



Deposited via The University of York.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/171296/>

Version: Accepted Version

---

**Article:**

Wang, Weining, Chen, Likai and Wu, Wei Biao (2021) Inference of Break-Points in High-Dimensional Time Series. *Journal of the American Statistical Association*. ISSN: 0162-1459

<https://doi.org/10.1080/01621459.2021.1893178>

---

**Reuse**

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

**Takedown**

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing [eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk) including the URL of the record and the reason for the withdrawal request.

# Inference of breakpoints in high-dimensional time series

Likai Chen

Department of Mathematics and Statistics, Washington University in St. Louis,

Weining Wang

Department of Economics and Related Studies, University of York, C.A.S.E.-

Center of Applied Statistics and Economics, Humboldt-Universität zu Berlin

and

Wei Biao Wu

Department of Statistics, University of Chicago

April 25, 2020

## Abstract

For multiple change-points detection of high-dimensional time series, we provide asymptotic theory concerning the consistency and the asymptotic distribution of the breakpoint statistics and estimated break sizes. The theory backs up a simple two-step procedure for detecting and estimating multiple change-points. The proposed two-step procedure involves the maximum of a MOSUM (moving sum) type statistics in the first step and a CUSUM (cumulative sum) refinement step on an aggregated time series in the second step. Thus, for a fixed time-point, we can capture both the biggest break across different coordinates and aggregating simultaneous breaks over multiple coordinates. Extending the existing high-dimensional Gaussian approximation theorem to dependent data with jumps, the theory allows us to characterize the size and power of our multiple change-point test asymptotically. Moreover, we can make inferences on the breakpoints estimates when the break sizes are small. Our theoretical setup incorporates both weak temporal and strong or weak cross-sectional

dependence and is suitable for heavy-tailed innovations. A robust long-run covariance matrix estimation is proposed, which can be of independent interest. An application on detecting structural changes of the U.S. unemployment rate is considered to illustrate the usefulness of our method.

*Keywords:* multiple change points detection; temporal and cross-sectional dependence; Gaussian approximation; inference of break locations

# 1 Introduction

Statistical inference of structural breaks in mean is an important subject to study, and involves estimating the trend functions, detecting and locating abnormal changes and making inferences on the break estimators. Breaks may arise in various applications in different fields, such as in network analysis, biology, engineering, economics and finance, among others. Specific examples are anomaly of network traffic data caused by attacks (Lévy-Leduc and Roueff (2009)), recurrent DNA copy number variants in multiple samples (Zhang et al. (2010)), abrupt changes in household electrical power consumption (Harlé et al. (2016)) and minimum wage policy changes analysis (Chen et al. (2017)), etc. In those data scenarios, temporal and cross-sectional dependence for large-dimensional data might pose challenges to statistical analysis.

To formulate our problem, we assume that observation vectors  $Y_1, Y_2, \dots, Y_n$  follow the model,

$$Y_t = \mu(t/n) + \epsilon_t, \quad t = 1, 2, \dots, n, \quad (1)$$

where  $(\epsilon_t)_t$  is a sequence of zero-mean  $p$ -dimensional stationary noise vectors and  $\mu(\cdot) = (\mu_1(\cdot), \mu_2(\cdot), \dots, \mu_p(\cdot))^\top : [0, 1] \rightarrow \mathbb{R}^p$  is a vector of unknown trend functions. In this way, the data generating process is trend stationary. We will model breaks occurring on the vector of trend functions  $\mu(t/n)$ . Notably, we assume that the trend function satisfies

$$\mu(u) = f(u) + \sum_{i=1}^{K_0} \gamma_i \mathbf{1}_{u \geq u_i}, \quad (2)$$

where  $K_0$  is an unknown integer representing the number of breaks;  $f(\cdot)$  ( $f(\cdot) = (f_1(\cdot), f_2(\cdot), \dots, f_p(\cdot))^\top : [0, 1] \rightarrow \mathbb{R}^p$ ) is a vector of smooth trend functions;  $u_k$ s with  $0 < u_1 < u_2 < \dots < u_{K_0} < 1$  are the time stamps of the change-points with  $|u_i - u_j| \gg b$ , where  $b$  is the

bandwidth parameter; and  $\gamma_k \in \mathbb{R}^p$  are the jump vectors with size  $|\gamma_k|_\infty$  ( $|\cdot|_\infty$  is the infinity norm) at point  $u_k$ . Note that the jump sizes are characterized in terms of the infinity norm; therefore, we do not require simultaneous jumps for all entities  $1 \leq j \leq p$ , and some coordinates of  $\gamma_k$  can be zero. Namely, we will focus on the largest jump (i.e.,  $|\gamma_k|_\infty$ ) happening in the cross-sectional dimension for any fixed time point  $k$  (cf. Theorem 2), and this is of particular interest when the jumps are sparse. In case many series jump at the same time, we further propose a refined method, which aggregates all the contemporaneous jumps (cf. Theorem 4). In most of the change-point settings, the smooth part of the trend functions is zero (i.e.,  $f \equiv 0$ ). This means that the trend functions are piecewise constant for each coordinate. In contrast, our model is more flexible and realistic, since we allow the mean functions to vary smoothly instead of staying at the same level between break-points.

The goal of this paper is to provide theory for structural break inference. We first detect the existence of breaks. We then deliver theorems to test for the existence of breaks, identify their change-point  $u_k$ , calibrate sizes of the breaks, i.e.  $|\gamma_k|_\infty$ ,  $1 \leq k \leq K_0$ , and construct confidence intervals for the estimated break points. Our theorem allows us to consider a multiple change-point test based on a threshold method on the maximum of generalized MOSUM statistics. We derive the asymptotic distribution of the test statistics including estimated breaks sizes, and the estimated breakpoint locations (cf. Theorem 3, 4 ii) ). The results provide solid foundations for conducting statistical inferences for multiple change-point estimation in high dimensional time series. Moreover, we consider a further aggregation step targeting at simultaneous breaks, and also this step gives us finer consistency rates of the break location estimation.

Multiple change-point detection can be classified into two categories, i.e. model selection and testing. The traditional model selection method, for example BIC, has the drawback of computational inefficiency, which can be improved by some recent LASSO

type penalization procedure, see for example, Killick et al. (2012), Tibshirani and Wang (2007), Li et al. (2016) and Lee et al. (2016). Regarding multiple change-point detection via testing, a classical method utilises an exhaustive search, which examines all the possible breakpoints combination. An exhaustive search is very time consuming and some dynamic technique and improved versions are invented, see for instance Bai and Perron (1998, 2003) and Jackson et al. (2005). A very popular approach is the binary segmentation introduced in Scott and Knott (1974). However its power might suffer for certain alternatives. This drawback can be handled by the wild binary segmentation algorithm developed in Fryzlewicz (2014) and Cho and Fryzlewicz (2015). Moreover, Fryzlewicz (2018) recently introduces a bottom-up algorithm to overcome the disadvantage of the classical binary segmentation.

Detection using the MOSUM statistics is another popular way for multiple change-point analysis; see, for example, Hušková and Slabý (2001) for i.i.d data; Wu and Zhao (2007) and Eichinger and Kirch (2018) for general temporal dependent data. Preuss et al. (2015) deal with multivariate time series for structural breaks in covariance. A MOSUM procedure has the advantage of computation simplicity and can avoid issues due to multiple testing in multiple break inference. A possible drawback is that MOSUM introduces a new bandwidth parameter. Such an issue can be dealt with through a multi-scale MOSUM, which uses multiple bandwidths; see, for instance Meier et al. (2019). Eichinger and Kirch (2018) provide a comprehensive theoretical analysis of multiple change-point detection using MOSUM analysis including the distribution theory of the estimated breakpoint. Our work can be viewed as a generalization of their work on the high-dimensional case as we adopt a MOSUM type of statistics in our first step.

Change-point detection for high-dimensional time series has recently drawn a lot of attention due to the increasing number of applications. In particular, we shall consider

the case of  $p \rightarrow \infty$ . Even in the simplest setup of a mean-shift model, large  $p$  may pose challenge to change-point detection. It is common to consider aggregation, either over the original time series or certain transformed statistics of individual time series and to convert the problem to a one-dimensional analysis. For instance, targeting at sparse breaks, Cho and Fryzlewicz (2015) propose a sparse binary segmentation which concerns an  $l_1$ -based aggregation with a hard threshold, and Wang and Samworth (2018) consider sparse singular value decomposition based on the CUSUM statistics. Moreover, there are a few other work looking at  $l_2$ -based aggregation of statistic: Bai (2010) evaluates the performance of a least square estimation of a single breakpoint with distribution theory on the break location estimates without assuming cross sectional dependency; Zhang et al. (2010) extend the method in Olshen et al. (2004); Enikeeva and Harchaoui (2019) and Liu et al. (2019) regard the detection of change-points in a high-dimensional mean vector as a minimax testing problem. For a single break point in time and targeting at sparse break coordinates, Jirak (2015) studies a CUSUM type statistic for each coordinate and then takes maximum of them, and asymptotic theory is provided to facilitate the simultaneous inferences of the breakpoint estimation. Cho (2016) proposes a double-CUSUM algorithm, etc. For a single change-point in time, distribution theory is still available in a few works, see for example Bai (2010). When it comes to multiple change points detection, the majority of the aforementioned work focus on developing novel algorithms, and a complete distribution theory is not readily available due to the complexity of the problem. An exception is Jirak (2015). Compared to Jirak (2015), we are taking a different path in terms of an algorithm using the MOSUM and an aggregation step with refined rates of estimator achieved. We thus provide a new angle to conduct inferences in multiple change-point analysis for high-dimensional time series.

It shall be noted that as there are already many novel algorithms developed, we do

not claim a total superiority of ours. The algorithm proposed here is a generalization or modification of the existing methods, which facilitates us to obtain a complete theory and good theoretical rates. Nevertheless, our aggregation step is different and complement to existing algorithms. For example, one main difference with the aggregation step is that we project based on the estimates in the first step. Cho and Fryzlewicz (2015) and Wang and Samworth (2018) use other approaches to find the best projection direction.

To summarize, we provide theory for a two-step multiple change-point procedure. We prove consistency results as well as distribution theorems for breakpoint location estimation, which is crucial for inference of breakpoints. The aggregation step can help us to achieve good rates of the breakpoint estimation. We deliver general theoretical results that allow heavy-tailed distribution and general spatial-temporal dependency assumption on the error term, and we do not require the mean function to be a piece-wise constant (i.e.  $f \equiv 0$ ). The detection procedure is not computationally expensive, as we only need to evaluate the statistic once for each point  $t$ . Additionally, we consider the estimation of the long-run covariance matrices. This paper is structured as follows. Section 2 construct a test and deliver its asymptotic performance for testing the existence of change-points. Section 3 introduces the two-step algorithm for inference on break estimation. The associated consistency and asymptotic distribution theorems are also covered in this section. Long-run covariance matrix estimation is derived in Section 4. Simulation results are in Section A in supplementary materials and an application on U.S. unemployment rate is given in Section 5. Detailed proofs are presented in Section B in the supplementary materials.

Notations: For a constant  $k > 0$  and a vector  $v = (v_1, \dots, v_d)^\top \in \mathbb{R}^d$ , we denote  $|v|_k = (\sum_{i=1}^d |v_i|^k)^{1/k}$ ,  $|v| = |v|_2$  and  $|v|_\infty = \max_{i \leq d} |v_i|$ . For a matrix  $A = (a_{ij})_{1 \leq i \leq m, 1 \leq j \leq n}$ , we define the spectral norm  $|A|_2 = \max_{|v|=1} |Av|$  and the max norm  $|A|_{\max} = \max_{i,j} |a_{i,j}|$ . For a function  $f$ , we denote  $|f|_\infty = \sup_x |f(x)|$ . We set  $(a_n)$  and  $(b_n)$  to be positive number

sequences. We write  $a_n = O(b_n)$  or  $a_n \lesssim b_n$  (resp.  $a_n \asymp b_n$ ) if there exists a positive constant  $C$  such that  $a_n/b_n \leq C$  (resp.  $1/C \leq a_n/b_n \leq C$ ) for all large  $n$ , and we denote  $a_n = o(b_n)$  (resp.  $a_n \sim b_n$ ), if  $a_n/b_n \rightarrow 0$  (resp.  $a_n/b_n \rightarrow 1$ ). For two sequences of random variables  $(X_n)$  and  $(Y_n)$ , we write  $X_n = o_{\mathbb{P}}(Y_n)$ , if  $X_n/Y_n \rightarrow 0$  in probability.

## 2 Testing the existence of change-points

In this section, we provide a test for the existence of breaks. Considering our observations generated by the model in (1) and (2), we would like to test the null hypothesis,

$$\mathcal{H}_0 : \quad \gamma_1 = \gamma_2 = \dots = \gamma_{K_0} = 0,$$

which corresponds to the case of no breaks, against the alternative of the existence of at least one break i.e.  $\mathcal{H}_A : \exists k \in 1, \dots, K_0, \quad \gamma_k \neq 0$ . It shall be noted that we do not need to assume the number of breaks ( $K_0$ ) to be bounded, but to rather restrict on the separation between breakpoints (c.f. Assumption 2.4).

In Subsection 2.1, we derive our test statistic. Its asymptotic property is given in Subsection 2.2. In Subsection 2.3, we derive the performance of the test based on Gaussian approximation, which provides the theoretical foundation for calculating the size and power of the test.

### 2.1 Test statistic

In this subsection, we introduce the test statistics and some intuition. Recall that our trend function  $\mu(u)$  can be disentangled into two parts, namely a smooth transition part  $f(u)$  and a jump part  $\gamma_i \{u \geq u_i\}$ . We can define the jump vector at point  $u$  as a gap between

the right-side function  $\mu^{(r)}(u)$  and the left-side function  $\mu^{(l)}(u)$ , which is

$$J(u) = \mu^{(r)}(u) - \mu^{(l)}(u), \quad \text{where we define } \mu^{(r)}(u) = \lim_{t \downarrow u} \mu(t) \quad \text{and} \quad \mu^{(l)}(u) = \lim_{t \uparrow u} \mu(t).$$

Due to the smoothness of the constituents of  $f(\cdot)$ , the gap function  $J(u) = 0$  when there is no jump, and  $J(u) = \gamma_k$  when  $u = u_k$ . A natural way to test the existence of change-points is to check whether the gap is zero (i.e.  $J(u) = 0$ ). To this end, we need  $\hat{\mu}^{(r)}(u)$  and  $\hat{\mu}^{(l)}(u)$ , which are estimates of  $\mu^{(r)}(u)$  and  $\mu^{(l)}(u)$ . We propose to adopt the local linear estimation technique, see Fan and Gijbels (1996).

The local linear estimates of  $\hat{\mu}^{(l)}(u)$  and  $\hat{\mu}^{(r)}(u)$  at the point  $u = i/n$  are of the following weighted form

$$\hat{\mu}_i^{(l)} := \hat{\mu}^{(l)}(i/n) = \sum_{t=i-bn}^{i-1} w_{i-t} Y_t \quad \text{and} \quad \hat{\mu}_i^{(r)} := \hat{\mu}^{(r)}(i/n) = \sum_{t=i+1}^{i+bn} w_{t-i} Y_t, \quad (3)$$

with weights

$$w_i = w_{i,b} = w_b(0, i/n), \quad i \geq 1, \quad w_0 = 0. \quad (4)$$

The weight functions are defined as

$$w_b(u, v) = \frac{K((v-u)/b)[S_2(u) - (u-v)S_1(u)]}{S_2(u)S_0(u) - S_1^2(u)}, \quad S_l(u) = \sum_{i=1}^n (u - i/n)^l K((i/n - u)/b), \quad (5)$$

where  $K(\cdot)$  is a kernel function and  $b$  is a bandwidth with  $b \rightarrow 0$  and  $bn \rightarrow \infty$ . It is worth noting that the estimator in (3) is equivalent to adopting a one-sided kernel function, i.e.  $K(u)\{u \geq 0\}$  to fix the boundary estimation issue for the kernel estimation method.

If there is no jump around the time point  $u = i/n$ , the gap estimate  $\hat{J}(i/n) = \hat{\mu}_i^{(l)} - \hat{\mu}_i^{(r)}$  would be small for all coordinates. Otherwise if for some entity  $1 \leq j \leq p$ , the gap estimate

$|\hat{J}_j(i/n)|$  is large, there might exist a jump around time  $i/n$  at coordinate  $j$ . Note that the test statistics is in fact of a MOSUM type, and we replace the uniform kernel for MOSUM by a local linear one to adapt for slowly varying trends  $f(u)$  in (2).

To conduct the breakpoint detection with  $p \rightarrow \infty$ , we consider the maximum of the gap statistics. Furthermore, we need to standardize our test statistics in order to get a regular limit distribution. To obtain the long-run variance matrix involved in the standardization, we need to specify the error process, as in model (1). We would like to make a general temporal and cross-sectional dependence assumption. This is a crucial issue, since for time series data, dependence is the rule rather than the exception. Specifically, we let

$$\epsilon_t = \sum_{k \geq 0} A_k \eta_{t-k}, \quad (6)$$

where  $\eta_t \in \mathbb{R}^{\tilde{p}}$  are independent and identically distributed (i.i.d.) random vectors with zero mean and a identity covariance matrix.  $A_k (k \geq 0)$ s are coefficient matrices in  $\mathbb{R}^{p \times \tilde{p}}$  such that  $\epsilon_t$  is a proper random vector, and  $p \leq \tilde{p} \leq c_p p$ , for some constant  $c_p > 1$ . If  $A_i = 0$  for all  $i \geq 1$ , then the noise sequences are temporally independent; if  $p = \tilde{p}$  and matrices  $A_i$  are diagonal, then the sequences become the model in Bai (2010), which is spatially independent. The VMA( $\infty$ ) process is very general and includes many important time series models such as a vector autoregressive moving averages (VARMA) model, i.e.

$$(1 - \sum_{j=1}^s \Theta_j \mathcal{B}^j) X_i = X_i - \sum_{j=1}^s \Theta_j X_{i-j} = \sum_{k=1}^t \Xi_k \eta_{i-k},$$

where  $\Theta_j$  and  $\Xi_k$  are real matrices such that  $\det(1 - \sum_{j=1}^s \Theta_j z)$  is not zero for all  $|z| \leq 1$  and  $\mathcal{B}$  is the backshift operator.

Correspondingly, we define the sum of the coefficient matrix to be  $S = \sum_{k \geq 0} A_k$ . The long-run covariance matrix for the error process is

$$\Sigma = S S^\top. \quad (7)$$

We denote  $\Sigma = (\sigma_{i,j})$ ,  $1 \leq i, j \leq p$ , and

$$\Lambda = \text{diag}(\sigma_{1,1}^{1/2}, \sigma_{2,2}^{1/2}, \dots, \sigma_{p,p}^{1/2}). \quad (8)$$

Following the previous intuition of the effect of jumps on the gap statistics  $\hat{J}(\cdot)$ , we consider the test statistic

$$T_n = \max_{bn+1 \leq i \leq n-bn} |V_i|_\infty, \quad \text{where } V_i = \Lambda^{-1}(\hat{\mu}_i^{(l)} - \hat{\mu}_i^{(r)}). \quad (9)$$

Note that we consider the normalized statistic as multiplying the jump estimates  $\hat{J}(i/n) = \hat{\mu}_i^{(l)} - \hat{\mu}_i^{(r)}$  by  $\Lambda^{-1}$  since the long-run variances  $\sigma_{j,j}$  for different coordinates  $1 \leq j \leq p$  can be very different.

## 2.2 Properties of the test statistics

We shall show the asymptotic properties of our test statistics  $T_n$  in (9) in this subsection. First we analyze the mean of the normalized jump estimators, i.e.  $\mathbb{E}V_i$ . Intuitively, we can decompose the level of our jump estimator  $\mathbb{E}V_i$  into two parts, one is the commonly encountered bias term for the nonparameteric kernel estimators of the smooth trend functions, and the other is induced by jumps on the deterministic trend, which is denoted as  $d_i$ . Recall  $\tau_k = nu_k$ , the definition of  $w_i$  in (4) for  $i = 1, 2, \dots, bn$ , and  $w_i = 0$  for  $i = 0$  and  $i > bn$ . We denote  $\Omega_i$  as a set of indices indicating the break locations within the  $bn$  neighborhood around time  $i$ , namely  $\Omega_i = \{k \mid |i - \tau_k| \leq bn, 1 \leq k \leq K_0\}$ . For a time point  $i$  with at least one break around  $\Omega_i \neq \emptyset$ , we define the weighted break size to be,

$$d_i = (1 - \sum_{t=1}^{|i-\tau_k|} w_t) \Lambda^{-1} \gamma_k, \quad k = \text{argmin}_{j \in \Omega_i} |i - \tau_j|, \quad (10)$$

and for the rest of locations  $i$ , let  $d_i = 0$ . We further stack  $d_i$  over all breakpoints that are of interest, which is denoted by  $\underline{d} = (d_{bn+1}^\top, d_{bn+2}^\top, \dots, d_{n-bn}^\top)^\top$ . It should be noted that under the null,  $\underline{d} = 0$ .

