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
Wei Biao Wu
Department of Statistics, University of Chicago

October 21, 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. (2020)), 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,

$$Y_t = \quad (t=n) + \quad _t; \quad t = 1; 2; \dots; n; \tag{1}$$

where $(\ _t)_t$ is a sequence of zero-mean p-dimensional stationary noise vectors and $\quad (\ ) = (\ _1(\ ); \ \ _2(\ ); \dots; \ \ _p(\ ))^{>} : [0; 1] \ \ ! \ \ \mathbb{R}^p$ 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 $\quad (t=n)$. Notably, we assume that the trend function satisfies

$$(u) = f(u) + \sum_{i=1}^{K_0} \ _i \mathbf{1}_{u \ \ u_i}; \tag{2}$$

where K_0 is an unknown integer representing the number of breaks; $f(\ )$ ($f(\ ) = (f_1(\ ); f_2(\ ); \dots, f_p(\ ))^{>} : [0; 1] \ \ ! \ \ \mathbb{R}^p$) is a vector of smooth trend functions; u_ks with $0 < u_1 < u_2 < \dots < u_{K_0} < 1$ are the time stamps of the change-points with $ju_i \quad u_j j \quad b$, where b is the

bandwidth parameter; and $_k$ 2 $\mathbb{R}^p$ are the jump vectors with size $j\ _k j_1$ ($j{:}j_1$ 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 j p, and some coordinates of $_k$ can be zero. Namely, we will focus on the largest jump (i.e., $j\ _k j_1$) 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 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. $j\ _k j_1$; 1 k 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 modified penaliza-

tion procedure, see for example, Killick et al. (2012), and LASSO (Least absolute shrinkage and selection operator) type penalization such as by 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 sparsified binary segmentation as in Cho and Fryzlewicz (2015). Moreover, Fryzlewicz (2018) recently introduces a bottom-up algorithm to overcome the disadvantage of the classical binary segmentation. Besides, Wu and Zhou (2019) propose multiscale abrupt change estimation under complex temporal dynamics.

Detection using the MOSUM (moving sum) statistics is another popular way for multiple change-point analysis; see, for example, Hušková and Slabỳ (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 $p \to \infty$: 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 (cumulative sum) 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). However, Bai (2010) is only concerning cross-sectional independent data. 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 our project is 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 piece-wise constant (i.e. $f \quad 0$). 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 constructs a test and delivers 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 $k \ 2 \ \mathsf{N}$ and a vector $v = (v_1; \vdots \vdots \vdots; v_d)^{>} \ 2 \ \mathsf{R}^d;$ we denote

$jvj_k = (\sum_{i=1}^{d} jv_i j^k)^{1=k}$, $jvj = jvj_2$ and $jvj_1 = \max_{i \ d} jv_i j$. For a matrix $A = (a_{ij})_{1 \ i \ m;1 \ j \ n}$, we define the spectral norm $jAj_2 = \max_{jvj=1} jAvj$ and the max norm $jAj_{\max} = \max_{i;j} ja_{i;j} j$. For a function $f;$ we denote $jf j_1 = \sup_x jf(x)j:$ We set (a_n) and (b_n) to be positive number sequences. We write $a_n = O(b_n)$ or $a_n \ . \ b_n$(resp. $a_n \quad b_n$) if there exists a positive constant C such that $a_n=b_n \quad C$(resp. $1=C \quad a_n=b_n \quad C$) for all large n, and we denote $a_n = o(b_n)$ (resp. $a_n \quad b_n$), if $a_n=b_n \ ! \ 0$ (resp. $a_n=b_n \ ! \ 1$). For two sequences of random variables (X_n) and (Y_n); we write $X_n = o_P(Y_n)$; if $X_n=Y_n \ ! \ 0$ 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,

$$H_0 : \quad {}_1 = \ {}_2 = \cdots = \ {}_{K_0} = 0;$$

which corresponds to the case of no breaks, against the alternative of the existence of at least one break i.e. $H_A : 9 \ k \ 2 \ f1; \quad ; K_0 g; \ s.t. \quad {}_k \ 6= 0$. 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 (u) can be disentangled into two parts, namely a smooth transition part $f(u)$ and a jump part $_i\mathbf{1}_{u\ u_i}$. We can define the jump vector at point u as a gap between the right-side function $^{(r)}(u)$ and the left-side function $^{(l)}(u)$, which is

$$J(u) = \ ^{(r)}(u) \quad ^{(l)}(u); \quad \text{where we define} \quad ^{(r)}(u) = \lim_{t\#u} \ (t) \quad \text{and} \quad ^{(l)}(u) = \lim_{t"u} \ (t):$$

Due to the smoothness of the constitutes of $f(:)$; the gap function $J(u) = 0$ when there is no jump, and $J(u) = \ _k$ 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 $\hat{}^{(r)}(u)$ and $\hat{}^{(l)}(u)$; which are estimates of $^{(r)}(u)$ and $^{(l)}(u)$: We propose to adopt the local linear estimation technique, see Fan and Gijbels (1996).

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

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

with weights

$$w_i = w_{i;b} = w_b(0; i=n); \quad i \quad 1; \quad w_0 = 0: \tag{4}$$

The weight functions are defined as

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

where $K(\cdot)$ is a kernel function and b is a bandwidth with $b \to 0$ and $bn \to \infty$. It is worth noting that the estimator in (3) is equivalent to adopting a one-sided kernel function, i.e. $K(u)\mathbf{1}_{u \geq 0}$ 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 $\hat{J}(i/n) = \hat{\ }_i^{(l)} - \hat{\ }_i^{(r)}$ 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 $p \to \infty$, 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

$$\epsilon_t = \sum_{k=0} A_k \eta_{t-k}, \tag{6}$$

where $\eta_t \in \mathbb{R}^{\tilde{p}}$ are independent and identically distributed (i.i.d.) random vectors with zero mean and an identity covariance matrix. $A_k, k \geq 0$, are coefficient matrices in $\mathbb{R}^{p \times \tilde{p}}$ such that ϵ_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(∞) process is very general and includes many important time series models such as a vector autoregressive moving averages (VARMA) model, i.e.

