Statistica Sinica

Adaptive Change Point Monitoring for High-Dimensional Data

Teng Wu, Runmin Wang, Hao Yan, Xiaofeng Shao

Department of Statistics, University of Illinois at Urbana-Champaign
Department of Statistical Science, Southern Methodist University
School of Computing Informatics & Decision Systems Engineering, Arizona State University

Abstract: In this paper, we propose a class of monitoring statistics for a mean shift in a sequence of high-dimensional observations. Inspired by the recent U-statistic based retrospective tests developed by Wang et al. (2019) and Zhang et al. (2020), we advance the U-statistic based approach to the sequential monitoring problem by developing a new adaptive monitoring procedure that can detect both dense and sparse changes in real time. Unlike Wang et al. (2019) and Zhang et al. (2020), where self-normalization was used in their tests, we instead introduce a class of estimators for q-norm of the covariance matrix and prove their ratio consistency. To facilitate fast computation, we further develop recursive algorithms to improve the computational efficiency of the monitoring procedure. The advantage of the proposed methodology is demonstrated via simulation studies and real data illustrations.

Key words and phrases: Change point detection, Sequential monitoring, Sequential testing, U-statistics.

1. Introduction

Change point detection problems have been extensively studied in many areas, such as statistics, econometrics and engineering, and there are wide applications to fields in physical science and engineering. The literature is huge and is still growing rapidly. For the low-dimensional data, it dates back to early work by Page (1954); MacNeill (1974); Brown et al. (1975), among others. More recent work that studied change point problems for low/fixed dimensional multivariate time series data can be found in Shao and Zhang (2010); Matteson and James (2014); Kirch et al. (2015); Bücher et al. (2019), among others. We refer to Perron (2006), Aue and Horváth (2013) and Aminikhanghahi and Cook (2017) for some excellent reviews on this topic.

The literature on change point detection can be roughly divided into two categories: retrospective testing and estimation of change points based on a complete data sequence offline and online sequential monitoring for change points based on some training data and sequentially arrived data. This paper is concerned with the sequential monitoring problem for temporally independent but cross-sectionally dependent high-dimensional data. There are two major lines of research for sequential change-point detection/monitoring. In one line, a huge body of work follows the paradigm set by pioneers in the field, such as Wald (1945), Page (1954) and Lorden (1971); see Lai (1995, 2001) and Polunchenko and Tartakovsky (2012) for comprehensive reviews. Most sequential detection methods along this line are optimized to have a minimal detection delay with a control of average run length under the null and also the existing procedures are mostly developed for low-dimensional data.

These methods often require both pre-change distribution and post-change distribution to be specified or some parametric assumption to be made. In the other line, Chu et al. (1996) assumed that there is a stretch of training data (without any change points) and sequential monitoring was applied to sequentially arrived testing data. They employ the invariance principle to control the type I error and their framework has been adopted by many other researchers in both parametric and nonparametric contexts. See Horváth et al. (2004); Aue et al. (2012); Wied and Galeano (2013); Fremdt (2015); Dette and Gösmann (2019). Along this line, it is typical to use size and power (plus average detection delay) to describe and compare operating characteristics of competing procedures. Our procedure falls into the second category. It seems to us that these two frameworks are in general difficult to compare, as they differ in terms of the model assumptions and evaluation criteria etc.

Nowadays, with the rapid improvement of data acquisition technology, high-dimensional data streams involving continuous sequential observations appear frequently in modern manufacturing and service industries, and the demand for efficient online monitoring tools for such data has never been higher. For example, in Yan et al. (2018), they proposed a method to monitor a multi-channel tonnage profile used for the forging process, which has thousands of attributes. Furthermore, image-based monitoring [Yan et al. (2014)] has become popular in the literature, which includes thousands of pixels per image. In Lévy-Leduc and Roueff (2009), they considered the problem of monitoring thousands of Internet traffic metrics provided by a French Internet service provider. This kind of high-dimensional data poses significant new challenges to the traditional multivariate statistical process control and monitoring, since when the dimension p is high and is comparable to the sample size n, most existing sequential monitoring methods constructed based on the fixed-dimension assumptions become invalid.

In this article, we propose a new class of sequential monitoring methodology to monitor a change in the mean of independent high-dimensional data based on (sequential) retrospective testing. Our proposal is inspired by recent work on retrospective testing of mean change in high-dimensional data by Wang et al. (2019) and Zhang et al. (2020). In Wang et al. (2019), the authors proposed a U-statistic based approach to target the L_2-norm of the mean difference by extending the U-statistic idea initiated in Chen and Qin (2010) from two-sample testing to the change point testing problem. In Zhang et al. (2020), they further extended the test in Wang et al. (2019) to a L_q-norm-based one mimicking He et al. (2018), where $q \in 2\mathbb{N}$, to capture sparse alternative. By combining L_2-norm-based test and L_q-norm-based one, the adaptive test statistic they proposed is shown to achieve high power for both dense and sparse alternatives. However, one of the limitations of these works is that these methods are designed for off-line analysis, which is not suitable to be applied to real-time online monitoring systems. Building on the works of Wang et al. (2019) and Zhang et al. (2020), we shall propose a new adaptive sequential monitoring procedure that can capture both sparse and dense alternatives. Instead of using the self-normalization scheme [Shao (2010); Shao and Zhang (2010); Shao (2015)], as done in Wang et al. (2019) and Zhang et al. (2020), we opt to use ratio-consistent estimators for $\|\Sigma\|_q^q$ based on the training data, where Σ is the common covariance matrix of the sequence of random vectors and provide a rigorous proof. Further, we develop recursive algorithms for fast implementation so that at each time the monitoring statistics can be efficiently computed. The resulting adaptive monitoring procedure via a combination of L_2 and L_q (say $q = 6$) based sequential tests are shown to be powerful against both dense and sparse alternatives via theory and simulations.

There is a growing literature on high-dimensional change point detection in the retrospective setting; see Horváth and Hušková (2012); Cho and Fryzlewicz (2015); Jirak (2015); Yu and Chen (2017); Wang and Samworth (2018); Yu and Chen (2019); Wang et al. (2019); Zhang et al. (2020); Wang and Shao (2020), among others. It is worth noting that Enikeeva and Harchaoui (2019) developed a test based on a combination of a linear statistic and a scan statistic, and their test can be adaptive to both sparse and dense alternatives. However, their Gaussian and independent components assumptions are also too restrictive. In addition, the literature for online monitoring of high-dimensional data streams has also been growing steadily in the literature of statistics and quality control in the last decade. In particular, Mei (2010) proposed a global monitoring scheme based on the sum of the cumulative sum monitoring statistics from each individual data stream. His method aims to minimize the delay time and control the global false alarm rate, which is based on the average run length under the null. This is different from the size and power analysis as done in our work. Note that the assumptions in Mei (2010) are quite restrictive in the sense that he assumed all data streams do not have cross-sectional dependence, and that both the pre-change and post-change distributions are known. See Wang and Mei (2015), Zou et al. (2015), Liu et al. (2019), and Li (2020) for several variants in the sense that they propose new ways of aggregating the local monitoring statistics. In Xie and Siegmund (2013), they proposed a mixture detection procedure based on a likelihood ratio statistic that takes into account the fraction of data streams being affected. They argue that the performance is good when the fraction of affected data streams are known and do not require to completely specify the post-change distribution. However, the mixture global log-likelihood they specified relies on that hypothesized affected fraction p_0, and they showed the robustness of different choices of p_0 only through numerical studies. The results they derived hold for data generated from a normal distribution or other exponential families of distributions. A common feature of all these works is that they assume the data streams do not have cross-sectional dependence, which may be violated in practice. As a matter of fact, our theory for the proposed monitoring statistic demonstrates the impact of the correlation/covariance structure of the multiple data streams, which seems not well appreciated in the above-mentioned literature.

The rest of the paper is structured as follows: In Section 2, we specify the change point monitoring framework we use and propose the monitoring statistic that targets the L_q-norm of the mean change. An adaptive monitoring scheme can be derived by combining the test statistic for different q's, $q \in 2\mathbb{N}$. Section 3 provides a ratio-consistent estimator for $\|\Sigma\|_q^q$, which is crucial when constructing the monitoring statistics. Section 4 provides simulation studies to examine the finite sample performance of the adaptive monitoring statistic. In Section 5, we apply the adaptive monitoring scheme to two real datasets. Section 6 concludes the paper. All the technical details can be found in the Appendix.

2. Monitoring Statistics

In this section, we specify the general framework we use to perform change point monitoring. We consider a closed-end change point monitoring scenario following Chu et al. (1996). Assume that we observe a sequence of temporally independent high dimensional observations $X_1, \ldots, X_n \in \mathbb{R}^p$, which are ordered in time and have constant mean $\boldsymbol{\mu}$ and covariance matrix Σ. We start the monitoring procedure from time $(n+1)$ to detect if the mean vector changes in the future. Throughout the analysis, we assume that all X_t's are independent over time. A decision is to be made at each of the time points, and we will signal an alarm when the monitoring statistic exceeds a certain boundary. The process ends at time nT regardless of whether a change point is detected, where T is a pre-specified number. The Type-I error of the monitoring procedure is controlled at α, which means the probability of signaling an alarm when there is no change within the period $[n+1, nT]$ is at most α.

Under the null hypothesis, no change occurs within the monitoring period and we have

$$E(X_t) = \mu \text{ for } t = 1, \ldots, nT.$$

against the alternative

Under the alternatives, the mean function changes at some time $t_0 > n$, and remains at the same level for the following observations. That is

$$E(X_t) = \begin{cases} \mu & 1 < t < t_0 \\ \mu + \Delta & t_0 \leq t \leq nT. \end{cases}$$

We propose a family of test statistics $T_{n,q}(k)$, which serves as the monitoring statistic targeting $\|\Delta\|_q$. The case $q = 2$ corresponds to dense alternatives, and larger values of q's correspond to sparser alternatives. We will discuss the formulation of our monitoring statistic for $q = 2$ and then extend to general q's in the following subsections.

2.1 L_2-norm-based monitoring statistics

In this section, we will first develop the L_2-norm-based monitoring statistic, which is especially useful to detect the dense alternative. Furthermore, we will discuss the asymptotic properties of the L_2-norm-based statistic. Finally, the recursive computational algorithm will be developed to allow efficient implementation.

2.1.1 Monitoring statistics

For a given time $k > n$, suppose we know a change point happens at the location m, where $n < m < k$. We can separate the observations into two independent samples: pre-break $X_1, \ldots, X_m$ and post-break $X_{m+1}, \ldots, X_k$. Consider using a two-sample U-statistic with kernel

$$h((X,Y),(X',Y')) = (X-Y)^T(X'-Y')$$

, where (X',Y') is an independent copy of (X,Y). Then we have

$$E[h((X,Y),(X',Y'))] = \|E(X)-E(Y)\|_2^2,$$

which estimates the squared L_2-norm of the mean difference. Indeed Wang et al. (2019) constructed a L_2-norm-based retrospective change point detection statistic by scanning over all possible m. For the online monitoring problem, we shall combine this idea with the approach in Dette and Gösmann (2019) to propose a monitoring statistic. To be more precise, at each time point k, we scan through all possible change point locations m ($n < m < k-2$), and perform a change point testing. We take the maximum of these U-statistics over m as our test statistics at time k. Suppose we can get a ratio-consistent estimator of $||\Sigma||_F$ learned from the training sample $\{X_1,\ldots,X_n\}$ denoted by $\widehat{\|\Sigma\|_F}$, our monitoring statistic at time $k = n+3,\ldots,nT$ is

$$\begin{aligned}
T_{n,2}(k) &= \frac{1}{n^3\widehat{\|\Sigma\|_F}} \max_{m=n+1\ldots,k-2} \sum_{l=1}^{p} \sum_{1\le i_1,i_2\le m}^{*} \sum_{m+1\le j_1,j_2\le k}^{*} (X_{i_1,l}-X_{j_1,l})(X_{i_2,l}-X_{j_2,l}) \\
&= \frac{1}{n^3\widehat{\|\Sigma\|_F}} \max_{m=n+1\ldots,k-2} G_k(m).
\end{aligned}$$

2.1.2 Asymptotic properties

To calibrate the size of the testing procedure, we need to obtain the asymptotic distribution of the test statistic under the null. The following conditions are imposed in Wang et al. (2019) to ensure the process convergence results.

Assumption 1. $tr(\Sigma^4) = o(\|\Sigma\|_F^4)$.

Assumption 2. Let $Cum(h) = \sum_{l_1,\ldots,l_h=1}^{p} cum^2(X_{1,l_1},\ldots,X_{1,l_h}) \le C||\Sigma||_F^h$ for $h = 2,3,4,5,6$ and some constant C. Here $cum(\cdot)$ is the joint cumulant. In general, for a sequence of random variable $Y_1,\ldots,Y_n$, their joint cumulant is defined as

$$cum(Y_1,\ldots,Y_n) = \sum_{\pi}(|\pi|-1)!(-1)^{|\pi|-1}\prod_{B\in\pi} E\left(\prod_{i\in B} Y_i\right),$$

where π runs through the list of all partitions of $\{1,\ldots,n\}$, B runs through the list of all blocks of partition π and and π is the number of parts in the partition.

Assumption 1 was also imposed in Chen and Qin (2010), who pioneered the use of U-statistic approach in the two-sample testing problem for high-dimensional data, and it can be satisfied by a wide range of covariance models. Assumption 2 can be viewed as some restrictions on the dependence structure, which holds under uniform bounds on moments and 'short-range' dependence type conditions on the entries of the vector $(X_{0,1}, ..., X_{0,p})$. See Wang et al. (2019) for more discussions about these two assumptions. Finally, under the null hypothesis and these assumptions, we provide the limiting distribution of the proposed monitoring statistic in Theorem 1.