If  $\tau_i - \tau_j = n(u_i - u_j) \asymp n$ , for any  $i, j$ , then for large  $n$ , the cardinality of  $\Omega_i$  is at most one, i.e.  $|\Omega_i| \leq 1$ . Actually this can be relaxed to  $\min_{1 \leq i \neq j \leq K_0} |\tau_i - \tau_j| \gg bn$ . We denote the smooth part of the local linear estimate as

$$\hat{f}_i^{(l)} = \sum_{t=i-bn}^{i-1} w_{i-t} f(t/n) \quad \text{and} \quad \hat{f}_i^{(r)} = \sum_{t=i+1}^{i+bn} w_{t-i} f(t/n).$$

By Fan and Gijbels (1996), under some smoothness conditions, the bias part of the estimated smooth functions would be of the order  $b^2$ , which goes to zero by assumption, i.e.

$$\max_{bn+1 \leq i \leq n-bn} |\Lambda^{-1}(\hat{f}_i^{(l)} - \hat{f}_i^{(r)})|_\infty = O(b^2). \quad (11)$$

Given the definition of our model  $Y_i = \mu(i/T) + \epsilon_i$ ,  $d_i$  can be expressed as

$$d_i = \mathbb{E}\{\Lambda^{-1}((\hat{\mu}_i^{(r)} - \hat{\mu}_i^{(l)}) - (\hat{f}_i^{(r)} - \hat{f}_i^{(l)}))\} = \mathbb{E}\{V_i - \Lambda^{-1}(\hat{f}_i^{(r)} - \hat{f}_i^{(l)})\}. \quad (12)$$

Combining (11) and (12),  $\mathbb{E}V_i$  can be approximated by the part induced by jumps  $\gamma_k$ s, as

$$|\mathbb{E}V_i - d_i|_\infty = |\Lambda^{-1}(\hat{f}_i^{(r)} - \hat{f}_i^{(l)})|_\infty = O(b^2). \quad (13)$$

Let us now consider the  $V_i - \mathbb{E}V_i$  part. We observe that the centered statistics can be expressed as a weighted sum of the error term, namely

$$V_i - \mathbb{E}V_i = \sum_{l=i-bn}^{i-1} w_{i-l} \Lambda^{-1} \epsilon_l - \sum_{l=i+1}^{i+bn} w_{l-i} \Lambda^{-1} \epsilon_l. \quad (14)$$

To approximate its distribution, we introduce a scaling matrix for variance of the limit distribution. Recall  $S = \sum_{k \geq 0} A_k$  and define a block matrix  $G^\circ = (G_{i,l}^\circ)_{bn+1 \leq i \leq n-bn, 1 \leq l \leq n} \in \mathbb{R}^{(n-2bn)p \times n\tilde{p}}$  with components as  $p \times \tilde{p}$  dimension matrices,

$$G_{i,l}^\circ = \begin{cases} w_{i-l} \Lambda^{-1} S, & \text{if } i - bn \leq l \leq i - 1, \\ -w_{l-i} \Lambda^{-1} S, & \text{if } i + 1 \leq l \leq i + bn, \end{cases} \quad (15)$$

and elsewhere zero. We let  $\underline{z}$  be a Gaussian vector in  $\mathbb{R}^{n\bar{p}}$  with zero mean and identity covariance matrix. We set  $G_{i,\cdot}^\diamond$  to be  $(G_{i,1}^\diamond, G_{i,2}^\diamond, \dots, G_{i,n}^\diamond)$ . It can be shown that  $G_{i,\cdot}^\diamond \underline{z}$  has a similar covariance structure as  $V_i - \mathbb{E}V_i$ . We shall use the distribution of  $|G_{i,\cdot}^\diamond \underline{z}|_\infty$  to approximate the distribution of  $|V_i - \mathbb{E}V_i|_\infty$ . Combining this approximation with the bias term in (13), we shall expect that for each time point  $i$ , our normalized break test statistics can be approximated by the maximum of a Gaussian vector centered at  $d_i$ , i.e.,

$$\mathbb{P}(|V_i|_\infty \leq u) \approx \mathbb{P}(|d_i + G_{i,\cdot}^\diamond \underline{z}|_\infty \leq u).$$

We now let the statistics go over all the time points, and recall  $T_n = \max_{bn+1 \leq i \leq n-bn} |V_i|_\infty$ . Then we shall expect

$$\mathbb{P}(T_n \leq u) \approx \mathbb{P}(|\underline{d} + G^\diamond \underline{z}|_\infty \leq u), \quad (16)$$

and equivalently

$$\mathbb{P}(T_n \leq u) \approx \mathbb{P}(|\underline{d} + Z|_\infty \leq u), \quad (17)$$

where  $Z = (Z_{bn+1}^\top, Z_{bn+2}^\top, \dots, Z_{n-bn}^\top)^\top$  and  $(Z_i)_{bn+1 \leq i \leq n-bn}$  is a sequence of centered Gaussian vectors in  $\mathbb{R}^p$  with covariance matrices  $Q_{i,j}$ ,

$$Q_{i,j} = \varpi_{i,j} \Lambda^{-1} \Sigma \Lambda^{-1} \quad \text{and} \quad \varpi_{i,j} = \sum_{l=1}^n w_{|i-l|} w_{|j-l|} \text{sign}(i-l) \text{sign}(j-l). \quad (18)$$

To see the equivalence between (16) and (17), let

$$Q = (Q_{i,j})_{bn+1 \leq i,j \leq n-bn} = G^\diamond G^{\diamond\top}.$$

Then  $Z$  is a Gaussian vector with zero mean and covariance matrix  $Q$ . Note that

$$Z_i \stackrel{d}{=} G_{i,\cdot}^\diamond \underline{z} \quad \text{and} \quad Z \stackrel{d}{=} G^\diamond \underline{z}. \quad (19)$$

This transformation from  $G^\diamond \underline{z}$  to  $Z$  is to show that the involved Gaussian process only depends on the long-run covariance matrix  $\Sigma$  and the weight functions.

The above argument will be rigorously formulated in Theorem 1 in the next subsection.

## 2.3 Gaussian approximation

In this subsection, we provide the formal theory supporting our test. We first present necessary assumptions. The following is to guarantee the smoothness of the trend functions  $\mu_j(u)$  when no break occurs.

ASSUMPTION 2.1. *Function  $f_j \in \mathcal{C}^2[0, 1]$  with  $\max_{1 \leq j \leq p} |f'_j|_\infty \leq c_f$ ,  $\max_{1 \leq j \leq p} |f''_j|_\infty \leq c_f$  for some constant  $c_f > 0$ .*

Additionally, to ensure the property of our kernel estimation, we need conditions on the kernel function.

ASSUMPTION 2.2. *The kernel  $K(\cdot) \geq 0$  is symmetric with support  $[-1, 1]$ ,  $|K|_\infty < \infty$  and  $\int_{-1}^1 K(x)dx = 1$ . Also assume  $K(x)$  has first-order derivative with  $|K'|_\infty < \infty$ . Let  $b \rightarrow 0$  and  $bn \rightarrow \infty$ .*

We also set conditions on the regularity of the long-run covariance matrix and the dependency strength of the noise sequence.

ASSUMPTION 2.3. *(Lower bound for the long run variance)  $\sigma_{j,j} \geq c_\sigma$ ,  $1 \leq j \leq p$  for some finite constant  $c_\sigma > 0$ .*

We need enough separation between adjacent breakpoints.

ASSUMPTION 2.4. *(Separation) Assume  $\min_{1 \leq i, j \leq K_0} |\tau_i - \tau_j| \gg bn$ .*

It is worth noting that Assumption 2.4 implies that the number of breaks  $K_0$  shall not exceed the order  $1/b$ .

ASSUMPTION 2.5. *(Dependence strength)  $\max_{1 \leq j \leq p} \sum_{k \geq i} |A_{k,j}|_2 / \sigma_{j,j}^{1/2} \leq c_s (i \vee 1)^{-\beta}$ , where  $\beta > 0$  is some constant and  $A_{k,j}$  is the  $j$ th row of  $A_k$ .*

Assumption 2.5 is a very general spatial and temporal dependence condition and embraces many interesting processes. It requires an algebraic decay rate of the temporal dependence. However, the cross-sectional dependence does not need to be weak; and in fact, it can be strong such that it has a factor structure. We provide an example as follows.

EXAMPLE 1. Assume that  $\eta_t, \eta'_t \in \mathbb{R}^p$  are i.i.d random vectors with zero mean and covariance matrix  $I_p$ . Let

$$\epsilon_t = F_t + Z_t, \text{ with } Z_t = \sum_{k \geq 0} \Lambda_k \eta_{t-k} \text{ and } F_t = \sum_{k \geq 0} v f_k^\top \eta'_{t-k}, \quad (20)$$

where  $\Lambda_k = \text{diag}(\lambda_{k,1}, \dots, \lambda_{k,p})$ ,  $v = (v_1, \dots, v_p)^\top$  and  $f_k = (f_{k,1}, \dots, f_{k,p})^\top$ . Here  $F_t$  is the factor term and  $Z_{t,j}$  are independent for different  $j$ . Then the long-run variances for  $Z_{t,j}$  and  $F_{t,j}$  are  $\sigma_{Z,j} = (\sum_{k \geq 0} \lambda_{k,j})^2$  and  $\sigma_{F,j} = |\sum_{k \geq 0} f_k|_2^2 v_j^2$ , respectively. If for some constant  $c > 0$ ,

$$\sum_{k \geq i} |\lambda_{k,j}| / \sigma_{Z,j}^{1/2} \leq ci^{-\alpha} \quad \text{and} \quad \sum_{k \geq i} |f_k|_2 |v_j| / \sigma_{F,j}^{1/2} \leq ci^{-\alpha}, \quad (21)$$

then Assumption 2.5 holds with  $\beta = \alpha$ . To see this, we note  $|A_{k,j,\cdot}|_2 = (\lambda_{k,j}^2 + |f_k|_2^2 v_j^2)^{1/2}$ , and  $\sigma_{j,j} = \sigma_{Z,j}^2 + \sigma_{F,j}^2$ . Hence,

$$\sum_{k \geq i} |A_{k,j,\cdot}|_2 \leq \sum_{k \geq i} (|\lambda_{k,j}| + |v_j| |f_k|_2) \leq ci^{-\alpha} (\sigma_{Z,j}^{1/2} + \sigma_{F,j}^{1/2}) \leq \sqrt{2} ci^{-\alpha} \sigma_{j,j}^{1/2}.$$

□

ASSUMPTION 2.6. (*Finite moment*) The innovations  $\eta_{i,j}$  are i.i.d. with  $\mu_q = \|\eta_{1,1}\|_q < \infty$  for some  $q \geq 4$ .

ASSUMPTION 2.7. (*Sub-exponential*) The innovations  $\eta_{i,j}$  are i.i.d. with  $\mu_e = \mathbb{E}e^{a_0|\eta_{1,1}|} < \infty$ , for some  $a_0 > 0$ .

Assumptions 2.6 and 2.7 put tail assumptions on the distribution of the noise sequences. Given the above-mentioned conditions, we provide the main Gaussian approximation theorem, which is essential for the asymptotic distribution of our test statistics  $T_n$ . Our theorem extends the Gaussian approximation theory in Chernozhukov et al. (2013a, 2017), which build on the Stein's method and the anti-concentration bounds. Markedly, our theory is developed for modeling dependent data. To this aim, one important technical non-triviality lies in handling the spatial-temporal dependency of the trend stationary high-dimensional processes. We derive the corresponding concentration inequalities based on  $m$ -dependence approximation of the underlying processes. Compared to the existing results on Gaussian approximation for time series, for example Zhang et al. (2017), our setting works for non-centered Gaussian approximation that accommodates our interest for time series with breaks.

**Theorem 1.** (*Gaussian approximation for the test statistics*) Under Assumptions 2.1-2.5 and  $b^5 n \log(np) = o(1)$ .

(i) If Assumption 2.6 holds and

$$np(bn)^{-q/2}(\log(np))^{3q/2} = o(1), \quad (22)$$

(ii) If Assumption 2.7 holds and

$$(bn)^{-1}(\log(np))^{\max\{7, 2(1+\beta)/\beta\}} = o(1), \quad (23)$$

then we have

$$\sup_{u \in \mathbb{R}} |\mathbb{P}(T_n \leq u) - \mathbb{P}(|\underline{d} + Z|_\infty \leq u)| \rightarrow 0. \quad (24)$$

REMARK 1. (Allowed dimension) One key theoretical insight is that we explicitly show the trade-off between the tail assumption of the innovations and the allowed dimension of the time series  $p$  relative to the sample size  $n$  in the above theorem. In particular, when we have exponential tail assumption on the distribution of the innovations, we allow an ultra high dimension setup indicating  $p$  to be at an exponential rate with respect to  $n$ . And when we have only finite moment assumptions, we can allow  $p$  to be at a polynomial order with respect to  $n$ . Specifically, for Theorem 1 case (i), we allow  $p$  to be of some polynomial order of  $n$ , and its order depends on the value of  $q$ . For some  $\nu_1 > 0$  and  $0 < \nu_2 < 1/2$ , assume  $p \asymp n^{\nu_1}$  and  $b \asymp n^{-\nu_2}$ . If  $\nu_1 + \nu_2 < q/2 - 1$  and  $\nu_2 > 1/5$ , then conditions in case (i) hold. It is easy to see that the bigger the moment  $q$  is, the larger the allowance of the dimension  $p$ . The moment condition 2.6 depends on  $q$  which characterizes the heavy tailedness of the noise, larger  $q$  means thinner tails. For case (ii), we can allow  $p$  to be exponential in  $n$ , i.e. the ultra high dimensional scenario. For instance, for some  $\nu_1 > 0$  and  $1/5 < \nu_2 < 1$ , we can set  $p \asymp e^{n^{\nu_1}}$  and  $b \asymp n^{-\nu_2}$ . If  $\nu_1 < 5\nu_2 - 1$  and  $\nu_1 \max\{7, 2(1 + \beta)/\beta\} < 1 - \nu_2$ , then conditions in case (ii) hold.  $\square$

It is not hard to understand the size and power implication of Theorem 1 to our test. Under the null hypothesis, we have  $\underline{d} = 0$ , then for any prefixed significant level  $\alpha \in (0, 1)$ , we have the critical value of our test as  $q_\alpha$  i.e. the quantile of the Gaussian limit distribution,

$$q_\alpha = \inf_{r \geq 0} \{r : \mathbb{P}(|Z|_\infty > r) \leq \alpha\}. \quad (25)$$

As from the Gaussian approximation result in (24), we have the approximated sizes of the test statistics,

$$\left| \mathbb{P}(T_n > q_\alpha) - \mathbb{P}(|Z|_\infty > q_\alpha) \right| \rightarrow 0.$$

We shall reject the null hypothesis at the significant level  $\alpha$ , if the test statistics exceed the critical value i.e.  $T_n > q_\alpha$ .

To evaluate our testing power, consider the alternative that if not all  $\gamma_k = 0$ , then  $\underline{d}$  is non-zero. We have the following corollary for the power, which is a straightforward consequence of Theorem 1.

**Corollary 1.** *(Power) Under conditions in Theorem 1 (i) or (ii). The testing power is*

$$\beta_\alpha - \mathbb{P}(|\underline{d} + Z|_\infty \geq q_\alpha) = o(1).$$

Thus, we can see that the power of our test would depend on the vector  $\underline{d}$ , the size of which is determined by the true jump sizes i.e.  $\gamma_k$ s.

### 3 Estimation and inference of breaks

In this section, we show how to estimate the number of change-points, the time stamps, the spatial coordinates and the sizes of the structural breaks. We summarize the key steps of the adopted two step procedure for the multiple change-point detection. The main reason for a two-step estimation is to achieve an optimal rate of consistency for our break estimation. The first step can be regarded as an extension of the MOSUM  $l_\infty$  aggregation. Namely, in our first step, we conduct a “rough” estimation through a MOSUM type statistic as in Equation (9), and we can draw a conclusion on the existence of a break. In case it exists, we obtain a “rough” estimate of the change-points locations. In the second step, we refine our jump estimates based on a one-dimensional aggregated time series. The aggregation can be viewed as a projection using information on the jump estimators from the first step. To be more specific, within each time region around the  $k$ th breakpoint, we can aggregate data by a weighted sum of different coordinates whose weights are determined by the first step jump size estimators ( $\hat{\gamma}_k$ ). Instead of looking at the biggest break at one time point, the aggregated change-point statistics carry more information regarding significant

jumps across contemporaneous locations, and would thus provide better precision. In the following, we introduce the first “rough” estimation step and its properties in Subsection 3.1. We further improve the first step in Subsection 3.1 through an aggregated statistics, which is proposed and analyzed in Subsection 3.2.

### 3.1 The “rough” estimation step

We define the sizes of the breakpoints at time  $k$  as

$$|\Lambda^{-1}\gamma_k|_\infty.$$

Here, we normalize  $\gamma_k$  by the long-run standard deviations for the same reason as  $V_i$  in (9). Intuitively, the noise fluctuation levels for different locations can be very different, and at one location, a break can be significant due to purely high noise level without normalization. We define the minimum size of breaks over time as

$$\delta^\diamond = \min_{1 \leq k \leq K_0} |\Lambda^{-1}\gamma_k|_\infty. \quad (26)$$

In the following, we outline the steps of our testing, detecting and estimation procedure.

**Step 1.** For significance level  $\alpha$ , we test the existence of jumps based on the critical value  $q_\alpha$  in (25). If we find no significant breaks, then we cannot reject the null  $\mathcal{H}_0$ . In case our test statistic exceeds the critical value, we reject  $\mathcal{H}_0$  and acknowledge the existence of breaks, then we proceed to step 2.

**Step 2.** To detect the change-points, we collect all the time stamps with the jump statistics  $|V_\tau|_\infty$  exceeding a threshold value  $w^\dagger$ , namely,  $\mathcal{A}_1 = \{bn + 1 \leq \tau \leq n - bn : |V_\tau|_\infty > w^\dagger\}$ , where  $V_\tau$  is defined in (9). Let  $\hat{\tau}_1$  be the time point  $\tau$  in  $\mathcal{A}_1$  that maximizes the test statistics  $|V_\tau|_\infty$ . We further eliminate a  $2bn$  neighborhood of time points around  $\hat{\tau}_1$  from  $\mathcal{A}_1$  to create  $\mathcal{A}_2$ . Then we find the next point in  $\mathcal{A}_2$  that maximize  $|V_\tau|_\infty$ , and repeat the

same operation until the set  $\mathcal{A}_k$  is empty. Namely, for  $k \geq 1$ , we let the  $k$ th estimated break point be denoted as  $\hat{\tau}_k = \operatorname{argmax}_{\tau \in \mathcal{A}_k} |V_\tau|_\infty$  and  $\mathcal{A}_{k+1} = \mathcal{A}_k \setminus \{\tau : |\tau - \hat{\tau}_k| \leq 2bn\}$ . We denote the maximum number of breakpoints as  $\hat{K}_0$ , with  $\hat{K}_0 = \max_{k \geq 1} \{k : \mathcal{A}_k \neq \emptyset\}$ .

**Step 3.** Given the detected breakpoints in Step 2, we calculate the break sizes over time. We denote the window size to be  $M = bn$ ,

$$\hat{\gamma}_k = \hat{\mu}_{\hat{\tau}_k - M}^{(l)} - \hat{\mu}_{\hat{\tau}_k + M}^{(r)} \quad \text{and} \quad \hat{\delta}^\diamond = \min_{1 \leq k \leq \hat{K}_0} |\Lambda^{-1} \hat{\gamma}_k|_\infty. \quad (27)$$

It is worth noting that in this algorithm, we only need to calculate the gap statistics  $|V_\tau|_\infty$  once for each point. Hence, it is not time consuming regardless of the true number of breakpoints. In Step 1, we test the existence of the breaks. In Step 2, we use the estimated  $|V_\tau|_\infty$  for all the points from  $bn + 1$  to  $n - bn$  and select the points that are beyond the threshold  $w$ . Intuitively, the points in  $\mathcal{A}_1$  would contain the break indices, as well as points in their neighborhood where estimates are contaminated by the breaks. Therefore in Step 2, we find the local maximums and discard points around them. In Step 3, we estimate the sizes of the change-points and calculate their minimum values.

In the following, we shall provide consistency results of estimates of the break numbers, locations and break sizes in Theorem 2; and derive asymptotic distribution of break sizes in Theorem 3.

We need to first impose the minimum jump size condition on the break size as

**ASSUMPTION 3.1.** *Let  $\delta^\diamond \gg \max \{ \sqrt{\log(pn)/(bn)}, b \}$ .*

It can be seen that the break size requirement is related to the dimensionality of the time series, the number of observations available and the bandwidth parameter. The larger the sample  $n$ , the smaller the requirement for  $\delta^\diamond$  due to the better approximation of the trends. In the following theorem, we show that we would asymptotically obtain the right

number of breaks. Moreover, we can bound the errors of the estimated break locations and the break sizes. The threshold  $\omega^\dagger$  shall be set as a high quantile of its limited Gaussian distribution to ensure the consistent estimation of the breaks.

**Theorem 2.** *We assume conditions in Theorem 1 (i) or (ii) hold, and Assumption 3.1. If  $\delta^\diamond/2 \geq \omega^\dagger \geq 2c'_w(bn)^{-1/2}\sqrt{\log(pn)}$ , where  $c'_w$  is the constant defined as  $(bn \sum_{i=0}^n w_i^2)^{1/2} \rightarrow c'_w$ , then*

$$(i) \mathbb{P}(\hat{K}_0 = K_0) \rightarrow 1.$$

$$(ii) |\hat{\tau}_k - \tau_{k^*}| = O_{\mathbb{P}}\{\log(np)/\delta^{\diamond 2}\}, \text{ where } k^* = \operatorname{argmin}_i |\hat{\tau}_k - \tau_i|.$$

$$(iii) |\Lambda^{-1}(\hat{\gamma}_k - \gamma_{k^*})|_\infty = O_{\mathbb{P}}((bn)^{-1/2}\log(np)^{1/2} + b), \text{ which indicates } |\hat{\delta}^\diamond - \delta^\diamond| = O_{\mathbb{P}}((bn)^{-1/2}\log(np)^{1/2} + b).$$

Result (i) indicates that the number of breaks can be consistently estimated, (ii) suggests that the estimated break dates  $u_k$  can be consistently determined in view of  $u_k = \tau_k/n$  and (iii) shows that the break sizes can be consistently recovered. The convergence rate of the break sizes is dependent on the bandwidth  $b$ , sample size  $n$  and the dimension of the time series  $p$ . It shall be noted that the consistency rate of  $\hat{\tau}_k$  depends on the break size  $\delta^\diamond$ , which depends only on the maximum break size for any fixed time. Therefore having several large breaks simultaneously would not improve the break size estimation.

Given the consistency of the breakpoints, we can obtain a distribution theory that facilitates us in making inferences on the break sizes. Let  $\tilde{Z}$  be a Gaussian vector in  $\mathbb{R}^p$  with zero mean and covariance matrix

$$\tilde{Q} := Q_{bn+1, bn+1} = 2 \sum_{t=1}^{bn} w_t^2 \Lambda^{-1} \Sigma \Lambda^{-1}. \quad (28)$$

**Theorem 3.** (*Break size inference*) Assume conditions in Theorem 2 and  $b^3 n \log(np) = o(1)$ . We have

$$\sup_{u \in \mathbb{R}} |\mathbb{P}(|\Lambda^{-1}(\hat{\gamma}_k - \gamma_{k^*})|_\infty \leq u) - \mathbb{P}(|\tilde{Z}|_\infty \leq u)| \rightarrow 0, \text{ where } k^* = \operatorname{argmin}_i |\hat{\tau}_k - \tau_i|.$$

This theorem indicates that the maximum of the difference between the estimated jump size  $\hat{\gamma}_k$  and the true jump size  $\gamma_k$  can be approximated by the maximum of a Gaussian random vector with the same asymptotic variance-covariance structure. Based on Theorem 2 (ii) and Theorem 3, we can construct the confidence interval for  $\gamma_k$ . We set

$$\alpha = \mathbb{P}(|\tilde{Z}|_\infty \geq q) \quad \text{and} \quad \theta = (\sigma_{1,1}^{1/2}, \sigma_{2,2}^{1/2}, \dots, \sigma_{p,p}^{1/2})^\top. \quad (29)$$

Then the confidence interval for vector  $\gamma_{k^*}$  at level  $\alpha$  is  $(\hat{\gamma}_k - q\theta, \hat{\gamma}_k + q\theta)$ .