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

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

Correspondingly, we define the sum of the coefficient matrix to be $S = \sum_k \ {}_0 A_k$. The long run covariance matrix for the error process is

$$\Sigma = SS^{>}: \tag{7}$$

We denote $\Sigma = (\ {}_{i;j}); 1 \quad i; j \quad p;$ and

$$\Lambda = \mathrm{diag}(\ {}^{1=2}_{1;1}; \ {}^{1=2}_{2;2}; \ldots; \ {}^{1=2}_{p;p}): \tag{8}$$

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

$$T_n = \max_{bn+1 \ i \ n \ bn} |V_i|_1; \quad \text{where} \quad V_i = \Lambda \ {}^{1}(\hat{}^{(l)}_i \quad \hat{}^{(r)}_i): \tag{9}$$

We adopt a supreme type of statistics as it shares good property under certain alternatives, see for example Bai and Saranadasa (1996). However, we do not claim the strict superiority of our test statistics. When the majority of locations exhibit simultaneous jumps, an l_2 type statistics tends to have better power.

We exclude Y_i in V_i, because that the weights in front of Y_i would be the same for the right side and the left side estimator, and will be canceled when taking the difference. Note that we consider the normalized statistic as multiplying the jump estimates $\hat{J}(i{=}n) = \hat{}^{(l)}_i \quad \hat{}^{(r)}_i$ by $\Lambda \ {}^{1}$ since the long-run variances $\ {}_{j;j}$ for different coordinates $1 \quad j \quad p$ can be very different. We refer to T_n as an infeasible test statistic since Λ is unknown. The estimation of Λ is deferred to Section 4.

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. EV_i. Intuitively, we can decompose the level of our jump estimator EV_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 the definition of w_i in (4) for $i = 1; 2; : : : ; bn$, and $w_i = 0$ for $i = 0$ and $i > bn$: We denote the location of breaks as $_k = nu_k$ and Ω_i as a set of indices indicating the break locations within the bn neighborhood around time i, namely $\Omega_i = \{k | ji \quad _kj \quad bn; 1 \quad k \quad K_0\}$: If $_i \quad _j = n(u_i \quad u_j) \quad n;$ for any $i; j;$ then for large $n;$ the cardinality of Ω_i is at most one, i.e. $j\Omega_i j \quad 1$: Actually this condition can be relaxed to $\min_{1 \quad i \neq j \quad K_0} j \ _i \quad _j j \quad bn$: For a time point i where $\Omega_i \neq ;$, we define the weighted break sizes to be,

$$d_i = (1 \quad \sum_{t=1}^{ji \quad kj} w_t)\Lambda^{\ 1} \ _k; \quad k = \text{argmin}_{j2 \ _i} ji \quad _j j; \tag{10}$$

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 $\underline{d} = (d_{bn+1}^{>}; d_{bn+2}^{>}; : : : ; d_{n \ bn}^{>})^{>}$. It should be noted that under the null, $\underline{d} = 0$.

We denote the smooth part of the local linear estimate as

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

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.

$$\max_{bn+1 \quad i \quad n \quad bn} j\Lambda^{\ 1}(\hat{f}_i^{(l)} \quad \hat{f}_i^{(r)})j_1 = O(b^2): \tag{11}$$

Given the definition of our model $Y_i = \quad (i=T) + \ _i$; d_i can be expressed as

$$d_i = \mathsf{E}\{\Lambda \ ^1((\hat{\ }_i^{(r)} \quad \hat{\ }_i^{(l)}) \quad (\hat{f}_i^{(r)} \quad \hat{f}_i^{(l)}))\} = \mathsf{E}\{V_i \quad \Lambda \ ^1(\hat{f}_i^{(r)} \quad \hat{f}_i^{(l)})\}: \tag{12}$$

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

$$jV_i \quad d_ij_1 = j\Lambda \ ^1(\hat{f}_i^{(r)} \quad \hat{f}_i^{(l)})j_1 = O(b^2): \tag{13}$$

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

$$V_i \quad \mathsf{E}V_i = \sum_{l=i \ \ bn}^{i \ \ 1} w_i \ _l\Lambda \ ^1 \ _l \quad \sum_{l=i+1}^{i+bn} w_l \ _i\Lambda \ ^1 \ _l: \tag{14}$$

To approximate its distribution, we introduce a scaling matrix for variance of the limit distribution. Recall $S = \sum_k \ _0 A_k$ and define a block matrix $G \ = (G_{i;l})_{bn+1 \ \ i \ \ n \ \ bn;1 \ \ l \ \ n} \ 2$ $\mathsf{R}^{(n \ \ 2bn)p \ \ np}$ with components as $p \quad \tilde{p}$ dimension matrices,

$$G_{i;l} = \begin{cases} w_i \ _l\Lambda \ ^1S; & \text{if} \quad i \quad bn \quad l \quad i \quad 1; \\ w_l \ _i\Lambda \ ^1S; & \text{if} \quad i+1 \quad l \quad i+bn; \end{cases} \tag{15}$$

and elsewhere zero. Let $\underline{z}$ be a Gaussian vector in R^{np} with zero mean and identity covariance matrix. We set $G_{i;}$ to be $(G_{i;1}; G_{i;2}; : : : ; G_{i;n})$: It can be shown that $G_{i;} \ \underline{z}$ has a similar covariance structure as $V_i \quad \mathsf{E}V_i$. We shall use the distribution of $jG_{i;} \ \underline{z}j_1$ to approximate the distribution of $jV_i \quad \mathsf{E}V_ij_1$: 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.,

$$\mathsf{P}(jV_ij_1 \quad u) \quad \mathsf{P}(jd_i + G_{i;} \ \underline{z}j_1 \quad u):$$