Theorem 1. *Under Assumptions 1 and 2, we have*

$$\max_{k=n+3}^{nT} T_{n,2}(k) \xrightarrow{D} \sup_{t\in[1,T]} \sup_{s\in[1,t]} G(s,t),$$

where

$$G(s,t) = t(t-s)Q(0,s) + stQ(s,t) - s(t-s)Q(0,t),$$

and Q is a Gaussian process whose covariance structure is the following

$$Cov(Q(a_1,b_1), Q(a_2,b_2)) = \begin{cases} (\min(b_1,b_2) - \max(a_1,a_2))^2 & if \quad \max(a_1,a_2) \le \min(b_1,b_2) \\ 0 & otherwise \end{cases}$$

In general, we can also consider some non-constant boundary function $w(t)$, that is,

$$\max_{k=n+1}^{nT} \frac{T_{n,2}(k)}{w(k/n-1)} \xrightarrow{D} \sup_{t\in[1,T]} \sup_{s\in[1,t]} \frac{G(s,t)}{w(t-1)}.$$

We take the double supremums here to control the familywise error rate. Therefore, we reject the null hypothesis if $T_{n,2}(k) > c_\alpha w(k/n-1)$ for some $k \in \{n+3, \dots, nT\}$. The size can be calibrated by choosing a c_α, such that

$$P\left(\sup_{t\in[1,T]} \sup_{s\in[1,t]} \frac{G(s,t)}{w(t-1)} > c_\alpha\right) = \alpha.$$

Different choices of $w(t)$ have been considered in Dette and Gösmann (2019).

- (T1) $w(t) = 1$,
- (T2) $w(t) = (t+1)^2$,
- (T3) $w(t) = (t+1)^2 \cdot \max\left\{\left(\frac{t}{t+1}\right)^{1/2}, 10^{-10}\right\}$.

These $w(t)$s are motivated by the law of iterated logarithm and are used to reduce the stopping delay under the alternative. Based on our simulation results and real data applications, the choice of $w(t)$ among the above three candidates does not seem to have a big impact on power and detection delay. So in practice, for closed-end procedure, any choice would work. The detailed comparisons are shown in the simulation studies in Section 4.

Remark The current method can be generalized to the open-end framework. For an open-end monitoring procedure, we are interested in testing

$$E(X_t) = \mu \text{ for } t = 1, 2, \ldots.$$

against the alternative

$$E(X_t) = \begin{cases} \mu & 1 < t < t_0 \\ \mu + \Delta & t > t_0. \end{cases}$$

for some $t_0 > n$. Suppose we use the same L_2 norm based monitoring statistic at time $k = n+3, \ldots$, i.e.,

$$T_{n,2}(k) = \frac{1}{n^3 \widehat{\|\Sigma\|_F}} \max_{m=n+1\ldots,k-2} G_k(m).$$

For a suitably chosen boundary function $w(\cdot)$, we expect that

$$\sup_{k=n+3}^{\infty} \frac{T_{n,2}(k)}{w(k/n-1)} \xrightarrow{D} \sup_{t\in[1,\infty)} \sup_{s\in[1,t]} \frac{G(s,t)}{w(t-1)},$$

as $n \to \infty$. The critical value can be determined by

$$P\left(\sup_{t\in[1,\infty)} \sup_{s\in[1,t]} \frac{G(s,t)}{w(t-1)} > c_\alpha \right) = \alpha.$$

We reject the null hypothesis if $T_{n,2}(k) > c_\alpha w(k/n-1)$ for some $k \in \{n+3, \ldots\}$. In practice, we can approximate critical values c_α based on the procedure we used for simulating the critical values in the closed-end procedure, by using a large T, say $T = 200$. Note that the boundary function used for open-end monitoring needs to satisfy certain smoothness and decay rate assumptions and the above three we used for the closed-end procedure are no longer applicable; see Assumption 2.4 in Gösmann et al. (2020) and related discussions.

The following theorem provides theoretical analysis for the power of the L_2-norm-based monitoring procedure.

Theorem 2. *Suppose that Assumptions 1 and 2 hold. Further assume that the change point location is at $\lfloor nr \rfloor$ for some $r \in (1,T)$, we have*

1. *When $\frac{n\Delta^T\Delta}{\|\Sigma\|_F} \to 0$,*

$$\max_{k=n+3,\ldots nT} T_{n,2}(k) \xrightarrow{D} \sup_{t\in[1,T]} \sup_{s\in[1,t]} G(s,t).$$

2. *When $\frac{n\Delta^T\Delta}{\|\Sigma\|_F} \to b \in (0+\infty)$,*

$$\max_{k=n+3,\ldots nT} T_{n,2}(k) \xrightarrow{D} \tilde{T}_2 = \sup_{t\in[1,T]} \sup_{s\in[1,t]} \left[G(s,t) + b\Lambda(s,t) \right],$$

where

$$\Lambda(s,t) = \begin{cases} (t-r)^2 s^2 & s \le r \\ r^2(t-s)^2 & s > r \\ 0 & otherwise \end{cases}.$$

3. *When* $\frac{n\Delta^T\Delta}{||\Sigma||_F} \to \infty$,

$$\max_{k=n+3,\dots nT} T_{n,2}(k) \xrightarrow{D} \infty.$$

Theorem 2 implies that, under the local alternative where $\frac{n\Delta^T\Delta}{||\Sigma||_F} \to 0$, the proposed monitoring procedure has trivial power. For the diverging alternative where $\frac{n\Delta^T\Delta}{||\Sigma||_F} \to +\infty$, the test has power converging to 1. When the strength corresponding to the change falls in between, the test has power between $(\alpha, 1)$.

2.1.3 Recursive computation

One challenge for the proposed monitoring statistic $T_{n,2}(k)$ is that it needs to be recomputed at each given time k. The brute force calculation of the test statistics has $O(n^4p)$ time complexity and O(1) space complexity. In this section, we develop a recursive algorithm to efficiently update the monitoring statistic, which greatly improves the computational efficiency for online monitoring. More specifically, we propose a recursive algorithm to update $G_k(m)$, which is a major component to compute the monitoring statistic $T_{n,2}(k)$ as follows:

$$\begin{aligned} G_k(m) = (k-m)(k-m-1) \sum_{1\le i<j\le m} X_i^T X_j + m(m-1) \sum_{m+1\le i<j\le k} X_i^T X_j \\ - (m-1)(k-m-1) \sum_{i=1}^{m} \sum_{j=m+1}^{k} X_i^T X_j. \end{aligned}$$

To compute $G_k(m)$, we need to keep track of two CUSUM processes

$$B_t = \sum_{i=1}^{t} X_i \text{ and } C_t = \sum_{i=1}^{t} X_i^T X_i,$$

where B_t's are still p-dimensional. The partial sum process $S(a,b) = \sum_{a\le i<j\le b} X_i^T X_j$ appeared in $G_k(m)$ can be expressed in terms of functions of B_t and C_t,

$$S(a,b) = \sum_{a\le i<j\le b} X_i^T X_j = \frac{1}{2}[(B_b - B_{a-1})^T (B_b - B_{a-1}) - (C_b - C_{a-1})].$$

The detailed algorithm is stated as follows,

1. **Initialization**: Start with the first pair $(m, k) = (n + 1, n + 3)$. Record the following quantities
$$B_{n+1}, B_{n+2}, B_{n+3}, C_{n+1}, C_{n+2}, C_{n+3}.$$
The first statistic can be calculated based on
$$\begin{aligned} G_{n+3}(n+1) = {} & 2 \cdot (B_{n+1}^T B_{n+1} - C_{n+1})/2 \\ & + (n+1)n[(B_{n+3} - B_{n+1})^T (B_{n+3} - B_{n+1}) \\ & - (C_{n+3} - C_{n+1})]/2 - nB_{n+1}^T (B_{n+3} - B_{n+1}). \end{aligned}$$

2. **Increase index from k to $k + 1$**: Fix index m, compute B_{k+1} and C_{k+1}:
$$B_{k+1} = B_k + X_{k+1}, C_{k+1} = C_k + X_{k+1}^T X_{k+1}.$$
The statistic for the pair $(m, k + 1)$ is
$$\begin{aligned} G_{k+1}(m) = {} & (k - m + 1)(k - m)(B_m^T B_m - C_m)/2 \\ & + m(m-1)[(B_{k+1} - B_m)^T (B_{k+1} - B_m)) \\ & - (C_{k+1} - C_m)]/2 - (m-1)(k-m) \sum_{i=1}^{m} B_m^T (B_{k+1} - B_m). \end{aligned}$$

3. **Increase index from m to $m + 1$**: For fixed index k, all B_i and C_i for $i = n \ldots, k$ are already recorded. The statistic for pair $(m + 1, k)$ is
$$\begin{aligned} G_k(m+1) = {} & (k - m - 1)(k - m - 2)(B_{m+1}^T B_{m+1} - C_{m+1})/2 \\ & + (m+1)m[(B_k - B_{m+1})^T (B_k - B_{m+1})) - (C_k - C_{m+1})]/2 \\ & - (k - m - 2)mB_{m+1}^T (B_k - B_{m+1}). \end{aligned}$$

The algorithm should start with $(m, k) = (n+1, n+3)$, increase the second index k first and then increase along the first index m. The recursive formulation reduces the time complexity to $O(n^2 p)$ with additional space complexity $O(np)$.

2.2 L_q-norm-based monitoring statistics

In this section, we generalize the monitoring statistic from L_2-norm to L_q-norm. As has been shown in the previous analysis, the power of the L_2-norm-based monitoring statistic depends on quantity $\|\Delta\|_2$, which is sensitive to dense alternatives. However, in real applications, we usually do not know a priori if the mean change is dense or not. As an approximation, we can consider a similar test statistic targeting $\|\Delta\|_q$, for $q \in 2\mathbb{N}$. When q is large, we are essentially testing against sparse alternatives. As a special case, if we let $q \to \infty$, $\lim_{q\to\infty} \|\Delta\|_q = \|\Delta\|_\infty$, we only target on the largest element (in absolute value) of the Δ.

2.2.1 Monitoring statistics

To define the monitoring statistics, we adopt the idea used in Zhang et al. (2020) without applying self-normalization. Self-normalization requires more extensive computation and can be avoided by using the Phase I data to obtain a ratio consistent estimator for $\|\Sigma\|_q$. Also, as pointed out by Shao (2015), self-normalization can result in a slight loss of power. Essentially, we can construct a L_q-norm-based test statistic at time $k = n+q+1, \dots, nT$,

$$
\begin{aligned}
T_{n,q}(k) &= \frac{1}{\sqrt{n^{3q}\widehat{\|\Sigma\|_q^q}}} \max_{m=n+1,\dots,k-q} \sum_{l=1}^{p} \sum_{1\le i_1,\dots,i_q\le m}^{*} \sum_{m+1\le j_1,\dots,j_q\le k}^{*} (X_{i_1,l}-X_{j_1,l})\cdots(X_{i_q,l}-X_{j_q,l}) \\
&= \frac{1}{\sqrt{n^{3q}\widehat{\|\Sigma\|_q^q}}} \max_{m=n+1,\dots,k-q} U_{n,q}(k,m),
\end{aligned}
$$

where $\widehat{\|\Sigma\|_q^q}$ is a ratio-consistent estimator of $\|\Sigma\|_q^q$.

2.2.2 Asymptotic properties

In this subsection, we study the asymptotic property of the L_q-norm-based test statistics. First, we impose the following conditions in Zhang et al. (2020) to facilitate the asymptotic analysis.

Assumption 3. Let $X_t = \mu + Z_t$. Suppose Z_t are i.i.d copies of Z_0 with mean 0 and covariance matrix Σ. There exists $c_0 > 0$ independent of n such that $\inf_{i=1\dots,p} Var(Z_{0,i}) \ge c_0$

Assumption 4. Z_0 has up to 8-th moments, with $\sup_{1\le j\le p} E[Z_{0,j}^8] \le C$, and for $h = 2\dots 8$ there exist constants C_h depending on h only and a constant $r > 2$ such that

$$
|cum(Z_{0,l_1}\dots, Z_{0,l_h})| \le C(1 \vee \max_{1\le i<j\le h} |l_i - l_j|)^{-r}.
$$

These assumptions appeared in Zhang et al. (2020), and Wang et al. (2019) showed that they imply Assumptions 1 and 2 for the case $q = 2$. Assumption 4 can be implied by geometric moment contraction [cf. Proposition 2 of Wu and Shao (2004)] or physical dependence measure proposed by Wu (2005) [cf. Section 4 of Shao and Wu (2007)], or α-mixing. It essentially requires weak cross-sectional dependence among the p components in the data.

Under the null hypothesis, to obtain the limiting distribution of monitoring statistic $T_{n,q}$, we utilize the limiting process in Zhang et al. (2020) and obtained the following theorem.

Table 1: Simulated critical values for L_q-norm-based test, $T = 2$

Boundary	T1		T2		T3	
Quantile	L_2	L_6	L_2	L_6	L_2	L_6
90%	0.756	3.235	0.204	0.867	0.141	0.592
95%	1.264	3.711	0.331	0.973	0.232	0.676
99%	2.715	4.635	0.706	1.196	0.485	0.837

Theorem 3. *Under Assumptions 3 and 4,*

$$\max_{k=n+q+1}^{nT} T_{n,q}(k) \xrightarrow{d} \tilde{T}_q := \sup_{t\in[1,T]} \sup_{s\in[1,t]} G_q(s,t),$$

where

$$G_q(s,t) = \sum_{c=0}^{q} (-1)^{q-c} \binom{q}{c} s^{q-c}(t-s)^c Q_{q,c}(s;[0,t]),$$

and $Q_{q,c}(r;[a,b])$ is a Gaussian process with covariance structure

$$cov(Q_{q,c_1}(r_1;[a_1,b_1]), Q_{q,c_2}(r_2;[a_2,b_2])) = \binom{C}{c} c!(q-c)!(r-A)^c(R-r)^{C-c}(b-R)^{q-C},$$

where $A = \max(a_1,a_2)$, $c = \min(c_1,c_2)$, $C = \max(c_1,c_2)$ and $b = \min(b_1,b_2)$. Two processes Q_{q_1,c_1} and Q_{q_2,c_2} are mutually independent if $q_1 \neq q_2 \in 2\mathbb{N}$.

The limiting null distribution is pivotal and its critical values can be simulated based on the following equation,

$$P\left(\sup_{t\in[1,T]} \sup_{s\in[1,t]} \frac{G_q(s,t)}{w(t-1)} > c_\alpha\right) = \alpha.$$

We reject the H_0 when $T_{n,q}(k) > c_\alpha w(k/n-1)$ for $k = n+q+1,\ldots,nT$. We tabulate the critical values for $T = 2$, $q = 2,6$ and different boundary functions in Table 1. Critical values under other settings are available upon request.

Finally, we study the power of the L_q-norm-based monitoring procedure in Theorem 4.

