Nonparametric Change Point Detection of Periodic Data

Lingzhe Guo* Reza Modarres†

Abstract

We consider detection of multiple changes in the distribution of periodic and autocorrelated data. We show that periodicity and autocorrelation degrade existing change detection methods since they blur the changes that these procedures aim to discover. To account for periodicity we transform the sequence of vector observations by embedding them in matrices and thereby producing a sequence of i.i.d. matrix observations. We propose methods of testing the equality of matrix distribution functions and offer change detection algorithms that can be applied to matrix observations. In particular, we use the E-divisive algorithm and apply clustering methods to a sample of observation matrices. Methods that ignore the periodicity have very low statistical power to detect changes in the mean or the variance of periodic data when the periodic effects overwhelm the actual changes, while the proposed methods detect such changes with high power. We illustrate the proposed methods by detecting changes in the total revenue for accounting, tax preparation, bookkeeping, and payroll services, provided by US Bureau of the Census.

Key Words: Matrix Distribution; Partition; Clustering, Homogeneity

1. Introduction

The goal of this paper to detect changes in the distribution of periodic data. Change point analysis concerns detection of changes in the distribution of the observations that are ordered by time or location. Suppose $\{\mathbf{Y}_i\}_{i=1}^{T}$ is a sequence of independent random vectors in $\Re^d$ with probability distribution functions F_i. The change point problem tests the null hypothesis, $H_0 : F_1 = F_2 = \ldots = F_T$ against the alternative

$$H_a : F_1 = \ldots = F_{\eta_1} \neq F_{\eta_1+1} = \ldots = F_{\eta_2} \neq F_{\eta_2+1} = \cdots = F_{\eta_s} \neq F_{\eta_s+1} \ldots = F_T$$

where $1 < \eta_1 < \eta_2 < \ldots < \eta_s < T$ are the respective unknown locations of change points and s is the unknown number of change points.

Investigators often assume that F_i have a common parametric family indexed by a (vector) valued parameter $\boldsymbol{\theta}$. Chen and Gupta (2012) present a lucid account of change point analysis and consider several distributions and models, including parametric, nonparametric, regression, times series, sequential, and Bayesian, among others. The goal of any change point analysis is twofold. One needs to detect if there are any significant changes and then to locate the change point(s). We consider the retrospective models that assume all observations in the sequence are available initially for detection of multiple change points. Change point methods finds numerous applications in stock market analysis, speech recognition, quality control, climate change, traffic accident rate, geological and genetics observations, among others. There has been considerable interest and progress in extending the methods of change point detection to multivariate data using both discrete and continuous distributions.

In financial analysis, Ross (2013) comments that the abnormal shifts in stock market is always worthy of detection. Medical condition monitoring involves trend detection in physiological variables such as heart rate and electroencephalogram for specific medical

*Department of Statistics, George Washington University, Washington, DC, USA
†Department of Statistics, George Washington University, Washington, DC, USA

issues. Bosc et al (2003) use magnetic resonance imaging to detect multiple sclerosis lesion evolution. Maboudou-Tchao and Hawkins (2013) develop a detection method for multivariate normally distributed data. Kim (1996) discusses the likelihood ratio test and proposes a method to detect a change in mean when observations are not independent. When a complaint of discrimination is made, an employer may respond by hiring more minorities. Freidlin and Gastwirth (2000) proposed cumulative-sum based procedures for the analysis of hiring data following the hypergeometric distribution. Applications are abundant in image analysis to detect abrupt events such as security breaches from video-based surveillance (Radke, 2005) and detection of credit card fraud (Bolton and Hand, 2002).

One can view change point detection as a process to partition the observation sequence into homogeneous adjacent segments. The observations within each segment are assumed unchanged. For example, Harnish et al. (2009) modify agglomerative clustering algorithms with a time-ordered constraint to locate the change points. Bahrampour et al. (2011) identify the change points by a weighted and constrained k–means clustering algorithm. Many investigators assume that the observations in each segment are independently and identically distributed (i.i.d) and tests for change points from models that have i.i.d errors are by now well understood. However, in many applications with periodic data, consideration of seasonal effects prompts us to modify the i.i.d. assumption of the identical segments. For example, the temperature of a region is a typical periodic observation affected by seasons. Hence, the distributions are allowed to vary within a fixed period of time. If we ignore the periodicity of the observations and apply the change detection procedures directly, the periodic effect will blur the changes we aim to detect. If one analyzes monthly temperature data to detect changes without considering the periodicity of the observations (different seasons), the changes in temperatures across years can be overwhelmed by the changes between seasons. Consequently, such change points are less likely to be detected. More details about the influence of periodicity are given in Section 5. To address these shortcomings, we combine the observations into blocks and compare the resulting blocks in order to account for the periodic effects.

Change detection methods are usually based on homogeneity tests. For example, Lung-Yut-Fong et al. (2011) use rank statistic to estimate change points. In order to perform the analysis on the blocks of observations, we offer a test of equality of distribution functions for matrix distributions. Our method generalizes the work of Szekely and Rizzo (2004), Baringhaus and Franz (2004), and Biswas and Ghosh (2014) on the equality of vector distribution functions to the equality of matrix distribution functions. The proposed method extends the change detection algorithms that are often applied to vector observations to matrix observations. Matteson and James (2014) propose a change detection algorithm based on the energy statistic of Szekely and Rizzo (2004). In particular, we use the E-divisive algorithm of Matteson and James (2014) and apply clustering methods to our basic data structure, which is a sample of observation matrices. Note that the assumption of independence still holds in this paper, which is necessary for derivations and proofs. The main strategy is to embed non-identical vectors in matrices and thereby producing a sequence of i.i.d. matrix observations.

A random matrix is a matrix whose entries are random variables. When the observations are drawn over time or in a sequence of repeated experiments, we can arrange the observations in a sequence of random matrices. For example, Banerjee et al. (2015) arranges a sequence of multivariate normal observations into well separated blocks of time in order to detect the change in the correlation structure. A change in the correlation matrix of the multivariate normal distribution is reflected in a change of the correlation matrix associated with each block. To model random matrices, many matrix-valued (discrete and continuous) distributions are presented in the literature. Gupta and Nagar (1999) discuss

matrix variate distributions systematically and present examples for matrix variate normal, Wishart, Dirichlet, and elliptical distributions, among others. Lovison (2006) proposes a matrix-valued Bernoulli distribution with extensions for categorical matrix data based on log-linear representation.

Using the proposed homogeneity test for change-point detection in matrices, a sequence of matrix observations is segmented into homogeneous groups by the change points. The change points are detected by optimizing the test statistic over the candidate positions of the segment boundaries. Conducting a homogeneity test on every possible segment (even after considering time constraints) is time consuming, especially when the number of change points are large. Work of Barry and Hartigan (1992, 1993) on product partition models and Loschi and Cruz (2005) on computing the probability of a change are particularly relevant in this direction. Vostrikova (1981) proposes binary segmentation procedure to detect the number of change points and their locations in a multidimensional random process. Matteson and James (2014) consider the change point problem as a constrained clustering problem and perform hierarchical divisive and agglomerative algorithms to determine the total number of change points and their locations. In this paper, we will use the hierarchical clustering algorithm for multiple change point detection when considering a sequence of matrix observations.

Assessing the significance of a candidate change point requires finding the null distribution of the optimized statistic. Since we optimize the test statistic with respect to the candidate positions, the null distribution of the optimized statistic is different from the null distribution of the test statistic for homogeneity test. In this direction, Yao and Davis (1986) obtain the asymptotic null distribution of the optimization based on the properties of Brownian bridge. Lung-Yut-Fong et al. (2011) use Monte Carlo experiments to obtain asymptotic p-value of the test for change points. Matteson and James (2014) determine the statistical significance of a change point based on permutation testing. In this paper, we also use permutation method to obtain the null distribution of the optimized test statistic.

The article is organized as follows. We discuss the methodology that transforms a periodic, but independent sequence of vectors into i.i.d. matrices in Section 2. In Section 3, we discuss tests for the homogeneity of two matrix variate distributions. In Section 4, we propose algorithms for detecting multiple change points based on the proposed statistics. A simulation study compares the proposed method with other change detection algorithms under three scenarios in Section 5. In Section 6, we use the proposed methods to detect changes in a real periodic data set. We provide summary and recommendations in the last Section. The proofs appear in the Appendix.

2. Transforming a Periodic Sequence

A sequence of independent random vectors $\{\mathbf{Y}_t\}_{t=1}^T$ is m-periodic if there is a positive integer m such that for any integer $t \in [1, m]$, the random vectors $\mathbf{Y}_t, \mathbf{Y}_{t+m}, \mathbf{Y}_{t+2m}, \cdots$ are i.i.d distributed. For example, suppose $m = 2$, the sequence $\mathbf{Y}_1, \mathbf{Y}_3, \mathbf{Y}_5 \cdots$ are i.i.d distributed and $\mathbf{Y}_2, \mathbf{Y}_4, \mathbf{Y}_6 \cdots$ are also i.i.d distributed. However, $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are independent but not identical. To maintain complete periods, we assume that T is a multiple of m. Let F be the joint distribution of $\{\mathbf{Y}_t\}_{t=1}^T$. A change point problem for a periodic sequence considers whether the joint distribution F changes over time.