We now let the statistics go over all the time points, and recall $T_n = \max_{bn+1 \ i \ n \ bn} jV_i j_1$. Then we shall expect

$$\mathsf{P}(T_n \quad u) \quad \mathsf{P}(j\underline{d} + G \ \underline{z} j_1 \quad u); \tag{16}$$

and equivalently

$$\mathsf{P}(T_n \quad u) \quad \mathsf{P}(j\underline{d} + Z j_1 \quad u); \tag{17}$$

where $Z = (Z_{bn+1}^{>}; Z_{bn+2}^{>}; : : : ; Z_n^{>} \ _{bn})^{>}$ and $(Z_i)_{bn+1 \ i \ n \ bn}$ is a sequence of centered Gaussian vectors in R^p with covariance matrices $\mathrm{Cov}(Z_i; Z_j) = Q_{i;j}$ of the following form:

$$Q_{i;j} = \$_{i;j} \Lambda \ ^1 \Sigma \Lambda \ ^1 \quad \text{and} \quad \$_{i;j} = \sum_{l=1}^{n} w_{ji} \ _{lj} w_{jj} \ _{lj} \mathrm{sign}(i \quad l) \mathrm{sign}(j \quad l): \tag{18}$$

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

$$Q = (Q_{i;j})_{bn+1 \ i;j \ n \ bn} = G \ G \ ^{>}:$$

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

$$Z_i \overset{d}{=} G_{i;} \underline{z} \quad \text{and} \quad Z \overset{d}{=} G \ \underline{z}: \tag{19}$$

This transformation from $G \ \underline{z}$ to Z is to show that the involved Gaussian process only depends on the long-run covariance matrix Σ and the weight functions. We note that Z are not element-wise independent, but with dependency governed by G . 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 $_j(u)$ when no break occurs.

ASSUMPTION 2.1. *Function* $f_j \in C^2[0,1]$ *with* $\max_{1 \le j \le p} \|f_j'\|_1 \le c_f$, $\max_{1 \le j \le p} \|f_j''\|_1 \le c_f$ *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* $K(\cdot) \ge 0$ *is symmetric with support* $[-1,1]$*, assume* $\|K\|_1 < 1$ *and* $\int_{-1}^{1} K(x)\mathrm{d}x = 1$*. Also assume* $K(x)$ *has first-order derivative with* $\|K'\|_1 < 1$ *on* $(0,1)$*. Let* $b \to 0$ *and* $bn \to 1$*. Denote* $\mu_i = \int_0^1 x^i K(x)\mathrm{d}x$*. Assume* $\mu_1^2 \ne \mu_2 \mu_0$.

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)* $\sigma_{j,j} \ge c$*,* $1 \le j \le p$ *for some finite constant* $c > 0$*:*

We need enough separation between adjacent breakpoints.

ASSUMPTION 2.4. *(Separation) Assume* $\min_{1 \le i,j \le K_0} |\gamma_i - \gamma_j| \ge bn$*:*

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)* $\max_{1 \le j \le p} \sum_k |A_{k,j,}|_2 = \sigma_{j,j}^{1/2} \le c_s(i-1)^{-}$*; where* > 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 $_t$; $^{0}_{t}$ 2 R^p are i.i.d random vectors with zero mean and covariance matrix I_p. Let

$$_t = F_t + Z_t; \text{ with } Z_t = \sum_{k\ 0} \Lambda_k\ _t\ _k \text{ and } F_t = \sum_{k\ 0} v f_k^{>}\ ^{0}_{t}\ _k; \tag{20}$$

where $\Lambda_k = \mathrm{diag}(\ _{k;1}; : : : ;\ _{k;p})$, $v = (v_1; : : : ; v_p)^{>}$ and $f_k = (f_{k;1}; : : : ; f_{k;p})^{>}$: 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 $\ _{Z;j} = (\sum_{k\ 0}\ _{k;j})^2$ and $\ _{F;j} = j\sum_{k\ 0} f_k j_2^2 v_j^2$, respectively. If for some constant $c > 0$;

$$\sum_{k\ i} j\ _{k;j} j = {}^{1=2}_{Z;j}\ \ ci \quad \text{and} \quad \sum_{k\ i} j f_k j_2 j v_j j = {}^{1=2}_{F;j}\ \ ci\ \ ; \tag{21}$$

then Assumption 2.5 holds with $\ =\ $. To see this, we note $jA_{k;j;}j_2 = (\ ^{2}_{k;j} + j f_k j_2^2 v_j^2)^{1=2}$; and $\ _{j;j} =\ ^{2}_{Z;j} +\ ^{2}_{F;j}$: Hence,

$$\sum_{k\ i} jA_{k;j;}j_2 \quad \sum_{k\ i}(j\ _{k;j}j + jv_j jj f_k j_2) \quad ci \quad (\ ^{1=2}_{Z;j} +\ ^{1=2}_{F;j}) \quad p\overline{2}ci \quad ^{1=2}_{j;j}:$$

ASSUMPTION 2.6. *(Finite moment) The innovations* $_{i;j}$ *are i.i.d. with* $_q = k\ _{1;1}k_q < 1$ *for some* $q\ \ 4$*:*

ASSUMPTION 2.7. *(Sub-exponential) The innovations* $_{i;j}$ *are i.i.d. with* $_e = \mathsf{E}e^{a_0 j\ _{1;1}j} <$ *1 ; 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)**.** *Under Assumptions 2.1-2.5,*

(i) if Assumption 2.6 holds, and $np(bn)^{q=2}(\log(np))^{3q=2} = o(1)$*; then for*

$$\Delta = (bn)^{1=6}\log^{7=6}(pn) + (bn)^{=(3+3)}\log^{2=3}(np) + b^{5=2}n^{1=2}\log^{1=2}(np);$$

we have

$$\sup_{u2R}\left|P(T_n \quad u) \quad P(j\underline{d} + Zj_1 \quad u)\right| \;.\; \Delta + ((np)^{2=q}=(bn))^{1=3}\log(pn);$$

(ii) if Assumption 2.7 holds, and $(bn)^{1}(\log(np))^{\max f7;2(1+)= g} = o(1)$*; we have*

$$\sup_{u2R}\left|P(T_n \quad u) \quad P(j\underline{d} + Zj_1 \quad u)\right| \;.\; \Delta;$$

where the constants in . are independent of $n; p; b$*.*

If in addition

$$b^5 n\log(np) = o(1); \tag{22}$$

then under both cases, we have

$$\sup_{u2R}\left|P(T_n \quad u) \quad P(j\underline{d} + Zj_1 \quad u)\right| \;!\; 0: \tag{23}$$

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 $_1 > 0$ and $0 < _2 < 1=2$; assume $p \quad n^{\ 1}$ and $b \quad n^{\ 2}$: If $_1 < (1 \quad _2)q=2 \quad 1$ and $_2 > 1=5$; then (23) holds. 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 $_1 > 0$ and $1=5 < _2 < 1$; we can set $p \quad e^{n^{\ 1}}$ and $b \quad n^{\ 2}$: If $_1 < 5 \ _2 \quad 1$ and $_1 \max\{7; 2(1 +)= g < 1 \quad _2$; then (23) holds.