### 3.2 The refined aggregation step

The estimation in the first step is only driven by  $|\gamma_k|_\infty$ , i.e. the maximum size of jumps at a time point  $\tau_k$ . Therefore it is only sensitive to the biggest jump across all the time series at the same time. In case there are multiple simultaneous time series jumps, it would be beneficial to modify our procedure to aggregate all of the series with a jump. This enlightens us to propose a two-stage method: first, we follow the steps in the previous subsections to detect the “rough” timing of the jumps and the estimated jump sizes; second, for each  $bn$  neighborhood of a change-point estimate  $\hat{\tau}_k$  obtained from step one, we update the change-point estimates according to a newly aggregated time series. The time series is calculated with a weighted sum of simultaneous observations corresponding to significant jump locations and the weights are based on the jump size estimates in the first step. The aggregation returns a one-dimensional time series with richer information on the cross-sectional jumps.

We denote  $\mathcal{S}_k$  to be the set of series that jump at location  $\tau_k$ , that is

$$\mathcal{S}_k = \{1 \leq j \leq p \mid \gamma_{k,j} \neq 0\}, \quad (30)$$

where  $\gamma_{k,j}$  is the  $j$ th coordinate of  $\gamma_k$ . Detailed steps of the aggregation are formulated as follows:

**Stage 1.** Apply Steps 1-3 in Subsection 3.1 to obtain  $\hat{\tau}_k$  and  $\hat{\gamma}_k$ ,  $k = 1, 2, \dots, \hat{K}_0$ . For some  $w^\dagger > 0$ , let the estimation of  $\mathcal{S}_k$  be

$$\hat{\mathcal{S}}_k = \{1 \leq j \leq p \mid |(\Lambda^{-1}\hat{\gamma}_k)_j| \geq w^\dagger\}. \quad (31)$$

In practice,  $w^\dagger$  can be chosen to be large enough to ensure that we can detect all the jumps with probability 1 as in Theorem 2.

**Stage 2.** For  $|t - \hat{\tau}_k| \leq 2bn$ , we let

$$X_t = \sum_{j \in \hat{\mathcal{S}}_k} (\Lambda^{-1}\hat{\gamma}_k)_j (\Lambda^{-1}Y_t)_j. \quad (32)$$

Note that after the modification, for all the jump locations, the new time series  $X_t$  would only contain positive sized jumps i.e.  $\sum_{j \in \hat{\mathcal{S}}_k} (\Lambda^{-1}\hat{\gamma}_k)_j^2$ . This step can be understood as a projection of the high-dimensional observations  $\Lambda^{-1}Y_t$  according to the direction of  $\Lambda^{-1}\hat{\gamma}_k (j \in \hat{\mathcal{S}}_k)$ . This is similar to the idea of Wang and Samworth (2018).

Based on the aggregated time series  $X_t$ , the refined change-point locations can be detected through a CUSUM type of test statistics, for  $k = 1, 2, \dots, \hat{K}_0$ ,

$$\tilde{\tau}_k = \operatorname{argmax}_{|t - \hat{\tau}_k| \leq bn} \left( \sum_{s=\hat{\tau}_k-2bn}^{\hat{\tau}_k+2bn} X_s \frac{t - \hat{\tau}_k + bn}{4bn + 1} - \sum_{s=\hat{\tau}_k-2bn}^{t-1} X_s \right) \sqrt{\frac{4bn + 1}{(t - (\hat{\tau}_k - 2bn) + 1)(\hat{\tau}_k + 2bn - t)}}. \quad (33)$$

After we update the break points estimation, we can construct confidence intervals for the updated breakpoints estimates  $\tilde{\tau}_k$ . We denote the long-run correlation matrix to be

$(\tilde{\sigma}_{i,j})_{i,j} = \Lambda^{-1}\Sigma\Lambda^{-1}$ , where  $\Sigma$  is the long run covariance matrix for  $\epsilon_t$ . We let  $\tilde{\Sigma}_k = (\tilde{\sigma}_{i,j})_{i,j \in \mathcal{S}_k}$  be the sub covariance matrix corresponding to coordinates in  $\mathcal{S}_k$  at time  $\tau_k$  and let the standardized significant break sizes  $\tilde{\gamma}_k = (\Lambda^{-1}\gamma_k)_{i \in \mathcal{S}_k}$ . We define two objects involved in the limit distributions of the breaks, i.e.,

$$a_k = |\tilde{\gamma}_k|_2^2 \quad \text{and} \quad \varsigma_k^2 = \tilde{\gamma}_k^\top \tilde{\Sigma}_k \tilde{\gamma}_k. \quad (34)$$

Then  $\varsigma_k^2$  is the long-run variance for the sequence  $\sum_{j \in \mathcal{S}_k} (\Lambda^{-1}\gamma_k)_j (\Lambda^{-1}\epsilon_t)_j$ . For the aggregated jump estimation, we alternatively define the minimum jump size across different locations and time points as

$$\delta^\dagger = \min_{1 \leq k \leq K_0} \min_{j \in \mathcal{S}_k} |(\Lambda^{-1}\gamma_k)_j|.$$

Then  $\delta^\dagger \leq \delta^\circ$  and it functions similarly as  $\delta^\circ$  to capture the identifiable jump size of the time series. We shall put the same assumption on  $\delta^\dagger$  as on  $\delta^\circ$ . It is worth noting that  $\delta^\dagger$  is the minimum jump size to ensure the consistency of our break estimation.

**ASSUMPTION 3.2.** *Let  $\delta^\dagger \gg \max\{\sqrt{\log(pn)/(bn)}, b\}$ .*

In the following corollary, we show that we can consistently recover the locations of the series with a jump for each change-point. It can be directly derived from Theorem 2 (iii).

**Corollary 2.** *We assume conditions in Theorem 1 (i) or (ii) hold, and Assumption 3.2. If  $\delta^\dagger/2 \geq w^\dagger \gg (bn)^{-1/2} \log(np)^{1/2} + b$ , then for any  $k$  we have*

$$\mathbb{P}(\hat{\mathcal{S}}_k = \mathcal{S}_k) \rightarrow 1.$$

In addition, we provide a theorem that allows us to make inference on the estimated break-dates  $\tilde{\tau}_k$ .

**Theorem 4.** (*Aggregated break estimation*) Assume conditions in Corollary 2, and that for some constants  $c_1, c_2 > 0$ ,

$$c_1 \leq \lambda_{\max}(\Lambda^{-1}\Sigma\Lambda^{-1})/\lambda_{\min}(\Lambda^{-1}\Sigma\Lambda^{-1}) \leq c_2. \quad (35)$$

Recall definition of  $a_k$  and  $\varsigma_k$  in (34). Then we have for any fixed  $k$ ,

(i)  $|\tilde{\tau}_k - \tau_{k^*}| = O_{\mathbb{P}}(\varsigma_k^2/a_k^2)$ .

(ii) In addition, if Assumption 2.5 holds with  $\beta > 1$ , then we have

$$\tilde{\tau}_k - \tau_{k^*} \xrightarrow{\mathcal{D}} (\varsigma_k/a_k)^2 \operatorname{argmax}_r (-2^{-1}|r| + \mathbb{W}(r)),$$

where  $\mathbb{W}(r)$  is a two-sided Brownian motion, that is  $\mathbb{W}(r) = \mathbb{W}_1(r)$ , if  $r > 0$ , and  $\mathbb{W}(r) = \mathbb{W}_2(-r)$ , if  $r \leq 0$ , and  $\mathbb{W}_1, \mathbb{W}_2$  are two independent Brownian motions.

REMARK 2. We shall note that the consistency rate of  $\tilde{\tau}_k$  is improved compared to the results for  $\hat{\tau}$  in Theorem 2 ii).  $a_k$  which is an  $l_2$  aggregation of simultaneous significant break sizes, plays a role in the rate of convergence of  $\tilde{\tau}_k$ . For instance, if we assume that there are  $s$  breaks which are of size  $\delta > 0$  in the cross-sectional dimensional, then  $a_k = s\delta^2$ . If moreover there is no cross-sectional correlation, i.e.,  $\tilde{\Sigma}_k = I$ , then we may expect  $\tilde{\tau}_k$  to be consistent so long that  $1/(s\delta^2) \rightarrow 0$ , while  $\hat{\tau}_k$  can be not consistent. Thus the rate of  $\tilde{\tau}_k$  will be better than  $\hat{\tau}$ . Moreover, the long-run variance also plays a critical role in the rate of convergence. For example, when the variance part of the limit distribution satisfies  $\varsigma_k^2 \leq |\tilde{\Sigma}_k|_2 a_k$ , if  $|\tilde{\Sigma}_k|_2/a_k = o(1)$  then by Theorem 4 (i), we have  $\tilde{\tau}_k \rightarrow \tau_k$  in probability. If  $\tilde{\Sigma}_k$  is a  $d$ -banded matrix,  $|\tilde{\Sigma}_k|_2 \leq (|\tilde{\Sigma}_k|_1 |\tilde{\Sigma}_k|_{\infty})^{1/2} \leq d$ . We can derive that  $|\tilde{\tau}_k - \tau_k| = O_{\mathbb{P}}(d/a_k)$ .  $\square$

To illustrate the insight of Remark 2, we compare the performance of a simple model with  $\mathcal{N}(0, 0.1)$  and one breakpoint placed at  $\tau_0 = 50$ . Figure 3.2 shows the histogram of

$\hat{\tau} - \tau_0$  and  $\tilde{\tau} - \tau_0$  respectively. The jump-size for all breaks are the same, with the value  $1.6\sqrt{\log(np)}(\tau_0)^{-1/3}$ . As the dimension  $p$  grows, we see the significant improvement of the performance of  $\tilde{\tau}$  relative to that of  $\hat{\tau}$ .

From Theorem 4, with estimates of  $a_k$  and  $\varsigma_k$ , we can construct a  $100(1-\alpha)\%$  confidence interval for  $\tilde{\tau}_k$ :

$$(\tilde{\tau}_k - \lfloor \hat{q}'_{1-\alpha/2} \rfloor - 1, \tilde{\tau}_k + \lfloor \hat{q}'_{\alpha/2} \rfloor + 1), \quad (36)$$

where  $q'_{1-\alpha/2}$  ( $q'_{\alpha/2}$ ) is  $1 - \alpha/2$  ( $\alpha/2$ )th quantile of the limit distribution of the break point  $\tilde{\tau}_k$ , i.e.  $\operatorname{argmax}_r \{-2^{-1}a_k|r| + \varsigma_k \mathbb{W}(r)\}$  and  $\hat{q}'_{\alpha/2}$  ( $\hat{q}'_{1-\alpha/2}$ ) are estimates of the quantiles.  $\lfloor \cdot \rfloor$  denotes the floor function.  $q'_{1-\alpha/2}$  ( $q'_{\alpha/2}$ ) can be calculated following Stryhn (1996). Alternatively, we can also simulate the critical values.

## 4 Long-run covariance matrix

In the previous sections, we assume that  $\Sigma$  is known. However, this is unrealistic in practice, as we mostly do not know the long-run covariance matrix. Thus, an estimation of the long-run covariance matrix is needed in Gaussian approximation. A simpler version of this problem was considered by Politis et al. (1999) and Lahiri (2003), who allow for a constant mean of the random vector. More generally, Chen and Wu (2017) consider the high-dimensional situation with smooth trends. However, this does not fit directly to our interest due to the possible existence of the breakpoints. We then propose a robust covariance matrix estimation motivating from the M-estimation method in Catoni et al. (2012). It is worth noting that due to the jumps, our method shall be different from the classical covariance matrix estimation. Our long-run variance-covariance matrix estimation is complementary to the recent article on high-dimensional robust matrix method under independence settings in Fan et al. (2017).

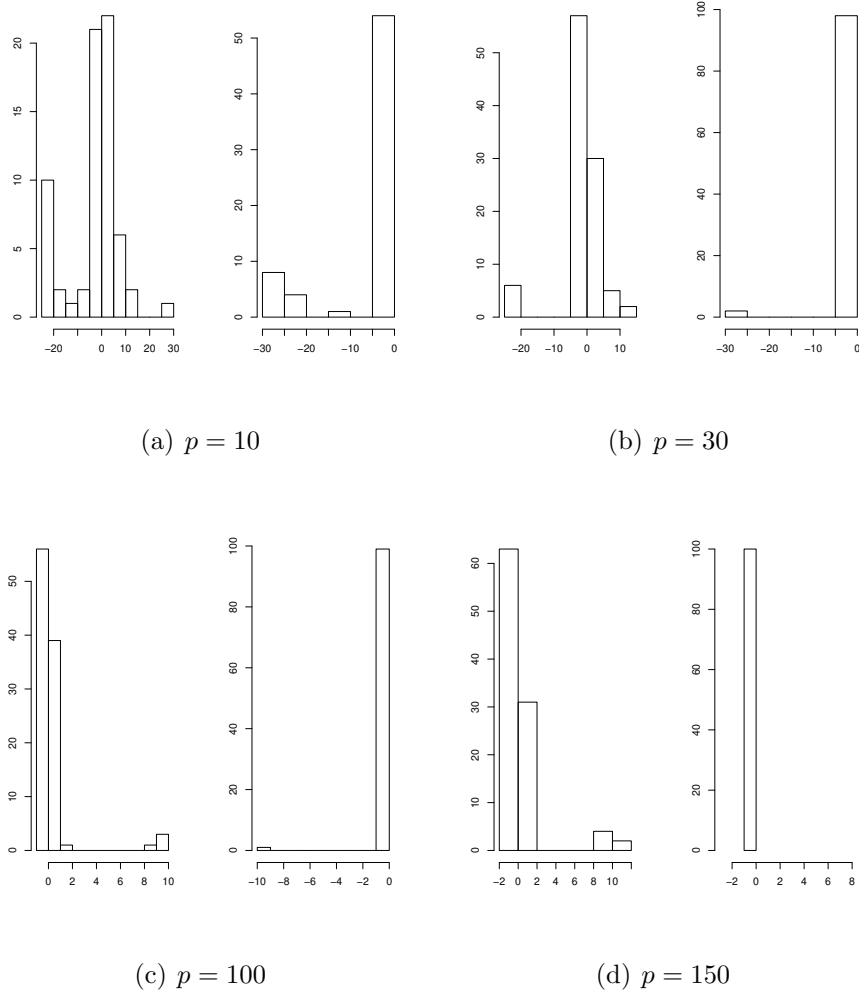


Figure 1: Histogram of  $\hat{\tau} - \tau_0$  (left) and  $\tilde{\tau} - \tau_0$  (right) for  $n = 100$ ,  $p = 10, 30, 100, 150$ ,  $K_0 = 1$ . The number of breaks in the cross-sectional dimension are  $s = 1, 5, 20, 30$  respectively, and there are 100 simulation samples. (a) describes the case with  $p = 10, s = 1$ ; (b) describes the case with  $p = 30, s = 5$ ; (c) describes the case with  $p = 100, s = 20$ ; and, (d) describes the case with  $p = 150, s = 30$ .

First of all, to account for temporal dependency, we group our observations into blocks of the same size  $m$ , for some  $m \in \mathbb{N}$ . We denote the number of blocks  $N_1 = \lfloor (n-m)/m \rfloor$ , and the observation indices within a block  $k$  is set to be  $\mathcal{A}_k = \{t \in \mathbb{N} : km+1 \leq t \leq (k+1)m\}$ , and we let

$$\xi_k = \sum_{t \in \mathcal{A}_k} Y_t/m,$$

be the average observations within the block  $\mathcal{A}_k$ . Without jumps, a natural estimate of the long-run covariance matrix is

$$\sum_{k=1}^{N_1} (m/2)(\xi_k - \xi_{k-1})(\xi_k - \xi_{k-1})^\top / N_1.$$

Note that we take the difference  $\xi_k - \xi_{k-1}$  to cancel out the trends, as the trend function  $\mu(\cdot)$  is smooth, and the aggregated difference between two consecutive blocks can be shown to be of order  $m/n$ , which vanishes when  $m/n \rightarrow 0$ . However, this estimator can be greatly contaminated by the jumps. Thus a robust covariance matrix estimation is needed. We borrow the framework of Catoni (2012), who considers a new robust  $M$ - estimation method. We extend the method for estimating our long run covariance matrix.

We denote  $\xi_k = (\xi_{k,1}, \xi_{k,2}, \dots, \xi_{k,p})^\top$  and let

$$\hat{\sigma}_{i,j,k} = m(\xi_{k,i} - \xi_{k-1,i})(\xi_{k,j} - \xi_{k-1,j})/2, \quad k = 1, 2, \dots, N_1. \quad (37)$$

For some  $\alpha_{i,j} > 0$ , we denote the  $M$ - estimation zero function of our variance-covariance matrix to be

$$h_{i,j}(u) = \sum_{k=1}^{N_1} \phi_{\alpha_{i,j}}(\hat{\sigma}_{i,j,k} - u)/N_1, \quad (38)$$

where  $\phi_\alpha(x) = \alpha^{-1}\phi(\alpha x)$  and

$$\phi(x) = \begin{cases} \log(2), & x \geq 1, \\ -\log(1 - x + x^2/2), & 0 \leq x \leq 1, \\ \log(1 + x + x^2/2), & -1 \leq x \leq 0, \\ -\log(2), & x \leq -1. \end{cases} \quad (39)$$

REMARK 3. Function  $|\phi(\cdot)|$  is bounded by  $\log(2)$  and is Lipschitz continuous with the Lipschitz constant bounded by 1. Also note that the function has envelopes of nice form,

$$-\log(1 - x + x^2/2) \leq \phi(x) \leq \log(1 + x + x^2/2). \quad (40)$$

□

We set the estimates of the components of the long-run covariance matrix  $\hat{\sigma}_{i,j}$  to be the solution to  $h_{i,j}(u) = 0$  (if more than one root, pick one of them). We can collect all the estimates of the variance and covariances and organize them into the variance covariance matrix,

$$\hat{\Sigma} = (\hat{\sigma}_{i,j})_{1 \leq i,j \leq p} \quad \text{and} \quad \hat{\Lambda} = (\hat{\sigma}_{1,1}^{1/2}, \hat{\sigma}_{2,2}^{1/2}, \dots, \hat{\sigma}_{p,p}^{1/2}). \quad (41)$$

We denote  $\bar{\sigma}_{i,i} = 2 \sum_{N_1/4 \leq k \leq 3N_1/4} \hat{\sigma}_{i,i,k} / N_1$  and let the  $\alpha_{i,j}$  in (38) be  $\bar{\sigma}_{i,i}^{1/2} \bar{\sigma}_{j,j}^{1/2} (m/bn)^{1/2}$ .

**Theorem 5.** (*Long-run variance precision*) *We assume that Assumption 2.5 holds with  $\beta \geq 1.5$  and let*

$$\varsigma = |\Lambda^{-1}(\hat{\Sigma} - \Sigma)\Lambda^{-1}|_{\max}.$$

*Then we have  $\varsigma \log(np)^2 \rightarrow 0$  in probability under either one of the following two conditions:*

(i) Assuming conditions in Theorem 1 (i),  $p \leq cn^v$  with  $v < q/4 - 1/2$  and some  $c > 0$ , we take  $m = \min\{n^{1-4v/(q-2)}, (bn)^{1/2}\}$ .

(ii) Assuming conditions in Theorem 1 (ii), we take  $m = \min\{b^{-1}, (bn)^{1/2}\}$ .

By the above theorem, for the diagonal values, we have  $\max_{1 \leq i \leq p} |\hat{\sigma}_{i,i} - \sigma_{i,i}|/\sigma_{i,i} = o_{\mathbb{P}}(1)$ . Let  $\hat{Q}$  be the same as  $Q$  in (18), with  $\Sigma$  therein replaced by  $\hat{\Sigma}$  in (41). We denote  $\hat{Z}$  as the Gaussian vector with covariance matrix  $\hat{Q}$ , then by Theorem 5 and Lemma 3,  $|\hat{Z} + d|_{\infty}$  converges to  $|Z + d|_{\infty}$  in distribution. Thus, all previous results are still valid with  $\hat{\Sigma}$  as well.

## 5 Application

As an application, we analyze the monthly the unemployment rate data in 20 U.S. states (namely, Alabama, Arizona, California, Colorado, Florida, Georgia, Illinois, Indiana, Kentucky, Michigan, Mississippi, New Jersey, New York, North Carolina, Ohio, Pennsylvania, Texas, Virginia, Washington and Wisconsin). The data time span is from January 1976 to September 2018, and the data source is Bureau of Labor Statistics from Department of Labor in the United States (<https://www.bls.gov/>). Figure 2 displays the 20 time series of unemployment rate. Although from a long time span and on an overall level, we do not see obvious abrupt structural changes, it would be still of great interest to consider detected changes induced by some well-known exogenous shocks, such as the subprime mortgage crisis in 2007-2008. It is understood that there will be likely a smooth cyclical trend associated with the unemployment time series, as they mostly rise during a recession and fall during periods of economics prosperity, following the business cycle. Further studies on whether the shock induced by recessions creates a significant structural change in the unemployment rate should be performed.

Figure 3 shows the estimated robust long-run correlation matrix using the method in Section 4. One sees some significant values in the correlations between residuals in different states. We can see that the correlations across different locations are not negligible, however our method is robust against the underlying spatial-temporal dependency.

Figure 4 plots the estimated breakpoints and the confidence intervals around them. We see that the estimated breaks  $\tilde{\tau}_k$  using the CUSUM statistics in Equation (33) pick up the breaks earlier than the estimates obtained from the non-aggregated method i.e.  $\hat{\tau}_k$ . We can see that our method can identify important dates such as the financial crisis period starting in Jan, 2009. Moreover,  $\tilde{\tau}_k$  tends to detect earlier dates of structure changes than the observed averaged peaks in the time series. Other time-points with significant jumps detected are January 1977, October 1981, January 1991 and October 2001. There are a few recession periods documented by the national bureau of Economics Research, namely November 1973 to March 1975, July 1981 to November 1982, July 1990 to March 1991, July 1981 to November 1982, July 1990 to March 1991 and March 2001 to November 2001. All the break-dates of the unemployment structure happen during or slightly before the recession periods, featuring a close relationship between the structure change of unemployment rate and the economic cycles. This implies that economic recessions indeed bring significant structural changes in unemployment rates across all the states.

