, we assume that the data follow a near epoch dependent process [10], which includes ARMA and GARCH processes, among others. We explore the validity of our method both theoretically and through simulation. We also illustrate its use with data on the S&P 500 index.

Keywords: Change point test, Functionals of absolutely regular process, Low-cost bootstrap, tail adaptive test, Two-sample U-statistics

1. Introduction

Change point analysis aims to detect distribution or parameter changes of sequential observations; it has many applications. In economics and finance, for instance, Andreou and Ghysels [1] and Perron [31] used it to detect structural changes of financial asset returns and volatility. In bioinformatics, Muggeo and Adelfio [29] applied it to identify damaged genes involving related diseases, and in Internet traffic, Lévy-Leduc and Roueff [27] employed it to detect Denial of Service (DoS). Other applications pertain, e.g., to climatology [35] and data classification [8]. The literature on change point problems dates back at least to [30]. For independent data, there is a large amount of theory and methods; see [14]. Articles dealing with the case of dependent data include [3, 25].

In this paper, we focus on location change point of multivariate data with arbitrary but fixed dimension d. We assume that there exists at most one change point (AMOC). In particular, let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be observations collected at successive points in time and $X_{i,k}$ be $\boldsymbol{X}_i$s kth coordinate with $k \in \{1, \ldots, d\}$. We wish to detect whether $\boldsymbol{\mu}_i = \mathrm{E}\boldsymbol{X}_i$ changes at some point $i \in \{1, \ldots, n\}$. Therefore, we consider the following hypotheses:

$$\mathcal{H}_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_n \quad \text{vs.} \quad \mathcal{H}_1 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_{k^*} \neq \boldsymbol{\mu}_{k^*+1} = \cdots = \boldsymbol{\mu}_n, \tag{1}$$

where $k^* \in \{1, \ldots, n-1\}$ is the unknown change point location.

For hypotheses (1) with $d = 1$, Csörgő and Horváth [15] studied the test statistic

$$\sup_{t\in[0,1]} \frac{1}{n^{3/2}} \left| \sum_{i=1}^{\lfloor nt \rfloor} \sum_{j=\lfloor nt \rfloor+1}^{n} h(X_i, X_j) \right|, \tag{2}$$

where $h : \mathbb{R}^2 \to \mathbb{R}$ is a kernel function. If $h(x, y) = y - x$, (2) is the cumulative sum (CUSUM) test statistic

$$\sup_{t\in[0,1]} \frac{1}{n^{3/2}} \left| \sum_{i=1}^{\lfloor nt \rfloor} \sum_{j=\lfloor nt \rfloor+1}^{n} (X_i - X_j) \right| = \sup_{t\in[0,1]} \frac{1}{\sqrt{n}} \left| \sum_{i=1}^{\lfloor nt \rfloor} X_i - \frac{\lfloor nt \rfloor}{n} \sum_{i=1}^{n} X_i \right|. \tag{3}$$

*Corresponding author.
Email address: zhoucheng-mike@163.com. (Cheng Zhou)