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 $\in (0; 1)$; we have the critical value of our test as q i.e. the quantile of the Gaussian limit distribution,

$$q = \inf_{r \ 0}\{r : \mathrm{P}(|Z|_1 > r) \quad g: \tag{24}$$

As from the Gaussian approximation result in (23), we have the approximated sizes of the test statistics,

$$\left|\mathrm{P}(T_n > q \ ) \quad \mathrm{P}(|Z|_1 > q \ )\right| \ ! \ 0:$$

We shall reject the null hypothesis at the significant level ; if the test statistics exceed the critical value i.e. $T_n > q$:

To evaluate our testing power, consider the alternative that if not all $_k = 0$; 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 satis es*

$$= \mathsf{P}(j\underline{d} + Zj_1 \quad q\ ) + o(1):$$

Thus, we can see that the power of our test would depend on the vector $\underline{d}$. The size of it is determined by the true jump sizes i.e. $_k$s. Since the covariance matrix for Z is $Q = (Q_{i;j})$; where $Q_{i;j} = \$_{i;j}\Lambda^{-1}\Sigma\Lambda^{-1}$ with $\$_{i;j}$ defined in (18). It can be calculated that $\$_{i;i} \quad (bn)^{-1}$, therefore $jZj_1 = O_{\mathsf{P}}((bn)^{-1=2}\log(np))$; which tends to zero by Assumption 2.6 and 2.7. Thus if $j\underline{d}j_1 \quad (bn)^{-1=2}\log(np)$; $\quad ! \ 1$ by Corollary 1.

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_1 aggregation. Namely, in our first step, we conduct a "rough" estimation though 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 kth breakpoint, we can aggregate data by a weighted sum of different coordinates whose weights are determined by the first step jump size estimators ($\hat{}_k$). 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

$$j\Lambda \ ^1 \ _k j_1 :$$

Here, we normalize $\ _k$ 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

$$= \min_{1 \ k \ K_0} j\Lambda \ ^1 \ _k j_1 : \tag{25}$$

In the following, we outline the steps of our testing, detecting and estimation procedure. **Step 1.** For significance level , we test the existence of jumps based on the critical value q in (24). If we find no significant breaks, then we cannot reject the null H_0. In case our test statistic exceeds the critical value, we reject 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 $jV\ j_1$ exceeding a threshold value w^y, namely, $A_1 = fbn + 1 \quad n \quad bn : jV\ j_1 > w^y g;$ where V is defined in (9). Let $\hat{}_1$ be the time point in A_1 that maximizes the test statistics $jV\ j_1$. We further eliminate a $2bn$ neighborhood of time points around $\hat{}_1$ from A_1 to create A_2. Then we find the next point in A_2 that maximize $jV\ j_1$, and repeat the same operation until the set A_k is empty. Namely, for $k \quad 1$, we let the kth estimated break point be denoted as $\hat{}_k = \text{argmax}_{\ 2A_k} jV\ j_1$ and $A_{k+1} = A_k\ n\ f \quad : j \quad \hat{}_k j \quad 2bng:$ We denote the maximum number of breakpoints as $\hat{K}_0$, with $\hat{K}_0 = \max_{k\ \ 1} fk : A_k\ 6= \ ;g:$ It is worth noting that we have chosen $2bn$ to exclude both bn neigborhood of and $\hat{}$.

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

$$\hat{}_k = \hat{}^{(l)}_{k\ \ M} \quad \hat{}^{(r)}_{k+M} \quad \text{and} \quad \hat{} = \min_{1\ \ k\ \ \hat{K}_0} j\Lambda\ \ ^1\hat{}_k j_1 : \tag{26}$$

It is worth noting that in this algorithm, we only need to calculate the gap statistics $jV\ j_1$ 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 $jV\ j_1$ for all the points from $bn + 1$ to $n \quad bn$ and select the points that are beyond the threshold w. Intuitively, the points in 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. *Assume the break size satis es* $\max\{\sqrt{\log(pn)=(bn)};\ b\}$:

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 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 $!^{y}$ 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), Assumption 3.1, and* (22) *hold. If* $\min f \quad !^{y}; !^{y}g \quad 2c_w^{0}\sqrt{\log(pn)=(bn)}$*; where* c_w^{0} *is the constant de ned as the limit* $(bn\sum_{i=0}^{n} w_i^2)^{1=2} \ ! \ c_w^{0}$*, then*

(i) $\mathrm{P}(\hat{K}_0 = K_0) \ ! \ 1$:

(ii) under Theorem 1 (i), $j\hat{}_k \quad {}_k j = O_{\mathrm{P}} f(np)^{2=q} = {}^{2}g$*; and under Theorem 1 (ii),* $j\hat{}_k \quad {}_k j = O_{\mathrm{P}} f\log^2(np) = {}^{2}g$*; uniformly over k; where* $k = \mathrm{argmin}_i j\hat{}_k \quad {}_i j$.