Consider a single hypothesized change point location τ. Let $\{\mathbf{Y}_t\}_{t=1}^T \in \Re^d$ be an independent sequence of time-ordered vectors and suppose $\mathscr{A}_\tau = \{\mathbf{Y}_1, \cdots, \mathbf{Y}_\tau\}$ and $\mathscr{B}_\tau = \{\mathbf{Y}_{\tau+1}, \cdots, \mathbf{Y}_T\}$ are two independent m-periodic random sequences with joint distribution functions F and G, respectively. We test the hypothesis $H_0 : F = G$ verses $H_a : F \neq G$. If H_0 is rejected, we conclude that there is a change point at τ, otherwise,

the changes only come from periodic effects, which are expected. Note that the marginal distributions of the vectors are not identical, even when no change point is detected.

A naive approach to handle periodic data is to obtain the weighted average of each period and regard the weighted averages as i.i.d observations. However, this reduction results in lose of information and the conclusions are influenced by the method of calculating the weighted averages. Another possible approach is to adjust the series for seasonality before checking for change points. However, the estimates of the seasonal parameters are biased when the mean shift due to the change point is large and ignored (Lund et al, 2007). To sidestep these difficulties one may transfer an m-periodic independent sequence of observations to an identical and independent sequence of observations. To achieve this, we combine the observations (vectors) into blocks (matrices). Each block contains the observations within the period.

Let $Q_j = [\mathbf{Y}_{(j-1)m+1}, \cdots, \mathbf{Y}_{mj}]$, for $j = 1 \cdots, n$, where $n = T/m$ is the number of matrices. If the number of the vectors is not a multiple of m, the remaining observations form the entries of last matrix. The index of $\mathbf{Y}_i$ and the index of corresponding matrix Q_j have the relationship $j = \lceil \frac{i}{m} \rceil$. With matrices, the periodicity is contained within blocks. Since we use blocks as the basic data structure, we assume that the change point does not exist within the blocks. After the arrangement of the vectors into matrices, we use the change detection methods where the observations are matrices. Therefore, under the null hypothesis of no change point, the observation matrices remain identical.

3. Homogeneity of Two Matrix Distributions

Maa, et al. (1996) construct a theoretical foundation for the use of the interpoint distances in comparison of high-dimensional data sets and show that the equality of the distributions of within and between sample interpoint vector distances is equivalent to $F = G$ for both discrete and continuous observations. Szekely and Rizzo (2004) and Baringhaus and Franz (2004) propose a distance statistic for testing the homogeneity of vector distributions. We extend their work to matrix-valued distributions using matrix norms.

Let $Q_1, \cdots, Q_\tau, Q_{\tau+1} \cdots, Q_n$ be an independent sequence of $d \times m$ random matrices such that $\mathscr{A}_\tau = \{Q_1, \cdots, Q_\tau\}$ follow matrix-valued distribution F and $\mathscr{B}_\tau = \{Q_{\tau+1}, \cdots, Q_n\}$ follow matrix-valued distribution G. We assume that τ is known and are interested in testing the null hypothesis $H_0' : F = G$ against the alternative $H_a' : F \neq G$. The Frobenius norm for a $d \times m$ matrix $A = (a_{ij})$ is defined as $\|A\|_{Fr} = (\sum_{i=1}^{d} \sum_{j=1}^{m} a_{ij}^2)^{1/2}$, where a_{ij} is the element in the ith row and jth column. The squared Frobenius norm is monotonically increasing function of the eigenvalues of a square matrix and invariant to orthogonal transformations. We use $\|\boldsymbol{u}\|_{Eu}$ to denote the Euclidean norm for a vector $\boldsymbol{u} = (u_1, \cdots, u_d)$, i.e. $\|\boldsymbol{u}\|_{Eu} = (\sum_{i=1}^{d} u_i^2)^{1/2}$. Thus, the Frobenius norm is the Euclidean norm on the space of matrices $\Re^{dm}$.

Theorem 1. *Let A_1, A_2, B_1, B_2 be independent $d \times m$ random matrices. Suppose A_1, A_2 are i.i.d from F, and B_1, B_2 are i.i.d from G. If the expectations $E\|A_1\|_{Fr}$ and $E\|B_1\|_{Fr}$ are finite, we have*

$$2E\|A_1 - B_1\|_{Fr} - E\|A_1 - A_2\|_{Fr} - E\|B_1 - B_2\|_{Fr} \geq 0,$$

with equality holding if and only if $F = G$.

Let $\mu_{FG} = E\|A_1 - B_1\|_{Fr}, \mu_{FF} = E\|A_1 - A_2\|_{Fr}$, and $\mu_{GG} = E\|B_1 - B_2\|_{Fr}$. Theorem 1 shows that instead of testing $H_0' : F = G$, we can consider an equivalent null

hypothesis $H_0' : 2\mu_{FG} - \mu_{FF} - \mu_{GG} = 0$ against the alternative $H_a' : 2\mu_{FG} - \mu_{FF} - \mu_{GG} > 0$. We estimate μ_{FG}, μ_{FF} and μ_{GG} by

$$\hat{\mu}_{FF}^{(\tau)} = \binom{\tau}{2}^{-1} \sum_{i=1}^{\tau-1} \sum_{j=i+1}^{\tau} \|Q_i - Q_j\|_{Fr}, \tag{1}$$

$$\hat{\mu}_{GG}^{(\tau)} = \binom{n-\tau}{2}^{-1} \sum_{i=\tau+1}^{n-1} \sum_{j=i+1}^{n} \|Q_i - Q_j\|_{Fr}, \tag{2}$$

$$\hat{\mu}_{FG}^{(\tau)} = (\tau(n-\tau))^{-1} \sum_{i=1}^{\tau} \sum_{j=\tau+1}^{n} \|Q_i - Q_j\|_{Fr}. \tag{3}$$

Thus, a statistic for testing $H_0' : F = G$ is

$$L_1(\mathscr{A}_\tau, \mathscr{B}_\tau) = 2\hat{\mu}_{FG}^{(\tau)} - \hat{\mu}_{FF}^{(\tau)} - \hat{\mu}_{GG}^{(\tau)}, \tag{4}$$

and we reject the null hypothesis for large values of $L_1(\mathscr{A}_\tau, \mathscr{B}_\tau)$. The limiting distribution of $L_1(\mathscr{A}_\tau, \mathscr{B}_\tau)$ is shown in the following theorem.

Theorem 2. *Suppose $\mathscr{A}_\tau = \{Q_1, \cdots, Q_\tau\}$ and $\mathscr{B}_\tau = \{Q_{\tau+1}, \cdots, Q_n\}$ are two sets of independent observations from matrix-valued distributions F. As $min(\tau, n-\tau) \to \infty$, the statistic $n \cdot L_1(\mathscr{A}_\tau, \mathscr{B}_\tau)$ is asymptotically distributed as $\sum_{i=1}^{\infty} \lambda_i (Z_i^2 - 1)$, where the constants λ_i depend on F and Z_i^2 are independent χ_1^2 random variables.*

Biswas and Ghosh (2014) derive necessary and sufficient conditions for the equality of vector-valued distributions. The next theorem extends their result to matrix distributions and shows that two matrix distributions are equal if and only if the mean Frobenius internorm of the within and between distances are equal.

Theorem 3. *Let A_1, A_2, B_1, B_2 be independent $d \times m$ random matrices. Suppose A_1, A_2 are identically distributed with the matrix-valued distribution F, and B_1, B_2 are identically distributed with the matrix-valued distribution G. If the expectations $E\|A_1\|_{Fr}$ and $E\|B_1\|_{Fr}$ are finite, we have $\mu_{FF} = \mu_{GG} = \mu_{FG}$ if and only if $F = G$.*

Based on Theorem 3, we propose $L_2(\mathscr{A}_\tau, \mathscr{B}_\tau)$ for testing equality of matrix-valued distributions where

$$L_2(\mathscr{A}_\tau, \mathscr{B}_\tau) = (\hat{\mu}_{FF}^{(\tau)} - \hat{\mu}_{FG}^{(\tau)})^2 + (\hat{\mu}_{GG}^{(\tau)} - \hat{\mu}_{FG}^{(\tau)})^2, \tag{5}$$

with $\hat{\mu}_{FF}^{(\tau)}, \hat{\mu}_{FG}^{(\tau)}$, and $\hat{\mu}_{GG}^{(\tau)}$ defined by equation (1)-(3). The limiting distribution of $L_2(\mathscr{A}_\tau, \mathscr{B}_\tau)$ is given in Theorem 4.