Figure 2: Plot of Unemployment rate of 20 U.S. states

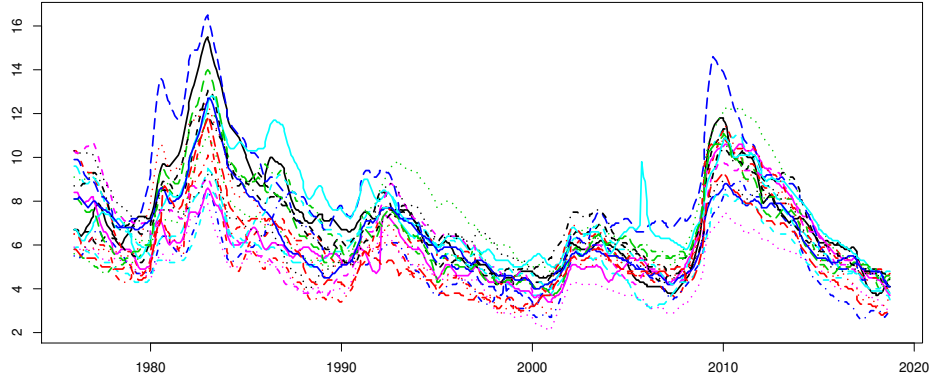


Figure 3: Plot of estimation of the robust long-run correlation matrix;  $m = 10$ .

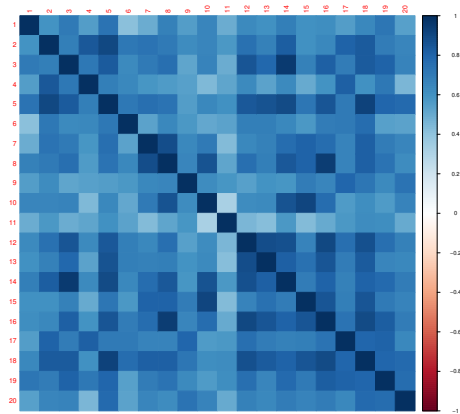
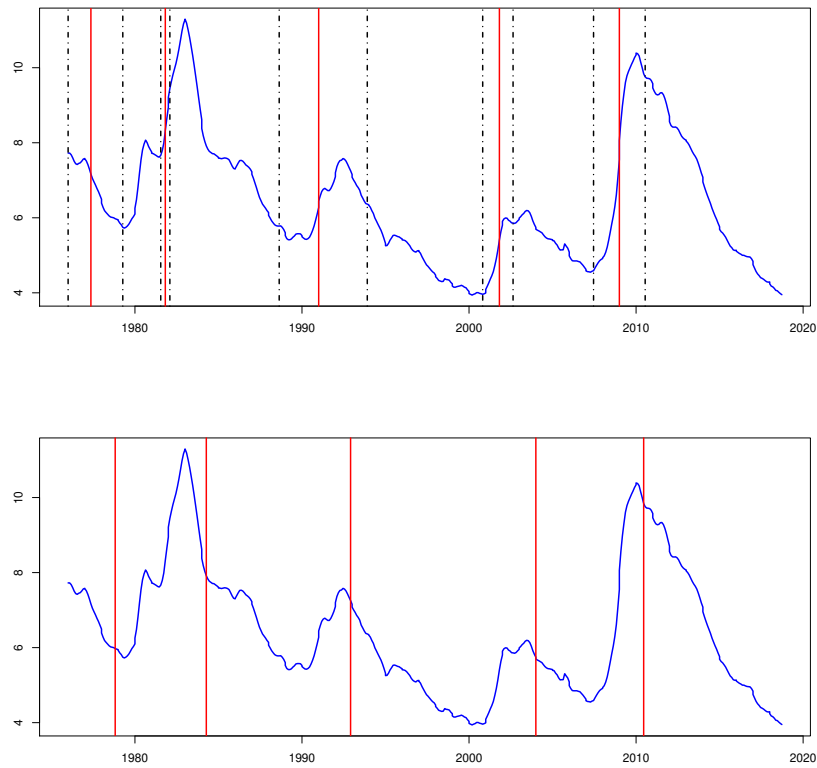


Figure 4: Plot of estimated breakpoints  $\tilde{\tau}_k(\hat{\tau}_k)$  (red lines) and their confidence intervals (dotted black lines).  $\tilde{\tau}_k$  (upper panel),  $\hat{\tau}_k$  (lower panel). The blue time series line represents the average unemployment rate over states under consideration.



## References

- Bai, J. (2010). Common breaks in means and variances for panel data. *J. Econometrics* 157(1), 78–92.
- Bai, J. and P. Perron (1998). Estimating and testing linear models with multiple structural changes. *Econometrica* 66(1), 47–78.
- Bai, J. and P. Perron (2003). Computation and analysis of multiple structural change models. *Journal of applied econometrics* 18(1), 1–22.
- Berkes, I., W. Liu, and W. B. Wu (2014). Komlós-Major-Tusnády approximation under dependence. *Ann. Probab.* 42(2), 794–817.
- Burkholder, D. L. (1988). Sharp inequalities for martingales and stochastic integrals. *Astérisque* (157-158), 75–94. Colloque Paul Lévy sur les Processus Stochastiques (Palaiseau, 1987).
- Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.* 48(4), 1148–1185.
- Catoni, O. et al. (2012). Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, Volume 48, pp. 1148–1185. Institut Henri Poincaré.
- Chen, L., W. Wang, and W. B. Wu (2017). Dynamic semiparametric factor model with structural breaks. *Preprint*.
- Chen, L. and W. B. Wu (2017). Testing for trends in high-dimensional time series. *preprint*.
- Chernozhukov, V., D. Chetverikov, and K. Kato (2013a). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.* 41(6), 2786–2819.
- Chernozhukov, V., D. Chetverikov, and K. Kato (2013b). Testing many moment inequalities. *arXiv preprint arXiv:1312.7614*.
- Chernozhukov, V., D. Chetverikov, and K. Kato (2015). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. *Probab. Theory Related Fields* 162(1-2), 47–70.

- Chernozhukov, V., D. Chetverikov, and K. Kato (2017). Central limit theorems and bootstrap in high dimensions. *Ann. Probab.* 45(4), 2309–2352.
- Cho, H. (2016). Change-point detection in panel data via double CUSUM statistic. *Electron. J. Stat.* 10(2), 2000–2038.
- Cho, H. and P. Fryzlewicz (2015). Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 77(2), 475–507.
- Eichinger, B. and C. Kirch (2018). A MOSUM procedure for the estimation of multiple random change points. *Bernoulli* 24(1), 526–564.
- El Machkouri, M., D. Volný, and W. B. Wu (2013). A central limit theorem for stationary random fields. *Stochastic Process. Appl.* 123(1), 1–14.
- Enikeeva, F. and Z. Harchaoui (2019). High-dimensional change-point detection under sparse alternatives. *Ann. Statist.* 47(4), 2051–2079.
- Fan, J. and I. Gijbels (1996). *Local polynomial modelling and its applications*, Volume 66 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.
- Fan, J., Q. Li, and Y. Wang (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 79(1), 247–265.
- Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *Ann. Statist.* 42(6), 2243–2281.
- Fryzlewicz, P. (2018). Tail-greedy bottom-up data decompositions and fast multiple change-point detection. *Ann. Statist.* 46(6B), 3390–3421.
- Harlé, F., F. Chatelain, C. Gouy-Pailler, and S. Achard (2016). Bayesian model for multiple change-points detection in multivariate time series. *IEEE Trans. Signal Process.* 64(16), 4351–4362.
- Hušková, M. and A. Slabý (2001). Permutation tests for multiple changes. *Kybernetika* 37(5), 605–622.

- Jackson, B., J. D. Scargle, D. Barnes, S. Arabhi, A. Alt, P. Gioumouisis, E. Gwin, P. Sangtrakulcharoen, L. Tan, and T. T. Tsai (2005). An algorithm for optimal partitioning of data on an interval. *IEEE Signal Processing Letters* 12(2), 105–108.
- Jirak, M. (2015). Uniform change point tests in high dimension. *Ann. Statist.* 43(6), 2451–2483.
- Killick, R., P. Fearnhead, and I. A. Eckley (2012). Optimal detection of changepoints with a linear computational cost. *J. Amer. Statist. Assoc.* 107(500), 1590–1598.
- Lahiri, S. N. (2003). *Resampling methods for dependent data*. Springer Series in Statistics. Springer-Verlag, New York.
- Lee, S., M. H. Seo, and Y. Shin (2016). The lasso for high dimensional regression with a possible change point. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78(1), 193–210.
- Lévy-Leduc, C. and F. Roueff (2009). Detection and localization of change-points in high-dimensional network traffic data. *Ann. Appl. Stat.* 3(2), 637–662.
- Li, D., J. Qian, and L. Su (2016). Panel data models with interactive fixed effects and multiple structural breaks. *Journal of the American Statistical Association* 111(516), 1804–1819.
- Liu, H., C. Gao, and R. J. Samworth (2019). Minimax rates in sparse, high-dimensional changepoint detection. *arXiv preprint arXiv:1907.10012*.
- Meier, A., C. Kirch, and H. Cho (2019). mosum: A package for moving sums in change-point analysis. *Preprint*.
- Nazarov, F. (2003). On the maximal perimeter of a convex set in  $\mathbb{R}^n$  with respect to a Gaussian measure. In *Geometric aspects of functional analysis*, Volume 1807 of *Lecture Notes in Math.*, pp. 169–187. Springer, Berlin.
- Olshen, A. B., E. Venkatraman, R. Lucito, and M. Wigler (2004). Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics* 5(4), 557–572.
- Politis, D. N., J. P. Romano, and M. Wolf (1999). *Subsampling*. Springer Series in Statistics. Springer-Verlag, New York.

- Preuss, P., R. Puchstein, and H. Dette (2015). Detection of multiple structural breaks in multivariate time series. *J. Amer. Statist. Assoc.* 110(510), 654–668.
- Rio, E. (2009). Moment inequalities for sums of dependent random variables under projective conditions. *J. Theoret. Probab.* 22(1), 146–163.
- Scott, A. J. and M. Knott (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, 507–512.
- Stryhn, H. (1996). The location of the maximum of asymmetric two-sided brownian motion with triangular drift. *Statistics & Probability Letters* 29(3), 279 – 284.
- Tibshirani, R. and P. Wang (2007). Spatial smoothing and hot spot detection for cgh data using the fused lasso. *Biostatistics* 9(1), 18–29.
- Wang, T. and R. J. Samworth (2018). High dimensional change point estimation via sparse projection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 80(1), 57–83.
- Wu, W.-B. and Y. N. Wu (2016). Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electron. J. Stat.* 10(1), 352–379.
- Wu, W. B. and Z. Zhao (2007). Inference of trends in time series. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 69(3), 391–410.
- Zhang, D., W. B. Wu, et al. (2017). Gaussian approximation for high dimensional time series. *The Annals of Statistics* 45(5), 1895–1919.
- Zhang, N. R., D. O. Siegmund, H. Ji, and J. Z. Li (2010). Detecting simultaneous changepoints in multiple sequences. *Biometrika* 97(3), 631–645. With supplementary data available online.

# SUPPLEMENTARY MATERIAL

## A Simulation

In this section, we conduct a simulation study to evaluate the accuracy of our method. The discrete version of the model can be written as:

$$y_{it} = u_{it} + \sum_{j=1}^{K_0} \gamma_{jit} \{t \geq \tau_j\} + \epsilon_{it}, \quad (42)$$

$i = 1, \dots, p, t = 1, \dots, n.$

We use cross validation to select the bandwidth and the block parameter. The detailed testing procedure is summarized as follows in line with the descriptions in Section 2.

**Step 1 (Long-run covariance estimation.)** We estimate the long-run covariance matrix  $\hat{\Sigma} = (\hat{\sigma}_{i,j})$  and its diagonal matrix  $\hat{\Lambda}$ . We first calculate  $\hat{\sigma}_{i,j,k}$  in (37) and we let  $\hat{\sigma}_{i,j}$  be the solution of  $h_{i,j}(u) = 0$  as in (38).

**Step 2 ( $Q$ -matrix relates to critical values.)** We construct the block matrix  $\hat{Q} = (\hat{Q}_{i,j})$ , where  $\hat{Q}_{i,j}$  is  $Q_{i,j}$  in (18), with  $\Sigma$  and  $\Lambda$  therein replaced by  $\hat{\Sigma}$  and  $\hat{\Lambda}$  respectively.

**Step 3 (Calculating critical values.)** We generate i.i.d. Gaussian vectors  $\hat{Z}^{(i)}$ ,  $i = 1, 2, \dots, N$ , with the covariance matrix  $\hat{Q}$ ; and we obtain  $\hat{q}_\alpha$  which is the empirical  $(1 - \alpha)$  quantile of the  $|\hat{Z}^{(i)}|_\infty$  over several samples and it can be viewed as an estimate of  $q_\alpha$  in (25).

**Step 4 (Testing the existence of jump.)** We construct  $\hat{T}_n$  as  $T_n$  in (9) with  $\Lambda$  replaced by  $\hat{\Lambda}$ . We reject the null hypothesis that there is no jump at level  $\alpha$  if  $\hat{T}_n$  is larger than  $\hat{q}_\alpha$ .

Step 5 (**Detecting significant break-points.**) Supposing that we reject the null in Step 4, we will continue with the following steps. To detect the significant jumps, we construct  $|\hat{V}_t|_\infty$  for  $t = bn + 1, bn + 2, \dots, n - bn$ , where  $\hat{V}_t$  is the same as  $V_t$  in (9) with  $\Lambda$  therein replaced by  $\hat{\Lambda}$ . Let  $\mathcal{A}_1 = \{\tau : |\hat{V}_\tau|_\infty > w^\dagger\}$ .  $w^\dagger$  can be set as  $\hat{q}_\alpha$  with  $\alpha$  to be small (e.g.  $\alpha = 0.0001$ ).

Step 6 (**Stamping multiple breaks**) In the case of multiple significant breaks in Step 5, we sequentially locate the multiple change-points following steps in Section 3.1. To be more specific, for  $k \geq 1$ , we let  $\hat{\tau}_k = \operatorname{argmax}_{\tau \in \mathcal{A}_k} |\hat{V}_\tau|_\infty$  and  $\mathcal{A}_{k+1} = \mathcal{A}_k \setminus \{\tau : |\tau - \hat{\tau}_k| \leq 2bn\}$ . Then the estimate of the number of breaks is  $\hat{K}_0 = \max_{k \geq 1} \{k : \mathcal{A}_k \neq \emptyset\}$ .

Step 7 (**Estimating the sizes of breaks**) We construct  $\hat{\gamma}_k$  as in Step 3 in Subsection 3.1. We set the estimates of the sizes of the jumps as  $\hat{\delta}_k = |\hat{\Lambda}^{-1} \hat{\gamma}_k|_\infty$  and their minimum as  $\hat{\delta}^\diamond = \min_{1 \leq k \leq \hat{K}_0} \hat{\delta}_k$ .

Step 8 (**Constructing confidence intervals for the sizes**) We construct  $\tilde{q}_\alpha$  as in (29).

Let

$\hat{\theta} = (\hat{\sigma}_{1,1}^{1/2}, \hat{\sigma}_{2,2}^{1/2}, \dots, \hat{\sigma}_{p,p}^{1/2})^\top$ . Then the confidence interval for vector  $\gamma_{k^*}$  at level  $2\alpha$  is  $(\hat{\gamma}_k - \tilde{q}_\alpha \hat{\theta}, \hat{\gamma}_k + \tilde{q}_\alpha \hat{\theta})$ .

Step 9 (**Aggregated jump location estimation and confidence interval construction**) Construct aggregated jump location estimates  $\tilde{\tau}_k$  as in (33). The confidence interval for  $\tau_k$  is  $(\tilde{\tau}_k - x, \tilde{\tau}_k + x)$ , where  $x$  is the  $1 - \alpha/2$  quantile of the distribution  $\operatorname{argmax}_r (-2^{-1} \hat{a}_k^2 |r| + \hat{\varsigma}_k \mathbb{W}(r))$  and  $\hat{a}_k$  (resp.  $\hat{\varsigma}_k$ ) is  $a_k$  (resp.  $\varsigma_k$ ) with  $\Lambda$ ,  $\Sigma$  and  $\gamma_k$  replaced by their estimations.

We first report a few results with a known variance-covariance matrix. We put it under rather simple settings for checking the performance of the algorithm with respect to different

$p = 20, 50, 100, \text{ and } 150, n = 100$ . Therefore, Step 2 of the above-mentioned algorithm is omitted. We also include the cases with strong cross-sectional dependence with factor structure and no cross sectional dependence. In particular, we consider different kinds of data generating processes. We choose a)  $f_i(u) = -i^2/p^2 + u^2$  and b)  $f_i(u) = \sin(2\pi u + i/p)$ . Let  $u_{it} = f_i(t/n)$ .  $\epsilon_t$  is taken to follow 1) an i.i.d. standard normal distribution, 2) a VAR(1) model, with a randomly simulated coefficient matrix (maximum eigenvalue smaller than 1) and 3) a factor structure together with a VAR(1) noise. In case 2), the factor loading and factors are generated with i.i.d.  $N(0, 1)$ . We set  $K_0 = 10$  breaks for all cases, and we increase the number of breaks in the cross-sectional dimension as the dimension increases, i.e.  $s = 1, 5, 20, 30$ . We set the break size to be  $3/\log(np)$ .

Figure 5 shows the simulated data with the model corresponding to the case a), 1), ii). We evaluate our simulation performance over 1000 samples. We report the averaged difference between the estimated number of breaks and the true break points (AD) ( $|\hat{K}_0 - K_0|$ ) as in Table 1. The averaged distances between the breaks  $\sum_{k=1}^{\hat{K}_0} |\tilde{\tau}_k - \tau_k^*|_\infty$  (AM) are shown in Table 2. We notice that as the dimension and the the number of breaks in the cross-sectional dimension grow, the estimation performance improves.

For the unknown covariance, we report results for the cases  $n = 500, 1000$  and  $p = 20, 30$ . The break-locations are selected to start at time-point 100 and are distanced by 100, and the break sizes are set to be either i)  $\gamma_{jit} = 0.05$  for  $i = 5, 10$  or ii)  $\gamma_{jit} = (\sqrt{t}/\log(pn))$ .

The estimation accuracy with covariance estimation is included in Tables 3 and 4. And the averaged coverage probabilities of the confidence interval for the breaks (AC) are in Table 5 at the confidence level of 90%. We do not consider  $p$  larger than  $n$ , as our long-run variance estimation has not been extended for the use of a high-dimensional setting. A further extension of our estimation to high-dimensional long -run variance estimation can be considered. As the sample sizes increase, the estimation precision is improved. We

Table 1: AD averaged over 1000 samples in different simulation scenarios, and their standard deviations in bracket.

	$p = 20, n = 100$		$p = 50, n = 100$	
	1)	3)	1)	3)
a)	0.152 (0.033)	0.187 (0.046)	0.121 (0.024)	0.125 (0.029)
b)	0.155 (0.041)	0.193 (0.054)	0.119 (0.025)	0.126 (0.028)
	$p = 100, n = 100$		$p = 150, n = 100$	
	1)	3)	1)	3)
a)	0.098 (0.015)	0.117 (0.021)	0.072 (0.013)	0.078 (0.014)
b)	0.093 (0.017)	0.124 (0.025)	0.084 (0.016)	0.090 (0.019)

Table 2: AM/n averaged over 1000 samples in different simulation scenarios, and their standard deviations in bracket.

	$p = 20, n = 100$		$p = 50, n = 100$	
	1)	3)	1)	3)
a)	0.046 (0.016)	0.057 (0.025)	0.033 (0.012)	0.041 (0.015)
b)	0.048 (0.019)	0.060 (0.028)	0.036 (0.014)	0.040 (0.018)
	$p = 100, n = 100$		$p = 150, n = 100$	
	1)	3)	1)	3)
a)	0.015 (0.007)	0.023 (0.009)	0.011 (0.003)	0.017 (0.008)
b)	0.019 (0.008)	0.028 (0.011)	0.012 (0.003)	0.020 (0.007)

Table 3: AD averaged over 1000 samples in different simulation scenarios, and their standard deviations in bracket.

2*		$p = 20, n = 500$		$p = 30, n = 1000$	
		1)	2)	1)	2)
2*a)	i)	0.035 (0.010)	0.046 (0.014)	0.029 (0.005)	0.030 (0.007)
	ii)	0.028 (0.011)	0.043 (0.015)	0.024 (0.004)	0.027 (0.006)
2*b)	i)	0.038 (0.012)	0.039 (0.012)	0.023 (0.004)	0.029 (0.003)
	ii)	0.023 (0.008)	0.032 (0.011)	0.026 (0.003)	0.028 (0.002)

can see that our method is robust against different data simulation scenarios, and we can achieve good level of accuracy with our method. In particular, the spatial and temporal dependency in the error term would not affect our estimation.

Figure 6 shows the plot of the estimated robust long-run covariance matrix (right) against the true one (left). On an overall level, we see that the true correlation matrix has been precisely recovered, as the patterns of these two plots look the same. We also report the distance between our robustly estimated variance-covariance matrix and the true one in Table 6. The estimation precision of the long-run variance-covariance matrix is maintained across different data-generating processes.

Table 4: AM/n averaged over 1000 samples in different simulation scenarios, and their standard deviations are in brackets.

$2^*$		$p = 20, n = 500$		$p = 30, n = 1000$	
		1)	2)	1)	2)
$2^*a)$	i)	0.044 (0.018)	0.056 (0.015)	0.033 (0.011)	0.038 (0.009)
	ii)	0.027 (0.014)	0.033 (0.013)	0.021 (0.008)	0.026 (0.007)
$2^*b)$	i)	0.045 (0.017)	0.057 (0.018)	0.028 (0.004)	0.035 (0.006)
	ii)	0.039 (0.012)	0.037 (0.014)	0.016 (0.003)	0.023 (0.004)

Table 5: AC in different simulation scenarios over all the estimated break-points and samples.

$2^*$		$p = 20, n = 500$		$p = 30, n = 1000$	
		1)	2)	1)	2)
$2^*a)$	i)	0.692	0.676	0.833	0.824
	ii)	0.719	0.708	0.856	0.833
$2^*b)$	i)	0.685	0.673	0.847	0.810
	ii)	0.741	0.715	0.889	0.872

Table 6: Averaged difference between the variance-covariance and the true matrix. ( $L_1$  norm divided by  $p(p-1)/2$ ).