(iii) $j\Lambda^{1}(\hat{}_k \quad {}_k)j_1 = O_{\mathrm{P}}((bn)^{1=2}\log(np)^{1=2} + b)$*; uniformly over k; which indicates* $j\hat{} \quad j = O_{\mathrm{P}}((bn)^{1=2}\log(np)^{1=2} + b)$:

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 $u_k = {}_k=n$ and (iii) shows that the break sizes can be consistently recovered. The convergence rate of the break sizes dependents on the bandwidth b, sample size n and the dimension of the time series p. It is worth noting that the bias is of order b in (iii), as the difference is taken with a gap of $2M$ as in equation (26). It shall be noted that the consistency rate of $\hat{}_k$ depends on the break size , 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. With respect to the condition $\min\{\quad !^{y};\ !^{y}g\quad 2c_w^{0}\sqrt{\log(pn)=(bn)};$ relative to the summary in the Table 1 in Cho (2016), our break size $\quad$ is comparable up to the weakest condition $\quad (nb)^{1=2}\ !\ 1$ up to a logarithmic factor. Moreover, when $p = 1$, our requirement of breaksize is similar to the rate as in Theorem 3.2 in Wu and Zhou (2019), namely $\quad p\overline{\log n}=p\overline{nb}.$

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 R^p with zero mean and covariance matrix

$$\tilde{Q} := Q_{bn+1;bn+1} = 2\sum_{t=1}^{bn} w_t^2 \Lambda^{\ 1}\Sigma\Lambda^{\ 1}: \tag{27}$$

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

$$\sup_{u2\mathsf{R}} j\mathsf{P}(j\Lambda^{\ 1}(\hat{}_k\quad {}_k)j_1\quad u)\quad \mathsf{P}(j\tilde{Z}j_1\quad u)j\ !\ 0;\ \text{where } k\ = \operatorname{argmin}_i j\hat{}_k\quad {}_i j:$$

This theorem indicates that the maximum of the difference between the estimated jump size $\hat{}_k$ and the true jump size ${}_k$ 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 joint confidence interval for ${}_{k;j}$. We set

$$= \mathsf{P}(j\tilde{Z}j_1\quad q)\quad \text{and}\quad = (\ {}^{1=2}_{1;1};\ {}^{1=2}_{2;2};\ldots;\ {}^{1=2}_{p;p})^{>};\tag{28}$$

some $\quad > 0$: Then as $n\ !\ 1$ with probability close to 1 $\quad$; we have

$$q\ {}^{1=2}_{j;j} + \hat{}_{k;j}\quad {}_{k;j}\quad q\ {}^{1=2}_{j;j} + \hat{}_{k;j}\quad 8j: \tag{29}$$

Theorem 3 can be extended to hold uniformly over k by stacking the statistics over all ks. In addition, we see that Theorem 2 and Theorem 3 are closely connected in the sense that we can reach the same threshold by stacking $\hat{}_k$ over all ks.

3.2 The refined aggregation step

The estimation in the first step is only driven by $j_{k}j_{1}$, i.e. the maximum size of jumps at a time point $_{k}$. Therefore it is only sensitive to the biggest jump across all the time series at the same time. The l_{1} type test potentially have more power for certain alternatives than the l_{2} type statistics, see for example Bai and Saranadasa (1996). However, if the majority or all of the entities exhibit simultaneous jumps, the supremum statistic tends to have lower power than the l_{2} statistic.

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 $\hat{}_{k}$ 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 S_{k} to be the set of series that jump at location $_{k}$; that is

$$S_{k} = f1 \quad j \quad p\, j \quad _{k;j} \; 6= 0g; \tag{30}$$

where $_{k;j}$ is the jth coordinate of $_{k}$: Detailed steps of the aggregation are formulated as follows:

Stage 1. Apply Steps 1-3 in Subsection 3.1 to obtain $\hat{}_{k}$ and $\hat{}_{k}$; $k = 1; 2; : : : ; \hat{K}_{0}$: For some $w^{y} > 0$; let the estimation of S_{k} be

$$\hat{S}_{k} = \{1 \quad j \quad p\,|j(\Lambda^{\ 1}\hat{}_{k})_{j}j \quad w^{y}\}: \tag{31}$$

In practice, w^{y} 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 $jt \quad \hat{}_k j \quad 2bn;$ we let

$$X_t = \sum_{j 2 \hat{S}_k} (\Lambda \ {}^{1} \hat{}_k)_j (\Lambda \ {}^{1} Y_t)_j : \tag{32}$$

Note that after the modification, for all the jump locations, the new time series X_t would only contain positive sized jumps i.e. $\sum_{j 2 \hat{S}_k} (\Lambda \ {}^{1} \hat{}_k)_j^2$.This step can be understood as a projection of the high-dimensional observations $\Lambda \ {}^{1} Y_t$ according to the direction of $\Lambda \ {}^{1} \hat{}_k (j \ 2 \ \hat{S}_k)$. 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 $k = 1; 2; : : : ; \hat{K}_0;$

$$\tilde{}_k = \underset{jt \ \hat{}_k j \ bn}{\operatorname{argmax}} \Big(\sum_{s=\hat{}_k \quad 2bn}^{\hat{}_k + 2bn} X_s \frac{t \quad \hat{}_k + 2bn}{4bn + 1} \quad \sum_{s=\hat{}_k \quad 2bn}^{t \ \ 1} X_s \Big) \sqrt{\frac{4bn + 1}{(t \quad (\hat{}_k \quad 2bn) + 1)(\hat{}_k + 2bn \quad t)}} : \tag{33}$$

After we update the break points estimation, we can construct confidence intervals for the updated breakpoints estimates $\tilde{}_k$. We denote the long-run correlation matrix to be $(\tilde{}_{i;j})_{i;j} = \Lambda \ {}^{1} \Sigma \Lambda \ {}^{1};$ where Σ is the long run covariance matrix for $\ {}_t$. We let $\tilde{\Sigma}_k = (\tilde{}_{i;j})_{i;j 2 S_k}$ be the sub covariance matrix corresponding to coordinates in S_k at time $\ {}_k$ and let the standardized significant break sizes $\tilde{}_k = (\Lambda \ {}^{1} \ {}_k)_{i 2 S_k} :$ We define two objects involved in the limit distributions of the breaks, i.e.,

$$a_k = j \tilde{}_k j_2^2 \quad \text{and} \quad \&_k^2 = \tilde{}_k^{>} \tilde{\Sigma}_k \tilde{}_k : \tag{34}$$

Then $\&_k^2$ is the long-run variance for the sequence $\sum_{j 2 S_k} (\Lambda \ {}^{1} \ {}_k)_j (\Lambda \ {}^{1} \ {}_t)_j$. For the aggregated jump estimation, we alternatively define the minimum jump size across different

locations and time points as

$$^{y} = \min_{1 \ k \ K_0} \min_{j 2 S_k} j(\Lambda^{\ 1}{}_{k})_j j:$$

Then ^{y} and it functions similarly as to capture the identifiable jump size of the time series. We shall put the same assumption on ^{y} as on . It is worth noting that ^{y} is the minimum jump size to ensure the consistency of our break estimation.