Theorem 4. *Suppose $\mathscr{A}_\tau = \{Q_1, \cdots, Q_\tau\}$ and $\mathscr{B}_\tau = \{Q_{\tau+1}, \cdots, Q_n\}$ are two sets of independent observations from matrix-valued distributions F and G, respectively. Suppose $\tau/n \to \lambda$ for some $\lambda \in (0,1)$ as $n \to \infty$. The statistic $n \cdot L_2(\mathscr{A}_\tau, \mathscr{B}_\tau)$ is asymptotically distributed as $\frac{2\sigma_0^2}{\lambda(1-\lambda)}\chi_1^2$, where $\sigma_0^2 = Var[E(\|Q_1 - Q_2\|_{Fr}|Q_1)]$.*

4. Detection of Change Points

4.1 Detection of a Single Change Point

We now consider the setting in which the position of the potential change points are unknown. We first assume that there is only one change point. Let $L(\tau)$ be the statistic

used, which can be $L_1(\tau)$ or $L_2(\tau)$ defined in equations (4) and (5), respectively. Suppose $Q_1, \cdots, Q_n \in \Re^{d\times m}$ is an independent sequence of matrix-valued observations, a single change point $\hat{\tau}$ is estimated with

$$\hat{\tau} = \arg\max_{\tau} L(\tau). \tag{6}$$

In practice, calculating $L(\tau)$ for every change candidate directly from the observations might be computationally expensive, especially when dimension and/or sample size is large. We propose the following computing forms for $L_1(\tau)$ and $L_2(\tau)$.

Lemma 1. *Let $\{Q_i\}_{i=1}^n$ be a sequence of matrix observations and suppose $\hat{\mu}_{FF}^{(\tau)}$, $\hat{\mu}_{GG}^{(\tau)}$, $\hat{\mu}_{FG}^{(\tau)}$, $\hat{\mu}_{FF}^{(\tau+1)}$, $\hat{\mu}_{GG}^{(\tau+1)}$ and $\hat{\mu}_{FG}^{(\tau+1)}$ are defined by equations (1)-(3). One can compute $\hat{\mu}_{FF}^{(\tau+1)}$, $\hat{\mu}_{GG}^{(\tau+1)}$ and $\hat{\mu}_{FG}^{(\tau+1)}$ based on $\hat{\mu}_{FF}^{(\tau)}$, $\hat{\mu}_{GG}^{(\tau)}$, and $\hat{\mu}_{FG}^{(\tau)}$ using*

$$\hat{\mu}_{FF}^{(\tau+1)} = \frac{\tau-1}{\tau+1}\hat{\mu}_{FF}^{(\tau)} + \frac{1}{\binom{\tau+1}{2}}\sum_{i=1}^{\tau}\|Q_i - Q_{\tau+1}\|_{Fr}, \tag{7}$$

$$\hat{\mu}_{GG}^{(\tau+1)} = \frac{n-\tau}{n-\tau-2}\hat{\mu}_{GG}^{(\tau)} - \frac{1}{\binom{n-\tau-1}{2}}\sum_{i=\tau+2}^{n}\|Q_i - Q_{\tau+1}\|_{Fr}, \tag{8}$$

$$\hat{\mu}_{FG}^{(\tau+1)} = \frac{1}{k}\{\tau(n-\tau)\hat{\mu}_{FG}^{(\tau)} - \sum_{i=1}^{\tau}\|Q_i - Q_{\tau+1}\|_{Fr} + \sum_{i=\tau+2}^{n}\|Q_i - Q_{\tau+1}\|_{Fr}\} \tag{9}$$

where $k = (\tau+1)(n-\tau-1)$.

Therefore, we only need to compute the distance between $Q_{\tau+1}$ and other points for updating the statistics $L_1(\tau+1)$ and $L_2(\tau+1)$. This short-cut form is of immense computational advantage since computing Frobenius norms are expensive, especially when d and/or m are large.

4.2 Detection of Multiple Change Points

Suppose that there are s change points in the sequence of matrix observations $\{Q_i\}_{i=1}^n$. We denote them as $\tau_1 < \cdots < \tau_s$. Matteson and James (2014) propose two hierarchical methods to estimate the locations of all the change points.

The change points can partition the observation sequence into $s+1$ adjacent clusters. A cluster is defined as $c(\tau_1, \tau_2) = \{Q_i \mid \tau_1 \le i \le \tau_2\}$ where $\tau_1 < \tau_2$. The observations within each cluster must remain consecutive in time order. Two clusters $c_1(\tau_1, \tau_2)$ and $c_2(\tau_3, \tau_4)$ where $\tau_2 \le \tau_3$ are adjacent if $\tau_3 = \tau_2 + 1$. In order to find the locations of the change points, we consider clustering the observation matrices into $s+1$ disjoint sets using various clustering algorithms and treat the starting observation of each cluster as a change point candidate. Hence, clustering is performed under time ordered constraints.

Kaufman and Rousseeuw (2009) describe divisive and agglomerative algorithms for hierarchical clustering. Divisive methods partition a single large cluster into smaller clusters. Suppose $k-1$ change points have been detected and partition the observation sequence into $k > 2$ clusters $\mathscr{C}_1, \cdots, \mathscr{C}_k$. Given these clusters, we apply the procedure for finding a single change point to the observations within each of the k clusters. Let $L(\hat{\tau}_i^*)$ be the criterion for the change point candidate $\hat{\tau}_i^*$ obtained from cluster $\mathscr{C}_i$. The kth change point $\hat{\tau}_k$ is estimated with

$$\hat{\tau}_k = \arg\max_{i\in\{1,\cdots,k\}} L(\hat{\tau}_i^*).$$

Starting with n individual clusters, agglomerative algorithm build new clusters by merging the closest clusters at each stage until one cluster remains. We can use L_1 or L_2 in equations (4) and (5) as measures of the dissimilarity or distance (linkage function) between two clusters. Suppose we have k ordered clusters as $\{\mathscr{C}_1, \cdots, \mathscr{C}_k\}$. For adjacent clusters $\mathscr{C}_i = \{Q_{h+1}, Q_{h+2}, \cdots, Q_{h+t}\}$ and $\mathscr{C}_{i+1} = \{Q_{h+t+1}, Q_{h+t+2}, \cdots, Q_{h+t+r}\}$, the linkages based on L_1 and L_2 are

$$D_1(\mathscr{C}_i, \mathscr{C}_{i+1}) = 2\hat{\mu}_{(i,i+1)} - \hat{\mu}_{(i,i)} - \hat{\mu}_{(i+1,i+1)}, \tag{10}$$

$$D_2(\mathscr{C}_i, \mathscr{C}_{i+1}) = (\hat{\mu}_{(i,i)} - \hat{\mu}_{(i,i+1)})^2 + (\hat{\mu}_{(i+1,i+1)} - \hat{\mu}_{(i,i+1)})^2, \tag{11}$$

where

$$\hat{\mu}_{(i,i)} = \binom{t}{2}^{-1} \sum_{\alpha=1}^{t-1} \sum_{\beta=\alpha+1}^{t} \|Q_{h+\alpha} - Q_{h+\beta}\|_{Fr},$$

$$\hat{\mu}_{(i+1,i+1)} = \binom{r}{2}^{-1} \sum_{\alpha=1}^{r-1} \sum_{\beta=\alpha+1}^{r} \|Q_{h+t+\alpha} - Q_{h+t+\beta}\|_{Fr},$$

$$\hat{\mu}_{(i,i+1)} = (tr)^{-1} \sum_{\alpha=1}^{t} \sum_{\beta=1}^{r} \|Q_{h+\alpha} - Q_{h+t+\beta}\|_{Fr},$$

and Q_{h+1} is the first matrix observation of $\mathscr{C}_i$, t and r are the number of observations of $\mathscr{C}_i$ and $\mathscr{C}_{i+1}$, respectively. For clusters that are not adjacent, the linkages are set to infinity. We merge the clusters with minimum linkage and find $k-1$ clusters or k change points. We continue the process until the number of clusters reach s. Since we set the linkage between distant clusters to infinity, only adjacent clusters can be merged. Therefore, the observations in the updated clusters remain consecutive in time order.

We will next show that once two clusters merge, the distances between the new cluster and other clusters can be computed from existing values. Suppose $\mathscr{A} = \{A_i\}_{i=1}^{n_1}, \mathscr{B} = \{B_i\}_{i=1}^{n_2}, \mathscr{C} = \{C_i\}_{i=1}^{n_3}$ are three clusters of matrices. The average inter matrix distances are

$$\hat{\mu}_{AA} = \binom{n_1}{2}^{-1} \sum_{i=1}^{n_1-1} \sum_{j=i+1}^{n_1} \|A_i - A_j\|_{Fr},$$

$$\hat{\mu}_{BB} = \binom{n_2}{2}^{-1} \sum_{i=1}^{n_2-1} \sum_{j=i+1}^{n_2} \|B_i - B_j\|_{Fr},$$

$$\hat{\mu}_{AB} = (n_1 n_2)^{-1} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \|A_i - B_j\|_{Fr},$$

$$\hat{\mu}_{CC} = \binom{n_3}{2}^{-1} \sum_{i=1}^{n_3-1} \sum_{j=i+1}^{n_3} \|C_i - C_j\|_{Fr},$$

$$\hat{\mu}_{AC} = (n_1 n_3)^{-1} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \|A_i - C_j\|_{Fr},$$

$$\hat{\mu}_{BC} = (n_2 n_3)^{-1} \sum_{i=1}^{n_2} \sum_{j=1}^{n_3} \|B_i - C_j\|_{Fr}.$$

Suppose clusters $\mathscr{A}$ and $\mathscr{B}$ are adjacent and merge to form a new cluster $\mathscr{K}$. Let $N =$