Theorem 4. *Suppose that Assumptions 3 and 4 hold and the change point location is at $\lfloor nr \rfloor$ for some $r \in (1,T)$,*

1. *When* $\frac{n^{q/2}\|\Delta\|_q^q}{\|\Sigma\|_q^{q/2}} \to 0$, $\max_{k=n+q+1}^{nT} T_{n,q}(k) \xrightarrow{D} \tilde{T}_q$;

2. *When* $\frac{n^{q/2}\|\Delta\|_q^q}{\|\Sigma\|_q^{q/2}} \to \gamma \in (0, +\infty)$,

$$\max_{k=n+q+1}^{nT} T_{n,q}(k) \xrightarrow{D} \sup_{t\in[1,T]} \sup_{s\in[1,t]} [G_q(s,t) + \gamma J_q(s;[0,t])],$$

where

$$J_q(s;[0,t]) = \begin{cases} r^q(t-s)^q & r \le s < t \\ s^q(t-r)^q & s \le r < t \\ 0 & otherwise \end{cases};$$

3. *When* $\frac{n^{q/2}\|\Delta\|_q^q}{\|\Sigma\|_q^{q/2}} \to \infty$, $\max_{k=n+q+1}^{nT} T_{n,q}(k) \xrightarrow{D} \infty$.

Analogous to the $q = 2$ case, the power of the test depends on $\|\Delta\|_q$. Therefore, for large q, the proposed test is sensitive to sparse alternatives.

2.2.3 Recursive computation

Similar to the L_2-based-test statistics, we would like to extend the recursive formulation to L_q-norm-based test statistic. According to Zhang et al.(2020), under the null, the process $U_{n,q}(k,m)$ can be simplified as

$$U_{n,q}(k,m) = \sum_{c=0}^{q} (-1)^{q-c} \binom{q}{c} P_{q-c}^{m-1-c} P_c^{k-m-q+c} S_{n,q,c}(m;1,k),$$

where $P_l^k = k!/(k-l)!$ and

$$S_{n,q,c}(m;s,k) = \sum_{l=1}^{p} \sum_{s\le i_1,\dots,i_c \le m}^{*} \sum_{m+1\le j_1,\dots,j_{q-c}\le k}^{*} \prod_{t=1}^{c} X_{i_t,l} \prod_{g=1}^{q-c} X_{j_g,l}.$$

Since $S_{n,q,c}(m;1,k)$'s are the major building blocks of our final test statistic and need to be computed at each time k, we need to find efficient ways to calculate them recursively. One key element is the sum of product terms like

$$B(c,m,l) := \sum_{1\le i_1,\dots,i_c\le m}^{*} \prod_{t=1}^{c} X_{i_t,l}, \qquad \text{and}$$

$$M(c,m,k,l) := \sum_{m\le j_1,\dots,j_c\le k}^{*} \prod_{g=1}^{c} X_{j_g,l}.$$

When we increase from m to $m+1$,

$$\sum_{1\le i_1,\ldots,i_c\le m+1}^{*}\prod_{t=1}^{c}X_{i_t,l}=\sum_{1\le i_1,\ldots,i_c\le m}^{*}\prod_{t=1}^{c}X_{i_t,l}+X_{m+1,l}\cdot\sum_{1\le i_1,\ldots,i_{c-1}\le m}^{*}\prod_{t=1}^{c-1}X_{i_t,l}.$$

We can derive the following recursive relationship for $B(c,k,l)$, i.e.

$$B(c,m+1,l)=B(c,m,l)+B(c-1,m,l)\cdot X_{k+1,l}. \tag{2.1}$$

There is similar recursive relationship for $M(c,m,k,l)$,

$$M(c,m+1,k,l)=M(c,m,k,l)+X_{m+1,l}M(c-1,m,k,l). \tag{2.2}$$

To enable the recursive computation, for each $c=0,\ldots,q$, we maintain a matrix to store the cumulative sums.

1. **Initialization**: Starting with $c=0$ and $c=1$, for all $l=1,\ldots,p$, we initialize $B(0,k+1,l),\ldots,B(0,k+q,l)=0$ and calculate

$$B(1,k+1,l)=\sum_{i=1}^{k+1}X_{i,l},\ldots,B(1,k+q,l)=\sum_{i=1}^{k+q}X_{i,l}.$$

   Then we can recursively calculate $B(c,i,l)$ for all $c=0,\ldots,q$ and $i\le k+q$ based on Equation 2.1.

2. **Update from** $B(c,k,l)$ **to** $B(c,k+1,l)$: We let $B(0,k+1,l)=B(0,k,l)+X_{k+1,l}$ and obtain the result for other $B(c,k+1,l)$ $(c\le q)$ using Equation 2.1.

3. **Update from** $M(c,m,k,l)$ **to** $M(c,m+1,k,l)$: Fix index k, for any $n+1\le m\le k-q$, $l=1,\ldots,p$, we let $M(0,m,k,l)=0$ and calculate

$$M(1,m,k,l)=\sum_{i=m}^{k}X_{i,l}.$$

   All other $M(c,m,k,l)$ where $c\le q$ and $n+1\le m\le k-q$, can be obtained via Equation 2.2. Construct the test statistic $T_{n,q}(k+1)$ using $B(c,k,l)$'s and $M(c,m,k,l)$'s and repeat from step 2.

The time complexity of the recursive formulation is $O(n^2pq)$ with space complexity $O(npq)$.

2.3 Combining multiple L_q-norm-based statistics

In this section, we propose to combine multiple L_q statistics to detect both dense and sparse alternatives. More specifically, based on theoretical results in Zhang et al. (2020), the U-process for different q's are asymptotically independent, which implies that $\{T_{n,q}\}_{k=n+q+1}^{nT}$ are asymptotically independent for $q \in 2\mathbb{N}$. Therefore, $\max_{k=n+q+1}^{nT} T_{n,q}(k)$ are asymptotically independent for $q \in I$, where $I \subset 2\mathbb{N}$, say $I = \{2, 6\}$. Thus we can combine the monitoring procedure for different q's and adjust for the asymptotic size. In general, if we want to combine a set of $q \in I$, we can adjust the size of each individual test to be $1-(1-\alpha)^{1/|I|}$ given the asymptotic independence and reject the null if any of the monitoring statistics exceeds its critical value. In Zhang et al.(2020) they provide a discussion on the power analysis for the identity covariance matrix case, and showed that the adaptive test enjoys good overall power.

In practice, there is this issue of which q to use. Based on the recommendation in Zhang et al. (2020), we set $q = 6$. As mentioned in the latter paper, using larger q leads to more trimming and more computational cost. As we demonstrate in the simulations, using $q = 6$ and combining with $q = 2$ show a very promising performance; see Section 4 for more details.

3. Ratio-consistent estimator for $\|\Sigma\|_q^q$

Notice that the test statistic $T_n(k)$ requires a ratio-consistent estimator for $\|\Sigma\|_q^q$. For example, when $q = 2$, this can be simplified to $\|\Sigma\|_F^2$. The ratio-consistent estimator for $\|\Sigma\|_F^2$ has been proposed in Chen and Qin (2010), but it seems difficult to generalize to $\|\Sigma\|_q^q$. In this section, we introduce a new class of ratio-consistent estimator for $\|\Sigma\|_q^q$ based on U-statistics. We first show the result when $q = 2$ and generalize it to $q \in 2\mathbb{N}$ later.

Assume $\{X_t\}_{t=1}^n \in \mathbb{R}^p$ are i.i.d. random vectors with mean $\boldsymbol{\mu}$ and variance Σ. Define

$$\widehat{\|\Sigma\|_F^2} = \frac{1}{4\binom{n}{4}} \sum_{1\le j_1<j_2<j_3<j_4\le n} tr\left((X_{j_1}-X_{j_2})(X_{j_1}-X_{j_2})^T(X_{j_3}-X_{j_4})(X_{j_3}-X_{j_4})^T\right), \quad (3.3)$$

as an estimator of $\|\Sigma\|_F^2$.

Theorem 5. *Under Assumption 1 and Cum(4)$\le C\|\Sigma\|_F^4$ in Assumption 2, $\widehat{\|\Sigma\|_F^2}$ is a ratio-consistent estimator of $\|\Sigma\|_F^2$. i.e. $\widehat{\|\Sigma\|_F^2}/\|\Sigma\|_F^2 \xrightarrow{p} 1$.*

Now we extend this idea to general $q \in 2\mathbb{N}$. We let

$$\widehat{\|\Sigma\|_q^q} = \frac{1}{2^q\binom{n}{2q}} \sum_{l_1,l_2=1}^{p} \sum_{1\le i_1<\cdots<i_q<j_1<\cdots<j_q\le n} \prod_{k=1}^{q} (X_{i_k,l_1}-X_{j_k,l_1})(X_{i_k,l_2}-X_{j_k,l_2}),$$

as an estimator for $\|\Sigma\|_q^q$, for any finite positive even number q. We can see that the proposed estimator is unbiased through the following proposition.

Proposition 1. *$\widehat{\|\Sigma\|_q^q}$ is an unbiased estimator of $\|\Sigma\|_q^q$.*

Proof of Proposition 1. Since $\{X_t\}_{t=1}^n$ are i.i.d.,

$$\begin{aligned}
\mathbb{E}[\widehat{\|\Sigma\|_q^q}] &= \frac{1}{2^q\binom{n}{2q}} \sum_{l_1,l_2=1}^{p} \sum_{1\le i_1<\cdots<i_q<j_1<\cdots<j_q\le n} \prod_{k=1}^{q} \mathbb{E}[(X_{i_k,l_1}-X_{j_k,l_1})(X_{i_k,l_2}-X_{j_k,l_2})] \\
&= \frac{1}{2^q\binom{n}{2q}} \sum_{l_1,l_2=1}^{p} \sum_{1\le i_1<\cdots<i_q<j_1<\cdots<j_q\le n} \prod_{k=1}^{q} (2\Sigma_{l_1,l_2}) \\
&= \frac{1}{2^q\binom{n}{2q}} \sum_{l_1,l_2=1}^{p} \binom{n}{2q} 2^q \Sigma_{l_1,l_2}^q = \|\Sigma\|_q^q.
\end{aligned}$$

This completes the proof. $\square$

The ratio consistency can be shown under the following assumption.

Assumption 5. We assume that

1. there exists $c>0$ such that $\inf_{i=1,\ldots,p}\Sigma_{i,i}>c$;
2. there exists $C>0$ and $r>2$ such that for $h=2,3,4$ and $1\le l_1\le\cdots\le l_h\le p$,

$$|cum(X_{0,l_1},\ldots,X_{0,l_h})| \le C(1\vee(l_h-l_1))^{-r}.$$

Notice that Assumption 5(2) is required for the ratio consistency, which is weaker than Assumption 4. The Assumptions 1-5 required for our theory do not state the explicit relationship between n and p. For example, when $\Sigma=I_p$, which means there is no cross-sectional dependence, all the assumptions are satisfied and (n,p) can go to infinity freely without any restrictions. When there is cross-sectional dependence, our assumptions may implicitly restrict the relative scale of n and p. In general, larger p is a blessing in our setting and it makes the asymptotic approximation more accurate and larger n is always preferred owing to large sample approximation. On the other hand, the computational cost increase when both the dimension and the sample size get large.

Theorem 6. *Under Assumption 5, $\widehat{\|\Sigma\|_q^q}$ is a ratio-consistent estimator of $\|\Sigma\|_q^q$, i.e., $\widehat{\|\Sigma\|_q^q}/\|\Sigma\|_q^q \xrightarrow{p} 1$.*

It is worth noting that implementing the above estimator may be time-consuming for large q. In practice, we can always take a random sample of all possible indices and form an incomplete U-statistic to approximate. The consistency of incomplete U-statistic can also be established but not pursued for simplicity.

4. Simulation Results

We compare the monitoring procedures for $q=2, q=6$ and $q=(2,6)$ combined. We consider $(n,p)=(100,50)$ with $T=2$, where the observations $X_i\sim N(\mu_i,\Sigma)$ are generated independently over time. We consider four possible choices of Σ,

$$\Sigma_{ij} = \rho^{|i-j|} \text{ for } \rho = 0, 0.2, 0.5, 0.8,$$

to evaluate the performance of the monitoring scheme for independent-components setting or under weak and strong dependence between components. All tests have nominal size $\alpha = 0.1$.

Under the null H_0, there is no change point, $\mu_i = 0$ for all i. For the alternative, we consider $\mu_i = \sqrt{\delta/r}(\mathbf{1}_r, \mathbf{0}_{p-r})$ for $i = (\lfloor 1.25n \rfloor + 1), \ldots, nT$. Under the dense alternative, we set $(\delta, r) = (1, p), (2, p)$. Under the sparse alternative, we set $(\delta, r) = (1, 3), (1, 1)$.

To illustrate the finite sample performance of our monitoring statistics, we compare with Mei (2010) (denoted as Mei) and Liu et al. (2019) (denoted as LZM), which are similar to the open-end scenario in Chu et al. (1996). Both methods do not require Phase I data and are originally designed to minimize the average run length. Therefore, they do not explicitly control the Type-I error. To make a fair comparison with the current methods, which are proposed under the closed-end monitoring framework, we generate n independent Gaussian sample from $N(\mathbf{0}, \boldsymbol{I}_{p\times p})$ and calculate Mei and LZM monitoring statistics. We empirically determine the critical value such that the empirical rejection rate is 10% based on 2500 simulated datasets. For Mei's methods, we need to specify the distribution after the change point, which we set it to be the distribution under the alternative $(\delta, r) = (1, p)$. For LZM's method, we use the same setting in Liu et al. (2019) and set $b = \log(10), \rho = 0.25, t = 4$ and $s = 1$.

Table 2 shows the size of the monitoring procedure for the benchmark methods and the proposed methods for three different boundary functions T1, T2, T3 introduced in Section 2.1 under different correlation coefficients ρ. Notice that the size is noticeably worse for $\rho = 0.8$. This is partially due to the poor performance in the ratio-consistent estimator since its variance increases as the cross-sectional dependence increases. Also, please note that the size seems to go in different directions for $q = 2$ and $q = 6$ as the correlation increases. The combined test, on the other hand, balances out such distortions. To make sure this is only a finite sample behavior, we increase (n, p) from $(100, 50)$ to $(200, 200)$, the size distortion for all tests improved noticeably for almost all settings. The additional results are available in the Supplementary Materials. By contrast, Mei and LZM only achieved correct size for the independent-component case, since we select the threshold under the exact same setting. However, when there is cross-sectional dependence between different data streams, the size is no longer controlled and the size distortion is much more severe than the L_q based tests.