$2^*$		$p = 20, T = 500$		$p = 30, T = 1000$	
		1)	2)	1)	2)
$2^*a)$	i)	0.005	0.008	0.003	0.007
	ii)	0.004	0.006	0.003	0.005
$2^*b)$	i)	0.006	0.009	0.003	0.005
	ii)	0.004	0.007	0.002	0.004

## B Proof

### B.1 Some useful Lemmas

LEMMA 1 (Basic properties of the weights). *We assume Assumption 2.2. We define  $\kappa_i = \int_0^1 x^i K(x) dx$  with finite  $\kappa_1, \kappa_2, \kappa_1^2 \neq \kappa_2 \kappa_0$ . Then by Fan and Gijbels (1996), the weights of the local linear estimator take the following form*

$$w_i = \frac{\kappa_2 - \kappa_1 i / (bn)}{\kappa_2 \kappa_0 - \kappa_1^2} \frac{K(i/(bn))}{bn} + O((bn)^{-2}).$$

*We have the following results which holds uniformly over  $i$ . There exist strictly positive constants  $c_w, c'_w, c''_w$  only depending on kernel  $K(\cdot)$ , such that*

$$\begin{aligned} bn \max_{0 \leq i \leq n} |w_i| &\leq c_w, \quad \max_{|i-j| \leq m} |w_i - w_j| \leq c_w \frac{m}{(bn)^2}, \\ (bn \sum_{i=0}^{bn} w_i^2)^{1/2} &\rightarrow c'_w, \quad \text{and} \quad \frac{bn}{k} \sum_{i=1}^k w_i \geq c''_w, k \leq bn. \end{aligned} \tag{43}$$

*Proof.* We only show the last one, since the rest are similar and easier. Note

$$bn \sum_{i=1}^k w_i/k = F(k/(bn)) + O((bn)^{-1}), \quad \text{where} \quad F(t) = \frac{\kappa_2 \int_0^t K(x)dx - \kappa_1 \int_0^t xK(x)dx}{(\kappa_2\kappa_0 - \kappa_1^2)t}.$$

Define the numerator function as  $g(t) = \kappa_2 \int_0^t K(x)dx - \kappa_1 \int_0^t xK(x)dx$ . We can see that  $g(0) = 0$ ,  $g(1) > 0$ , and the derivative function  $g'(t) = (\kappa_2 - \kappa_1 t)K(t)$ , which is strictly larger than 0 before  $\kappa_2/\kappa_1$  and less than 0 afterwards. Therefore we have  $F(x) > 0$  on  $(0, 1]$ . In addition, we note  $F(0+) = \kappa_2 K(0)/(\kappa_2\kappa_0 - \kappa_1^2) > 0$  and  $F(1) = 1$ . Thus  $\inf_{t \in (0,1]} F(t) > 0$  in view of  $F(t)$  is a continuous function.  $\square$

LEMMA 2 (Burkholder (1988), Rio (2009)). *Let  $q > 1$ ,  $q' = \min\{q, 2\}$ . Let  $M_T = \sum_{t=1}^T \xi_t$ , where  $\xi_t \in \mathcal{L}^q$  are martingale differences. Then*

$$\|M_T\|_q^{q'} \leq K_q^{q'} \sum_{t=1}^T \|\xi_t\|_q^{q'}, \quad \text{where } K_q = \max((q-1)^{-1}, \sqrt{q-1}).$$

## B.2 Asymptotic results for Gaussian vector

LEMMA 3 (Comparison). *Let  $X = (X_1, X_2, \dots, X_v)^\top$  and  $Y = (Y_1, Y_2, \dots, Y_v)^\top$  be two centered Gaussian vectors in  $\mathbb{R}^v$  and let  $d = (d_1, d_2, \dots, d_v)^\top \in \mathbb{R}^v$ . We denote  $\Delta = \max_{1 \leq i, j \leq v} |\sigma_{i,j}^X - \sigma_{i,j}^Y|$ , where we define  $\sigma_{i,j}^X = \mathbb{E}(X_i X_j)$  (resp.  $\sigma_{i,j}^Y = \mathbb{E}(Y_i Y_j)$ ). Assume that  $Y_i$ s have the same variance  $\sigma^2 > 0$ . Then we have*

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}(|X + d|_\infty \leq x) - \mathbb{P}(|Y + d|_\infty \leq x) \right| \lesssim \Delta^{1/3} \log(v)^{2/3}, \quad (44)$$

where the constant involved in  $\lesssim$  only depends on  $\sigma$ .

*Proof.* It suffices to show for any  $d \in \mathbb{R}^v$ ,

$$\sup_x \left| \mathbb{P}(\max_{1 \leq i \leq v} (X_i + d_i) \leq x) - \mathbb{P}(\max_{1 \leq i \leq v} (Y_i + d_i) \leq x) \right| \lesssim \Delta^{1/3} \log(v)^{2/3}.$$

To this end, we define

$$F_\beta^*(z) = \beta^{-1} \log \left( \sum_{j=1}^v \exp(\beta(z_j + d_j)) \right).$$

Replace the  $F_\beta(\cdot)$  in the proof of Theorem 2 in Chernozhukov et al. (2015) by  $F_\beta^*(z)$ . Then by the argument in equation (10) in Chernozhukov et al. (2015), we have

$$\mathbb{P} \left( \max_{1 \leq i \leq v} (X_i + d_i) \leq x \right) \leq \mathbb{P} \left( \max_{1 \leq i \leq v} (Y_i + d_i) \leq x + \delta + \beta^{-1} \log(v) \right) + c(\delta^{-2} + \beta \delta^{-1}) \Delta,$$

where  $c$  is some absolute constant. Then by Lemma 4, we have

$$\mathbb{P} \left( \max_{1 \leq i \leq v} (X_i + d_i) \leq x \right) - \mathbb{P} \left( \max_{1 \leq i \leq v} (Y_i + d_i) \leq x \right) \lesssim (\delta + \beta^{-1} \log(v)) \sqrt{\log(v)} + (\delta^{-2} + \beta \delta^{-1}) \Delta,$$

where the constant in  $\lesssim$  only depending on  $\sigma$ . Take  $\beta = \delta^{-1} \log(v)$  and  $\delta = \log(v)^{1/6} \Delta^{1/3}$ . Same argument can be applied in the other direction, and the desired result follows.  $\square$

LEMMA 4 (Nazarov (2003)). *Let  $X = (X_1, X_2, \dots, X_v)^\top$  be a centered Gaussian vector in  $\mathbb{R}^v$ . Assume  $\mathbb{E}(X_i^2) \geq b$  for some  $b > 0$  and all  $1 \leq i \leq v$ . Then for any  $e > 0$  and  $d \in \mathbb{R}^v$ ,*

$$\sup_{x \in \mathbb{R}} \mathbb{P} \left( \left| \|X + d\|_\infty - x \right| \leq e \right) \leq ce \sqrt{\log(v)}, \quad (45)$$

where  $c$  is some constant depending only on  $b$ .

### B.3 Proof of Theorem 1

The proof of Theorem 1 is quite involved. We shall first provide some intuitive ideas of the proof strategy. We define

$$I_\epsilon := \max_{bn+1 \leq i \leq n-bn} \left| \sum_{t=i-bn}^{i-1} w_{i-t} \Lambda^{-1} \epsilon_t - \sum_{t=i+1}^{i+bn} w_{t-i} \Lambda^{-1} \epsilon_t + d_i \right|_\infty. \quad (46)$$

By (13) and (14) we have

$$|T_n - I_\epsilon| \leq \max_{bn+1 \leq i \leq n-bn} |\mathbb{E}V_i - d_i|_\infty = O(b^2). \quad (47)$$

Thus we only need to work on  $I_\epsilon$ . For some  $m > 0$ , let a truncated version of the error term be defined as

$$\epsilon_{t,m} = \sum_{k=0}^{m-1} A_k \eta_{t-k}.$$

Consider the  $m$ -dependent approximation  $I_{\epsilon,m}$  of  $I_\epsilon$ , where  $I_{\epsilon,m}$  is  $I_\epsilon$  with  $\epsilon_t$  replaced by  $\epsilon_{t,m}$ . Then we have  $I_\epsilon \approx I_{\epsilon,m}$  for large  $m$ . Let  $I_{z,m}$  be  $I_{\epsilon,m}$  with  $\eta_t$  therein replaced by  $z_t$ , where  $(z_t, t \in \mathbb{Z})$  are i.i.d. Gaussian vectors with zero mean and identity covariance matrix in  $\mathbb{R}^{\bar{p}}$ . Since  $I_{\epsilon,m}$  can be rewritten into the format of the maximum of summation of independent vectors, by the Gaussian approximation theorem in Chernozhukov et al. (2017), the distributions of  $I_{\epsilon,m}$  and  $I_{z,m}$  are close. We complete the proof by showing that the distributions of  $I_{z,m}$  and  $|Z + \underline{d}|_\infty$  are close, and the continuity of the maximum of a non-centered Gaussian distribution.

*Proof.* We now proceed with the formal argument. We shall first focus on case (i). Let  $m = (bn)^{1/(\beta+1)}$ , for any  $\alpha > 0$ ,

$$\mathbb{P}((bn)^{1/2}T_n \leq u) \leq \mathbb{P}((bn)^{1/2}|T_n - I_{\epsilon,m}| \geq \alpha) + \mathbb{P}((bn)^{1/2}I_{\epsilon,m} \leq u + \alpha)$$

and

$$\mathbb{P}((bn)^{1/2}|Z + \underline{d}|_\infty \leq u) = \mathbb{P}((bn)^{1/2}|Z + \underline{d}|_\infty \leq u + \alpha) - \mathbb{P}(u < (bn)^{1/2}|Z + \underline{d}|_\infty \leq u + \alpha).$$

Hence

$$\begin{aligned} & \sup_{u \in \mathbb{R}} \left[ \mathbb{P}((bn)^{1/2}T_n \leq u) - \mathbb{P}((bn)^{1/2}|Z + \underline{d}|_\infty \leq u) \right] \\ & \leq \mathbb{P}\left((bn)^{1/2}|T_n - I_{\epsilon,m}| \geq \alpha\right) + \sup_{u \in \mathbb{R}} \left| \mathbb{P}(I_{\epsilon,m} \leq u) - \mathbb{P}(|Z + \underline{d}|_\infty \leq u) \right| \\ & \quad + \sup_{u \in \mathbb{R}} \mathbb{P}\left(|(bn)^{1/2}|Z + \underline{d}|_\infty - u| \leq \alpha\right) =: I_1 + I_2 + I_3. \end{aligned}$$

For the  $I_1$  part,  $|T_n - I_{\epsilon,m}| \leq |T_n - I_\epsilon| + |I_\epsilon - I_{\epsilon,m}|$ . Recall (47), then  $|T_n - I_\epsilon| \leq c_0 b^2$  for some constant  $c_0 > 0$ . We define  $\alpha' = 2c_1 \max\{(bn)^{-1/2}(np)^{1/q}, 1\} m^{-\beta+\beta/q}$ , where the constant  $c_1$  is the one to be defined in Lemma 5. Then by Lemma 5, we have

$$\mathbb{P}((bn)^{1/2}|I_\epsilon - I_{\epsilon,m}| \geq \alpha') = o(1).$$

Hence for  $\alpha = \alpha' + c_0(bn)^{1/2}b^2$ ,  $I_1 = o(1)$ .

For the  $I_2$  part, we note that

$$I_2 \leq \sup_{u \in \mathbb{R}} |\mathbb{P}(I_{\epsilon,m} \leq u) - \mathbb{P}(I_{z,m} \leq u)| + \sup_{u \in \mathbb{R}} |\mathbb{P}(I_{z,m} \leq u) - \mathbb{P}(|Z + \underline{d}|_\infty \leq u)| =: I_{21} + I_{22}.$$

By Lemma 7 (1), we have  $I_{21} = o(1)$ . By Lemma 8,  $I_{22} = o(1)$ . Hence  $I_2 = o(1)$ .

For the  $I_3$  part, the diagonal entities in  $bnQ$  take the same value i.e.  $\sigma^{\diamond 2} = 2bn \sum_{i=1}^{bn} w_i^2$ , which by (43), converges to  $2c'_w{}^2 > 0$ , where  $c'_w$  is a finite constant. By Lemma 4

$$I_3 \lesssim \alpha \log(np)^{1/2} = o(1).$$

The desired result follows by combining the  $I_1$ - $I_3$  parts and a similar argument for the other side of the inequality.

For case (ii), we have the same decomposition  $I_1$ - $I_3$ . For the  $I_1$  part, we define  $\alpha = c_1 \log(np)^{1/2} m^{-\beta} + c_0(bn)^{1/2}b^2$ , for some constant  $c_1 > 0$ . Then by Lemma 6,  $I_1 = o(1)$ . For  $I_2$  part, by Lemma 7 (2) and Lemma 8, we have  $I_2 = o(1)$ . For  $I_3$ , same argument can be applied. Combining the rates of  $I_1$ - $I_3$ , we obtain the desired result.  $\square$

Lemma 5 and 6 give us concentration inequalities for the  $m$ -dependent approximation of  $I_\epsilon$ .

LEMMA 5 ( $m$ -dependent approximation for polynomial case). *Assume conditions in Theorem 1 (i). For some  $m > 0$  and  $u > 0$ , we have*

$$\mathbb{P}((bn)^{1/2}|I_\epsilon - I_{\epsilon,m}| \geq u + c_1(bn)^{-1/2}(np)^{1/q}m^{-\beta}) \leq c_2(e^{-c_3u^2m^{2\beta}} + npm^{-q\beta}(bn)^{-q/2}u^{-q}),$$

where  $c_1, c_2, c_3$  are some positive constants only depending on  $q, c_p, c_w, c_s, \mu_q$ .

*Proof.* We note that  $I_\epsilon - I_{\epsilon,m}$  can be bounded by

$$|I_\epsilon - I_{\epsilon,m}| \leq \max_{bn+1 \leq i \leq n-bn} \left( \left| \sum_{t=i-bn}^{i-1} w_{i-t} \Lambda^{-1}(\epsilon_t - \epsilon_{t,m}) \right|_\infty + \left| \sum_{t=i+1}^{i+bn} w_{t-i} \Lambda^{-1}(\epsilon_t - \epsilon_{t,m}) \right|_\infty \right) =: I_1 + I_2.$$

We let  $E_{i,l} = \sum_{t=(i-bn) \vee (l+m)}^{i-1} w_{i-t} \Lambda^{-1} A_{t-l}$ , then  $I_1$  can be rewritten into

$$I_1 = \max_{\substack{bn+1 \leq i \leq n-bn \\ 1 \leq j_1 \leq p}} \left| \sum_{\substack{l \leq i-m-1 \\ 1 \leq j_2 \leq \tilde{p}}} E_{i,l,j_1,j_2} \eta_{l,j_2} \right|, \quad (48)$$

where  $E_{i,l,j_1,j_2}$  is the  $(j_1, j_2)$ th entity of matrix  $E_{i,l}$  and  $\eta_{l,j_2}$  is the  $j_2$ th entity of  $\eta_l$ . Since  $\eta_{l,j_2}$ s are independent for different  $(l, j_2)$ , by Lemma A.2 in Chernozhukov et al. (2013b), for  $u > 0$ ,

$$\mathbb{P}(\sqrt{bn}I_1 \geq 2\sqrt{bn}\mathbb{E}I_1 + u) \leq e^{-u^2/(3\sigma^2)} + K_q u^{-q} H_q, \quad (49)$$

where  $K_q$  is some constant only depending on  $q$ ,

$$\sigma^2 = bn \max_{\substack{bn+1 \leq i \leq n-bn \\ 1 \leq j_1 \leq p}} \sum_{\substack{l \leq i-m-1 \\ 1 \leq j_2 \leq \tilde{p}}} \mathbb{E}(E_{i,l,j_1,j_2} \eta_{l,j_2})^2,$$

and

$$H_q = (bn)^{q/2} \sum_{\substack{l \leq n-bn-m-1 \\ 1 \leq j_2 \leq \tilde{p}}} \mathbb{E} \left( \max_{\substack{(bn+1) \vee (l+m+1) \leq i \leq n-bn \\ 1 \leq j_1 \leq p}} |E_{i,l,j_1,j_2} \eta_{l,j_2}|^q \right).$$

Then we start to analyze the rates of the objects involved in (49). We define  $E_{i,l,j_1,\cdot}$  to be the  $j_1$ th row of  $E_{i,l}$ . For the  $\sigma^2$  part, by Assumption 2.5 and (43),

$$|E_{i,l,j_1,\cdot}|_2 \leq \sum_{t \geq l+m} w_{i-t} \sigma_{j_1,j_1}^{-1/2} |A_{t-l,j_1,\cdot}|_2 \leq c_w c_s m^{-\beta} / (bn), \quad (50)$$

and therefore

$$\sum_{l \leq n-bn-m-1} |E_{i,l,j_1,\cdot}|_2 \leq \sum_{t=i-bn}^{i-1} \sum_{l \leq t-m} w_{i-t} \sigma_{j_1,j_1}^{-1/2} |A_{t-l,j_1,\cdot}|_2 \leq c_w c_s m^{-\beta}. \quad (51)$$

Combining the above arguments and recall that  $\mathbb{E} \eta_{i,j}^2 = 1$ , we have

$$\sigma^2 \leq bn \max_{\substack{bn+1 \leq i \leq n-bn \\ 1 \leq j_1 \leq p}} \left( \sum_{l \leq i-m-1} |E_{i,l,j_1,\cdot}|_2 \max_{l \leq n-bn-m-1} |E_{i,l,j_1,\cdot}|_2 \right) \leq (c_w c_s)^2 m^{-2\beta}. \quad (52)$$

For the  $H_q$  part, by Assumption 2.5 and (43),  $\max_{i,j_1,j_2} |E_{i,l,j_1,j_2}| \leq c_w c_s ((1-l) \vee m)^{-\beta} / (bn)$ . Recall that  $\tilde{p} \leq c_p p$ . Then we have

$$\begin{aligned} H_q &\leq (bn)^{q/2} \sum_{\substack{l \leq n-bn-m-1 \\ 1 \leq j_2 \leq \tilde{p}}} [c_w c_s ((1-l) \vee m)^{-\beta} / (bn)]^q \mu_q^q \\ &\leq (c_w c_s)^q \mu_q^q (bn)^{-q/2} \tilde{p} \left( \sum_{-m \leq l \leq n-bn-m} m^{-\beta q} + \sum_{l < -m} (1-l)^{-\beta q} \right) \leq c_0 (bn)^{-q/2} n p m^{-\beta q}, \end{aligned} \quad (53)$$

where  $c_0 = 3c_p (c_w c_s)^q \mu_q^q$ .

For  $\mathbb{E} I_1$  part, note that  $\mathbb{E} I_1 \leq \|I_1\|_q$ . By Lemma 2, we have

$$\mathbb{E} I_1 \leq \left( \sum_{i,j_1} \mathbb{E} \left( \left| \sum_{l,j_2} E_{i,l,j_1,j_2} \eta_{l,j_2} \right|^q \right) \right)^{1/q} \leq \left( \sum_{i,j_1} \left( (q-1) \sum_l |E_{i,l,j_1,\cdot}|_2^2 \mu_q^2 \right)^{q/2} \right)^{1/q}.$$

Thus by (50) and (51) we have

$$\mathbb{E}I_1 \lesssim (bn)^{-1/2}m^{-\beta}(np)^{1/q}, \quad (54)$$

where the constant in  $\lesssim$  only depends on  $c_w, c_s, \mu_q, q$ . Our conclusions follows by applying (52), (53) and (54) into (49) and a similar argument for  $I_2$ .  $\square$

LEMMA 6 (m-dependent approximation for exponential case). *We assume conditions in Theorem 1 (ii). We have*

$$\mathbb{P}((bn)^{1/2}|I_\epsilon - I_{\epsilon,m}| \geq u) \leq \begin{cases} 2npe^{-a_1m^{2\beta}u^2}, & \text{if } u < a_2(bn)^{1/2}m^{-\beta}, \\ 2npe^{-a_3m^\beta(bn)^{1/2}u}, & \text{if } u \geq a_2(bn)^{1/2}m^{-\beta}, \end{cases}$$

where  $a_1, a_2, a_3$  are some positive constants only depending on  $a_0, c_w, c_s, \mu_e$ .