$(n_1+n_2)(n_1+n_2-1)$, the updated inter matrix distance are

$$\begin{aligned}\hat{\mu}_{KK} &= \binom{n_1+n+2}{2}^{-1}\sum_{i=1}^{n_1-1}\sum_{j=i+1}^{n_1}\|A_i-A_j\|_{Fr}+\sum_{i=1}^{n_2-1}\sum_{j=i+1}^{n_2}\|B_i-B_j\|_{Fr}+\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\|A_i-B_j\|_{Fr},\\ &= \frac{1}{N}\{n_1(n_1-1)\hat{\mu}_{AA}+n_2(n_2-1)\hat{\mu}_{BB}+n_1n_2\hat{\mu}_{AB}\}\\ \hat{\mu}_{KC} &= ((n_1+n_2)n_3)^{-1}\sum_{j=1}^{n_3}(\sum_{i=1}^{n_1}\|A_i-C_j\|_{Fr}+\sum_{i=1}^{n_2}\|B_i-C_j\|_{Fr}),\\ &= \frac{n_1}{(n_1+n_2)}\hat{\mu}_{AC}+\frac{n_2}{(n_1+n_2)}\hat{\mu}_{BC}.\end{aligned}$$

Hence, if the clusters $\mathscr{K}$ and $\mathscr{C}$ are adjacent, the linkages between them is

$$\begin{aligned}D_1(\mathscr{K},\mathscr{C}) =&2\hat{\mu}_{KC}-\hat{\mu}_{KK}-\hat{\mu}_{CC}\\ =&\frac{2n_1}{(n_1+n_2)}\hat{\mu}_{AC}+\frac{2n_2}{(n_1+n_2)}\hat{\mu}_{BC}-\frac{n_1(n_1-1)}{N}\hat{\mu}_{AA}-\frac{n_2(n_2-1)}{N}\hat{\mu}_{BB}\\ &-\frac{n_1n_2}{N}\hat{\mu}_{AB}-\hat{\mu}_{CC}\\ D_2(\mathscr{K},\mathscr{C}) =&(\hat{\mu}_{KC}-\hat{\mu}_{KK})^2+(\hat{\mu}_{KC}-\hat{\mu}_{CC})^2\\ =&(\frac{n_1}{n_1+n_2}\hat{\mu}_{AC}+\frac{n_2}{n_1+n_2}\hat{\mu}_{BC}-\frac{n_1(n_1-1)}{N}\hat{\mu}_{AA}-\frac{n_2(n_2-1)}{N}\hat{\mu}_{BB}\\ &-\frac{n_1n_2}{N}\hat{\mu}_{AB})^2+(\frac{n_1}{n_1+n_2}\hat{\mu}_{AC}+\frac{n_2}{n_1+n_2}\hat{\mu}_{BC}-\hat{\mu}_{CC})^2\end{aligned}$$

Therefore, we do not need to store the original data after the initialization of the cluster linkage. After each merge, the linkage between the new cluster and its adjacent clusters can be computed efficiently. This method provides computational and storage advantages.

4.3 Discovering the Significance of Change Points

The previous sections have proposed a procedure for estimating the locations of the change points. We use a permutation method to determine the significance of a change point. Suppose that $k-1$ change points have been estimated, resulting in k clusters, and $\hat{\tau}_k$ is the newly proposed change point candidate with test statistic $L(\hat{\tau}_k)$. To determine whether $\hat{\tau}_k$ is significant, one needs to obtain the behavior of $L(\hat{\tau}_k)$ under the null hypothesis that the observations in the clusters separated by $\hat{\tau}_k$ are homogeneous. The null distribution is different from that in homogeneity test due to the optimization in equation (6). A possible calibration approach relies on permutation tests.

Under the null hypothesis that $\hat{\tau}_k$ is not a change point, we conduct a permutation test as follows. Consider the cluster where $\hat{\tau}_k$ locates, we reorder the observations within the cluster under block constrain and construct a new sequence. The units that are reordered are matrices instead of vectors. The new sequence will still be periodic with same period size. We then apply the single change detection method described in section 4.1 to the new sequence, and denote the test statistic by $L^{(r)}(\hat{\tau}_k)$ for the rth permutation. When the null hypothesis of no additional change points is true, the distribution of $L^{(r)}(\hat{\tau}_k)$ is the same as the distribution of $L(\hat{\tau}_k)$. Thus, we can obtain the empirical distribution of $L(\hat{\tau}_k)$ based on the values of the statistics $L^{(r)}(\hat{\tau}_k)$.

The permutation test will result in exact critical values if we consider all possible permutations. However, unless the sample size is small, this is not computationally feasible. Instead, we obtain an approximate value by performing a sequence of R random permutations. The critical value c_α is approximated using the $(1-\alpha)$th quantile of the sequence

$\{L^{(r)}(\hat{\tau}_k)\}_{r=1}^{R}$. If the test statistic $L(\hat{\tau}_k)$ is larger than the critical value c_α, we say that the change point $\hat{\tau}_k$ is significant.

5. Simulation

In this section, we present simulation results that compare the statistics L_1 and L_2 with E-divisive method (Matteson and James, 2014), Bayesian method (Erdman and Emerson, 2007), regression model (Zeileis et.al., 2001), and empirical distribution function method (Kojadinovic, 2017).

Barry and Hartigan (1993) propose the product partition models for change point problems. This Bayesian methodology assigns probability distributions to the change points and the parameters of each cluster. The point that maximizes the posterior probability is considered as a change candidate. If the posterior probability of a change candidate reaches a specified threshold, then it is labeled a change point. Erdman and Emerson (2007) have implemented this method in the R package bcp. We refer to this method as Bayesian.

The R package strucchange is used for testing structural changes in linear regression models. The package is constructed by Zeileis et. al. (2001) and contains tools for both online and offline change points detection. A change point location is estimated when it minimizes the residual sum of square of the regression model, and BIC is used for multiple change points detection. We refer to this method as Regression.

The R package npcp provides non-parametric CUSUM tests for detecting changes in possibly serially dependent univariate or multivariate observations. Kojadinovic (2017) implemented this routine based on the works from Holmes et. al. (2013), Bucher and Kojadinovic (2016), and Bucher et. al. (2017). We refer to this method as Empirical.

The detection procedures are performed for periodic sequences under two scenarios: 1) change in the mean and 2) change in the variance. The period of the sequences is $m = 2$ and the effect of periodicity is a shift of the mean. Each simulation applies the change point detection methods to 1000 independent replicates. Within these simulations, we consider hypotheses of changes in the mean and changes in the variance of periodic sequences. Suppose $\{\mathbf{X}_t\}_{t=1}^{n_1}$ and $\{\mathbf{Y}_t\}_{t=1}^{n_2}$ are observations drawn from bivariate normal distributions. We choose different mean vectors and covariance matrices for different simulation scenarios.

5.1 Change in Mean for Periodic Sequence

Consider the scenario that the changes only occur in the mean of the periodic sequence. The covariance matrices of $\{\mathbf{X}_t\}_{t=1}^{n_1}$ and $\{\mathbf{Y}_t\}_{t=1}^{n_2}$ are the same, denoted as Σ, while the means are different. Specifically, the means of $\mathbf{X}_1, \mathbf{X}_3, \cdots, \mathbf{X}_{n_1-1}$ are $\boldsymbol{\mu}_1 = (0,0)'$, whereas the means of $\mathbf{X}_2, \mathbf{X}_4, \cdots, \mathbf{X}_{n_1}$ are $\boldsymbol{\mu}_1 + \boldsymbol{\xi} = (\xi, \xi)'$, the means of $\mathbf{Y}_1, \mathbf{Y}_3, \cdots, \mathbf{Y}_{n_2-1}$ are $\boldsymbol{\mu}_2 = (\mu_2, \mu_2)'$, and the means of $\mathbf{Y}_2, \mathbf{Y}_4, \cdots, \mathbf{Y}_{n_2}$ are $\boldsymbol{\mu}_2 + \boldsymbol{\xi} = (\mu_2 + \xi, \mu_2 + \xi)'$. In the simulations, the measure of periodic effect varies in $\xi \in \{0, 10, 100\}$ and there are three different choices of μ_2 in $\{0, 1, 10\}$. The sample sizes are $n_1 = n_2 = 50$. The covariance matrix Σ is the identity matrix.

Consider the null hypothesis of no change points. Table 1 reports the probability of rejecting the null hypothesis, which is the type-I error when $\mu_2 = 0$ and power when $\mu_2 \geq 0$. When there is no periodic effect ($\xi = 0$), all methods have similar type-I errors close to the nominal-level of 0.05. When periodic effect is small, the power of E-divisive method is slightly larger than L_1 and L_2. Bayesian method seems biased at the power falls below the nominal level. Regression method shows low power when the mean difference is small, while the Empirical method performs similarly to L_1 and L_2. When periodic effect exists, the E-divisive method seldom rejects the null hypothesis. When sample size is

Table 1: The probability of rejecting the null hypothesis of no change when the change occurs in the mean. The sample sizes are $n_1 = n_2 = 50$.