ASSUMPTION 3.2. *Let* ^{y} $\max\left\{\sqrt{\log(pn)=(bn)}; b\right\}$:

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* ^{y}=2 w^{y} $(bn)^{\ 1=2}\log(np)^{1=2} + b$; *then we have*

$$\mathsf{P}(\hat{S}_k = S_k; \ 1 \quad k \quad K_0) \ ! \ 1:$$

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

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

$$c_1 \quad \max(\Lambda^{\ 1}\Sigma\Lambda^{\ 1}) = \min(\Lambda^{\ 1}\Sigma\Lambda^{\ 1}) \quad c_2: \tag{35}$$

Recall de nition of a_k *and* $\&_k$ *in* (34). *Then we have for any xed* $1 \quad k \quad K_0$;

(i) $j\tilde{}_k \quad {}_k j = O_{\mathsf{P}}(\&_k^2=a_k^2)$:

(ii) *In addition, if Assumption 2.5 holds with* > 1, *and* $1 \quad \&_k^2=a_k^2 \quad bn$; *then we have*

$$\tilde{}_k \quad {}_k \ \overset{D}{!} \ (\&_k=a_k)^2 \mathrm{argmax}_r(\ 2\ ^{1}jrj + \mathsf{W}(r));$$

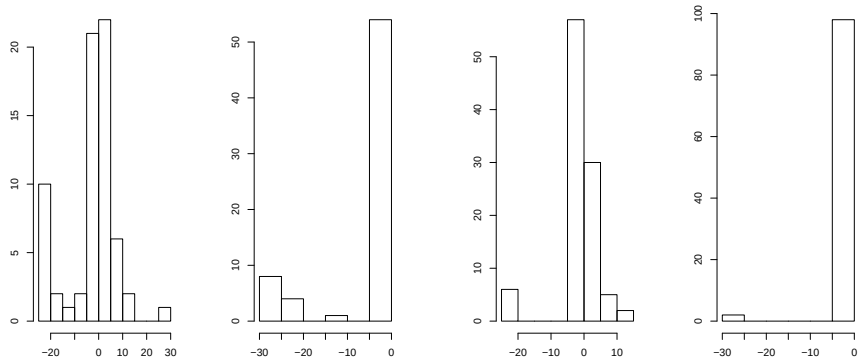
where $\mathsf{W}(r)$ *is a two-sided Brownian motion. That is* $\mathsf{W}(r) = \mathsf{W}_1(r)$; *if* $r > 0$, *and* $\mathsf{W}(r) = \mathsf{W}_2(\ r)$; *if* $r \quad 0$. $\mathsf{W}_1; \mathsf{W}_2$ *are two independent Brownian motions.*

REMARK 2. We shall note that the consistency rate of $\tilde{}_k$ is improved compared to the results for $\hat{}$ 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 $\tilde{}_k$. For instance, if we assume that there are s breaks which are of size > 0 in the cross-sectional dimensional, then $a_k = s^{2}$. If moreover there is no cross-sectional correlation, i.e., $\tilde{\Sigma}_k = I$, then we may expect $\tilde{}_k$ to be consistent so long that $1=(s^{2}) \to 0$, while $\hat{}_k$ can be not consistent. Thus the rate of $\tilde{}_k$ will be better than $\hat{}$. 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 $\&_k^2 \quad j\tilde{\Sigma}_k j_2 a_k$, if $j\tilde{\Sigma}_k j_2 = a_k = o(1)$ then by Theorem 4 (i), we have $\tilde{}_k \to {}_k$ in probability. This corresponds to the insight of Bai (2010) and Hansen (2000). We can also see that when the breaks are truly sparse in the cross sectional dimension or the break size for each time series is very small, the l^2 aggregation cannot improve the performance compared to the previous step. Also when there are strong cross-sectional dependence l_2 aggregation will not improve the break estimation performance. Moreover, we also need the aggregated breaksize to shrink to zero $(1 \quad \&_k^2 = a_k^2)$ to obtain the limit distribution. If $\tilde{\Sigma}_k$ is a d-banded matrix, $j\tilde{\Sigma}_k j_2 \quad (j\tilde{\Sigma}_k j_1 j\tilde{\Sigma}_k j_1)^{1=2} \quad d$: We can derive that $j\tilde{}_k \quad {}_k j = O_P(d=a_k)$:

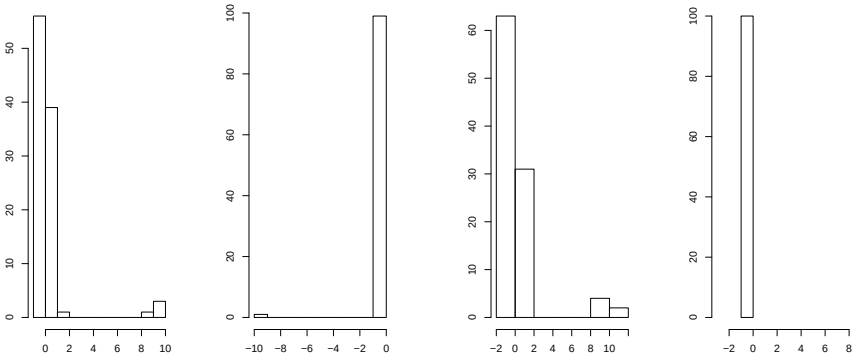
To illustrate the insight of Remark 2, we compare the performance of a simple model with $N(0; 0{:}1)$ and one breakpoint placed at ${}_0 = 50$. Figure 3.2 shows the histogram of $\hat{} \quad {}_0$ and $\tilde{} \quad {}_0$ respectively. The jump-size for all breaks are the same, with the value $1{:}6\sqrt{\log(np)}({}_0)^{1=3}$. As the dimension p grows, we see the significant improvement of the performance of $\tilde{}$ relative to that of $\hat{}$.