*Proof.* Recall the definition of  $I_1$  and  $I_2$  in the proof of Lemma 5. Let  $e^* = c_w c_s m^{-\beta} / (bn)$  and  $c^* = a_0 / e^*$ . Then by (50),  $\mathbb{E}e^{cE_{i,l,j_1,j_2}\eta_{l,j_2}} < \infty$ , for any  $0 < c \leq c^*$ , and we have

$$\begin{aligned} \mathbb{E}(e^{cI_1}) &\leq \sum_{\substack{bn+1 \leq i \leq n-bn \\ 1 \leq j_1 \leq p}} \mathbb{E} \left( \exp \left\{ c \sum_{\substack{l \leq n-bn-m-1 \\ 1 \leq j_2 \leq \bar{p}}} E_{i,l,j_1,j_2} \eta_{l,j_2} \right\} + \exp \left\{ -c \sum_{\substack{l \leq n-bn-m-1 \\ 1 \leq j_2 \leq \bar{p}}} E_{i,l,j_1,j_2} \eta_{l,j_2} \right\} \right) \\ &=: I_{11} + I_{12}. \end{aligned}$$

Since  $\mathbb{E}\eta_{i,j} = 0$ , for  $E_{i,l,j_1,j_2} \neq 0$ , we have

$$\begin{aligned} \mathbb{E}(e^{cE_{i,l,j_1,j_2}\eta_{l,j_2}}) &= 1 + \frac{\mathbb{E}(e^{cE_{i,l,j_1,j_2}\eta_{l,j_2}} - 1 - cE_{i,l,j_1,j_2}\eta_{l,j_2})}{c^2 E_{i,l,j_1,j_2}^2} c^2 E_{i,l,j_1,j_2}^2 \\ &\leq 1 + \frac{\mathbb{E}(e^{c^*e^*|\eta_{l,j_2}|} - 1 - c^*e^*|\eta_{l,j_2}|)}{(c^*e^*)^2} c^2 E_{i,l,j_1,j_2}^2 \\ &\leq 1 + \frac{\mu_e}{a_0^2} c^2 E_{i,l,j_1,j_2}^2, \end{aligned}$$

where the first inequality is because that for any  $x > 0$ , the function  $g_x(t) = (e^{tx} - 1 - tx)/t^2$  increases on  $t \in (0, \infty)$ , and  $e^t - t \leq e^{|t|} - |t|$ . We define  $c' = \mu_e/a_0^2$ , and the rate of  $I_{11}$  is derived as follows,

$$\begin{aligned} I_{11} &\leq \sum_{i,j_1} \prod_{l,j_2} (1 + c'c^2 E_{i,l,j_1,j_2}^2) \leq \sum_{\substack{bn+1 \leq i \leq n-bn \\ 1 \leq j_1 \leq p}} \exp\left\{c'c^2 \sum_{\substack{l \leq n-bn-m-1 \\ 1 \leq j_2 \leq \bar{p}}} E_{i,l,j_1,j_2}^2\right\}, \\ &\leq np \exp\{c^2 c_1 m^{-2\beta}/(bn)\}, \end{aligned}$$

where  $c_1 = c'c_w^2 c_s^2$ , the second inequality is due to  $1 + x \leq e^x$  for any  $x \geq 0$ , and the last inequality is by (52). Same bound can be derived for  $I_{12}$ . We note that

$$\mathbb{P}(I_1 \geq u) \leq e^{-cu} \mathbb{E}(e^{cI_1}) \leq e^{-cu} (I_{11} + I_{12}).$$

We define  $c^\diamond = bnm^{2\beta}u/(2c_1)$ . Hence if  $c^\diamond < c^*$ , then  $\mathbb{P}(I_1 \geq u) \leq 2npe^{-u^2 m^{2\beta} bn/(4c_1)}$ ; if  $c^\diamond \geq c^*$ , then  $\mathbb{P}(I_1 \geq u) \leq 2npe^{\mu_e bn - a_0/(c_w c_s) bnm^\beta u}$ . The proof for  $I_2$  is similar and therefore omitted.  $\square$

LEMMA 7. *Let  $m \rightarrow \infty$  and  $m/(bn) \rightarrow 0$ .*

(1) *Assume conditions in Theorem 1 (i), we have*

$$\begin{aligned} &\sup_{u \in \mathbb{R}} |\mathbb{P}(I_{\epsilon,m} \leq u) - \mathbb{P}(I_{z,m} \leq u)| \\ &\lesssim (bn)^{-1/6} \log^{7/6}(pn) + ((np)^{2/q}/(bn))^{1/3} \log(pn), \end{aligned}$$

where the constant in  $\lesssim$  only depends on  $c_w, c'_w, c_s$  and  $\mu_q$ .

(2) *Assume conditions in Theorem 1 (ii), we have*

$$\sup_{u \in \mathbb{R}} |\mathbb{P}(I_{\epsilon,m} \leq u) - \mathbb{P}(I_{z,m} \leq u)| \lesssim (bn)^{-1/6} \log(pn)^{7/6},$$

where the constant in  $\lesssim$  only depends on  $c_w, c'_w, c_s, a_0$  and  $\mu_e$ .

*Proof.* First we consider the case of (1). We denote

$$D_{i,l} = \sum_{t=(i-bn)\vee l}^{(i-1)\wedge(m+l-1)} w_{i-t}\Lambda^{-1}A_{t-l}, \quad \text{and} \quad D_{i,l}^* = \sum_{t=(i+1)\vee l}^{(i+bn)\wedge(m+l-1)} w_{t-i}\Lambda^{-1}A_{t-l}. \quad (55)$$

Then  $I_{\epsilon,m}$  can be rewritten into

$$I_{\epsilon,m} = \max_{bn+1 \leq i \leq n-bn} \left| \sum_{i-m+1-bn \leq l \leq i-1} D_{i,l}\eta_l - \sum_{i-m+2 \leq l \leq i+bn} D_{i,l}^*\eta_l + d_i \right|_{\infty}.$$

Let  $N_0 = (n - 2bn)p$  and  $N_1 = (n + m - 1)\tilde{p}$ . Let  $G = (G_{i,l})_{i,l}$ ,  $bn + 1 \leq i \leq n - bn$ ,  $2 - m \leq l \leq n$ , be a block matrix in  $\mathbb{R}^{N_0 \times N_1}$  with

$$G_{i,l} = \begin{cases} D_{i,l} & \text{if } i - m + 1 - bn \leq l \leq i - m + 1, \\ D_{i,l} - D_{i,l}^* & \text{if } i - m + 2 \leq l \leq i - 1, \\ -D_{i,l}^* & \text{if } i \leq l \leq i + bn, \end{cases} \quad (56)$$

and elsewhere zero. We define  $d_{i,j_1}$  to be the  $j_1$ th entity of  $d_i$ ,  $N_2 = bnN_1$  and  $G_{i,l,j_1,j_2}$  be the  $(j_1, j_2)$ th entity of  $G_{i,l}$ . Then

$$N_2^{1/2}I_{\epsilon,m} = \max_{\substack{bn+1 \leq i \leq n-bn \\ 1 \leq j_1 \leq p}} \left| \sum_{\substack{2-m \leq l \leq n \\ 1 \leq j_2 \leq \tilde{p}}} g_{i,l,j_1,j_2} + N_2^{1/2}d_{i,j_1} \right|, \quad \text{where } g_{i,l,j_1,j_2} = N_2^{1/2}G_{i,l,j_1,j_2}\eta_{j_2}. \quad (57)$$

For any  $r \leq q$ , we denote

$$M_r := \max_{\substack{bn+1 \leq i \leq n-bn \\ 1 \leq j_1 \leq p}} \theta_{ij_1,r}, \quad \text{where} \quad \theta_{ij_1,r}^r := \sum_{\substack{2-m \leq l \leq n \\ 1 \leq j_2 \leq \tilde{p}}} \mathbb{E}|g_{i,l,j_1,j_2}|^r / N_1 = \sum_{l=2-m}^n |G_{i,l,j_1,\cdot}|_r^r \mu_r^r N_2^{r/2} / N_1.$$

By Assumption 2.5 and (43), for any  $r \geq 2$ ,

$$|D_{i,l,j_1,\cdot}|_r \leq |D_{i,l,j_1,\cdot}|_2 \leq c_w c_s / (bn), \quad \text{and similarly } |D_{i,l,j_1,\cdot}^*|_r \leq c_w c_s / (bn). \quad (58)$$

Then by (56),  $\max_{i,l,j_1} |G_{i,l,j_1,\cdot}|_r \leq 2c_w c_s / (bn)$ . Since  $G_{i,l}$  is zero for  $l < i - m + 1 - bn$  or  $l > i + bn$ ,

$$M_r \leq (4c_w c_s \mu_r)(N_1/bn)^{1/2-1/r}. \quad (59)$$

Especially, for  $r = 2$ ,  $M_r \leq c_1$  where  $c_1 = 4c_w c_s \mu_2$ . By (55), for  $i - bn \leq l \leq i - m$ , and  $S_m = \sum_{k=0}^{m-1} A_k$ ,

$$G_{i,l} = D_{i,l} = \sum_{t=l}^{m+l-1} w_{i-t} \Lambda^{-1} A_{t-l} = w_{i-l} \Lambda^{-1} S + w_{i-l} \Lambda^{-1} (S_m - S) + \sum_{t=l}^{l+m-1} (w_{i-t} - w_{i-l}) \Lambda^{-1} A_{t-l}.$$

Therefore by Assumption 2.5 and (43), we can derive

$$\begin{aligned} \left| |G_{i,l,j_1,\cdot}|_2 - w_{i-l} \right| &= \sigma_{j_1,j_1}^{-1/2} w_{i-l} \sum_{k \geq m} |A_{k,j_1,\cdot}|_2 + \sigma_{j_1,j_1}^{-1/2} \sum_{t=l}^{l+m-1} |w_{i-t} - w_{i-l}| |A_{t-l,j_1,\cdot}|_2 \\ &\leq c_w c_s m^{-\beta} / (bn) + c_w c_s m / (bn)^2. \end{aligned}$$

Note  $\sum_{i=1}^m w_i^2 bn = O(m/(bn)) = o(1)$ . Thus by (43),

$$\min_{\substack{bn \leq i \leq n-bn \\ 1 \leq j_1 \leq p}} \theta_{ij_1,2} \geq \min_{\substack{bn \leq i \leq n-bn \\ 1 \leq j_1 \leq p}} \left( \sum_{l=i-bn}^{i-m} |G_{i,l,j_1,\cdot}|_2^2 \mu_2^2 N_2 / N_1 \right)^{1/2} \geq c'_w \mu_2 - o(1) \geq c_1, \quad (60)$$

some constant  $c_1 > 0$ . Since  $\max_{i,l,j_1,j_2} |G_{i,l,j_1,j_2}| \leq 2c_w c_s / (bn)$ ,

$$\max_{l,j_2} \mathbb{E}(\max_{i,j_1} |g_{i,l,j_1,j_2}|^q) = \max_{i,l,j_1,j_2} |G_{i,l,j_1,j_2} N_2^{1/2}|^q \leq (2c_w c_s)(N_1/(bn))^{q/2}.$$

Note

$$B_n := \max \left\{ M_3^3, M_4^2, \left( \max_{l,j_2} \mathbb{E}(\max_{i,j_1} |g_{i,l,j_1,j_2}|^q) \right)^{1/q} \right\} \lesssim (N_1/(bn))^{1/2},$$

where the constant in  $\lesssim$  only depends on  $c_w, c_s, \mu_q$ . By Proposition 2.1 in Chernozhukov et al. (2017) we have

$$\begin{aligned} & \sup_{u \in \mathbb{R}} |\mathbb{P}(N_2^{1/2} I_{\epsilon, m} \leq u) - \mathbb{P}(N_2^{1/2} I_{z, m} \leq u)| \\ & \lesssim (B_n^2 \log^7(pn)/N_1)^{1/6} + (B_n^2 \log^3(pn)/N_1^{1-2/q})^{1/3} \\ & \lesssim (bn)^{-1/6} \log^{7/6}(pn) + ((np)^{2/q}/(bn))^{1/3} \log(pn). \end{aligned}$$

For part (2), let  $M = \log_2(\mu_e) \vee 1$ , and  $B'_n = (2c_w c_s M/a_0)(N_1/(bn))^{1/2}$ . Since  $\max_{i, l, j_1, j_2} |G_{i, l, j_1, j_2}| \leq 2c_w c_s/(bn)$ , we have

$$\max_{i, l, j_1, j_2} \mathbb{E}(e^{g_{i, l, j_1, j_2}/B'_n}) \leq 2.$$

Note

$$B_n := \max\{M_3^3, M_4^2, B'_n\} \lesssim (N_1/(bn))^{1/2}.$$

Apply the same argument as for part (1) with this new  $B_n$ , then Proposition 2.1 in Chernozhukov et al. (2017) leads to

$$\sup_{u \in \mathbb{R}} |\mathbb{P}(N_2^{1/2} I_{\epsilon, m} \leq u) - \mathbb{P}(N_2^{1/2} I_{z, m} \leq u)| \lesssim (B_n^2 \log^7(pn)/N_1)^{1/6} \lesssim (bn)^{-1/6} \log^{7/6}(pn).$$

□

LEMMA 8. *Assume conditions in Theorem 1 (i) or (ii), for  $m \rightarrow \infty$ ,  $m/(bn) \rightarrow 0$ ,*

$$\sup_{u \in \mathbb{R}} |\mathbb{P}(I_{z, m} \leq u) - \mathbb{P}(|Z + \underline{d}|_\infty \leq u)| \lesssim (m/(bn) + m^{-\beta})^{1/3} \log(np)^{2/3},$$

where the constant in  $\lesssim$  only depends on  $c_w, c'_w$  and  $c_s$ .

*Proof.* We recall that  $D_{i,l}, D_{i,l}^*$  in (55),  $G = (G_{i,l})$  in (56) and  $G^\diamond$  in (15). It is not hard to see that the covariance matrix for  $I_{z,m}$  is  $GG^\top$  and the covariance matrix for  $Z$  is  $Q = G^\diamond G^{\diamond\top}$ . We let

$$H^0 = (G_{i,l})_{\substack{2-m \leq l \leq 0 \\ bn+1 \leq i \leq n-bn}}, \quad \text{and} \quad H^1 = (G_{i,l})_{\substack{1 \leq l \leq n \\ bn+1 \leq i \leq n-bn}}.$$

Then  $G = (H^0, H^1)$  and

$$\begin{aligned} |GG^\top - G^\diamond G^{\diamond\top}|_{\max} &\leq |H^0 H^{0\top}|_{\max} + 2|(H^1 - G^\diamond)G^{\diamond\top}|_{\max} + |(H^1 - G^\diamond)(H^1 - G^\diamond)^\top|_{\max} \\ &=: I_1 + I_2 + I_3. \end{aligned}$$

By (58),  $\max_{i,l,j} |G_{i,l,j,\cdot}|_2 \leq 2c_w c_s / (bn)$ . Therefore

$$(bn)I_1 \leq (bn) \max_{i_1, i_2, j_1, j_2} \sum_{l=2-m}^0 |G_{i_1, l, j_1, \cdot}|_2 |G_{i_2, l, j_2, \cdot}|_2 \leq (2c_w c_s)^2 m / (bn).$$

Denote  $\Delta_{i,l} = G_{i,l} - G_{i,l}^\diamond$ . For  $i - m + 1 - bn \leq l < i - bn$ ,  $\Delta_{i,l} = D_{i,l}$ , and thus  $|\Delta_{i,l,j,\cdot}|_2 \leq c_w c_s / (bn)$ . For  $i - bn \leq l \leq i - m + 1$ , we have

$$\Delta_{i,l} = D_{i,l} - w_{i-l} \Lambda^{-1} S = \sum_{t=l}^{m+l-1} (w_{i-t} - w_{i-l}) \Lambda^{-1} A_{t-l} - w_{i-l} \Lambda^{-1} \sum_{t \geq m} A_t. \quad (61)$$

Hence  $|\Delta_{i,l,j,\cdot}|_2 \leq c_w c_s m / (bn)^2 + c_w c_s m^{-\beta} / (bn)$ . For  $i - m + 1 \leq l \leq i - 1$ ,  $\Delta_{i,l} = D_{i,l} - D_{i,l}^* - w_{i-l} \Lambda^{-1} S$ . Then  $|\Delta_{i,l,j,\cdot}|_2 \leq 3c_w c_s / (bn)$ . Similarly we can bound  $|\Delta_{i,l,j,\cdot}|_2$  for  $i \leq l \leq i + bn$ . For the rest  $l$ ,  $\Delta_{i,l} = 0$ . We note that  $|G_{i,l,j,\cdot}^\diamond| \leq c_w c_s / (bn)$ . Consequently,

$$(bn)I_2 \leq (bn) \max_{i_1, i_2, j_1, j_2} \sum_{l=1}^n |\Delta_{i_1, l, j_1, \cdot}|_2 |G_{i_2, l, j_2, \cdot}^\diamond|_2 \lesssim m / (bn) + m^{-\beta},$$

where the constant in  $\lesssim$  only depends on  $c_w, c_s$ . Similarly we have  $(bn)I_3 \lesssim m / (bn) + m^{-\beta}$ .

Combining  $I_1$ - $I_3$ ,

$$(bn)|GG^\top - G^\diamond G^{\diamond\top}|_{\max} \lesssim m / (bn) + m^{-\beta}.$$

By (43), for any  $j$  we have  $bnQ_{j,j} = 2bn \sum_{i=1}^n w_i^2 \rightarrow 2c_w'^2$ . Then the desired result follows from Lemma 3.  $\square$

## B.4 Proof of Theorem 2

*Proof of (i).* Note  $1 - \Phi(x) \leq (2\pi)^{-1/2}x^{-1}e^{-x^2/2}$ , where  $\Phi(\cdot)$  is the cumulative distribution function of a standard normal distribution. Recall  $G_{i,l}^\diamond$  in (15). Let  $G_{i,\cdot}^\diamond = (G_{i,1}^\diamond, G_{i,2}^\diamond, \dots, G_{i,n}^\diamond)$  and  $\underline{z}$  be a Gaussian vector in  $\mathbb{R}^{np}$  with zero mean and identity covariance matrix. Let  $G_{i,\cdot,j}^\diamond$  be the  $j$ th row of  $G_{i,\cdot}^\diamond$ , then

$$\mathbb{P}(|G^\diamond \underline{z}|_\infty \geq u) \leq \sum_{i=bn}^{n-bn} \sum_{j=1}^p \mathbb{P}(|G_{i,\cdot,j}^\diamond \underline{z}| \geq u) \leq np(2\pi)^{-1/2}(\sigma/u)e^{-u^2/(2\sigma^2)}, \quad (62)$$

where  $\sigma = |G_{i,\cdot,j}^\diamond|_2 = (2 \sum_{t=1}^{bn} w_t^2)^{1/2}$ , which by (43) converges to  $(bn/2)^{-1/2}c'_w > 0$ . Thus

$$\mathbb{P}(|G^\diamond \underline{z}|_\infty \geq 2c'_w \log(np)^{1/2} (bn)^{-1/2}) \rightarrow 0. \quad (63)$$

Let  $\mathcal{S} := \{1 \leq i \leq n : |i - \tau_k| > bn, \text{ for all } 1 \leq k \leq K_0\}$ . For any  $i \in \mathcal{S}$ ,  $d_i = 0$ . Hence by Theorem 1,

$$\sup_{u \in \mathbb{R}} \left| \mathbb{P}\left(\max_{i \in \mathcal{S}} |V_i|_\infty \geq u\right) - \mathbb{P}\left(\max_{i \in \mathcal{S}} |G_{i,\cdot}^\diamond \underline{z}|_\infty \geq u\right) \right| \rightarrow 0. \quad (64)$$

Since  $\max_{i \in \mathcal{S}} |G_{i,\cdot}^\diamond \underline{z}|_\infty \leq |G^\diamond \underline{z}|_\infty$ , by (63) and (64) we have  $\mathbb{P}(\max_{i \in \mathcal{S}} |V_i|_\infty \geq \omega^\dagger) \rightarrow 0$ . Thus we obtain

$$\lim_{n \rightarrow \infty} \mathbb{P}(\forall t \in \mathcal{A}_1, \exists 1 \leq k \leq K_0, |t - \tau_k| \leq bn) = 1. \quad (65)$$

Recall that  $d_{\tau_k} = \Lambda^{-1}\gamma_k$ . Since  $|d_{\tau_k} + G_{\tau_k,\cdot}^\diamond \underline{z}|_\infty \geq |d_{\tau_k}|_\infty - |G_{\tau_k,\cdot}^\diamond \underline{z}|_\infty$ , we have

$$\begin{aligned} \mathbb{P}\left(\min_{1 \leq k \leq K_0} |d_{\tau_k} + G_{\tau_k,\cdot}^\diamond \underline{z}|_\infty \leq \omega^\dagger\right) &\leq \mathbb{P}\left(\max_{1 \leq k \leq K_0} |G_{\tau_k,\cdot}^\diamond \underline{z}|_\infty \geq \min_{1 \leq k \leq K_0} |d_{\tau_k}|_\infty - \omega^\dagger\right) \\ &\leq \mathbb{P}(|G^\diamond \underline{z}|_\infty \geq \delta^\diamond - \omega^\dagger). \end{aligned}$$

Since  $\delta^\diamond \geq 2\omega^\dagger$ ,  $\mathbb{P}(\min_{1 \leq k \leq K_0} |d_{\tau_k} + G_{\tau_k,\cdot}^\diamond \underline{z}|_\infty \leq \omega^\dagger) \rightarrow 0$ . Subsequently the break statistics will be bigger than the threshold at the points of break with probability approach 1,

$\mathbb{P}(\min_{1 \leq k \leq K_0} |V_{\tau_k}|_\infty \leq \omega^\dagger) \rightarrow 0$  in view of

$$\sup_{u \in \mathbb{R}} \left| \mathbb{P}(|V_{\tau_k}|_\infty \leq u) - \mathbb{P}(|d_{\tau_k} + G_{\tau_k, \cdot, \cdot}^\circ|_\infty \leq u) \right| \rightarrow 0.$$

Therefore we have

$$\mathbb{P}(\tau_k \in \mathcal{A}_1, 1 \leq k \leq K_0) \rightarrow 1. \quad (66)$$

Let  $\mathcal{B}(\tau, r) = \{t : |t - \tau| \leq r\}$ . By (65) and (66), we have

$$\lim_n \mathbb{P} \left( \{\tau_1, \tau_2, \dots, \tau_{K_0}\} \subseteq \mathcal{A}_1 \subseteq \cup_{1 \leq k \leq K_0} \mathcal{B}(\tau_k, bn) \right) = 1.$$

Since for  $k_1 \neq k_2$ ,  $|\tau_{k_1} - \tau_{k_2}| \gg bn$ , for any  $k_1 \neq k_2$  and  $t \in \mathcal{B}(\tau_{k_1}, bn)$ , for all large  $n$ ,  $\mathcal{B}(t, 2bn) \cap \mathcal{B}(\tau_{k_2}, 2bn) = \emptyset$ . Thus we complete the proof.  $\square$

*Proof of (ii).* Let  $\mu_i^{(l)}$  (resp.  $U_i^{(l)}$ ) be  $\hat{\mu}_i^{(l)}$  with  $Y_i$  therein replaced by  $\mu(i/n)$  (resp.  $\epsilon_i$ ). Similarly we can define  $\mu_i^{(r)}$  and  $U_i^{(r)}$ . Let  $\Delta\mu_i = \mu_i^{(l)} - \mu_i^{(r)}$  and  $\Delta U_i = U_i^{(l)} - U_i^{(r)}$ . Let  $\Delta f_i$  be  $\Delta\mu_i$  with  $\mu$  replaced by  $f$ .

Let  $\tau \in \{\tau_1, \dots, \tau_{K_0}\}$  be some break point associated with jump  $\gamma$ . For any  $t$  such that  $|t - \tau| \leq bn$ , we have  $\Delta\mu_t = (1 - \sum_{i=1}^{|t-\tau|} w_i)\gamma + \Delta f_t$ . Hence

$$\begin{aligned} V_t &= \Lambda^{-1} \Delta\mu_t + \Lambda^{-1} \Delta U_t \\ &= (1 - \sum_{i=1}^{|t-\tau|} w_i) \Lambda^{-1} \gamma + \Lambda^{-1} \Delta f_t + \Lambda^{-1} \Delta U_\tau + (\Lambda^{-1} \Delta U_t - \Lambda^{-1} \Delta U_\tau). \end{aligned} \quad (67)$$

Let  $\hat{\tau} = \operatorname{argmax}_{\{t: |t-\tau| \leq bn\}} |V_t|_\infty$ . The proceeding proof contains three steps.