ξ	μ_2	L_1	L_2	E-divisive	Bayesian	Regression	Empirical
0	0	0.051	0.041	0.057	0	0.032	0.034
	0.5	0.733	0.138	0.854	0.001	0.494	0.711
	1	1	0.918	1	0.008	0.979	0.821
	10	1	1	1	1	1	1
10	0	0.051	0.043	**0**	**0**	**0**	**0**
	0.5	0.710	0.133	**0**	**0**	**0**	**0**
	1	1	0.911	**0.001**	**0**	**0**	**0**
	10	1	1	1	**0.107**	1	**0.367**
100	0	0.049	0.052	**0**	**0**	**0**	**0**
	0.5	0.730	0.154	**0**	**0**	**0**	**0**
	1	1	0.925	**0**	**0**	**0**	**0**
	10	1	1	**0.445**	**0**	**0**	**0.104**

Table 2: The power of different methods for detecting change of the mean of periodic data when $\mu_2 = 0.5$ with different sample sizes.

ξ	$n_1 = n_2$	L_1	L_2	E-divisive	Bayesian	Regression	Empirical
0	50	0.750	0.151	0.868	0.001	0.472	0.689
	80	0.929	0.207	0.980	0	0.672	0.929
	100	0.976	0.278	0.994	0.001	0.737	0.961
	120	0.992	0.360	0.999	0	0.827	0.990
10	50	0.747	0.125	**0**	**0**	**0**	**0**
	80	0.941	0.246	**0**	**0**	**0**	**0.002**
	100	0.971	0.290	**0**	**0**	**0**	**0.021**
	120	0.994	0.391	**0**	**0**	**0**	**0.011**

$n_1 = n_2 = 50$, this method only reject 44.5% of the time, even though the mean difference $\mu_2 - \mu_1$ is as large as 10. Bayesian, Regression and empirical methods perform worse and have nearly no power when the periodic effect is large. When $\mu_2 \geq 0$, L_1 performs better than L_2, especially when the mean difference of the two samples is small. In addition, increasing of periodic effect (ξ) has little impact on the performance of L_1 and L_2.

Table 2 shows the power of different methods when sample size changes. The mean difference is $\mu_2 = 0.5$ and the sample sizes vary in $n_1 = n_2 \in \{50, 80, 100, 120\}$. In addition, we consider both cases when there is no periodic effect ($\xi = 0$) or there is periodic effect ($\xi = 10$). In Table 2, we can see that if there is no periodic effect, E-divisive method performs better than L_1 and L_2. However, this gap is narrowed when the sample sizes n_1 and n_2 increase. This is reasonable because combining vectors into matrices decreases the sample size, which will degrade permutation methods, especially when the origin sample size is not large. The Empirical method performs similarly to L_1 and works better when sample size increases. The Bayesian, L_2, and Regression methods do not perform well when the mean difference is as small as 0.5. When periodic effect occurs ($\xi > 0$), L_1 performs the best. Moreover, increasing the sample size improves the performance of L_1 and L_2.

Table 3: The probability of rejecting the null hypothesis of no change when the change occurs in variance and $n_1 = n_2 = 50$.

ξ	σ	L_1	L_2	E-divisive	Bayesian	Regression	Empirical
0	1	0.048	0.058	0.051	0	0.032	0.038
	3	0.708	0.893	0.701	0.108	0.058	0.153
	4	0.942	0.991	0.962	0.211	0.074	0.231
	5	0.988	0.997	0.993	0.369	0.080	0.363
10	1	0.051	0.037	**0**	**0**	**0**	**0**
	3	0.733	0.896	**0**	**0**	**0**	**0**
	4	0.927	0.982	**0**	**0**	**0**	**0**
	5	0.982	0.998	**0**	**0**	**0**	**0**
100	1	0.045	0.043	**0**	**0**	**0**	**0**
	3	0.706	0.912	**0**	**0**	**0**	**0**
	4	0.951	0.994	**0**	**0**	**0**	**0**
	5	0.990	0.998	**0**	**0**	**0**	**0**

5.2 Change in Variance for Periodic Sequence

We consider the scenario where unexpected changes occur in the variance of the periodic sequence, while the expected periodic changes are still a shift of the mean. Thus, the means of $\mathbf{X}_1, \mathbf{X}_3, \cdots, \mathbf{X}_{n_1-1}$ and $\mathbf{Y}_1, \mathbf{Y}_3, \cdots, \mathbf{Y}_{n_2-1}$ are the same as $\boldsymbol{\mu}_1 = (0, 0)'$, whereas the means of $\mathbf{X}_2, \mathbf{X}_4, \cdots, \mathbf{X}_{n_1}$ and $\mathbf{Y}_2, \mathbf{Y}_4, \cdots, \mathbf{Y}_{n_2}$ are $\boldsymbol{\mu}_1 + \boldsymbol{\xi} = (\xi, \xi)'$. The covariance matrices of $\{\mathbf{X}_t\}_{t=1}^{n_1}$ are Σ, while the covariance matrices of $\{\mathbf{Y}_t\}_{t=1}^{n_2}$ are $\sigma\Sigma$. For the simulations, the measure of periodic effect varies in $\xi \in \{0, 10, 100\}$. The matrix Σ is identity matrix and we consider different choice of σ in $\{1, 3, 4, 5\}$. The sample sizes are considered as $n_1 = n_2 = 50$.

Table 3 reports the type-I error under the null hypothesis of no change and the power when the alternative hypothesis is change in the variance of the periodic sequence. When there is periodic effect ($\xi = 0$), all three statistics maintain their nominal-level. Moreover, the statistic L_2 shows more power than all other methods. Bayesian, Regression and Empirical method seldom detect the variance change even when no periodic effect occurs. When there is periodic effect ($\xi > 0$), the existing methods do not reject the null hypothesis of homogeneity at all whereas L_1 and L_2 show very good detection power. The power of L_2 is larger than L_1, especially when the variance changes are small.

Table 4 displays the power for detecting variance changes by these methods, with increasing sample size. The sample sizes vary in $n_1 = n_2 \in \{50, 80, 100, 120\}$. Clearly, L_2 perform better than all other methods even when there is no periodic effect. This is because the statistic proposed by Biswas and Ghosh (2014) is more sensitive for variance difference than the statistic proposed by Szekely and Rizzo (2004) and Baringhaus and Franz (2004). In addition, both L_1 and L_2 performs better when the sample size increases.

5.3 Autocorrelation Effects

We have thus far considered the effects of periodicity on the existing change detection methods with time series data. In practice, the observations are not only influenced by periodic effects, but they are also correlated. In this section, we consider the performance of the proposed methods when the observations are correlated. To isolate and study the effects of autocorrelation, we assume no periodicity. Hence, the distributions of the observations are identical if no changes occur. The dependence structure we consider is assumed exists

Table 4: The power of different methods for detecting variance change of periodic data when $\sigma = 3$ with different sample size

ξ	$n_1 = n_2$	L_1	L_2	E-divisive	Bayesian	Regression	Empirical
0	50	0.685	0.882	0.714	0.080	0.073	0.125
	80	0.933	0.999	0.960	0.089	0.044	0.287
	100	0.989	1	0.997	0.101	0.041	0.451
	120	0.994	1	0.995	0.116	0.030	0.583
10	50	0.712	0.917	**0**	**0**	**0**	**0**
	80	0.946	0.998	**0**	**0**	**0**	**0**
	100	0.982	1	**0**	**0**	**0**	**0**
	120	0.993	1	**0**	**0**	**0**	**0**

only within specific time periods. For example, suppose observations are gathered hourly and we are interested in daily changes in the mean. While hourly observations within a day are correlated, the observations of the next day are assumed to be independent of those of the previous days.

Consider a sequence of observations $\mathbf{Y}_1, \cdots, \mathbf{Y}_m, \mathbf{Y}_{m+1}, \cdots, \mathbf{Y}_{2m}, \cdots, \mathbf{Y}_n$, where n/m is an integer. We assume that the correlations occur within $\mathbf{Y}_{km+1}, \cdots, \mathbf{Y}_{(k+1)m}$, for $k = 0, \cdots, \frac{n}{m} - 1$. Let $A_{k+1} = [\mathbf{Y}_{km+1}, \cdots, \mathbf{Y}_{(k+1)m}]$ be the matrix constructed by correlated observations. The observation sequence $\mathbf{Y}_1, \cdots, \mathbf{Y}_n$ is then transformed into the block sequence $A_1, \cdots, A_{n/m}$. The observations in different blocks are independent. Since we only consider the effect of the correlations, the observations within each block are identically distributed, i.e. the marginal distribution of $\mathbf{Y}_{km+1}, \cdots, \mathbf{Y}_{(k+1)m}$ are the same for $k = 0, \cdots, \frac{n}{m} - 1$.

The matrix normal is by far the most studied matrix-valued distribution and is commonly used in applications. Gupta and Nagar (1999) give the definition of matrix normal distributions with the notations $\mathcal{MN}(M, \Sigma, \Psi)$, where M is the expected value, Σ is the covariance of the variables, and Ψ is the covariance of the observation vectors. Since the marginal distributions of the observations (columns) in each block (matrix) are the same, the expected matrix M should be under the constrain that the entries are the same within each row. The type of dependence considered is constant correlations, e.g. Ψ is an $m \times m$ matrix with 1s on the main diagonal and ψ on the off-diagonal. The observations are independent if and only if $\psi = 0$.