Table 3 provides the power result (left column) and ADT (right column) for different tests under dense alternatives. As expected, the L_2-based test demonstrates higher power compared to that of the L_6-based test. The power of the combined test falls in between and is closer to the power of L_2-based test. As the correlation increases, the powers of all tests decrease due to reduced signal. Among three different boundary functions, T2 seems to be the one with the shortest average delay time (ADT) with a slight sacrifice in power. Mei's method is only better than the L_6 based test when there is no strong cross-sectional dependence, and is generally worse than all other methods and have relatively longer delay

Table 2: Size of different monitoring procedures

			T1			T2			T3		
$\alpha = 0.1$	Mei	LZM	L_2	L_6	Comb	L_2	L_6	Comb	L_2	L_6	Comb
$\rho = 0$	0.094	0.105	0.086	0.048	0.067	0.093	0.045	0.071	0.097	0.045	0.070
$\rho = 0.2$	0.058	0.125	0.083	0.048	0.057	0.082	0.045	0.055	0.083	0.046	0.051
$\rho = 0.5$	0.002	0.176	0.103	0.048	0.084	0.104	0.048	0.082	0.108	0.048	0.080
$\rho = 0.8$	0.000	0.409	0.135	0.028	0.085	0.145	0.027	0.093	0.137	0.026	0.086

even when the distribution under the alternative is correctly specified. Notice that when $\rho = 0.8$, Mei's method lost the power completely. LZM in general has the slightly shorter detection delay but at the cost of a much lower power compared with L_2 based test and the combined test. This means the LZM is quicker in signaling an alarm when a change point is detected. Although LZM showed good power for the strong cross-sectional dependence case compared with the combined test, it comes at the price of much distorted size. This is related to the fact that LZM assumes all data streams are independent.

Table 4 provides power of different tests under sparse alternatives. The L_6-based test and the combined test are very comparable in power and L_2-based test exhibits inferior power in most settings as expected. One interesting observation is that for the case $(\delta, r) = (1, 3)$, the L_2-based test still shows slightly higher power than the L_6-based test when $\rho = 0.2$, which means that for this particular setting, the change is not "sparse" enough. As the correlation increases, we observe a noticeable drop in power, which is similar to the dense alternative setting and is again attributed to the reduced signal size. Among three different boundary functions, T2 still has shortest average delay time with a slight power loss compared to other two boundary functions. Mei's method has worse power since it is designed for the dense signals and the distribution under the alternative is misspecified. By comparison, LZM gives consistently good power and short delay time across all settings. However, the good power under strong cross-sectional dependence is still offset by the severe size distortion under the null.

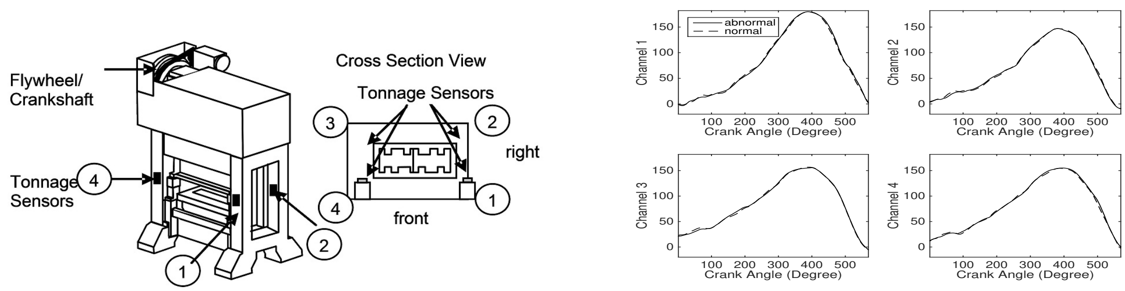
Apart from evaluating the size and power of the monitoring procedure, we also compare the computational cost of the recursive formulation versus the brute force approach. For the case of $(n, p) = (100, 50)$, the average run-time of the brute force approach is 12.89 times of the recursive algorithm under H_0, and the average run-time of the brute force approach is 13.07 times of that for the recursive algorithm under the alternative. The codes are implemented in R. This demonstrates the substantial efficiency gain from the recursive computational algorithm.

Table 3: Power under dense alternatives

Power $\alpha = 0.1$	(δ, r)	Mei power	Mei ADT	LZM power	LZM ADT	$w(t)$	L_2 power	L_2 ADT	L_6 power	L_6 ADT	Comb power	Comb ADT
$\rho = 0.0$	$(1, p)$	0.852	72.9	0.628	38.0	T1	0.958	51.9	0.295	64.6	0.926	55.0
						T2	0.951	44.3	0.284	63.0	0.921	47.7
						T3	0.953	46.8	0.286	63.4	0.921	50.2
	$(2, p)$	0.999	69.3	1.000	15.1	T1	1.000	27.5	0.919	56.2	1.000	29.5
						T2	1.000	20.4	0.919	54.3	1.000	21.9
						T3	1.000	22.9	0.920	54.9	1.000	24.7
$\rho = 0.2$	$(1, p)$	0.740	73.3	0.675	38.2	T1	0.935	51.8	0.302	64.4	0.907	54.9
						T2	0.930	44.2	0.291	62.9	0.906	47.7
						T3	0.933	46.7	0.294	63.5	0.903	50.3
	$(2, p)$	1.000	69.9	1.000	15.6	T1	1.000	28.0	0.884	56.6	1.000	30.0
						T2	1.000	20.8	0.884	54.8	1.000	22.3
						T3	1.000	23.4	0.883	55.3	1.000	25.2
$\rho = 0.5$	$(1, p)$	0.243	74.1	0.715	34.3	T1	0.844	52.9	0.274	63.3	0.796	55.8
						T2	0.843	45.2	0.267	61.5	0.787	47.9
						T3	0.847	47.9	0.267	62.0	0.792	50.7
	$(2, p)$	0.932	72.2	1.000	15.7	T1	1.000	30.7	0.864	55.9	1.000	33.0
						T2	1.000	23.1	0.861	54.2	1.000	24.8
						T3	1.000	25.7	0.861	54.8	1.000	27.8
$\rho = 0.8$	$(1, p)$	0.000	NA	0.803	29.0	T1	0.632	54.6	0.165	62.5	0.560	56.8
						T2	0.637	46.4	0.162	60.9	0.575	48.6
						T3	0.642	49.4	0.162	61.4	0.568	51.8
	$(2, p)$	0.001	74.0	0.997	16.1	T1	0.990	38.3	0.666	56.0	0.984	40.8
						T2	0.990	30.1	0.663	54.2	0.983	32.1
						T3	0.990	32.7	0.666	54.9	0.983	35.4

Table 4: Power under sparse alternatives

Power		Mei		LZM			L_2		L_6		Comb	
$\alpha = 0.1$	(δ, r)	power	ADT	power	ADT	$w(t)$	power	ADT	power	ADT	power	ADT
$\rho = 0.0$	$(1, 3)$	0.422	74.0	0.990	27.4	T1	0.976	51.5	0.999	37.8	0.999	40.5
						T2	0.967	43.8	0.999	35.9	0.999	38.0
						T3	0.972	46.4	0.999	36.6	0.999	39.0
	$(1, 1)$	0.400	73.9	1.000	23.4	T1	0.961	51.5	0.951	51.0	0.976	52.2
						T2	0.958	44.1	0.953	49.5	0.974	46.3
						T3	0.959	46.4	0.953	50.0	0.976	48.7
$\rho = 0.2$	$(1, 3)$	0.274	74.1	0.990	29.1	T1	0.946	52.2	0.937	51.6	0.955	52.6
						T2	0.939	44.6	0.935	50.0	0.955	47.1
						T3	0.943	47.1	0.936	50.5	0.954	49.2
	$(1, 1)$	0.268	74.1	1.000	23.9	T1	0.961	52.6	0.998	37.3	0.999	40.2
						T2	0.951	45.3	0.998	35.4	0.999	37.7
						T3	0.957	47.6	0.998	36.0	0.999	38.6
$\rho = 0.5$	$(1, 3)$	0.048	74.5	0.972	28.2	T1	0.871	54.7	0.881	51.5	0.887	53.4
						T2	0.856	47.1	0.878	49.8	0.884	48.7
						T3	0.860	49.9	0.880	50.4	0.886	50.6
	$(1, 1)$	0.036	74.3	1.000	23.2	T1	0.880	55.9	0.997	38.0	0.997	40.7
						T2	0.871	49.1	0.997	36.1	0.997	38.2
						T3	0.879	51.2	0.997	36.8	0.997	39.2
$\rho = 0.8$	$(1, 3)$	0.000	NA	0.971	24.7	T1	0.621	58.9	0.800	52.9	0.808	55.3
						T2	0.610	50.6	0.801	51.3	0.802	51.5
						T3	0.614	53.7	0.803	51.9	0.807	53.2
	$(1, 1)$	0.000	NA	1.000	21.5	T1	0.602	61.1	0.998	38.8	0.997	41.7
						T2	0.588	53.6	0.998	36.8	0.997	39.3
						T3	0.601	56.8	0.998	37.5	0.997	40.2



(a) The setup of the forging machine

(b) Data collected in the forging machine

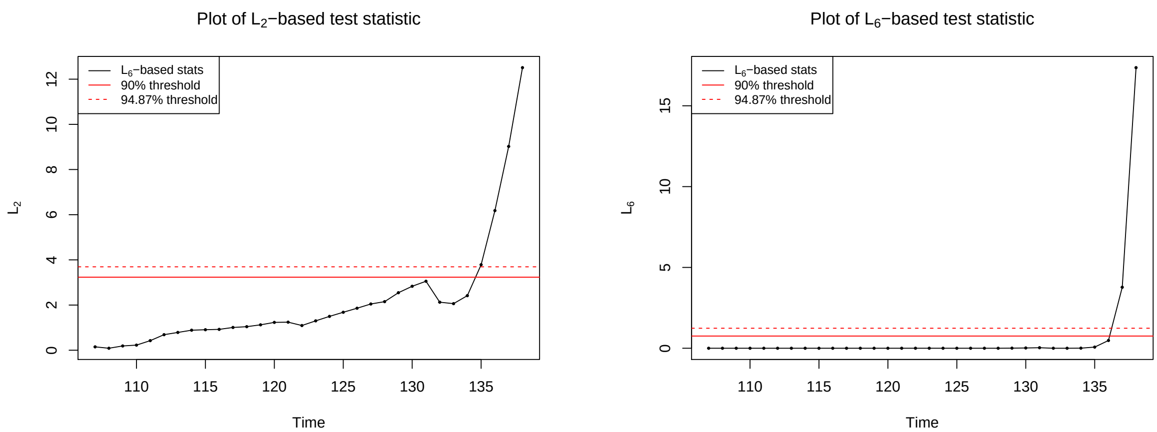
Figure 1: Forging machine setup and the collected tonnage dataset

5. Data Illustration

5.1 Tonnage dataset

We first propose to apply the proposed methodology to monitor the multi-channel tonnage profile collected in a forging process in (Lei et al., 2010), where four different strain gauge sensors are mounted at each column of the forging machine, measuring the exerted force of the press. The setup of the process is shown in Figure 1a. The four strain gauge sensors represent the signature of the product and are used for process monitoring and change detection in Lei et al. (2010). For example, 10 examples of the signals before the changes and after the changes are shown in Figure 1b. As mentioned by Lei et al. (2010); Yan et al. (2018), a missing part only affects a small region of the signals, which makes it very hard to detect, as shown in Figure 1b.

We select a subset of the data with $n = 200$, where the first 130 observations are from the normal tonnage sample, and the samples after 130 are abnormal. We project the data to 20-dimensional space by training an anomaly basis on a holdout sample as has been done in Yan et al. (2018). The first 100 observations are treated as a Phase I stage without any changes and we learn the 2-norm and q-norm of the covariance matrix from them. The monitoring scheme started at observation 107 (trimming due to $q = 6$). The L_6-based test stopped at time 137, and estimated the possible change point location at time 128 by performing a retrospective test at time 137. The L_2 based test signaled slightly earlier at time 135 and also estimated the change at 128. The combined test signaled an alarm at time 135 with the same estimated location. The trajectory of the L_2 and the L_6 test statistics are shown in Figure 2a and 2b, respectively. Notice that, when we set the size of individual test to be $\alpha^* = (1-0.1)^{1/2} = 5.13\%$, the size of the combined test will be $\alpha = 1 - \alpha^{*2} = 0.1$. We signal an alarm when at least one test statistic exceeds the corresponding threshold.



(a) L_2 based test for tonnage data (b) L_6 based test for tonnage data

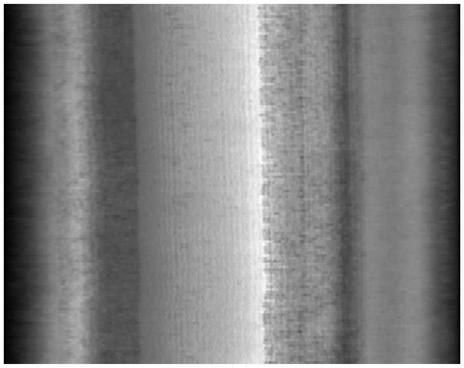
Figure 2: Testing Statistics for tonnage data

5.2 Rolling dataset

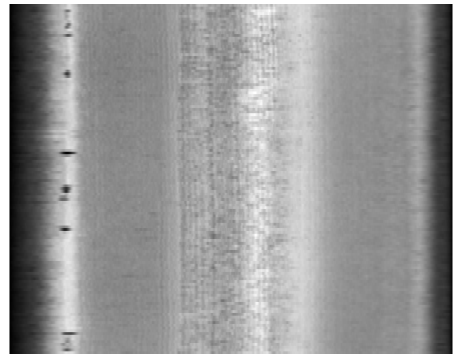
We then consider the process monitoring in a steel rolling manufacturing process. Surface defects such as seam defects can result in stress concentration on the bulk and may cause failures if the steel bar is used in the product. However, the rolling process is a high-speed process with the rolling bar moving around 200 miles per hour, providing real-time online anomaly detection for the high-speed rolling process is very important to prevent further the product damage.

The dataset is collected in such high-speed rolling process. Here, we have selected a segment near the end of the rolling bar, which contains 100 images of the rolling process. To remove the trend, we have applied a smooth background remover and further downsample the image to only 16×64 pixels. An example of the normal image and the abnormal image are shown in Figure 3a and 3b, respectively.