Step 1. Let  $j_\tau = \operatorname{argmax}_j |V_{\hat{\tau}, j}|$ , where  $V_{\hat{\tau}, j}$  is the  $j$ th entity of  $V_{\hat{\tau}}$ . This step shows

$$\liminf_n |(\Lambda^{-1} \gamma)_{j_\tau}| / \delta^\circ \geq 1.$$

We shall show by contradiction. By (47),  $|\Delta f_t|_\infty = O(b^2)$ . If  $|(\Lambda^{-1} \gamma)_{j_\tau}| \leq c\delta^\circ$ , for some  $c < 1$ , then by (67),  $|V_{\hat{\tau}}|_\infty \leq c\delta^\circ + O(b^2) + |\Lambda^{-1} \Delta U_{\hat{\tau}}|_\infty$ . Let  $\tilde{U}_t$  be  $U_t$  with  $\eta_{i,j}$  replaced

by  $z_{i,j}$  where  $z_{i,j}$  are i.i.d standard normal random variables. Then  $\max_t |\Lambda^{-1} \Delta \tilde{U}_t|_\infty = O_{\mathbb{P}}((bn)^{-1/2} \log(np)^{1/2})$ . Then by Gaussian approximation Theorem 1,

$$\max_t |\Lambda^{-1} \Delta U_t|_\infty = O_{\mathbb{P}}((bn)^{-1/2} \log(np)^{1/2}).$$

Since  $\delta^\circ \gg (bn)^{-1/2} \log(np)^{1/2}$ , we have  $|V_{\hat{\tau}}|_\infty \leq c\delta^\circ(1 + o_{\mathbb{P}}(1))$ . On the other hand, by (67),  $|V_\tau|_\infty \geq \delta^\circ + O(b^2) - |\Lambda^{-1} \Delta U_\tau|_\infty = \delta^\circ(1 + o_{\mathbb{P}}(1))$ . These imply  $\mathbb{P}(V_{\hat{\tau}} < V_\tau) \rightarrow 1$ , which is a contradiction.

Step 2. This step shows

$$\max_t |\Lambda^{-1} \Delta U_\tau - \Lambda^{-1} \Delta U_t|_\infty / |t - \tau|^{1/2} = O_{\mathbb{P}}\{\log(np)^{1/2} / (bn)\}. \quad (68)$$

Let  $t \leq \tau$ , the other direction can be similarly dealt with. Let  $(z_{i,j})$  be i.i.d. standard Gaussian random variables. Define  $\Delta \tilde{U}_t$  (resp.  $\tilde{\epsilon}_t$ ) to be  $\Delta U_t$  (resp.  $\epsilon_t$ ) with  $\eta_{i,j}$  therein replaced by  $z_{i,j}$ . Then by Gaussian approximation Theorem 1, it suffices to show (68) with  $U_t$  replaced by  $\tilde{U}_t$ . We note that

$$\begin{aligned} \Delta \tilde{U}_t - \Delta \tilde{U}_\tau &= \sum_{i=t-bn}^{\tau-bn-1} w_{t-i} \tilde{\epsilon}_i + \sum_{i=\tau-bn}^{t-1} (w_{t-i} - w_{\tau-i}) \tilde{\epsilon}_i - \sum_{i=t+1}^{\tau-1} (w_{i-t} + w_{\tau-i}) \tilde{\epsilon}_i \\ &- \sum_{i=\tau+1}^{t+bn} (w_{i-t} - w_{i-\tau}) \tilde{\epsilon}_i - \sum_{i=t+bn+1}^{\tau+bn} w_{i-\tau} \tilde{\epsilon}_i + w_{\tau-t} \tilde{\epsilon}_t - w_{\tau-t} \tilde{\epsilon}_\tau =: \sum_{k=1}^7 r_k. \end{aligned}$$

For  $r_1$ , we have

$$|\Lambda^{-1} r_1|_\infty = \max_{1 \leq j_1 \leq p} \left| \sum_{l \leq \tau-bn-1, 1 \leq j_2 \leq p} E_{l,j_1,j_2} z_{l,j_2} \right|, \quad \text{where } E_l = \sum_{i=(t-bn) \vee l}^{\tau-bn-1} w_{t-i} \Lambda^{-1} A_{i-l},$$

with  $E_{l,j_1,j_2}$  as the  $(j_1, j_2)$ th entity of matrix  $E_l$ . Then

$$\max_t (|\Lambda^{-1} r_1|_\infty / |t - \tau|^{1/2}) = O_{\mathbb{P}}\{\log(np)^{1/2} / (bn)\}.$$

A similar argument leads to the same bound for  $r_3$  and  $r_5$ . For  $r_2$ , we can rewrite

$$|\Lambda^{-1}r_2|_\infty = \max_{1 \leq j_2 \leq p} \left| \sum_{l \leq t-1} E'_{l,j_1,j_2} z_{l,j_2} \right|, \quad \text{where} \quad E'_l = \sum_{i=(\tau-bn) \vee l}^{t-1} (w_{t-i} - w_{\tau-i}) \Lambda^{-1} A_{i-l}.$$

Then similarly we have  $\max_t |\Lambda^{-1}r_2|_\infty / |\tau - t|^{3/2} = O_{\mathbb{P}}\{\log(np)^{1/2}/(bn)^2\}$ . The same bound can be derived for  $r_4$  as well. We obtain (68) by summing the above bounds up.

Step 3. Without loss of generality, assume  $\gamma_{j_\tau} > 0$ . Then by the argument in step 1, with probability tending to 1,  $V_{\hat{\tau},j_\tau} > 0$ . By Assumption 2.1, we have  $|\Delta f_t - \Delta f_\tau|_\infty = O(|t - \tau|/n)$ .

With probability tending to 1, by (67),

$$|V_\tau|_\infty - |V_{\hat{\tau}}|_\infty \geq V_{\tau,j_\tau} - V_{\hat{\tau},j_\tau} \geq \sum_{i=1}^{|\hat{\tau}-\tau|} w_i (\Lambda^{-1}\gamma)_{j_\tau} - O(|\hat{\tau} - \tau|/n) - |\Lambda^{-1}\Delta U_\tau - \Lambda^{-1}\Delta U_{\hat{\tau}}|_\infty.$$

By (43), we have  $\sum_{i=1}^{|\hat{\tau}-\tau|} w_i \geq c''_w |t - \tau|/(bn)$ . Hence by Step 1 and Step 2 we further derive

$$|V_\tau|_\infty - |V_{\hat{\tau}}|_\infty \geq c''_w |\hat{\tau} - \tau| \delta^\circ / (bn) - O(|\hat{\tau} - \tau|/n) - O_{\mathbb{P}}(|\tau - \hat{\tau}|^{1/2} \log(np)^{1/2} / (bn)).$$

Since  $|V_\tau|_\infty < |V_{\hat{\tau}}|_\infty$ , we have

$$|\tau - \hat{\tau}| = O_{\mathbb{P}}\{\log(np)/\delta^{\circ 2}\}.$$

□

*Proof of (iii).* Recall the definition of  $\mu_t^{(l)}$ ,  $\mu_t^{(r)}$ ,  $U_t^{(l)}$  and  $U_t^{(r)}$  in the proof of (ii) and  $M = bn$ . Since  $M \gg \log(np)/\delta^{\circ 2}$ ,

$$|\mu_{\hat{\tau}_k - M}^{(l)} - \mu((\hat{\tau}_k - M)/n)|_\infty = |f_{\hat{\tau}_k - M}^{(l)} - f((\hat{\tau}_k - M)/n)|_\infty = O(b).$$

Similarly  $|\mu_{\hat{\tau}_k + M}^{(r)} - \mu((\hat{\tau}_k + M)/n)|_\infty = O(b)$ . Since  $\max_{1 \leq j \leq p} |f'_j|$  is bounded,

$$|\mu((\hat{\tau}_k + M)/n) - \mu((\hat{\tau}_k - M)/n) - \gamma_{k^*}|_\infty = |f((\hat{\tau}_k + M)/n) - f((\hat{\tau}_k - M)/n)|_\infty = O(M/n).$$

Hence

$$\begin{aligned} |\Lambda^{-1}(\hat{\gamma}_k - \gamma_{k^*})|_\infty &= |\Lambda^{-1}(\mu_{\hat{\tau}_k+M}^{(r)} - \mu_{\hat{\tau}_k-M}^{(l)} - \gamma_{k^*}) + \Lambda^{-1}U_{\hat{\tau}_k+M}^{(r)} - \Lambda^{-1}U_{\hat{\tau}_k-M}^{(l)}|_\infty \\ &\leq O(b + M/n) + |\Lambda^{-1}U_{\hat{\tau}_k-M}^{(l)} - \Lambda^{-1}U_{\hat{\tau}_k+M}^{(r)}|_\infty. \end{aligned} \quad (69)$$

By Gaussian approximation and (63) we have

$$\mathbb{P}(|\Lambda^{-1}U_{\hat{\tau}_k-M}^{(l)} - \Lambda^{-1}U_{\hat{\tau}_k+M}^{(r)}|_\infty \geq 2c'_w \log(np)^{1/2}/(bn)^{1/2}) \rightarrow 0.$$

Inserting the above equation into (69) and we obtain the desired result.  $\square$

## B.5 Proof of Theorem 3

*Proof.* We recall  $M = bn$ . Similar to (69), we have  $|\Lambda^{-1}(\hat{\gamma}_k - \gamma_{k^*}) - \Lambda^{-1}(U_{\hat{\tau}_k+M}^{(r)} - U_{\hat{\tau}_k-M}^{(l)})|_\infty \leq cM/n$ . Therefore

$$\begin{aligned} &\sup_{u \in \mathbb{R}} \left| \mathbb{P}((bn)^{1/2} |\Lambda^{-1}(\hat{\gamma}_k - \gamma_{k^*})|_\infty \leq u) - \mathbb{P}((bn)^{1/2} |\tilde{Z}|_\infty \leq u) \right| \\ &\leq \sup_{u \in \mathbb{R}} \mathbb{P}(|(bn)^{1/2} |\tilde{Z}|_\infty - u| \leq c(bn)^{1/2} M/n) \\ &+ \sup_{u \in \mathbb{R}} \left| \mathbb{P}(|\Lambda^{-1}(U_{\hat{\tau}_k-M}^{(l)} - U_{\hat{\tau}_k+M}^{(r)})|_\infty \leq u) - \mathbb{P}(|\tilde{Z}|_\infty \leq u) \right| = \text{I}_1 + \text{I}_2. \end{aligned}$$

We note that  $(bn)^{1/2} \tilde{Z}_j$  are i.i.d with variance  $2(bn) \sum_{t=1}^{bn} w_t^2$ , which by (43) converges to  $2c'_w{}^2 > 0$ . Therefore by Lemma 4,

$$\text{I}_1 = O\{(bn)^{1/2}(M/n)\log(np)^{1/2}\} = o(1).$$

Let  $\tilde{G} = (\tilde{G}_1, \tilde{G}_2, \dots, \tilde{G}_n)$ , where  $\tilde{G}_l = w_{\hat{\tau}_k-M-l} \Lambda^{-1} S$ , if  $\hat{\tau}_k - M - bn \leq l \leq \hat{\tau}_k - M - 1$ , and  $\tilde{G}_l = w_{l-(\hat{\tau}_k+M)} \Lambda^{-1} S$ , if  $\hat{\tau}_k + M + 1 \leq l \leq \hat{\tau}_k + M + bn$  and elsewhere zero. Let  $\underline{z}$  be Gaussian vector in  $\mathbb{R}^{n\bar{p}}$  with zero mean and identity covariance matrix. Then  $\tilde{G}\underline{z} \stackrel{d}{=} \tilde{Z}$ . By

the same argument as in Theorem 1 with  $d_i = 0$  we have

$$\sup_{u \in \mathbb{R}} |\mathbb{P}(|\Lambda^{-1}(U_{\hat{\tau}_k - M}^{(l)} - U_{\hat{\tau}_k + M}^{(r)})|_\infty \leq u) - \mathbb{P}(|\tilde{G}\underline{z}|_\infty \leq u)| = o(1).$$

Thus  $I_2 = o(1)$  and we complete the proof.  $\square$

## B.6 Proof of Theorem 4

*Proof of (i).* We shall condition on the event where  $\hat{\mathcal{S}}_k = \mathcal{S}_k$  and  $|\hat{\tau}_k - \tau_k| \ll bn$ . By Theorem 2 and Corollary 2, the event would take place with probability tending to 1. Denote  $\varepsilon_t = \sum_{j \in \mathcal{S}_k} (\Lambda^{-1} \hat{\gamma}_k)_j (\Lambda^{-1} \varepsilon_t)_j$ , and  $\hat{a}_k = \sum_{j \in \mathcal{S}_k} (\Lambda^{-1} \gamma_k)_j (\Lambda^{-1} \hat{\gamma}_k)_j$ . Then we have

$$X_t = \hat{a}_k \mathbf{1}_{t \geq \tau_k} + \sum_{j \in \mathcal{S}_k} f_j(t/n) (\Lambda^{-1} \hat{\gamma}_k)_j + \varepsilon_t.$$

Let  $l(t) = \sqrt{(4bn + 1)/(t(4bn + 1 - t))}$ ,  $t_0 = \hat{\tau}_k - 2bn$ ,  $t_1 = \hat{\tau}_k + 2bn$  and

$$D_t = \left( \sum_{s=t_0}^{t_1} X_s \frac{t - t_0}{4bn + 1} - \sum_{s=t_0}^{t-1} X_s \right) l(t).$$

For any  $r > 0$ , we have

$$\begin{aligned} D_{\tau_k + r} - D_{\tau_k} &= \left( \sum_{s=t_0}^{t_1} X_s \frac{\tau_k - t_0}{4bn + 1} - \sum_{s=t_0}^{\tau_k - 1} X_s \right) (l(\tau_k + r) - l(\tau_k)) \\ &\quad + \left( \sum_{s=t_0}^{t_1} X_s \frac{r}{4bn + 1} - \sum_{s=\tau_k}^{\tau_k + r - 1} X_s \right) l(\tau_k + r) = I_1 + I_2. \end{aligned}$$

Denote  $I_i(f)$  (resp.  $I_i(\varepsilon)$ ,  $I_i(a)$ ) to be  $I_i$  with  $X_t$  therein replaced by  $f(t/n)$  (resp.  $\varepsilon_t$ ,  $\hat{a}_k \mathbf{1}_{t \geq \tau_k}$ ),  $i = 1, 2$ .

Firstly, consider the  $f$  part. Note  $|l(\tau_k + r) - l(\tau_k)| \lesssim (bn)^{-3/2} r$ . Thus by the continuity

of  $f_j$ , for  $I_1$  part, we have

$$\begin{aligned} \max_j \left| \sum_{s=t_0}^{t_1} f_j(s/n) \frac{\tau_k - t_0}{4bn + 1} - \sum_{s=t_0}^{\tau_k - 1} f_j(s/n) \right| |l(\tau_k + r) - l(\tau_k)| &= O((b^2n) |l(\tau_k + r) - l(\tau_k)|) \\ &= O(br(bn)^{-1/2}). \end{aligned}$$

Similarly we can handle the  $f$  part in  $I_2$  and therefore

$$|I_1(f) + I_2(f)| = O(br(bn)^{-1/2} |\tilde{\gamma}_k|_1).$$

Secondly, let us consider the drift part.

$$\begin{aligned} &I_1(a) + I_2(a) \\ &= (t_1 - \tau_k + 1) \hat{a}_k \frac{\tau_k - t_0}{4bn + 1} (l(\tau_k + r) - l(\tau_k)) + \left( (t_1 - \tau_k + 1) \hat{a}_k \frac{r}{4bn + 1} - r \hat{a}_k \right) l(\tau_k + r) \\ &= - \frac{\hat{a}_k (\tau_k - t_0) (4bn + 1)^{1/2} r}{(\tau_k + r - t_0)(\tau_k - t_0) (\sqrt{(t_1 - \tau_k - r)(\tau_k - t_0)} + \sqrt{(\tau_k - t_0 + r)(t_1 - \tau_k)}}) \\ &\geq - (bn)^{-1/2} \hat{a}_k / 2. \end{aligned}$$

Tirdly, let us focus on the  $\varepsilon$  part. By Theorem 2 (iii),  $|(\Lambda^{-1} \hat{\gamma}_k)_{j \in \mathcal{S}_k}|_2 = |\tilde{\gamma}_k|_2 (1 + o_{\mathbb{P}}(1))$  and thus  $\hat{a}_k = |\tilde{\gamma}_k|_2 (1 + o_{\mathbb{P}}(1))$ . Then together with (35), we obtain that the long run variance for  $\varepsilon_k$  is  $\varsigma_k^2 (1 + o_{\mathbb{P}}(1))$ . Hence by Theorem 1 in El Machkouri et al. (2013),

$$\left| \sum_{s=\hat{\tau}_k - 2bn}^{\hat{\tau}_k + 2bn} \varepsilon_s \right| = O_{\mathbb{P}}((bn)^{1/2} \varsigma_k \log(bn)^{1/2}) \quad \text{and} \quad \left| \sum_{s=\tau_k}^{\tau_k + r} \varepsilon_s \right| = O_{\mathbb{P}}(r^{1/2} \varsigma_k).$$

Therefore

$$I_1(\varepsilon) = O_{\mathbb{P}}((bn)^{1/2} \varsigma_k \log(bn)^{1/2} |l(\tau_k + r) - l(\tau_k)|) = O_{\mathbb{P}}((bn)^{-1} r \varsigma_k \log(bn)^{1/2}).$$

For  $I_2$  part, we have

$$\begin{aligned} I_2(\varepsilon) &= \left( \sum_{s=t_0}^{t_1} \varepsilon_s \frac{r}{4bn+1} - \sum_{s=\tau_k}^{\tau_k+r-1} \varepsilon_s \right) l(\tau_k+r) \\ &= O_{\mathbb{P}}\left( (bn)^{-1} r \varsigma_k \log(bn)^{1/2} + r^{1/2} \varsigma_k (bn)^{-1/2} \right). \end{aligned}$$

Combining all the previous parts we have

$$\begin{aligned} & D_{\tau_k+r} - D_{\tau_k} \\ & \geq - (bn)^{-1/2} a_k r (1 + o_{\mathbb{P}}(1)) / 2 + O(br(bn)^{-1/2} |\tilde{\gamma}_k|_1) + O_{\mathbb{P}}\left( (bn)^{-1} r \varsigma_k \log(bn)^{1/2} + r^{1/2} \varsigma_k (bn)^{-1/2} \right) \end{aligned} \quad (70)$$

Note  $|\tilde{\gamma}_k|_1 \leq |\mathcal{S}_k|^{1/2} |\tilde{\gamma}_k|_2$ , thus

$$|\tilde{\gamma}_k|_1 b \leq |\mathcal{S}_k|^{1/2} |\tilde{\gamma}_k|_2 b \ll |\mathcal{S}_k|^{1/2} |\tilde{\gamma}_k|_2 \delta^\dagger \leq |\tilde{\gamma}_k|_2^2 = a_k.$$

Therefore in (70), we have  $O(br(bn)^{-1/2} |\tilde{\gamma}_k|_1) = o((bn)^{-1/2} a_k r)$ .

Since  $\tilde{\Sigma}_k$  is a covariance matrix with diagonal entities 1,  $|\tilde{\Sigma}_k|_2 \leq |\mathcal{S}_k|$ . Note  $\varsigma_k^2 \leq |\tilde{\Sigma}_k|_2 a_k$ , thus

$$\varsigma_k \leq |\mathcal{S}_k|^{1/2} a_k^{1/2}.$$

Then we have

$$\varsigma_k (bn)^{-1/2} \log(bn)^{1/2} \leq |\mathcal{S}_k|^{1/2} a_k^{1/2} (bn)^{-1/2} \log(bn)^{1/2} \ll |\mathcal{S}_k|^{1/2} a_k^{1/2} \delta^\dagger \leq a_k,$$

where the last inequality is because  $a_k = |\tilde{\gamma}_k|_2^2 \geq \delta^{\dagger 2} |\mathcal{S}_k|$ . Therefore  $O_{\mathbb{P}}((bn)^{-1} r \varsigma_k \log(bn)^{1/2}) = o_{\mathbb{P}}((bn)^{-1/2} a_k r)$  in (70). Inserting the above equations into (70) leads to

$$(bn)^{1/2} (D_{\tau_k+r} - D_{\tau_k}) \geq -a_k r (1/2 + o_{\mathbb{P}}(1)) + O_{\mathbb{P}}(\varsigma_k r^{1/2}).$$

Since  $D_{\tilde{\tau}_k}$  is the maximum,  $D_{\tau_k+r} - D_{\tau_k} > 0$ . Therefore  $r = O_{\mathbb{P}}(\varsigma_k^2 / a_k^2)$ . By a similar argument for the  $r < 0$  part, the desired result follows.  $\square$

*Proof of (ii).* Let  $\mathcal{F}_t$  be the  $\sigma$ -field generated by  $\{\eta_{s,j}, s \leq t, 1 \leq j \leq p\}$ . Denote the projection operator  $\mathcal{P}_t \cdot = \mathbb{E}(\cdot | \mathcal{F}_t) - \mathbb{E}(\cdot | \mathcal{F}_{t-1})$ . Let  $\bar{\gamma}_{k,j} = (\Lambda^{-1} \gamma_k)_j$ , if  $j \in \mathcal{S}_k$ ,  $\bar{\gamma}_{k,j} = 0$  if  $j \notin \mathcal{S}_k$ . Let  $(\eta'_t)$  be an i.i.d copy of  $(\eta_t)$ . Then

$$c_t := \|\mathcal{P}_0 \varepsilon_t\|_4 \leq \|\bar{\gamma}_k^\top \Lambda^{-1} A_t (\eta_0 - \eta'_0)\|_4 \lesssim |\bar{\gamma}_k^\top \Lambda^{-1} A_t|_2 \mu_4,$$

where the last inequality is by Lemma 2 and that  $\eta_{0,j}$ ,  $1 \leq j \leq p$ , are i.i.d. By Assumption 2.5,

$$\sum_{s \geq m} c_s \lesssim \sum_{s \leq m} |\bar{\gamma}_k^\top \Lambda^{-1} A_s|_2 \leq \sum_{j=1}^p \sum_{s \geq m} \bar{\gamma}_{k,j} \sigma_{j,j}^{-1/2} |A_{s,j \cdot}|_2 \lesssim |\bar{\gamma}_k|_1 m^{-\beta} = |\tilde{\gamma}_k|_1 m^{-\beta}.$$

Thus by Corollary 2.1 in Berkes et al. (2014), strong invariance principle holds for  $\sum_{s \leq t} \varepsilon_s$ .