Suppose the changes happen in the mean vector. Each simulation applies the change point detection methods to 1000 independent replicates and computes the probability of detecting the change. Within these simulations, the observations are generated by blocks (matrices). Suppose $\{A_i\}_{i=1}^{n1}$ are drawn from matrix normal distribution $\mathcal{MN}(M_1, \Sigma, \Psi)$, and $\{B_i\}_{i=1}^{n2}$ are drawn from matrix normal distribution $\mathcal{MN}(M_2, \Sigma, \Psi)$. The dimension and the block size are set as $d = 2$ and $m = 4$, correspondingly. The Σ is $d \times d$ identity matrix, while the Ψ is a $m \times m$ constant correlation matrix where ψ varies in $\{0, 0.1, 0.3, 0.5\}$. The block sample sizes are $n_1 = n_2 = 25$. The entries of M_1 are the same as $\mu_1 = 0$, while M_2 is filled by μ_2 that varies in $\{0, 0.5, 1\}$.

Consider the null hypothesis of no change points. Table 5 reports the probability of rejecting the null hypothesis, which is the type-I error when $\mu_2 = 0$ and power when $\mu_2 \geq 0$. When the observations are independent ($\psi = 0$), most methods have similar type-I errors, close to the nominal-level of 0.05. The power of L_1, E-divisive and Empirical methods are competitive in this case. When the observations are not independent ($\psi > 0$), the E-divisive method has large type-I errors. When $\psi = 0.1$, E-divisive method's type-I

Table 5: The probability of rejecting the null hypothesis of no change point when the changes occur in the mean and the matrix Ψ has constant correlation ψ.

ψ	μ_2	L_1	L_2	E-divisive	Bayesian	Regression	Empirical
0	0	0.049	0.027	0.054	0	0.022	0.034
	0.5	0.952	0.440	0.945	0	0.745	0.977
	1	1	0.917	1	0.015	1	0.984
0.1	0	0.045	0.041	**0.203**	0.005	0.058	0.046
	0.5	0.925	0.398	0.937	0.006	0.770	0.943
	1	0.999	0.829	1	0.019	0.999	0.978
0.3	0	0.046	0.043	**0.53**	**0.279**	**0.223**	0.047
	0.5	0.841	0.325	0.966	0.243	0.787	0.813
	1	0.988	0.681	1	0.183	0.998	0.944
0.5	0	0.043	0.052	**0.858**	**0.908**	**0.364**	0.042
	0.5	0.780	0.252	0.991	0.899	0.823	0.661
	1	0.942	0.518	1	0.858	0.997	0.926

error is as large as 0.203. The power values are dubious when a method does not maintain its nominal level at $\mu_2 = 0$. The L_1, L_2 and Empirical methods still have nominal type-I error around 0.05. The L_1 and Empirical methods also have competitive powers when correlations occur. The Bayesian method does not work well when the mean difference is small.

6. Application

We illustrate the proposed methods for detecting changes to the total revenue for accounting, tax preparation, bookkeeping, and payroll services in a data set provided by US Bureau of the Census. The data is recorded quarterly from 2004 to 2017, and is displayed in Figure 1. The periodicity of the series can clearly be seen from the plot. Within a year, the revenue is the highest in the first quarter and the lowest in the third or fourth quarter because nearly all tax-related activities happen in the first quarter of the year. We would like to analyze the data to determine whether the revenue changes across years and detect any change point. We used E-divisive method (Matteson and James, 2014) with significance level $\alpha = 0.05$, and this method did not detect any change point.

We used the proposed L_1 and L_2 methods of periodic change point detection with period size $m = 4$, $R = 500$ permutations and $\alpha = 0.05$. Both methods estimate a significant change point in the 4th quarter of 2012. The change point appears with a vertical line in Figure 1. The detected change point is significant with a p-value of less than 0.001 when using either L_1 or L_2 methods. We also apply Bayesian, Regression, and Empirical methods to the data set. Bayesian model can not detect any change point in the data set. The maximum value of the posterior probability is only 0.582. The regression model detects three change points located in 1st quarter of 2007, 1st quarter of 2012, and 1st quarter of 2015. The BIC of this separation reaches the minimum value 1126. Moreover, the empirical model detects the change point in the 1st quarter of 2012 with p-value less than 0.001.

The change point in 4th quarter of 2012 corresponds to the American Taxpayer Relief Act of 2012, passed by the United States Congress on January 1, 2013. This Act gives higher tax rate at upper income levels and establishes caps on tax deductions and credits for those taxpayers at upper income levels. Consequently, people or companies at upper income levels sought more elaborate and extensive tax services in order to cushion the

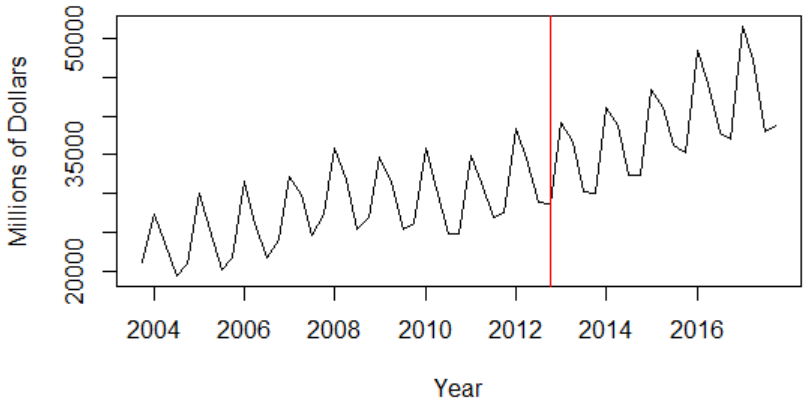

Figure 1: Total Revenue for Accounting, Tax Preparation, Bookkeeping, and Payroll Services. (Source: https://fred.stlouisfed.org/series/REV5412TAXABL144QNSA)

impact from the Act. As a result, the revenue for accounting, tax preparation, bookkeeping, and payroll services increased from the 4th quarter of 2012. Examination of the related metadata does not reveal any clear reasons for the changes in 2007 and 2015 detect by regression method. This example supports our findings from the simulation study and provides a cautionary note on the routine application of change point detection methods when the underlying assumptions are not satisfied.

7. Summary and Recommendations

We have proposed methods of detecting multiple change points in periodic data. Popular methods of change detection are designed to work with independent and identically distributed vectors of observations. When seasonal effects or autocorrelations exist in the observed vectors, they can no longer be identically distributed. Ignoring the periodic effects will blur the changes that detection procedures aim to discover. To take the periodic effect into account while maintaining iid data structure, we transform the sequence of observation vectors by embedding them in matrices. We propose methods of testing the equality of matrix distribution functions, generalizing the work of Szekely and Rizzo (2004), Baringhaus and Franz (2004), and Biswas and Ghosh (2014) on the equality of vector distribution functions to the equality of matrix distribution functions. The proposed methods extend the change detection algorithms that are often applied to vector observations to matrix observations. We also obtain short-cut computing formulas based on the Frobenius norms to accelerate the detection process.

For multiple change point detection, we use hierarchical divisive and agglomerative algorithms. We derive a combinatorial solution for the linkage function of the agglomerative algorithm. To determine the significance of a newly proposed change point, we use permutation permutations under block constrain. We compare the proposed detection methods with an existing one that ignores the periodicity. A simulation study considers detection of changes in the mean and the variance of autocorrelated or periodic data. The results show that detection methods that ignore the periodicity of the data have very low statistical power to detect changes in the mean or the variance when periodic effects overwhelm the actual changes, while our methods detect such changes with high power. The proposed methods

perform well with autocorrelated observations.

REFERENCES

American Taxpayer Relief Act of 2012. Public Law 112-240, 126 Stat. 2313. $https:\backslash\backslash www.gpo.gov\backslash fdsys\backslash pkg\backslash PLAW-112publ240$.

Bahrampour, S., Moshiri, B., and Salahshoor, K. (2011). Weighted and constrained possibilistic C-means clustering for online fault detection and isolation. Applied Intelligence, 35(2), 269-284.

Banerjee, T., Firouzi, H., and Hero, A. O. (2015). Non-parametric quickest change detection for large scale random matrices. IEEE International Symposium on Information Theory (ISIT), 146-150.

Baringhaus, L., and Franz, C. (2004). On a new multivariate two-sample test. J. of multivariate analysis, 88(1), 190-206.

Barry, D. and Hartigan, J. (1992). Product partition models for change point problems. The Annals of Statistics, 20(1), 260-279.

Barry, D. and Hartigan, J. A. (1993). A Bayesian Analysis for Change Point Problems. J. of the American Statistical Association, 88(421), 309-319.

Biswas, M., and Ghosh, A. K. (2014). A nonparametric two-sample test applicable to high dimensional data. J. of Multivariate Analysis, 123, 160-171.

Bolton, R. J., and Hand, D. J. (2002). Statistical fraud detection: A review. Statistical Science, 235-249.

Bosc, M., Heitz, F., Armspach, J. P., Namer, I., Gounot, D., and Rumbach, L. (2003). Automatic change detection in multimodal serial MRI: application to multiple sclerosis lesion evolution. Neuroimage, 20(2), 643-656.