We treated the first 50 observations as training set and obtained ratio-consistent estimators $\widehat{\|\Sigma\|_q^q}$. After performing the change point monitoring procedure, the L_6-norm-based test signals an alarm at the time 97 and estimated that the possible change point location is at time 89 based on the retrospective test. On the other hand, the L_2 based test fails to detect the change within the finite time horizon. The combined test also signals the alarm at time 97. We present the rolling image at time 91 in Figure 3b. This shows that after downsampling, the change is still quite sparse. The adaptive monitoring procedure is still powerful as long as one test has power. We also present the trajectory of the test statistic at each time point in Figure 4a and 4b. Notice that there is a downshift in the L_2-based monitoring statistic right after the estimated change. This is due to the fact that the signal is very sparse, and the construction of our proposed statistic may admit negative values for a short period of time. The negative values here should not be a major concern as the test



(a) Normal rolling image



(b) Abnormal rolling image

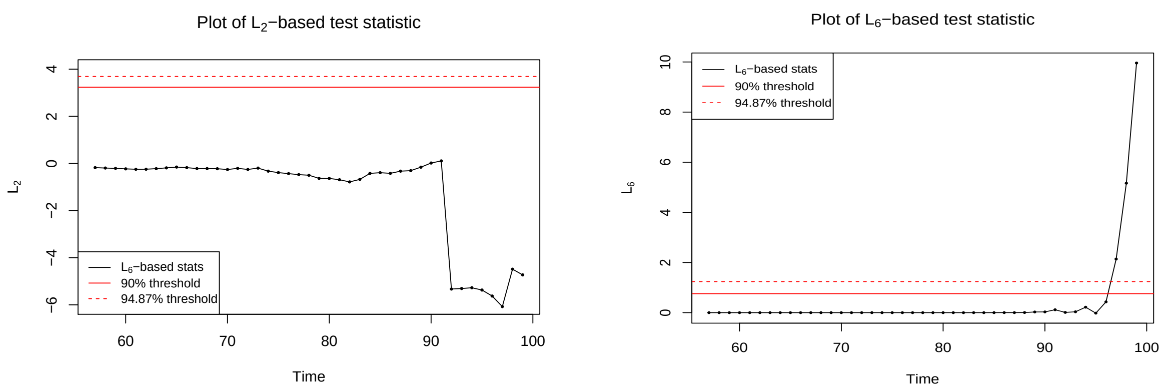
Figure 3: Examples of the rolling images

statistic should admit positive values in probability under the alternatives. We confirmed this by adding an artificial dense change in the data. On the other hand, L_6-based monitoring statistics detect the change efficiently due to the ability to capture the sparse change in the system.

6. Summary and Conclusion

In this article, we propose a new methodology to monitor a mean shift in temporally independent high-dimensional observations. Our change point monitoring method targets at the L_q-norm of the mean change for $q = 2, 4, 6, \cdots$. By combining the monitoring statistics for different values of $q \in 2\mathbb{N}$, the adaptive procedure achieves overall satisfactory power against both sparse and dense changes in the mean. This can be very desirable from a practitioner's viewpoint as often we do not have the knowledge about the type of alternatives. Compared with the recently developed methods for monitoring large-scale data streams [e.g., Mei (2010), Xie and Siegmund (2013), Liu et al. (2019)], our method is fully nonparametric and does not require strong distributional assumptions. Furthermore, our method allows for certain cross-sectional dependence across data streams, which could naturally arise in many applications.

To conclude, we mention a few interesting future directions. Firstly, our focus in this paper is on the mean change, and it is natural to ask whether the method can be extended to monitor a change in the covariance matrix. Secondly, many streaming data have weak dependence over time due to its sequential nature, and how to accommodate weak temporal dependence will be of interest. Thirdly, in the current implementation, the ratio-consistent estimators are learned from the training data and do not change as more observations are available. In practice, if the monitoring scheme runs for a long time without signaling an alarm, it might be helpful to periodically update ratio-consistent estimators to gain efficiency,



(a) L_2 based test for rolling data

(b) L_6 based test for rolling data

Figure 4: Examples of the rolling images

especially when the initial training sample is short. However, it may be impractical to update this estimator for each k since there seems no easy recursive way to update this estimator and the associated computational cost is high. The user might need to determine how often to update it based on the actual computational resources. Fourthly, even though the proposed algorithm can detect the sparse change, in many applications, it is also an important problem to identify which individual data stream actually experiences a change, which will be left for future research.

Supplementary Materials

The supplementary materials contains technical proofs for the theoretical results as well as additional simulation results.

Proof of Theorem 1. We can directly apply the results shown in Wang et al. (2019) for the partial sum process

$$S_n(a,b) = \sum_{i=\lfloor na \rfloor+1}^{\lfloor nb \rfloor-1} \sum_{j=\lfloor na \rfloor+1}^{i} X_{i+1}^T X_j.$$

The partial sum process

$$\left\{ \frac{\sqrt{2}}{n||\Sigma||_F} S_n(a,b) \right\}_{(a,b)\in[0,T]^2} \rightsquigarrow Q \quad in \quad l^\infty([0,T]^2)$$

where Q is a Gaussian process whose covariance structure is the following

$$Cov(Q(a_1, b_1), (a_2, b_2)) = \begin{cases} (\min(b_1, b_2) - \max(a_1, a_2))^2 & if \quad \max(a_1, a_2) \leq \min(b_1, b_2) \\ 0 & otherwise \end{cases}$$

The test statistic is a continuous transformation of the Gaussian process and the results stated follows. □

Proof of Theorem 2. We now analyze the power of the first proposed test. Suppose the change point is at k^*, where $k^*/n \to r$ for some constant $r \in (1, T)$. This assures that the change point does not occur extremely early or late in the monitoring period. Under the alternative hypothesis, define a new sequence of random vectors Y_i,

$$Y_i = \begin{cases} X_i & i = 1, \dots, k^* \\ X_i - \Delta & i = k^* + 1, \dots, n \end{cases}.$$

This sequence does not have a change point. Without loss of generosity, assume Y_i's are centered.

Suppose that

$$\frac{n\Delta^T \Delta}{||\Sigma||_F} \to b \in [0 + \infty).$$

When $m < k < k^*$, $G_k(m)$ statistic will not be affected. It suffices to consider the case $m < k^* < k$ and $k^* < m < k$. Following the decomposition in Wang et al. (2019), under the fixed alternative when $k^* > m$,

$$\begin{aligned} G_k(m) = & G_k^Y(m) + (k - k^*)(k - k^* - 1)m(m-1)||\Delta||_2^2 \\ & - 2(k - k^*)(k - m - 2)(m-1)\sum_{j=1}^{m} Y_j^T \Delta \\ & - 4(m-1)(m-2)(k - k^*)\sum_{j=m+1}^{k^*} Y_j^T \Delta. \end{aligned}$$

$G_n^Y(m)$ is the statistic calculated for the Y_i sequence. Let $s_n(k) = \sum_{j=1}^{k} Y_j^T \Delta$. Then

$$\sup_{1 \leq l \leq k \leq nT} |\sum_{j=l}^{k} Y_j^T \Delta| \leq 2 \sup_{1 \leq k \leq nT} |s_n(k)| = O_p(n^{1/2}(\Delta^T \Sigma \Delta)^{1/2}).$$

The last part is obtained by Kolmogorov's inequality. This implies that when $k^* > m$,

$$\frac{1}{n^3\|\Sigma\|_F} G_k(m) = \frac{1}{n^3\|\Sigma\|_F} G_k^Y(m) + \frac{(k - k^*)(k - k^* - 1)m(m-1)}{n^3} \frac{||\Delta||_2^2}{||\Sigma||_F} + O_p(\frac{n^{1/2}(\Delta^T \Sigma \Delta)^{1/2}}{||\Sigma||_F}).$$

Similarly, we can show when $k^* > m$

$$\frac{1}{n^3\|\Sigma\|_F}G_k(m) = \frac{1}{n\|\Sigma\|_F}G_k^Y(m)+\frac{k^*(k^*-1)(k-m)(k-m-1)}{n^3}\frac{||\Delta||_2^2}{||\Sigma||_F}+O_p(\frac{n^{1/2}(\Delta^T\Sigma\Delta)^{1/2}}{||\Sigma||_F}).$$

The last part is converging to 0 in probability. Therefore, the test statistic T_n can be viewed as an extension to the original process. The second terms are also a process depend on m and k^*. Under the fixed alternative, the $G_k(m)$ converge to the process

$$\frac{1}{n^3\|\Sigma\|_F}\{G_{\lfloor nt\rfloor}(\lfloor ns\rfloor)\}_{s\in[0,1]} \to G(s,t)+b\Lambda(s,t),$$

where

$$\Lambda(s,t) = \begin{cases} (t-r)^2s^2 & s \leq r \\ r^2(t-s)^2 & s > r \\ 0 & otherwise \end{cases}.$$

This implies that, when $b = 0$, the process is the same with the null process, and the proposed monitoring scheme will have trivial power. When the b is not zero, since the remainder term is positive, we will have non -trivial power.

When

$$\frac{n\Delta^T\Delta}{||\Sigma||_F} \to \infty.$$

Following above decomposition, we have

$$\max_k T_n(k) \geq T_n(k^*) = \frac{1}{n\|\Sigma\|_F}D_{nT}^Y(k^*) + O(\frac{n||\Delta||_2^2}{||\Sigma||_F}) \to \infty$$

Since the first term is pivotal and is bounded in probability, the test have power converging to 1. □

Proof of Theorem 3. We can directly apply the results in Theorem 2.1 and 2.2 in Zhang et al.(2020), which stated that for

$$S_{n,q,c}(r;[a,b]) = \sum_{l=1}^{p}\sum^{*}_{\lfloor na\rfloor+1\leq i_1,\ldots,i_c\leq\lfloor nr\rfloor}\sum^{*}_{\lfloor nr\rfloor+1\leq j_1,\ldots,j_{q-c}\leq\lfloor nb\rfloor}\prod_{t=1}^{c}X_{i_t,l}\prod_{g=1}^{q-c}X_{j_g,l},$$

we have

$$\frac{1}{\sqrt{n^q\|\Sigma\|_q^q}}S_{n,q,c}(r;[a,b]) \rightsquigarrow Q_{q,c}(r;[a,b]),$$

where $Q_{q,c}$ is the Gaussian process stated in Theorem 4. The monitoring statistic is a continuous transformation of process $S_{n,q,c}$'s and the asymptotic result follows. □

Proof of Theorem 4. We first discuss the case when $\frac{n^{q/2}\|\Delta\|_q^q}{\|\Sigma\|_q^{q/2}} \to \gamma \in [0, +\infty)$ and the true change point is at location $k^* = \lfloor nr \rfloor$. Here we adopt the process convergence results in Theorem 2.3 of Zhang et al.(2020), which stated that for $(k, m) = (\lfloor ns \rfloor, \lfloor nt \rfloor)$,

$$\frac{1}{\sqrt{n^{3q}\|\Sigma\|_q^q}} D_{n,q}(s; [0,b]) = \frac{1}{\sqrt{n^{3q}\|\Sigma\|_q^q}} \sum_{l=1}^{p} \sum_{0 \le i_1, \ldots, i_q \le k}^{*} \sum_{k+1 \le j_1, \ldots, j_q \le m}^{*} (X_{i_1,l} - X_{j_1,l}) \cdots (X_{i_q,l} - X_{j_q,l}),$$
$$\rightsquigarrow G_q(s,t) + \gamma J_q(s; [0,t])$$

where

$$J_q(s; [0,t]) = \begin{cases} r^q (t-s)^q & r \le s < t \\ s^q (t-r)^q & s \le r < t \\ 0 & otherwise \end{cases}$$

Therefore, by continuous mapping theorem, when $\gamma \in [0, +\infty)$, the results in the theorem hold.

For the case $\frac{n^{q/2}\|\Delta\|_q^q}{\|\Sigma\|_q^{q/2}} \to +\infty$

$$\max_k T_{n,q}(k) \ge T_{n,q}(k^*) = \frac{1}{n\|\Sigma\|_F} D_{nT}^Y(k^*) + C \frac{n^{q/2}\|\Delta\|_q^q}{\|\Sigma\|_q^{q/2}} \to \infty$$

□

Proof of Theorem 5. By straightforward calculation, we have

$$\begin{aligned}
\widehat{\|\Sigma\|_F^2} &= \frac{1}{4\binom{n}{4}} \sum_{1 \le j_1 < j_2 < j_3 < j_4 \le n} tr\left((X_{j_1} - X_{j_2})(X_{j_1} - X_{j_2})^T (X_{j_3} - X_{j_4})(X_{j_3} - X_{j_4})^T\right) \\
&= \frac{1}{4\binom{n}{4}} \sum_{1 \le j_1 < j_2 < j_3 < j_4 \le n} [(X_{j_1} - X_{j_2})^T (X_{j_3} - X_{j_4})]^2 \\
&= \frac{1}{4\binom{n}{4}} \sum_{1 \le j_1 < j_2 < j_3 < j_4 \le n} [(X_{j_1}^T X_{j_3})^2 + (X_{j_2}^T X_{j_3})^2 + (X_{j_2}^T X_{j_4})^2 + (X_{j_1}^T X_{j_4})^2] \\
&- \frac{2}{4\binom{n}{4}} \sum_{1 \le j_1 < j_2 < j_3 < j_4 \le n} [X_{j_1}^T X_{j_3} X_{j_1}^T X_{j_4} + X_{j_2}^T X_{j_3} X_{j_2}^T X_{j_4} + X_{j_1}^T X_{j_3} X_{j_2}^T X_{j_3} + X_{j_1}^T X_{j_4} X_{j_2}^T X_{j_4}] \\
&+ \frac{2}{4\binom{n}{4}} \sum_{1 \le j_1 < j_2 < j_3 < j_4 \le n} [X_{j_1}^T X_{j_3} X_{j_2}^T X_{j_4} + X_{j_2}^T X_{j_3} X_{j_1}^T X_{j_4}] \\
&= I_{n,1} + I_{n,2} + I_{n,3} + I_{n,4} - (I_{n,5} + I_{n,6} + I_{n,7} + I_{n,8}) + (I_{n,9} + I_{n,10}).
\end{aligned}$$

For $I_{n,1}$,

$$\mathbb{E}[I_{n,1}] = \frac{1}{4\binom{n}{4}} \sum_{1\le j_1<j_2<j_3<j_4\le n} \mathbb{E}[(X_{j_1}^T X_{j_3})^2] = \frac{1}{4} tr(\mathbb{E}[X_{j_3}X_{j_3}^T X_{j_1}X_{j_1}^T]) = \|\Sigma\|_F^2/4.$$

Thus $\mathbb{E}[I_{n,1}/\|\Sigma\|_F^2] = 1/4$. By similar arguments, it is obvious to see that $\mathbb{E}[I_{n,i}/\|\Sigma\|_F^2] = 1/4$ for $i = 1, 2, 3, 4$, and $\mathbb{E}[I_{n,i}/\|\Sigma\|_F^2] = 0$ for $i = 5, ..., 10$.