Thus similar to (70), we have

$$D_{\tau_k+r} - D_{\tau_k} = -2^{-1} a_k r l(\tau_k) (1 + o_{\mathbb{P}}(1)) - \sum_{s=\tau_k}^{\tau_k+r-1} \varepsilon_s l(\tau_k + r) \xrightarrow{\mathcal{D}} (-2^{-1} a_k r + \varsigma_k \mathbb{B}(r)) l(\tau_k).$$

The  $r < 0$  part can be similarly dealt with. □

## B.7 Proof of Theorem 5

*Proof of (i).* The main idea follows the proof of Proposition 2.4 in Catoni (2012), however due to the dependence and the break points, our result is much more involved. Let

$$\mathcal{S} = \{k \mid \mathcal{A}_k \text{ or } \mathcal{A}_{k-1} \text{ contains break points}\}.$$

Then by assumption  $|\mathcal{S}| \leq 2K_0$ . We look at estimators without the break point first.

$$\bar{h}_{i,j}(u) = \sum_{k \notin \mathcal{S}} \phi_{\alpha_{i,j}}(\hat{\sigma}_{i,j,k} - u) / N_2, \quad \text{where } N_2 = N_1 - |\mathcal{S}|.$$

Let

$$\tilde{\sigma}_{i,j} = \sum_{k \notin \mathcal{S}} \mathbb{E} \hat{\sigma}_{i,j,k} / N_2 \quad \text{and} \quad v_{i,j}^2 = \sum_{k \notin \mathcal{S}} \mathbb{E} \hat{\sigma}_{i,j,k}^2 / N_2 - \tilde{\sigma}_{i,j}^2.$$

Define functions

$$\begin{aligned} B_{i,j}^+(u, x) &= \tilde{\sigma}_{i,j} - u + \alpha_{i,j} [(\tilde{\sigma}_{i,j} - u)^2 + v_{i,j}^2] / 2 + x, \\ B_{i,j}^-(u, x) &= \tilde{\sigma}_{i,j} - u - \alpha_{i,j} [(\tilde{\sigma}_{i,j} - u)^2 + v_{i,j}^2] / 2 - x. \end{aligned}$$

The proof contains four steps.

Step 1. This step shows that function  $\mathbb{E} \bar{h}_{i,j}(u)$  for any  $i, j$  satisfies, the expected loss functions have upper and lower envelope functions,

$$B_{i,j}^-(u, 0) \leq \mathbb{E} \bar{h}_{i,j}(u) \leq B_{i,j}^+(u, 0).$$

By (40),  $\phi(x) \leq x + x^2/2$  and thus

$$\mathbb{E} \bar{h}_{i,j}(u) \leq \sum_{k \notin \mathcal{S}} (\mathbb{E}(\hat{\sigma}_{i,j,k} - u) + \alpha_{i,j} \mathbb{E}(\hat{\sigma}_{i,j,k} - u)^2 / 2) / N_2 = B_{i,j}^+(u, 0).$$

Similarly we can bound the other side.

Step 2. This step shows for any  $x > 0$ , the estimated influence function  $\bar{h}_{i,j}(u)$  is highly concentrated around its mean, for  $C_0 > 0$  and  $x \gtrsim (N_2 \log(N_2))^{1/2}$ ,

$$\sum_{i,j=1}^p \mathbb{P} \left( \sup_{|u - \sigma_{i,j}| \leq C_0} |\bar{h}_{i,j}(u) - \mathbb{E} \bar{h}_{i,j}(u)| \geq x (\sigma_{i,i}^{1/2} \sigma_{j,j}^{1/2}) / N_2 \right) \lesssim p^2 (N_2 \log(n))^{q/4} x^{-q/2} + e^{-x^2/(cN_2)}, \quad (71)$$

where  $c$  and the constant in  $\lesssim$  are independent of  $n, p$ .

First introduce some notation. For any random variable  $X$ , denote  $\mathbb{E}_0 X = X - \mathbb{E} X$ , the centering operator. Let  $\mathcal{F}_k = (\eta_t, t \in \cup_{s \leq k} \mathcal{A}_s)$  and  $\mathcal{F}_{k, \{s\}}$ ,  $s \leq k$ , be  $\mathcal{F}_k$  with  $\eta_t, t \in \mathcal{A}_s$

therein replaced by  $\eta'_t$ , where  $\eta'_t$  are i.i.d copy of  $\eta_t$ . For any random variable  $X = h(\mathcal{F}_k)$ , let  $X_{\{i\}} = h(\mathcal{F}_{k,\{i\}})$ . Denote  $\Delta\xi_k = \xi_k - \xi_{k-1}$ . We now show that the temporal dependence measure decays with polynomial rate. Let  $\zeta_{i,j,k}(u) = \phi_{\alpha_{i,j}}(\hat{\sigma}_{i,j,k} - u)$ . Since  $|\phi'|_\infty \leq 1$ , we have for any  $s \in \mathbb{N}$  and any  $u$ ,

$$\begin{aligned} & \left\| \sup_u |\zeta_{i,j,k}(u) - \zeta_{i,j,k,\{k-s\}}(u)| \right\|_{q/2} \leq \left\| \hat{\sigma}_{i,j,k} - \hat{\sigma}_{i,j,k,\{k-s\}} \right\|_{q/2} \\ & \leq 2^{-1}m \left( \left\| \mathbb{E}_0 \Delta\xi_{k,i} (\Delta\xi_{k,j} - \Delta\xi_{k,j,\{k-s\}}) \right\|_{q/2} + \left\| \mathbb{E}_0 (\Delta\xi_{k,i} - \Delta\xi_{k,i,\{k-s\}}) \Delta\xi_{k,j,\{k-s\}} \right\|_{q/2} \right) \\ & = : 2^{-1}m(I_1 + I_2). \end{aligned}$$

Let  $U_{k,i}$  (resp.  $f_{k,i}$ ) be  $\xi_{k,i}$  with  $Y_t$  replaced by  $\epsilon_t$  (resp.  $f(t/n)$ ). Then  $\xi_{k,i} = \epsilon_{k,i} + f_{k,i}$  when there is no break. Let  $\Delta U_{k,i} = U_{k,i} - U_{k-1,i}$  and  $\Delta f_{k,i} = f_{k,i} - f_{k-1,i}$ . Then we have

$$I_1 \leq \left\| \mathbb{E}_0 \Delta f_{k,i} (\Delta U_{k,j} - \Delta U_{k,j,\{k-s\}}) \right\|_{q/2} + \left\| \mathbb{E}_0 \Delta U_{k,i} (\Delta U_{k,j} - \Delta U_{k,j,\{k-s\}}) \right\|_{q/2} =: I_{11} + I_{12}. \quad (72)$$

Since  $\max_j |f_j|_\infty < f^*$ ,

$$\max_{1 \leq j \leq p} |\Delta f_{k,j}| \leq f^* m/n. \quad (73)$$

Let  $E_{k,l,i,\cdot} = \sum_{t=(km+1)\vee l}^{(k+1)m} A_{t-l,i,\cdot}$ , where  $A_{t-l,i,\cdot}$  is the  $i$ th row of matrix  $A_{t-l}$ . Then for  $s \geq 1$ ,

$$U_{k,i} = \sum_{l \leq (k+1)m} E_{k,l,i,\cdot} \eta_l / m \quad \text{and} \quad \Delta U_{k,i} - \Delta U_{k,i,\{k-s\}} = \sum_{l \in \mathcal{A}_{k-s}} (E_{k,l,i,\cdot} - E_{k-1,l,i,\cdot}) (\eta_l - \eta'_l) / m. \quad (74)$$

Since  $\eta_l$  are i.i.d, by Lemma 2, (73) and (74),

$$\left\| \Delta f_{k,i} (\Delta U_{k,j} - \Delta U_{k,j,\{k-s\}}) \right\|_{q/2} \leq 2f^* c_q \left( \sum_{l \in \mathcal{A}_{k-s}} (|E_{k,l,i,\cdot}|_2 + E_{k-1,l,i,\cdot}|_2)^2 \right)^{1/2} \mu_{q/2}/n. \quad (75)$$

By Assumption 2.5, we have for any  $s > 1$ ,

$$\sum_{l \in \mathcal{A}_{k-s}} |E_{k,l,i}|_2^2 \lesssim m(m(s-1))^{-2\beta} \sigma_{i,i}, \quad \text{and} \quad \sum_{l \leq (k+1)m} |E_{k,l,i}|_2^2 \lesssim m\sigma_{i,i}, \quad (76)$$

where the constant in  $\lesssim$  only depending on  $\beta, c_s$ . Hence by (75) and (76),

$$I_{11} \lesssim m^{1/2} n^{-1} (m(s-1))^{-\beta} \sigma_{i,i}^{1/2}, \quad (77)$$

where the constant in  $\lesssim$  only depends on  $\beta, c_s, q, \mu_q, f^*$ . By Lemma 2 and (76)

$$\begin{aligned} \|\mathbb{E}_0 U_{k,i}(U_{k,j} - U_{k,j,\{k-s\}})\|_{q/2} &= \left\| \mathbb{E}_0 \left( \sum_{l_1 \leq (k+1)m} E_{k,l_1,i} \eta_{l_1} \sum_{l_2 \in \mathcal{A}_{k-s}} E_{k,l_2,j} (\eta_{l_2} - \eta'_{l_2}) \right) \right\|_{q/2} m^{-2} \\ &\lesssim m^{-2} \left( \sum_{l_1 \leq (k+1)m} \sum_{l_2 \in \mathcal{A}_{k-s}} |E_{k,l_1,i}|_2^2 |E_{k,l_2,j}|_2^2 \right)^{1/2} \\ &\lesssim m^{-1} (m(s-1))^{-\beta} \sigma_{i,i}^{1/2} \sigma_{j,j}^{1/2}, \end{aligned} \quad (78)$$

where the constant in  $\lesssim$  only depends on  $\mu_q, q, c_s$ . Thus  $I_{12} \lesssim m^{-1} (m(s-1))^{-\beta} \sigma_{i,i}^{1/2} \sigma_{j,j}^{1/2}$ .

By combining the bounds for  $I_{11}$  and  $I_{12}$  and a similar argument for  $I_2$ , we have

$$\delta_s := \max_k \left\| \sup_u |\zeta_{i,j,k}(u) - \zeta_{i,j,k,\{k-s\}}(u)| \right\|_{q/2} \lesssim ((ms)^{-\beta} \mathbf{1}_{s>1} + \mathbf{1}_{s \leq 1}) \sigma_{i,i}^{1/2} \sigma_{j,j}^{1/2}, \quad (79)$$

where the constant in  $\lesssim$  only depends on  $\mu_q, c_s, c_\sigma, \beta, q, f^*$ .

Let  $\delta := \sigma_{i,i}^{1/2} \sigma_{j,j}^{1/2} x / (2N_2)$  and  $A_n$  be the  $\delta$  net for  $\{u : |u - \sigma_{i,j}| \leq C_0\}$ . Denote  $f(u) = \bar{h}_{i,j}(u) - \mathbb{E} \bar{h}_{i,j}(u)$ . Then by  $|\phi'|_\infty \leq 1$ ,

$$\sup_{|v - \sigma_{i,j}| \leq C_0} \min_{u \in A_n} |f(u) - f(v)| \leq \delta.$$

Therefore  $|A_n| = 2C_0/\delta = O(n)$  and

$$\mathbb{P} \left( \sup_{|u - \sigma_{i,j}| \leq C_0} |\bar{h}_{i,j}(u) - \mathbb{E} \bar{h}_{i,j}(u)| \geq x(\sigma_{i,i}^{1/2} \sigma_{j,j}^{1/2}) / N_2 \right) \leq \mathbb{P} \left( \max_{u \in A_n} |\bar{h}_{i,j}(u) - \mathbb{E} \bar{h}_{i,j}(u)| \geq x(\sigma_{i,i}^{1/2} \sigma_{j,j}^{1/2}) / (2N_2) \right).$$

Desired result follows from Lemma 5.8 in Zhang et al. (2017).

Step 3. This step shows for the estimator

$$\max_{1 \leq i, j \leq p} |\tilde{\sigma}_{i,j} - \sigma_{i,j}| = O(m^{-\beta/(\beta+1)} \sigma_{i,i}^{1/2} \sigma_{j,j}^{1/2} + m^3/n^2), \quad \text{and} \quad v_{i,j}^2 = O(\sigma_{i,i} \sigma_{j,j}), \quad (80)$$

where the convergence is uniform for  $1 \leq i, j \leq p$ .

Let  $\hat{\epsilon}_{i,j,k}$  be  $\hat{\sigma}_{i,j,k}$  with  $Y_t$  replaced by  $\epsilon_t$  and let  $\sigma_{i,j}^\diamond = \mathbb{E}\hat{\epsilon}_{i,j,1}$ . Then by (73),

$$|\tilde{\sigma}_{i,j} - \sigma_{i,j}^\diamond| \leq m \sum_{k \notin \mathcal{S}} |\Delta f_{k,i}| |\Delta f_{k,j}| / (2N_2) = O(m^3/n^2). \quad (81)$$

Note the convergence in above  $O(\cdot)$  and all the followings are uniform for  $i, j$ . Let  $\rho_{i,j,k} = \mathbb{E}(\epsilon_{0,i} \epsilon_{k,j})$ . Then for any  $L < m$ ,

$$|m\mathbb{E}(U_{1,i}U_{1,j}) - \sigma_{i,j}| = \left| m^{-1} \sum_{-m < k < m} (m - |k|) \rho_{i,j,k} - \sigma_{i,j} \right| = O\left( \sum_{|k| \geq L} |\rho_{i,j,k}| + Lm^{-1} \sum_{k \in \mathbb{Z}} |\rho_{i,j,k}| \right).$$

By Assumption 2.5,  $\sum_{|k| \geq L} |\rho_{i,j,k}| \leq \sum_{t \in \mathbb{Z}, |k| \geq L} |A_{t,i}|_2 |A_{t+k,j}|_2 = O(L^{-\beta} \sigma_{i,i}^{1/2} \sigma_{j,j}^{1/2})$ . Take  $L = m^{1/(\beta+1)}$ , then  $|m\mathbb{E}(U_{1,i}U_{1,j}) - \sigma_{i,j}| = O(m^{-\beta/(\beta+1)} \sigma_{i,i}^{1/2} \sigma_{j,j}^{1/2})$ . And similarly  $|m\mathbb{E}(U_{1,i}U_{2,j})| = O(m^{-\beta/(\beta+1)} \sigma_{i,i}^{1/2} \sigma_{j,j}^{1/2})$ . Hence

$$\sigma_{i,j}^\diamond = m(\mathbb{E}(U_{1,i}U_{1,j}) + \mathbb{E}(U_{2,i}U_{2,j}) - \mathbb{E}(U_{1,i}U_{2,j}) - \mathbb{E}(U_{2,i}U_{1,j}))/2 = \sigma_{i,j} + O(m^{-\beta/(\beta+1)} \sigma_{i,i}^{1/2} \sigma_{j,j}^{1/2}).$$

Together with (81) we obtain the first part in (80).

Since  $\hat{\sigma}_{i,j,k} = m(\Delta f_{k,i} + \Delta U_{k,i})(\Delta f_{k,j} + \Delta U_{k,j})/2$ , we have  $\mathbb{E}\hat{\sigma}_{i,j,k} = m\Delta f_{k,i}\Delta f_{k,j}/2 + \sigma_{i,j}^\diamond$ . By (73) and (81),

$$v_{i,j}^2 = \sum_{k \notin \mathcal{S}} \mathbb{E}\hat{\sigma}_{i,j,k}^2 / N_2 - \tilde{\sigma}_{i,j}^2 = \sum_{k \notin \mathcal{S}} \text{Var}(\hat{\sigma}_{i,j,k}) / N_2 + O(m^3 n^{-2} \sigma_{i,i}^{1/2} \sigma_{j,j}^{1/2}).$$

Note by (74) and (76) we have

$$m^2 \text{Var}(\Delta f_{k,i} \Delta U_{k,j}) = O(m^3 n^{-2} \sigma_{j,j}), \quad \text{and} \quad m^2 \text{Var}(\Delta U_{k,i} \Delta U_{k,j}) = O(\sigma_{i,i} \sigma_{j,j}).$$

Thus  $\text{Var}(\hat{\sigma}_{i,j,k}) = O(\sigma_{i,i}\sigma_{j,j})$  and the second part in (80) holds.

Step 4. Since  $|\mathcal{S}| \leq 2K_0$ , for any  $i, j$ , and  $|\phi|_\infty \leq \log(2)$ ,

$$|N_1 h_{i,j}(u)/N_2 - \bar{h}_{i,j}(u)| \leq 2\log(2)K_0/(\alpha_{i,j}N_2). \quad (82)$$

Combining (82), Step 1 and step 2 with  $x = N_2/\log^3(np)$ , then with probability tending 1, for all  $1 \leq i, j \leq p$ , and  $|u - \sigma_{i,j}| \leq C_0$ ,

$$B_{i,j}^-(u, \Delta) \leq N_1 N_2^{-1} h_{i,j}(u) \leq B_{i,j}^+(u, \Delta), \quad (83)$$

where

$$\Delta = h\sigma_{i,i}^{1/2}\sigma_{j,j}^{1/2} + 2\log(2)K_0/(\alpha_{i,j}N_2) \quad \text{and} \quad h = \log(np)^{-3}. \quad (84)$$

Note if

$$\alpha_{i,j}^2 v_{i,j}^2 + 2\alpha_{i,j}\Delta \leq 1, \quad (85)$$

then  $B_{i,j}^+(u, \Delta)$  exists real roots. Denote the smaller one as  $u^+$ , which satisfies  $u^+ \leq \bar{\sigma}_{i,j} + \alpha_{i,j}v_{i,j}^2 + 2\Delta$ . Take  $\alpha_{i,j}^* = \alpha_{i,j}\sigma_{i,i}^{1/2}\sigma_{j,j}^{1/2}$ . By Step 3 and Assumption 2.3, if (85), then

$$\sigma_{i,i}^{-1/2}\sigma_{j,j}^{-1/2}(u^+ - \sigma_{i,j}) = O\left\{m^{-\beta/(\beta+1)} + m^3/n^2 + \alpha_{i,j}^* + h + mK_0/(\alpha_{i,j}^*n)\right\}. \quad (86)$$

Similar bound can be obtained for  $u^-$  as well. When (83) holds,  $u^- \leq \hat{\sigma}_{i,j} \leq u^+$ . Take  $\alpha_{i,j}^* = (m/(bn))^{1/2}$ , then with probability greater than  $1 - \log(n)^{-1}$ ,

$$\sigma_{i,i}^{-1/2}\sigma_{j,j}^{-1/2}|\hat{\sigma}_{i,j} - \sigma_{i,j}| \lesssim m^{-\beta/(\beta+1)} + m^3/n^2 + (m/(bn))^{1/2} + \log(np)^{-3},$$

where the convergence is uniform for all  $1 \leq i, j \leq p$ . Since there exists some constant  $c_1, c_2 > 0$ , such that  $c_1 \leq \bar{\sigma}_{j,j}/\sigma_{j,j} \leq c_2$ , and any  $i, j$  with probability tending to 1. Thus the desired result follows.  $\square$

*Proof of (ii).* Same argument as for (i), except that we need to replace Step 2 by Step 2' with  $x \ll N_2/\log(np)^{2.5}$ . Then we obtain the desired result.

Step 2'. This step shows

$$\sum_{i,j=1}^p \mathbb{P}\left(\sup_{|u-\sigma_{i,j}|\leq C_0} |\bar{h}_{i,j}(u) - \mathbb{E}\bar{h}_{i,j}(u)| \geq x(\sigma_{i,i}^{1/2}\sigma_{j,j}^{1/2})/N_2\right) \lesssim p^2 n e^{-cxN_2^{-1/2}}, \quad (87)$$

where  $c$  and the constant in  $\lesssim$  are independent of  $n, p, i, j$ .

The proof follows similar argument as in Step 2 and Theorem 3 in Wu and Wu (2016). □

Figure 5: Visualization of one sample of simulated data with jump in case a),2),ii).

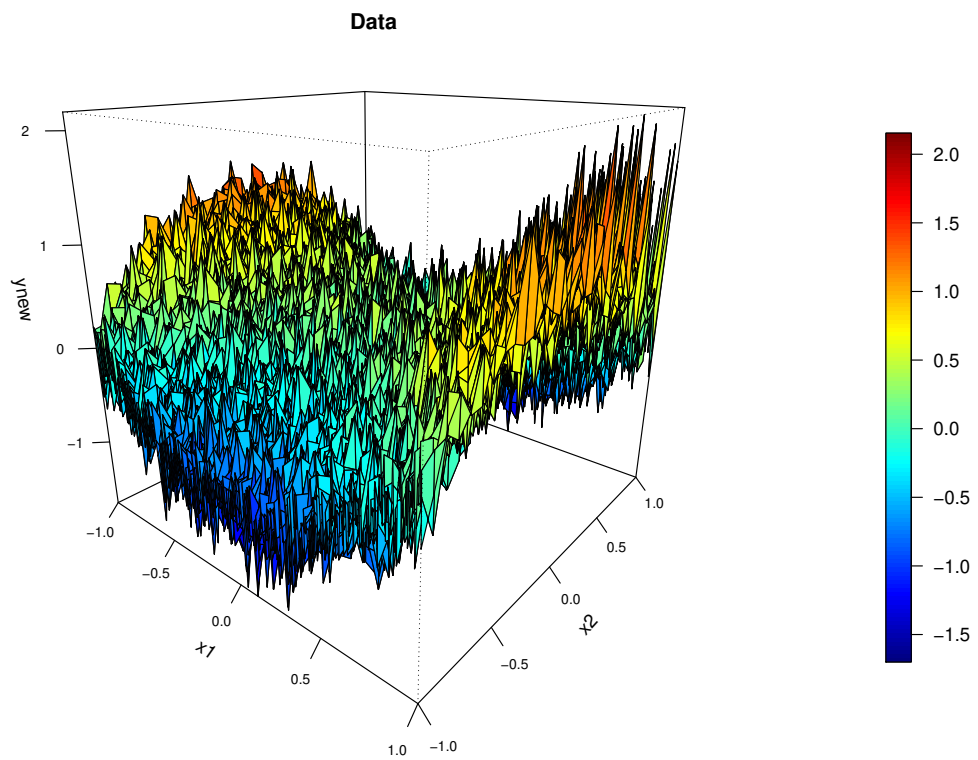


Figure 6: Visualization of the real (left) and estimated correlation matrix (using the robust estimation method).