Bcher, A. and Kojadinovic, I. (2016), A dependent multiplier bootstrap for the sequential empirical copula process under strong mixing, Bernoulli 22:2, pages 927-968.

Bcher, A., Fermanian, J.-D., and Kojadinovic, I., (2017), Combining cumulative sum change-point detection tests for assessing the stationarity of continuous univariate time series.

Chen, J., and Gupta, A. K. (2012). Parametric statistical change point analysis: with applications to genetics, medicine, and finance. Springer Science and Business Media.

Erdman, C., and Emerson, J. W. (2007). bcp: an R package for performing a Bayesian analysis of change point problems. J. of Statistical Software, 23(3), 1-13.

Freidlin, B. and and Gastwirth, J. L. (2000). Changepoint tests designed for the analysis of hiring data arising in employment discrimination cases. J. of Business & Economic Statistics, Vol. 18, No. 3, 315-322.

Gupta, A. K., and Nagar, D. K. (1999). Matrix variate distributions (Vol. 104). CRC Press.

Harnish, P., Nelson, B., and Runger, G. (2009). Process partitions from time-ordered clusters. J. of Quality Technology, 41(1), 3-17.

Holmes, M., Kojadinovic, I. and Quessy, J-F., (2013), Nonparametric tests for change-point detection la Gombay and Horvth, J. of Multivariate Analysis 115, pages 16-32.

Kaufman, L., and Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis (Vol. 344). John Wiley and Sons.

Kim, H. J. (1996). Change-point detection for correlated observations. Statistica Sinica, 275-287.

Kojadinovic, M. I. (2017). Package npcp. Econometric Reviews, 23(1), 53-70.

Loschi, R. H. and Cruz, F. R. B. (2005). Extension to the product partition model: computing the probability of a change. Computational Statistics and Data Analysis, 48(2), 255-268.

Lovison, G. (2006). A matrix-valued Bernoulli distribution. J. of Multivariate Analysis, 97(7), 1573-1585.

Lund, R., and Reeves, J. (2002). Detection of undocumented changepoints: A revision of the two-phase regression model. J. of Climate, 15(17), 2547-2554.

Lund, R., Wang, X. L., Lu, Q., Reeves, J., Gallagher, C., and Feng, Y. (2007). Changepoint Detection in Periodic and Autocorrelated Time Series, J. of Climate, Vol. 20, 5178-5190.

Lung-Yut-Fong, A., Lvy-Leduc, C., and Capp, O. (2011). Homogeneity and change-point detection tests for multivariate data using rank statistics. arXiv preprint arXiv:1107.1971.

Maa, J. F., Pearl, D. K. and Bartoszyński, R. (1996). Reducing multidimensional two-sample data to one-dimensional interpoint comparisons. Annals of Statistics, 24, 1069-1074.

Maboudou-Tchao, E. M., and Hawkins, D. M. (2013). Detection of multiple change-points in multivariate data. J. of Applied Statistics, 40(9), 1979-1995.

Matteson, D. S., and James, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. J. of the American Statistical Association, 109(505), 334-345.

Radke, R. J., Andra, S., Al-Kofahi, O., and Roysam, B. (2005). Image change detection algorithms: a systematic survey. IEEE transactions on image processing, 14(3), 294-307.

Ross, G. J. (2013). Modeling financial volatility in the presence of abrupt changes. Physica A: Statistical Mechanics and its Applications, 392(2), 350-360.

Szekely, G. J., and Rizzo, M. L. (2004). Testing for equal distributions in high dimension. InterStat, 5.

Szekely, G. J., and Rizzo, M. L. (2005). Hierarchical clustering via joint between-within distances: Extending

Ward's minimum variance method. J. of classification, 22(2), 151-183.
Vostrikova, L.J. (1981). Detecting disorder in multidimensional random processes. Soviet Mathematics Doklady, 24, 55-59
Yao, Y.C. and Davis, R.A. (1986). The asymptotic behavior of the likelihood ratio statistics for testing shift in mean in a sequence of independent normal variates. Sankhya: The Indian J. of Statistics, Series A, 48(3), 339-353.
Zeileis, A., Leisch, F., Hornik, K., and Kleiber, C. (2002). strucchange. An R package for testing for structural change in linear regression models. J. of Statistical Software, Vol 7, 1-38.

A. Appendix:

The following Proposition defines a projection method for vectorization of matrices. The proof of the Proposition when $m = 1$ appears in Baringhaus and Franz (2004).

Proposition 1. *Suppose $\boldsymbol{u} \in \Re^{dm}$ is a d-dimensional vector and let S^{dm-1} be the surface of the unit sphere in $\Re^{dm}$. The Euclidean norm of $\boldsymbol{u}$ can be represent as*

$$\|\boldsymbol{u}\|_{Eu} = \gamma_{dm} \int_{S^{dm-1}} |a'\boldsymbol{u}| d\mu(a),$$

where μ is the uniform distribution on S^{dm-1}, and

$$\gamma_{dm} = \frac{\sqrt{\pi}(dm-1)\Gamma((dm-1)/2)}{2\Gamma(dm/2)}.$$

Proof of Theorem 1

Proof. To simplify notation, we use μ_{FF}, μ_{GG} and μ_{FG} to represent the expected value of inter-norm within or between two distribution:

$$\mu_{FF} = E\|A_1 - A_2\|_{Fr}, \;\; \mu_{GG} = E\|B_1 - B_2\|_{Fr}, \;\; \mu_{FG} = E\|A_1 - B_1\|_{Fr}. \tag{12}$$

Let $vec(\cdot)$ be an operator that maps a matrix to a vector by stacking the columns of the matrix on top of one another. Let $\boldsymbol{u}_1 = vec(A_1)$, $\boldsymbol{u}_2 = vec(A_2)$, $\boldsymbol{v}_1 = vec(B_1)$ and $\boldsymbol{v}_2 = vec(B_2)$.

Based on Proposition 1, one can show that

$$\|\boldsymbol{u}_1 - \boldsymbol{u}_2\|_{Eu} = \gamma_{dm} \int_{S^{dm-1}} |a'(\boldsymbol{u}_1 - \boldsymbol{u}_2)| d\mu(a),$$

where μ is the uniform distribution on $S^{dm-1} = \{\boldsymbol{x} \in \Re^{dm} : \|\boldsymbol{x}\|_{Eu} = 1\}$, the surface of the unit sphere in $\Re^{dm}$ and γ_{dm} in Proposition 1. Similarly, we have

$$\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_{Eu} = \gamma_{dm} \int_{S^{dm-1}} |a'(\boldsymbol{v}_1 - \boldsymbol{v}_2)| d\mu(a),$$
$$\|\boldsymbol{u}_1 - \boldsymbol{v}_1\|_{Eu} = \gamma_{dm} \int_{S^{dm-1}} |a'(\boldsymbol{u}_1 - \boldsymbol{v}_1)| d\mu(a).$$

Following Baringhaus and Franz (2004), one can show that for each $a \in S^{dm-1}$,

$$2E|a'(\boldsymbol{u}_1 - \boldsymbol{v}_1)| - E|a'(\boldsymbol{u}_1 - \boldsymbol{v}_2)| - E|a'(\boldsymbol{v}_1 - \boldsymbol{v}_2)| \geq 0. \tag{13}$$

The equality holds if and only if the distribution of $a'\boldsymbol{u}_1$ and $a'\boldsymbol{v}_1$ coincide. Integrating with respect to μ on both sides of inequality (13), we have

$$2E\|\boldsymbol{u}_1 - \boldsymbol{v}_1\|_{Eu} - E\|\boldsymbol{u}_1 - \boldsymbol{u}_2\|_{Eu} - E\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_{Eu} \geq 0. \tag{14}$$

The equality holds if and only if for almost all $a \in S^{dm-1}$, the distribution of $a'\boldsymbol{u}_1$ and $a'\boldsymbol{v}_1$ coincide. Since for each $t \in \Re$, the function $E(e^{ita'u_1})$ and $E(e^{ita'v_1})$ are continuous, the equality in (14) holds if and only if $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$ have the same Fourier transform, or $F = G$.