The outline of the proof is as following. We will show that $4I_{n,i}/\|\Sigma\|_F^2 \to_p 1$ for $i = 1, 2, 3, 4$, and $I_{n,i}/\|\Sigma\|_F^2 \to_p 0$, for $i = 5, ..., 10$. Since some of the $I_{n,i}$ share very similar structures, we will only present the proof for (1) $4I_{n,1}/\|\Sigma\|_F^2 \to_p 1$ and (2) $I_{n,5}/\|\Sigma\|_F^2 \to_p 0$. Other terms can be proved by similar arguments.

To show (1), it suffices to show that $\mathbb{E}[16I_{n,1}^2/\|\Sigma\|_F^4] \to 1$. To see this,

$$\begin{aligned}
\mathbb{E}[16I_{n,1}^2/\|\Sigma\|_F^4] &= \frac{1}{\binom{n}{4}^2\|\Sigma\|_F^4} \sum_{1\le j_1<j_2<j_3<j_4\le n} \sum_{1\le j_5<j_6<j_7<j_8\le n} \mathbb{E}[(X_{j_1}^T X_{j_3})^2 (X_{j_5}^T X_{j_7})^2] \\
&= \frac{1}{\binom{n}{4}^2\|\Sigma\|_F^4} \sum_{1\le j_1<j_2<j_3<j_4\le n} \sum_{1\le j_5<j_6<j_7<j_8\le n} \sum_{l_1,l_2,l_3,l_4=1}^{p} \mathbb{E}[X_{j_1,l_1}X_{j_3,l_1}X_{j_1,l_2}X_{j_3,l_2}X_{j_5,l_3}X_{j_7,l_3}X_{j_5,l_4}X_{j_7,l_4}].
\end{aligned}$$

As we know that the expectation of a product of random variables can be expressed in terms of joint cumulants, we have

$$\mathbb{E}[X_{j_1,l_1}X_{j_3,l_1}X_{j_1,l_2}X_{j_3,l_2}X_{j_5,l_3}X_{j_7,l_3}X_{j_5,l_4}X_{j_7,l_4}] = \sum_{\pi}\prod_{B\in\pi} cum(X_{j,l} : (j,l) \in B),$$

where π runs through the list of all partitions of $\{(j_1, l_1), (j_1, l_2), ..., (j_7, l_3), (j_7, l_4)\}$ and B runs through the list of all blocks of the partition π. Since $j_1 < j_3$ and $j_5 < j_7$, it is impossible to have three or more indices in $\{j_1, j_3, j_5, j_7\}$ such that they are identical. Thus for the right hand side of the above expression, we only need to take the sum over all partitions with all block sizes smaller than 5, because for joint cumulants with order greater than 5, it must contain at least 3 indices from j_1, j_3, j_5, j_7 and at least one is not identical to the other two. And the joint cumulants will equal to zero since it involves two or more independent random variables.

Also since the mean of all random variables included in the left hand side of the above expression are all zero, we do not need to consider the partition with block size 1. Thus the expression can be simplified as

$$\begin{aligned}
&\mathbb{E}[X_{j_1,l_1}X_{j_3,l_1}X_{j_1,l_2}X_{j_3,l_2}X_{j_5,l_3}X_{j_7,l_3}X_{j_5,l_4}X_{j_7,l_4}] \\
=&C_1^{(j_1,j_3,j_5,j_7)}\mathbb{E}[X_{0,l_1}X_{0,l_2}X_{0,l_3}X_{0,l_4}]^2 + C_2^{(j_1,j_3,j_5,j_7)}\mathbb{E}[X_{0,l_1}X_{0,l_2}X_{0,l_3}X_{0,l_4}]\Sigma_{l_1,l_2}\Sigma_{l_3,l_4} \\
&+ \Sigma_{l_1,l_2}^2\Sigma_{l_3,l_4}^2,
\end{aligned}$$

where $C_1^{(j_1,j_3,j_5,j_7)}$, $C_2^{(j_1,j_3,j_5,j_7)}$ are finite positive constants purely based on the value of

j_1, j_3, j_5, j_7. $C_1^{(j_1,j_3,j_5,j_7)}$ can only be nonzero if $j_1 = j_5$ and $j_3 = j_7$, and $C_2^{(j_1,j_3,j_5,j_7)}$ is nonzero if at least two of (j_1, j_3, j_5, j_7) are equal. This implies that

$$\sum_{1 \le j_1 < j_2 < j_3 < j_4 \le n} \sum_{1 \le j_5 < j_6 < j_7 < j_8 \le n} C_1^{(j_1,j_3,j_5,j_7)} = o(n^8),$$

and

$$\sum_{1 \le j_1 < j_2 < j_3 < j_4 \le n} \sum_{1 \le j_5 < j_6 < j_7 < j_8 \le n} C_2^{(j_1,j_3,j_5,j_7)} = o(n^8).$$

Furthermore, according to Assumption 2, $\sum_{l_1,l_2,l_3,l_4=1}^{p} cum(X_{0,l_1}, X_{0,l_2}, X_{0,l_3}, X_{0,l_4})^2 \le C\|\Sigma\|_F^4$. It can be verified that

$$\begin{aligned}\sum_{l_1,l_2,l_3,l_4=1}^{p} \mathbb{E}[X_{0,l_1}X_{0,l_2}X_{0,l_3}X_{0,l_4}]^2 \lesssim& \sum_{l_1,l_2,l_3,l_4=1}^{p} cum(X_{0,l_1}, X_{0,l_2}, X_{0,l_3}, X_{0,l_4})^2 \\ &+ \sum_{l_1,l_2,l_3,l_4=1}^{p} \Sigma_{l_1,l_2}^2 \Sigma_{l_3,l_4}^2 \\ &\lesssim \|\Sigma\|_F^4,\end{aligned}$$

and by using the Cauchy-Schwartz inequaility,

$$\begin{aligned}&\sum_{l_1,l_2,l_3,l_4=1}^{p} \mathbb{E}[X_{0,l_1}X_{0,l_2}X_{0,l_3}X_{0,l_4}]\Sigma_{l_1,l_2}\Sigma_{l_3,l_4} \\ \le&\sqrt{\sum_{l_1,l_2,l_3,l_4=1}^{p} \mathbb{E}[X_{0,l_1}X_{0,l_2}X_{0,l_3}X_{0,l_4}]^2}\sqrt{\sum_{l_1,l_2,l_3,l_4=1}^{p} \Sigma_{l_1,l_2}^2 \Sigma_{l_3,l_4}^2} \le \sqrt{C}\|\Sigma\|_F^4. \end{aligned} \tag{6.4}$$

This indicates that

$$\begin{aligned}&\mathbb{E}[16 I_{n,1}^2/\|\Sigma\|_F^4] \\ =&\frac{1}{\binom{n}{4}^2\|\Sigma\|_F^4} \sum_{1 \le j_1 < j_2 < j_3 < j_4 \le n} \sum_{1 \le j_5 < j_6 < j_7 < j_8 \le n} C_1^{(j_1,j_3,j_5,j_7)} \sum_{l_1,l_2,l_3,l_4=1}^{p} \mathbb{E}[X_{0,l_1}X_{0,l_2}X_{0,l_3}X_{0,l_4}]^2 \\ +&\frac{1}{\binom{n}{4}^2\|\Sigma\|_F^4} \sum_{1 \le j_1 < j_2 < j_3 < j_4 \le n} \sum_{1 \le j_5 < j_6 < j_7 < j_8 \le n} C_2^{(j_1,j_3,j_5,j_7)} \sum_{l_1,l_2,l_3,l_4=1}^{p} \mathbb{E}[X_{0,l_1}X_{0,l_2}X_{0,l_3}X_{0,l_4}]\Sigma_{l_1,l_2}\Sigma_{l_3,l_4} \\ +&\frac{1}{\binom{n}{4}^2\|\Sigma\|_F^4} \sum_{1 \le j_1 < j_2 < j_3 < j_4 \le n} \sum_{1 \le j_5 < j_6 < j_7 < j_8 \le n} \sum_{l_1,l_2,l_3,l_4=1}^{p} \Sigma_{l_1,l_2}^2 \Sigma_{l_3,l_4}^2 = o(1) + o(1) + 1 \to 1.\end{aligned}$$

Thus, $4I_{n,1}/\|\Sigma\|_F^2 \to_p 1$, and (1) is proved. By similar arguments, $4I_{n,i}/\|\Sigma\|_F^2 \to_p 1$ holds for $i = 2, 3, 4$.

To show (2), we need to prove $\mathbb{E}[I_{n,5}^2/\|\Sigma\|_F^4] \to 0$. To see this,

$$\mathbb{E}[I_{n,5}^2/\|\Sigma\|_F^4] = \frac{1}{4\binom{n}{4}^2\|\Sigma\|_F^4} \sum_{1\le j_1<j_2<j_3<j_4\le n} \sum_{1\le j_5<j_6<j_7<j_8\le n} \mathbb{E}[(X_{j_1}^T X_{j_3} X_{j_1}^T X_{j_4})(X_{j_5}^T X_{j_7} X_{j_5}^T X_{j_8})]$$

$$= \frac{1}{\binom{n}{4}^2\|\Sigma\|_F^4} \sum_{1\le j_1<j_2<j_3<j_4\le n} \sum_{1\le j_5<j_6<j_7<j_8\le n} \sum_{l_1,l_2,l_3,l_4=1}^{p} \mathbb{E}[X_{j_1,l_1} X_{j_3,l_1} X_{j_1,l_2} X_{j_4,l_2} X_{j_5,l_3} X_{j_7,l_3} X_{j_5,l_4} X_{j_8,l_4}].$$

By similar arguments for the joint cumulants we provided in the the proof for (1), it can be proved that

$$\begin{aligned}&\mathbb{E}[X_{j_1,l_1} X_{j_3,l_1} X_{j_1,l_2} X_{j_4,l_2} X_{j_5,l_3} X_{j_7,l_3} X_{j_5,l_4} X_{j_8,l_4}]\\ =&C_1^{(j_1,j_3,j_4,j_5,j_7,j_8)}\mathbb{E}[X_{0,l_1}X_{0,l_2}X_{0,l_3}X_{0,l_4}]\Sigma_{l_1,l_3}\Sigma_{l_2,l_4} + C_2^{(j_1,j_3,j_4,j_5,j_7,j_8)}\Sigma_{l_1,l_2}\Sigma_{l_3,l_4}\Sigma_{l_1,l_3}\Sigma_{l_2,l_4}.\end{aligned}$$

If $C_1^{(j_1,j_3,j_4,j_5,j_7,j_8)} \neq 0$, then $j_1 = j_5$. And if $C_2^{(j_1,j_3,j_4,j_5,j_7,j_8)} \neq 0$, $j_3 = j_5$ and $j_4 = j_8$. These two properties guarantee that

$$\sum_{1\le j_1<j_2<j_3<j_4\le n} \sum_{1\le j_5<j_6<j_7<j_8\le n} C_1^{(j_1,j_3,j_4,j_5,j_7,j_8)} = o(n^8),$$

and

$$\sum_{1\le j_1<j_2<j_3<j_4\le n} \sum_{1\le j_5<j_6<j_7<j_8\le n} C_2^{(j_1,j_3,j_4,j_5,j_7,j_8)} = o(n^8).$$

Furthermore we have shown the bound for $\sum_{l_1,l_2,l_3,l_4=1}^{p} \mathbb{E}[X_{0,l_1}X_{0,l_2}X_{0,l_3}X_{0,l_4}]\Sigma_{l_1,l_2}\Sigma_{l_3,l_4}$ in (6.5). And

$$\begin{aligned}&\sum_{l_1,l_2,l_3,l_4=1}^{p} \Sigma_{l_1,l_2}\Sigma_{l_3,l_4}\Sigma_{l_1,l_3}\Sigma_{l_2,l_4} = \sum_{l_1,l_4=1}^{p}\left(\sum_{l_2=1}^{p}\Sigma_{l_1,l_2}\Sigma_{l_2,l_4}\right)\left(\sum_{l_3=1}^{p}\Sigma_{l_1,l_3}\Sigma_{l_3,l_4}\right)\\ &= \sum_{l_1,l_4=1}^{p}[(\Sigma^2)_{l_1,l_4}]^2 = tr(\Sigma^4) = o(\|\Sigma\|_F^4),\end{aligned}$$

by Assumption 1. Thus,

$$
\begin{aligned}
&\mathbb{E}[I_{n,5}^2/\|\Sigma\|_F^4]\\
=&\frac{1}{4\binom{n}{4}^2\|\Sigma\|_F^4}\sum_{1\le j_1<j_2<j_3<j_4\le n}\sum_{1\le j_5<j_6<j_7<j_8\le n}C_1^{(j_1,j_3,j_4,j_5,j_7,j_8)}\sum_{l_1,l_2,l_3,l_4=1}^{p}\mathbb{E}[X_{0,l_1}X_{0,l_2}X_{0,l_3}X_{0,l_4}]\Sigma_{l_1,l_3}\Sigma_{l_2,l_4}\\
+&\frac{1}{4\binom{n}{4}^2\|\Sigma\|_F^4}\sum_{1\le j_1<j_2<j_3<j_4\le n}\sum_{1\le j_5<j_6<j_7<j_8\le n}C_2^{(j_1,j_3,j_4,j_5,j_7,j_8)}\sum_{l_1,l_2,l_3,l_4=1}^{p}\Sigma_{l_1,l_2}\Sigma_{l_3,l_4}\Sigma_{l_1,l_3}\Sigma_{l_2,l_4}\\
=&o(1)+o(1)\to 1.
\end{aligned}
$$

This indicates $I_{n,5}/\|\Sigma\|_F^2 \to_p 0$. And by similar arguments we can prove that $I_{n,i}/\|\Sigma\|_F^2 \to_p 0$, for all $i = 6, ..., 10$. Combine the above results, we have $\widehat{\|\Sigma\|_F^2}/\|\Sigma\|_F^2 \to_p 1$. This completes the proof.