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

$$\left(\tilde{}_k \quad b\hat{q}'_1 \quad {}_{=2}c \quad 1;\ \tilde{}_k + b\hat{q}' \quad {}_{=2}c + 1\right); \tag{36}$$



(a) $p = 10$ (b) $p = 30$



(c) $p = 100$ (d) $p = 150$

Figure 1: Histogram of $\hat{\quad}_0$(left) and $\tilde{\quad}_0$(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 $p = 150; s = 30$.

where $q_{1 \quad =2}^{\emptyset}$ ($q^{\emptyset}{}_{=2}$) is 1 $\quad$ =2 (=2)th quantile of the limit distribution of the break point $\tilde{}_k$, i.e. $\text{argmax}_r f \;\; 2 \;\; {}^1 a_k jrj + \&_k W(r) g$ and $\hat{q}^{\emptyset}{}_{=2}(\hat{q}_{1 \quad =2}^{\emptyset})$ are estimates of the quantiles. $b\, c$ denotes the floor function. $q_{1 \quad =2}^{\emptyset}$ ($q^{\emptyset}{}_{=2}$) 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 Σ 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 (2019) 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 (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).

First of all, to account for temporal dependency, we group our observations into blocks of the same size m, for some $m \; 2 \; \mathbb{N}$. We denote the number of blocks $N_1 = b(n \quad m)=mc;$ and the observation indices within a block k is set to be $A_k = f t \; 2 \; \mathbb{N} : km+1 \quad t \quad (k+1)mg$, and we let

$$_k = \sum_{t2A_k} Y_t=m;$$

be the average observations within the block A_k: Without jumps, a natural estimate of the

long-run covariance matrix is

$$\sum_{k=1}^{N_1}(m=2)(\ _k \quad \ _{k\ 1})(\ _k \quad \ _{k\ 1})^{>}=N_1:$$

Note that we take the difference $\ _k \quad \ _{k\ 1}$ to cancel out the trends, as the trend function $(\ )$ is smooth, and the aggregated difference between two consecutive blocks can be shown to be of order $m=n$, which vanishes when $m=n\ !\ 0$. However, this estimator can be greatly contaminated by the jumps, as jumps are not smooth and cannot be canceled out by taking difference. 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 $\ _k = (\ _{k;1};\ _{k;2};:::;\ _{k;p})^{>}$ and let

$$\hat{}_{i;j;k} = m(\ _{k;i} \quad \ _{k\ 1;i})(\ _{k;j} \quad \ _{k\ 1;j})=2; \quad k = 1;2;:::;N_1: \tag{37}$$

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

$$h_{i;j}(u) = \sum_{k=1}^{N_1} \quad _{i;j}(\hat{}_{i;j;k} \quad u)=N_1; \tag{38}$$

where $(x) = \ ^1\ (\ x)$ and

$$(x) = \begin{cases} \log(2); & x \quad 1; \\ \log(1 \quad x + x^2=2); & 0 \quad x \quad 1; \\ \log(1 + x + x^2=2); & 1 \quad x \quad 0; \\ \log(2); & x \quad 1: \end{cases} \tag{39}$$

REMARK 3. Function $j\ (\ )j$ 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,

$$\log(1 \quad x + x^2=2) \qquad (x) \quad \log(1 + x + x^2=2): \tag{40}$$

We set the estimates of the components of the long-run covariance matrix $\hat{}_{i;j}$ 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,

$$\hat{\Sigma} = (\hat{}_{i;j})_{1 \ \ i;j \ \ p} \quad \text{and} \quad \hat{\Lambda} = \text{diag}(\hat{}^{1=2}_{1;1}; \hat{}^{1=2}_{2;2}; \ldots; \hat{}^{1=2}_{p;p}): \tag{41}$$

We denote $^{-}_{i;i} = 2\sum_{N_1=4 \ \ k \ \ 3N_1=4} \hat{}_{i;i;k}=N_1$ and let the $\ _{i;j}$ in (38) be $^{-1=2}_{i;i}\,^{-1=2}_{j;j}(m=n)^{1=2}$:

Theorem 5. *(Long-run variance precision) We assume that Assumption 2.5 holds with* 1:5 *and let*

$$\& = j\Lambda \ ^1(\hat{\Sigma} \quad \Sigma)\Lambda \ ^1j_{\max}:$$

Then for K_0 nite, we have $\& = O_P(n \ ^{1=4}\log(np))$ under either one of the following two conditions:

(i) Assuming conditions in Theorem 1 (i), $p \quad cn^v$ with $v < q=8 \quad 1=2$ and some $c > 0$; we take $m = \min fn^{1 \ \ 8v=(q \ \ 4)}; n^{1=2}g$:

(ii) Assuming conditions in Theorem 1 (ii), we take $m = n^{1=2}$:

By the above theorem, for the diagonal values, we have $\max_{1 \ \ i \ \ p} j\hat{}_{i;i} \quad _{i;i}j= \ _{i;i} = o_P(1)$: Let $\hat{Q}$ be the same as Q in (18), with Σ 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, $j\hat{Z} + dj_1$ converges to $jZ + dj_1$ 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 ($n = 513$), 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. We select b according to a cross validation method, $m = 10$, and $_{i;j} = {}^{-1=2}_{i;i}{}^{-1=2}_{j;j}(m=bn)^{1=2}$ which varies over different i,j. We have used the estimated 0:999 quantile of the maximum of the Gaussian random variables (as in equation (19) with correlation matrix replaced by its estimator), which in our case is estimated as 2:10. We refer to the guidance of the selection of tuning parameters as in Remark 4 in the Supplementary materials.