By the definition of Frobenius and Euclidean norms, we have

$$\begin{aligned}\mu_{FF} &= E\|A_1 - A_2\|_{Fr} = E\|\boldsymbol{u}_1 - \boldsymbol{u}_2\|_{Eu},\\ \mu_{GG} &= E\|B_1 - B_2\|_{Fr} = E\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_{Eu},\\ \mu_{FG} &= E\|A_1 - B_1\|_{Fr} = E\|\boldsymbol{u}_1 - \boldsymbol{v}_1\|_{Eu}.\end{aligned}$$

Thus, we have $2\mu_{FG} - \mu_{FF} - \mu_{GG} \geq 0$, and the equality holds if and only if $F = G$. □

Proof of Theorem 2

Proof. We use $\{A_1, \cdots, A_{n_1}\}$ to represent $\{Q_1, \cdots, Q_\tau\}$, and $\{B_1, \cdots, B_{n_2}\}$ to represent $\{Q_{\tau+1}, \cdots, Q_n\}$, where $n_1 = \tau, n_2 = n - \tau$. Let $h(A_i, A_j; B_p, B_q)$ be a real-valued function such that

$$h(A_i, A_j; B_p, B_q) = \|A_i - B_p\|_{Fr} + \|A_j - B_q\|_{Fr} + \|A_i - A_j\|_{Fr} + \|B_p - B_q\|_{Fr}. \quad (15)$$

It is clear that h is symmetric within each argument (A_i, A_j) and (B_p, B_q). One can show that L_1 is the U-statistic corresponding to the kernel function h. That is,

$$L_1 = \binom{n_1}{2}^{-1}\binom{n_2}{2}^{-1}\sum_{i=1}^{n_1}\sum_{i=1}^{n_1-1}\sum_{j=i+1}^{n_1}\sum_{p=1}^{n_2-1}\sum_{q=p+1}^{n_2} h(A_i, A_j; B_p, B_q). \quad (16)$$

If the distribution F and G are identical, by Theorem 1, we have $E(h(A_1, A_2; B_1, B_2)) = 0$. In addition, since $E[A_1, A_2; B_1, B_2|A_1 = \mathbb{A}_1, B_1 = \mathbb{B}_1] = 0$ for almost all matrix realization $(\mathbb{A}_1, \mathbb{B}_1)$, L_1 is a degenerate kernel U-statistic. The asymptotic distribution of L_1 can be inferred from the work of Hoeffding (1948) for the case $k = 1$, which shows that $n \cdot L_1$ has a non-degenerate limiting distribution $\sum_{i=1}^{\infty}\lambda_i(Z_i^2 - 1)$ where the constants λ_i depend on F and Z_i^2 are independent χ_1^2 random variables. □

Proof of Theorem 3

Proof. If $\mu_{FF} = \mu_{GG} = \mu_{FG}$, we have $2\mu_{FG} - \mu_{FF} - \mu_{GG} = 0$, which implies $F = G$ by Theorem 1. Suppose $F = G$, the distributions of A_1, A_2, B_1 and B_2 are equal. Therefore, the distributions of $\|A_1 - A_2\|_{Fr}$, $\|B_1 - B_2\|_{Fr}$ and $\|A_1 - B_1\|_{Fr}$ are also equal, implying the fact that $\mu_{FF} = \mu_{GG} = \mu_{FG}$. □

Proof of Theorem 4

Proof. Following the work of Biswas and Ghosh (2014) for vector distributions, note that $n \cdot L_2(\tau)$ can be expressed as

$$n \cdot L_2(\tau) = \frac{1}{2}([\sqrt{n}(\hat{\mu}_{FF}^{(\tau)} - \hat{\mu}_{GG}^{(\tau)})]^2 + [\sqrt{n}L_1(\tau)]^2),$$

where $\hat{\mu}_{FF}^{(\tau)}$ and $\hat{\mu}_{GG}^{(\tau)}$ are given in equations (1)-(3). From Theorem 2, we have $nL_1(\tau) = O_p(1)$, and hence $\sqrt{n}L_1(\tau) \xrightarrow{p} 0$, as $n \to \infty$.

Let $\mu_{FF} = E\|Q_1 - Q_2\|_{Fr}$ and $\mu_{GG} = E\|Q_{\tau+1} - Q_{\tau+2}\|_{Fr}$. Under null hypothesis, we have $\mu_{FF} = \mu_{GG}$, and hence $\sqrt{n}(\hat{\mu}_{FF}^{(\tau)} - \hat{\mu}_{GG}^{(\tau)}) = \sqrt{n}[(\hat{\mu}_{FF}^{(\tau)} - \mu_{FF}) - (\hat{\mu}_{GG}^{(\tau)} - \mu_{GG})]$. Note that

$$\hat{\mu}_{FF}^{(\tau)} - \mu_{FF} = \binom{\tau}{2}^{-1} \sum_{i=1}^{\tau-1} \sum_{j=i+1}^{\tau} (\|Q_i - Q_j\|_{Fr} - \mu_{FF})$$

is a U-statistic with symmetric kernel function $h(Q_i, Q_j) = \|Q_i - Q_j\|_{Fr} - \mu_{FF}$. Therefore, we have

$$R_1 = \sqrt{\tau}(\hat{\mu}_{FF}^{(\tau)} - \mu_{FF}) \xrightarrow{d} N(0, 4\sigma_0^2),$$

where $\sigma_0^2 = Var[E(\|Q_1 - Q_2\|_{Fr}|Q_1)]$. Similarly, we have

$$R_2 = \sqrt{n-\tau}(\hat{\mu}_{GG}^{(\tau)} - \mu_{GG}) \xrightarrow{d} N(0, 4\sigma_0^2).$$

Since $\hat{\mu}_{FF}^{(\tau)}$ and $\hat{\mu}_{GG}^{(\tau)}$ are independent, one can show

$$\sqrt{n}(\hat{\mu}_{FF}^{(\tau)} - \hat{\mu}_{GG}^{(\tau)}) = \sqrt{n/\tau}R_1 - \sqrt{n/(n-\tau)}R_2 \xrightarrow{d} N(0, (\frac{1}{\lambda} + \frac{1}{1-\lambda})4\sigma_0^2).$$

Therefore, as $min(\tau, n-\tau) \to \infty$, we obtain

$$n \cdot L_2(\tau) = \frac{1}{2}([\sqrt{n}(\hat{\mu}_{FF}^{(\tau)} - \hat{\mu}_{GG}^{(\tau)})]^2 + [\sqrt{n}L_1(\tau)]^2) \xrightarrow{d} \frac{2\sigma_0^2}{\lambda(1-\lambda)}\chi_1^2.$$

□

Proof of Lemma 1

Proof. Using equations (1)-(3) for $\hat{\mu}_{FF}^{(\tau+1)}, \hat{\mu}_{GG}^{(\tau+1)}$ and $\hat{\mu}_{FG}^{(\tau+1)}$, that define the Frobenius inter-norm of the within and between matrices, we obtain

$$\begin{aligned}
\hat{\mu}_{FF}^{(\tau+1)} &= \binom{\tau+1}{2}^{-1} \sum_{i=1}^{\tau} \sum_{j=i+1}^{\tau+1} \|Q_i - Q_j\|_{Fr} \\
&= \binom{\tau+1}{2}^{-1} \sum_{i=1}^{\tau-1} \sum_{j=i+1}^{\tau} \|Q_i - Q_j\|_{Fr} + \sum_{i=1}^{\tau} \|Q_i - Q_{\tau+1}\|_{Fr} \\
&= \frac{\binom{\tau}{2}}{\binom{\tau+1}{2}} \hat{\mu}_{FF}^{(\tau)} + \frac{1}{\binom{\tau+1}{2}} \sum_{i=1}^{\tau} \|Q_i - Q_{\tau+1}\|_{Fr}, \\
&= \frac{\tau-1}{\tau+1} \hat{\mu}_{FF}^{(\tau)} + \frac{1}{\binom{\tau+1}{2}} \sum_{i=1}^{\tau} \|Q_i - Q_{\tau+1}\|_{Fr}, \\
\hat{\mu}_{GG}^{(\tau+1)} &= \binom{n-\tau-1}{2}^{-1} \sum_{i=\tau+2}^{n-1} \sum_{j=i+1}^{n} \|Q_i - Q_j\|_{Fr}, \\
&= \binom{n-\tau-1}{2}^{-1} (\sum_{i=\tau+1}^{n-1} \sum_{j=i+1}^{n} \|Q_i - Q_j\|_{Fr} - \sum_{i=\tau+2}^{n} \|Q_i - Q_{\tau+1}\|_{Fr}) \\
&= \frac{\binom{n-\tau}{2}}{\binom{n-\tau-1}{2}} \hat{\mu}_{GG}^{(\tau)} - \frac{1}{\binom{n-\tau-1}{2}} \sum_{i=\tau+2}^{n} \|Q_i - Q_{\tau+1}\|_{Fr}, \\
&= \frac{n-\tau}{n-\tau-2} \hat{\mu}_{GG}^{(\tau)} - \frac{1}{\binom{n-\tau-1}{2}} \sum_{i=\tau+2}^{n} \|Q_i - Q_{\tau+1}\|_{Fr},
\end{aligned}$$

$$\begin{aligned}
\hat{\mu}_{FG}^{(\tau+1)} &= \frac{1}{k}\sum_{i=1}^{\tau+1}\sum_{j=\tau+2}^{n}\|Q_i - Q_j\|_{Fr} \\
&= \frac{1}{k}\{\sum_{i=1}^{\tau}\sum_{j=\tau+1}^{n}\|Q_i - Q_j\|_{Fr} - \sum_{i=1}^{\tau}\|Q_i - Q_{\tau+1}\|_{Fr} + \sum_{i=\tau+2}^{n}\|Q_i - Q_{\tau+1}\|_{Fr}\} \\
&= \frac{1}{k}\{\tau(n-\tau)\hat{\mu}_{FG}^{(\tau)} - \sum_{i=1}^{\tau}\|Q_i - Q_{\tau+1}\|_{Fr} + \sum_{i=\tau+2}^{n}\|Q_i - Q_{\tau+1}\|_{Fr}\}
\end{aligned}$$

□