□

Proof of Theorem 6. We can rewrite $\widehat{\|\Sigma\|_q^q}$ as

$$
\begin{aligned}
&\widehat{\|\Sigma\|_q^q} = \frac{1}{2^q\binom{n}{2q}}\sum_{l_1,l_2=1}^{p}\sum_{1\le i_1<\cdots<i_q<j_1<\cdots<j_q\le n}\prod_{k=1}^{q}(X_{i_k,l_1}X_{i_k,l_2}+X_{j_k,l_1}X_{j_k,l_2}-X_{i_k,l_1}X_{j_k,l_2}-X_{j_k,l_1}X_{i_k,l_2})\\
=&\frac{1}{2^q\binom{n}{2q}}\sum_{1\le i_1<\cdots<i_q<j_1<\cdots<j_q\le n}\sum_{t_1,s_1\in\{i_1,j_1\}}\cdots\sum_{t_q,s_q\in\{i_q,j_q\}}\sum_{l_1,l_2=1}^{p}\prod_{k=1}^{q}(-1)^{\mathbf{1}\{t_k\neq s_k\}}X_{t_k,l_1}X_{s_k,l_2}\\
=&\frac{1}{2^q\binom{n}{2q}}\sum_{1\le i_1<\cdots<i_q<j_1<\cdots<j_q\le n}\sum_{t_1\in\{i_1,j_1\}}\cdots\sum_{t_q\in\{i_q,j_q\}}\sum_{l_1,l_2=1}^{p}\prod_{k=1}^{q}X_{t_k,l_1}X_{t_k,l_2}\\
&+\frac{1}{2^q\binom{n}{2q}}\sum_{1\le i_1<\cdots<i_q<j_1<\cdots<j_q\le n}\sum_{t_1,s_1\in\{i_1,j_1\}}\cdots\sum_{t_q,s_q\in\{i_q,j_q\}}\\
&\qquad\sum_{l_1,l_2=1}^{p}\mathbf{1}\{\cup_{k=1}^{q}\{t_k\neq s_k\}\}\prod_{k=1}^{q}(-1)^{\mathbf{1}\{t_k\neq s_k\}}X_{t_k,l_1}X_{s_k,l_2}.
\end{aligned}
$$

The second equality in the above expression is by calculating the cross products among q brackets, and the third equality is splitting the terms based on different values of t_k, s_k for $k = 1, ..., q$. The first term in the third equality contains all products with $t_k = s_k$ for all $k = 1, ..., q$, and the second term contains products with at least one $k = 1, ..., q$ such that $t_k \neq s_k$.

The outline of the proof is as follows. We want to show:

1. for every $t_1 \in \{i_1, j_1\}, ..., t_q \in \{i_q, j_q\}$,

$$I(t_1, ..., t_q) = \frac{1}{\binom{n}{2q}\|\Sigma\|_q^q} \sum_{1\le i_1<\cdots<i_q<j_1<\cdots<j_q\le n} \sum_{l_1,l_2=1}^{p} \prod_{k=1}^{q} X_{t_k,l_1}X_{t_k,l_2} \to_p 1;$$

2. for every $t_1, s_1 \in \{i_1, j_1\}, ..., t_q, s_q \in \{i_q, j_q\}$ and there exists at least one $k = 1, ..., q$ such that $t_k \neq s_k$,

$$J(t_1, s_1, ..., t_q, s_q) = \frac{1}{\binom{n}{2q}\|\Sigma\|_q^q} \sum_{1\le i_1<\cdots<i_q<j_1<\cdots<j_q\le n} \sum_{l_1,l_2=1}^{p} \prod_{k=1}^{q} X_{t_k,l_1}X_{s_k,l_2} \to_p 0.$$

And it is easy to see that if these two results hold, then $\widehat{\|\Sigma\|_q^q}/\|\Sigma\|_q^q \to_p 1$. As we observe that most of terms are structurally very similar, we shall only present the proof for $I(i_1, ..., i_q) \to_p 1$ and a general proof for (2).

It is trivial to see that $\mathbb{E}[\sum_{l_1,l_2=1}^{p}\prod_{k=1}^{q} X_{t_k,l_1}X_{t_k,l_2}/\|\Sigma\|_q^q] = 1$. This indicates that to show (1), it suffices to show that $\mathbb{E}[I(i_1, ..., i_q)^2] \to 1$. To show this,

$$\begin{aligned}\mathbb{E}[I(i_1, ..., i_q)^2] =& \frac{1}{\binom{n}{2q}^2\|\Sigma\|_q^{2q}} \sum_{1\le i_1<\cdots<i_q<j_1<\cdots<j_q\le n} \sum_{1\le i'_1<\cdots<i'_q<j'_1<\cdots<j'_q\le n} \\ & \sum_{l_1,l_2,l_3,l_4=1}^{p} \mathbb{E}\left[\prod_{k=1}^{q} X_{i_k,l_1}X_{i_k,l_2}X_{i'_k,l_3}X_{i'_k,l_4}\right].\end{aligned}$$

Due to the special structure of our statistic,

$$\mathbb{E}\left[\prod_{k=1}^{q} X_{i_k,l_1}X_{i_k,l_2}X_{i'_k,l_3}X_{i'_k,l_4}\right] = \sum_{m=0}^{q} C_m E(X_{0,l_1}X_{0,l_2}X_{0,l_3}X_{0,l_4})^m(\Sigma_{l_1,l_2}\Sigma_{l_3,l_4})^{q-m},$$

where $C_m = C_m(i_1, ..., i_q, i'_1, ..., i'_q) \ge 0$ is a function of all indices for all $m = 1, 2, ..., q$. $C_m = 1$ if there are exact m indices in $\{i_1, ..., i_q\}$ which equal to m indices in $\{i'_1, ..., i'_q\}$, and $C_m = 0$ otherwise. These events are mutually exclusive which indicates that $\sum_{m=0}^{q} C_m = 1$. This indicates that for all $m = 1, ..., q$,

$$\sum_{1\le i_1<\cdots<i_q<j_1<\cdots<j_q\le n} \sum_{1\le i'_1<\cdots<i'_q<j'_1<\cdots<j'_q\le n} C_m(i_1, ..., i_q, i'_1, ..., i'_q) = o(n^{4q}),$$

and

$$\frac{1}{\binom{n}{2q}^2} \sum_{1\le i_1<\cdots<i_q<j_1<\cdots<j_q\le n} \sum_{1\le i'_1<\cdots<i'_q<j'_1<\cdots<j'_q\le n} C_0(i_1, ..., i_q, i'_1, ..., i'_q) \to 1.$$

Furthermore, for any $m = 1, ..., q$, by Hölder's inequaility for vector spaces, we have

$$\begin{aligned}
&\sum_{l_1,l_2,l_3,l_4=1}^{p} |E(X_{0,l_1}X_{0,l_2}X_{0,l_3}X_{0,l_4})|^m |\Sigma_{l_1,l_2}\Sigma_{l_3,l_4}|^{q-m} \\
\leq &\left(\sum_{l_1,l_2,l_3,l_4=1}^{p} |E(X_{0,l_1}X_{0,l_2}X_{0,l_3}X_{0,l_4})^m|^{q/m}\right)^{m/q} \left(\sum_{l_1,l_2,l_3,l_4=1}^{p} (|\Sigma_{l_1,l_2}\Sigma_{l_3,l_4}|^{q-m})^{q/(q-m)}\right)^{(q-m)/q} \\
= &\left(\sum_{l_1,l_2,l_3,l_4=1}^{p} |E(X_{0,l_1}X_{0,l_2}X_{0,l_3}X_{0,l_4})|^q\right)^{m/q} \left(\sum_{l_1,l_2,l_3,l_4=1}^{p} |\Sigma_{l_1,l_2}|^q |\Sigma_{l_3,l_4}|^q\right)^{(q-m)/q} \\
\leq &C\|\Sigma\|_q^{2m}\|\Sigma\|_q^{2(q-m)} = C\|\Sigma\|_q^{2q},
\end{aligned}$$

where the last inequality is due to Assumption 5, and to see this,

$$\begin{aligned}
&\sum_{l_1,l_2,l_3,l_4=1}^{p} |E(X_{0,l_1}X_{0,l_2}X_{0,l_3}X_{0,l_4})|^q \\
\leq &C\sum_{l_1,l_2,l_3,l_4=1}^{p} |cum(X_{0,l_1}, X_{0,l_2}, X_{0,l_3}, X_{0,l_4})|^q + |\Sigma_{l_1,l_2}\Sigma_{l_3,l_4}|^q + |\Sigma_{l_1,l_3}\Sigma_{l_2,l_4}|^q + |\Sigma_{l_1,l_4}\Sigma_{l_2,l_3}|^q \\
\leq &C\sum_{1\leq l_1\leq l_2\leq l_3\leq l_4\leq p} (1 \vee (l_4 - l_1))^{-2rq} + 3C\|\Sigma\|_q^{2q} \leq Cp^2 + 3C\|\Sigma\|_q^{2q} \leq C\|\Sigma\|_q^{2q},
\end{aligned}$$

for some generic positive constant C, since $\|\Sigma\|_q^{2q} = (\sum_{i,j=1}^{p} \Sigma_{i,j}^q)^2 \geq Cp^2$ under Assumption

(5.1). Therefore,

$$
\begin{aligned}
&\mathbb{E}[I(i_1,...,i_q)^2]\\
=&\frac{1}{\binom{n}{2q}^2\|\Sigma\|_q^{2q}}\sum_{1\le i_1<\cdots<i_q<j_1<\cdots<j_q\le n}\sum_{1\le i_1'<\cdots<i_q'<j_1'<\cdots<j_q'\le n}\sum_{l_1,l_2,l_3,l_4=1}^{q}\mathbb{E}\left[\prod_{k=1}^{p}X_{i_k,l_1}X_{i_k,l_2}X_{i_k',l_3}X_{i_k',l_4}\right]\\
=&\frac{1}{\binom{n}{2q}^2\|\Sigma\|_q^{2q}}\sum_{1\le i_1<\cdots<i_q<j_1<\cdots<j_q\le n}\sum_{1\le i_1'<\cdots<i_q'<j_1'<\cdots<j_q'\le n}\\
&\qquad\sum_{l_1,l_2,l_3,l_4=1}^{p}\sum_{m=0}^{q}C_m\mathbb{E}(X_{0,l_1}X_{0,l_2}X_{0,l_3}X_{0,l_4})^m(\Sigma_{l_1,l_2}\Sigma_{l_3,l_4})^{q-m}\\
=&\frac{1}{\binom{n}{2q}^2\|\Sigma\|_q^{2q}}\sum_{1\le i_1<\cdots<i_q<j_1<\cdots<j_q\le n}\sum_{1\le i_1'<\cdots<i_q'<j_1'<\cdots<j_q'\le n}C_0(i_1,...,i_q,i_1',...,i_q')\sum_{l_1,l_2,l_3,l_4=1}^{p}\Sigma_{l_1,l_2}^q\Sigma_{l_3,l_4}^q\\
+&\frac{1}{\binom{n}{2q}^2\|\Sigma\|_q^{2q}}o(n^{4q})\sum_{l_1,l_2,l_3,l_4=1}^{p}\sum_{m=1}^{q}\mathbb{E}(X_{0,l_1}X_{0,l_2}X_{0,l_3}X_{0,l_4})^m(\Sigma_{l_1,l_2}\Sigma_{l_3,l_4})^{q-m}\\
=&1+o(1)\to 1.
\end{aligned}
$$

This completes the proof for (1). To show (2), it suffices to show that $\mathbb{E}[J(t_1,s_1,...,t_q,s_q)^2]\to 0$. Specifically,

$$
\begin{aligned}
\mathbb{E}[J(t_1,s_1,...,t_q,s_q)^2]=&\frac{1}{\binom{n}{2q}^2\|\Sigma\|_q^{2q}}\sum_{1\le i_1<\cdots<i_q<j_1<\cdots<j_q\le n}\sum_{1\le i_1'<\cdots<i_q'<j_1'<\cdots<j_q'\le n}\\
&\sum_{l_1,l_2,l_3,l_4=1}^{p}\mathbb{E}\left[\prod_{k=1}^{q}X_{t_k,l_1}X_{s_k,l_2}X_{t_k',l_3}X_{s_k',l_4}\right],
\end{aligned}
$$

for $t_1,s_1\in\{i_1,j_1\},...,t_q,s_q\in\{i_q,j_q\},t_1',s_1'\in\{i_1',j_1'\},...,t_q',s_q'\in\{i_q',j_q'\}$, and there exists at least one $k=1,...,q$ such that $t_k\ne s_k$ ($t_k'\ne s_k'$). Since the expectation of a product of random variables can be expressed in terms of joint cumulants, we have

$$
\mathbb{E}\left[\prod_{k=1}^{q}X_{t_k,l_1}X_{s_k,l_2}X_{t_k',l_3}X_{s_k',l_4}\right]=\sum_{\pi}\prod_{B\in\pi}cum(X_{i,l}:(i,l)\in B),
$$

where π runs through the list of all partitions of $\{(t_1,l_1),(s_1,l_2),...,(t_q',l_3),(s_q',l_4)\}$ and B runs thorough the list of all blocks of the partition π. Due to the special structure of our statistic, there is a set of partitions $\mathcal{S}$ such that for every $\pi\in\mathcal{S}$, the product of joint cumulants over all $B\in\pi$ is zero. And for each $\pi\in\mathcal{S}^c$ there are nice properties related to the blocks $B\in\pi$. Here we shall illustrate these properties as follows. To be clear, since we are dealing with a double indexed array $X_{i,l}$, we call "i" as the temporal index and "l" as the spatial index. For $\forall \pi \in \mathcal{S}^c$,

1. **The size of every block $B \in \pi$ cannot exceed 4.** Since $i_1, ..., i_q, j_1, ..., j_q$ are all distinct, and $i'_1, ..., i'_q, j'_1, ..., j'_q$ are all distinct, it is impossible to have any three indices in $\{i_1, ..., i_q, j_1, ..., j_q, i'_1, ..., i'_q, j'_1, ..., j'_q\}$ that are equal. And any joint cumulants of order greater than or equal to 5 will include at least three indices and they cannot be all equal.

2. **There are no blocks with size 1.** This is because the cumulant of a single random variable with mean zero is also zero.

3. **Every $B \in \pi$ must contain only one distinct temporal index.** Otherwise $\prod_{B\in\pi} cum(X_{i,l} : (i,l) \in B) = 0$.

The above properties imply that for $\forall \pi \in \mathcal{S}^c$ and $\forall B \in \pi$, $cum(X_{i,l} : (i,l) \in B)$ has to be one of the followings: $cum(X_{0,l_1}, X_{0,l_2}, X_{0,l_3}, X_{0,l_4})$, $cum(X_{0,l_1}, X_{0,l_2}, X_{0,l_3})$, $cum(X_{0,l_1}, X_{0,l_2}, X_{0,l_4})$, $cum(X_{0,l_1}, X_{0,l_3}, X_{0,l_4})$, $cum(X_{0,l_2}, X_{0,l_3}, X_{0,l_4})$, Σ_{l_1,l_2}, Σ_{l_1,l_3}, Σ_{l_1,l_4}, Σ_{l_2,l_3}, Σ_{l_2,l_4}, Σ_{l_3,l_4}.

If we assume $l_1 \le l_2 \le l_3 \le l_4$, it can be shown that

$$\prod_{B\in\pi} cum(X_{i,l} : (i,l) \in B) \le C(1 \vee (l_2 - l_1))^{-r}(1 \vee (l_4 - l_3))^{-r}, \tag{6.5}$$

for some generic positive constant C and any partition π. To see this, we notice that at least one $k = 1, ..., q$, say k_0, such that $t_{k_0} \neq s_{k_0}$ and $t'_{k_0} \neq s'_{k_0}$. For every $\pi \in \mathcal{S}^c$ there exists $B_1, B_2 \in \pi$ such that $(t_{k_0}, l_1) \in B_1$ and $(s'_{k_0}, l_4) \in B_2$. Based on the third property above, all other elements in B_1 must have the same temporal index as t_{k_0}. And because of the first property above, all i_k, j_k for $k \neq k_0$ and s_{k_0} are different from t_{k_0}. This implies that the spatial indices for all other elements in B_1 have to be either l_3 or l_4, not l_1 and l_2. For the same reason, the spatial indices for all other elements in B_2 can only be either l_1 or l_2. Therefore,

$$cum(X_{i,l} : (i,l) \in B_1) \in \{cum(X_{0,l_1}, X_{0,l_3}, X_{0,l_4}), \Sigma_{l_1,l_3}, \Sigma_{l_1,l_4}\},$$

and

$$cum(X_{i,l} : (i,l) \in B_2) \in \{cum(X_{0,l_1}, X_{0,l_2}, X_{0,l_4}), \Sigma_{l_1,l_4}, \Sigma_{l_2,l_4}\}.$$

Under Assumption (5.2), $cum(X_{i,l} : (i,l) \in B_1) \le C(1 \vee (l_2 - l_1))^{-r}$ and $cum(X_{i,l} : (i,l) \in B_2) \le C(1 \vee (l_4 - l_3))^{-r}$. And the joint cumulants are uniformly bounded above for those $B \in \pi \setminus \{B_1, B_2\}$. Thus Equation 6.5 is proved.

Furthermore, define $Ind(t_1, s_1, ..., t_q, s_q, t'_1, s'_1, ..., t'_q, s'_q)$ as the indicator function corresponding to the event that for every $k = 1, 2, .., q$ that $t_k \neq s_k$, there exists $k' = 1, ..., q$ such that $t_k = t'_{k'}$ or $t_k = s'_{k'}$, then $\mathbb{E}\left[\prod_{k=1}^{p} X_{t_k,l_1} X_{s_k,l_2} X_{t'_k,l_3} X_{s'_k,l_4}\right] \neq 0$ only if

$$Ind(t_1, s_1, ..., t_q, s_q, t'_1, s'_1, ..., t'_q, s'_q) = 1.$$

It is also easy to see that

$$\sum_{1\le i_1<\cdots<i_q<j_1<\cdots<j_q\le n}\ \sum_{1\le i'_1<\cdots<i'_q<j'_1<\cdots<j'_q\le n} Ind(t_1,s_1,...,t_q,s_q,t'_1,s'_1,...,t'_q,s'_q)=o(n^{4q}).$$

Combining all the results above, we have

$$\begin{aligned}
&\mathbb{E}[J(t_1,s_1,...,t_q,s_q)^2]\\
=&\frac{1}{\binom{n}{2q}^2\|\Sigma\|_q^{2q}}\sum_{1\le i_1<\cdots<i_q<j_1<\cdots<j_q\le n}\ \sum_{1\le i'_1<\cdots<i'_q<j'_1<\cdots<j'_q\le n}\ \sum_{l_1,l_2,l_3,l_4=1}^{p}\mathbb{E}\left[\prod_{k=1}^{q}X_{t_k,l_1}X_{s_k,l_2}X_{t'_k,l_3}X_{s'_k,l_4}\right],\\
\le&\frac{C}{\binom{n}{2q}^2\|\Sigma\|_q^{2q}}\sum_{1\le i_1<\cdots<i_q<j_1<\cdots<j_q\le n}\ \sum_{1\le i'_1<\cdots<i'_q<j'_1<\cdots<j'_q\le n}\\
&\sum_{1\le l_1\le l_2\le l_3\le l_4\le p} Ind(t_1,s_1,...,t_q,s_q,t'_1,s'_1,...,t'_q,s'_q)(1\vee(l_2-l_1))^{-r}(1\vee(l_4-l_3))^{-r}\\
\le&\frac{o(n^{4q})}{\binom{n}{2q}^2\|\Sigma\|_q^{2q}}\left(\sum_{1\le l_1\le l_2\le p}(1\vee(l_2-l_1))^{-r})\right)^2\lesssim\frac{p^2}{\|\Sigma\|_q^{2q}}o(1)=o(1)\to 0,
\end{aligned}$$

where the last equality is because $\|\Sigma\|_q^{2q}=(\sum_{i,j=1}^{p}\Sigma_{i,j}^q)^2\gtrsim p^2$.

This completes the proof of (2), as well as the whole proof.

□

Table 5 shows additional simulation results for the size of the proposed monitoring statistics for $n=200$. The size distortion problem has improved for almost all settings.

Table 5: Size of different monitoring procedures

$(n,p)=(200,200)$	T1			T2			T3		
size $\alpha=0.1$	L2	L6	Comb	L2	L6	Comb	L2	L6	Comb
$\rho=0.2$	0.104	0.072	0.074	0.097	0.072	0.073	0.102	0.071	0.073
$\rho=0.5$	0.105	0.064	0.091	0.107	0.064	0.085	0.104	0.065	0.087
$\rho=0.8$	0.127	0.037	0.089	0.133	0.038	0.099	0.131	0.039	0.099

Acknowledgements

Xiaofeng Shao acknowleges partial support from NSF grants DMS-1807032 and DMS-2014018. Hao Yan acknowleges partial support from NSF grants DMS-1830363 and CMMI-1922739.

References

Aminikhanghahi, S. and D. J. Cook (2017). A survey of methods for time series change point detection. *Knowledge and Information Systems 51*(2), 339–367.

Aue, A., S. Hörmann, L. Horváth, M. Hušková, and J. G. Steinebach (2012). Sequential testing for the stability of high-frequency portfolio betas. *Econometric Theory 28*(4), 804–837.

Aue, A. and L. Horváth (2013). Structural breaks in time series. *Journal of Time Series Analysis 34*(1), 1–16.

Brown, R. L., J. Durbin, and J. M. Evans (1975). Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 37*(2), 149–163.

Bücher, A., J.-D. Fermanian, and I. Kojadinovic (2019). Combining cumulative sum change-point detection tests for assessing the stationarity of univariate time series. *Journal of Time Series Analysis 40*(1), 124–150.

Chen, S. and Y. Qin (2010). A two sample test for high dimensional data with application to gene-set testing. *The Annals of Statistics 38*, 808–835.

Cho, H. and P. Fryzlewicz (2015). Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 77*(2), 475–507.

Chu, C.-S. J., M. Stinchcombe, and H. White (1996). Monitoring structural change. *Econometrica: Journal of the Econometric Society 64*, 1045–1065.

Dette, H. and J. Gösmann (2019). A likelihood ratio approach to sequential change point detection for a general class of parameters. *Journal of the American Statistical Association*, 1–17.

Enikeeva, F. and Z. Harchaoui (2019). High-dimensional change-point detection under sparse alternatives. *The Annals of Statistics 47*(4), 2051–2079.

Fremdt, S. (2015). Page's sequential procedure for change-point detection in time series regression. *Statistics 49*(1), 128–155.

Gösmann, J., T. Kley, and H. Dette (2020). A new approach for open-end sequential change point monitoring. *Journal of Time Series Analysis, to appear*.

He, Y., G. Xu, C. Wu, and W. Pan (2018). Asymptotically independent u-statistics in high-dimensional testing. *arXiv preprint arXiv:1809.00411*.

Horváth, L. and M. Hušková (2012). Change-point detection in panel data. *Journal of Time Series Analysis 33*(4), 631–648.

Horváth, L., M. Hušková, P. Kokoszka, and J. Steinebach (2004). Monitoring changes in linear models. *Journal of Statistical Planning and Inference 126*(1), 225–251.

Jirak, M. (2015). Uniform change point tests in high dimension. *The Annals of Statistics 43*(6), 2451–2483.

Kirch, C., B. Muhsal, and H. Ombao (2015). Detection of changes in multivariate time series with application to eeg data. *Journal of the American Statistical Association 110*(511), 1197–1216.

Lai, T. L. (1995). Sequential changepoint detection in quality control and dynamical systems. *Journal of Royal Statistical Society, Series B (Statistical Methodology) 57*, 613–658.

Lai, T. L. (2001). Sequential analysis: Some classical problems and new challenges. *Statistica Sinica 11*, 303–408.

Lei, Y., Z. Zhang, and J. Jin (2010). Automatic tonnage monitoring for missing part detection in multi-operation forging processes. *Journal of Manufacturing Science and Engineering 132*(5).

Lévy-Leduc, C. and F. Roueff (2009). Detection and localization of change-points in high-dimensional network traffic data. *The Annals of Applied Statistics 3*(2), 637–662.

Li, J. (2020). Efficient global monitoring statistics for high-dimensional data. *Quality Reliability Engineering International 36*, 18–32.

Liu, K., R. Zhang, and Y. Mei (2019). Scalable sum-shrinkage schemes for distributed monitoring large-scale data streams. *Statistica Sinica 29*, 1–22.

Lorden, G. (1971). Procedures for reacting to a change in distribution. *Annals of Mathematical Statistics 42*, 1897–1908.

MacNeill, I. B. (1974). Tests for change of parameter at unknown times and distributions of some related functionals on brownian motion. *The Annals of Statistics 2*(5), 950–962.

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

Mei, Y. (2010). Efficient scalable schemes for monitoring a large number of data streams. *Biometrika 97*(2), 419–433.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika 41*(1/2), 100–115.

Perron, P. (2006). Dealing with structural breaks. *Palgrave Handbook of Econometrics Vol. 1: Econometric Theory, K. Patterson and T.C. Mills (eds.), Palgrave Macmillan*, 278–352.

Polunchenko, A. S. and A. G. Tartakovsky (2012). State-of-the-art in sequential change-point detection. *Methodology and computing in applied probability 14*(3), 649–684.

Shao, X. (2010). A self-normalized approach to confidence interval construction in time series. *Journal of Royal Statistical Society, Series B 72*, 343–366.

Shao, X. (2015). Self-normalization for time series: a review of recent developments. *Journal of the American Statistical Association 110*, 1797–1817.

Shao, X. and W. B. Wu (2007). Asymptotic spectral theory for nonlinear time series. *The Annals of Statistics 35*(4), 1773–1801.

Shao, X. and X. Zhang (2010). Testing for change points in time series. *Journal of the American Statistical Association 105*, 1228–1240.

Wald, A. (1945). Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics 16*, 117–186.

Wang, R. and X. Shao (2020). Dating the break in high-dimensional data. *Available at https://arxiv.org/pdf/2002.04115.pdf*.

Wang, R., S. Volgushev, and X. Shao (2019). Inference for change points in high dimensional data. *arXiv preprint arXiv:1905.08446*.

Wang, T. and R. J. Samworth (2018). High dimensional change point estimation via sparse projection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80*(1), 57–83.

Wang, Y. and Y. Mei (2015). Large-scale multi-stream quickest change detection via shrinkage post-change estimation. *IEEE Transactions on Information Theory 61*(12), 6926–6938.

Wied, D. and P. Galeano (2013). Monitoring correlation change in a sequence of random variables. *Journal of Statistical Planning and Inference 143*(1), 186–196.

Wu, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences 102*(40), 14150–14154.

Wu, W. B. and X. Shao (2004). Limit theorems for iterated random functions. *Journal of Applied Probability 41*(2), 425–436.

Xie, Y. and D. Siegmund (2013). Sequential multi-sensor change-point detection. *The Annals of Statistics 41*(2), 670–692.

Yan, H., K. Paynabar, and J. Shi (2014). Image-based process monitoring using low-rank tensor decomposition. *IEEE Transactions on Automation Science and Engineering 12*(1), 216–227.

Yan, H., K. Paynabar, and J. Shi (2018). Real-time monitoring of high-dimensional functional data streams via spatio-temporal smooth sparse decomposition. *Technometrics 60*(2), 181–197.

Yu, M. and X. Chen (2017). Finite sample change point inference and identification for high-dimensional mean vectors. *arXiv preprint arXiv:1711.08747*.

Yu, M. and X. Chen (2019). A robust bootstrap change point test for high-dimensional location parameter. *arXiv preprint arXiv:1904.03372*.

Zhang, Y., R. Wang, and X. Shao (2020). Adaptive change point inference for high-dimensional data. *Preprint*.

Zou, C., Z. Wang, X. Zi, and W. Jiang (2015). An efficient online monitoring method for high-dimensional data streams. *Technometrics 57*(3), 374–387.

Teng Wu, Department of Statistics, University of Illinois at Urbana-Champaign

E-mail: tengwu2@illinois.edu

Runmin Wang, Department of Statistical Science, Southern Methodist University

E-mail: runminw@smu.edu

Hao Yan, School of Computing Informatics & Decision Systems Engineering, Arizona State University

E-mail: haoyan@asu.edu

Xiaofeng Shao, Department of Statistics, University of Illinois at Urbana-Champaign

E-mail: xshao@illinois.edu