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 $\tilde{}_k$ using the CUSUM statistics in Equation (33) pick up the breaks earlier than the estimates obtained from the non-aggregated method i.e. $\hat{}_k$. We can see that our method can identify important dates such as the financial crisis period starting in Jan, 2009. Moreover, $\tilde{}_k$ 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 1975, 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

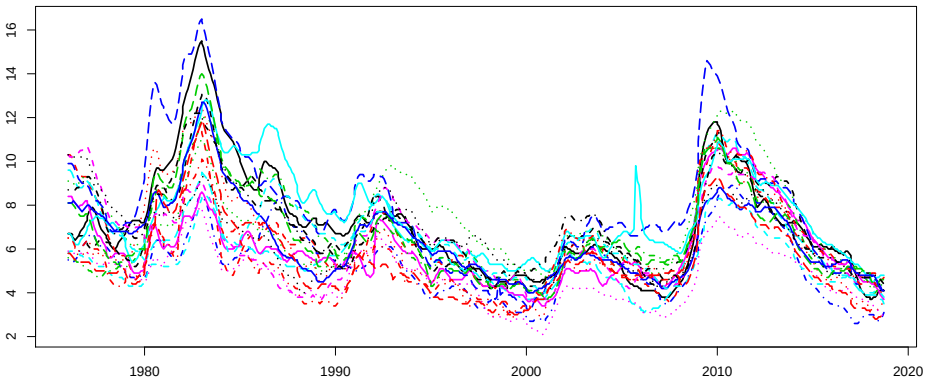


Figure 3: Plot of estimation of the robust long-run correlation matrix; $m = 10$.

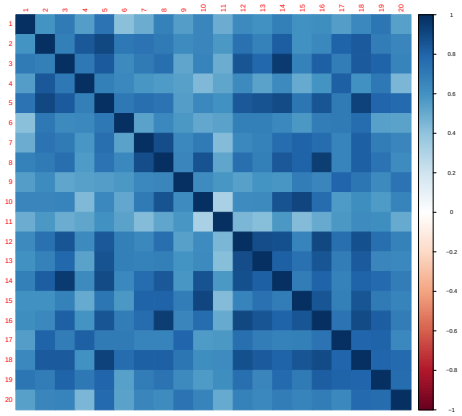
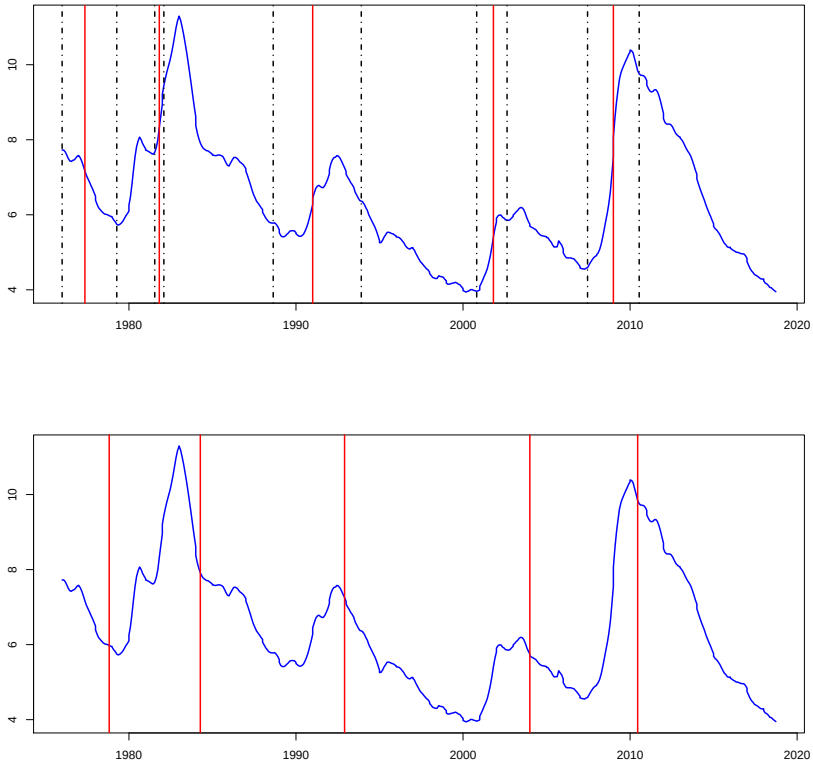


Figure 4: Plot of estimated breakpoints $\tilde{}_k(\hat{}_k)$ (red lines) and their confidence intervals (dotted black lines). $\tilde{}_k$ (upper panel), $\hat{}_k$ (lower panel). The blue time series line represents the average unemployment rate over states under consideration.



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.

Bai, Z. and H. Saranadasa (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica*, 311–329.

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. *Asterisque* (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 Poincare Probab. Stat. 48* (4), 1148–1185.

Chen, L., W. Wang, and W. B. Wu (2020). Dynamic semiparametric factor model with structural breaks. *Journal of Business & Economic Statistics*, 1–15.

Chen, L. and W. B. Wu (2019). Testing for trends in high-dimensional time series. *Journal of the American Statistical Association 114* (526), 869–881.

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.

Hansen, B. E. (2000). Sample splitting and threshold estimation. *Econometrica 68* (3), 575–603.

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. Slabỳ (2001). Permutation tests for multiple changes. *Kybernetika 37* (5), 605–622.

Imbens, G. W. and T. Lemieux (2008). Regression discontinuity designs: A guide to practice. *Journal of econometrics 142* (2), 615–635.

Jackson, B., J. D. Scargle, D. Barnes, S. Arabhi, A. Alt, P. Gioumousis, 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 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. and Z. Zhou (2019). Mace: Multiscale abrupt change estimation under complex temporal dynamics. *arXiv preprint arXiv:1909.06307* .

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. We define $u_{it} = \ _i(t{=}T)$. The discrete version of the model can be written as:

$$y_{it} = u_{it} + \sum_{j=1}^{K_0} \ _{jit}\mathbf{1}_{t \quad _j} + \ _{it}; \tag{42}$$

$i = 1; \quad ; p,\ t = 1; \quad ; n.$

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 covairance estimation.**) We estimate the long-run covariance matrix $\hat{\Sigma} = (\hat{}_{i;j})$ and its diagonal matrix $\hat{\Lambda}$: We first calculate $\hat{}_{i;j;k}$ in (37) and we let $\hat{}_{i;j}$ 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 Σ and Λ therein replaced by $\hat{\Sigma}$ and $\hat{\Lambda}$ respectively.

Step 3 (**Calculating critical values.**) We generate i.i.d. Gaussian vectors $\hat{Z}^{(i)};\ i = 1; 2; : : : N;$ with the covariance matrix $\hat{Q}$; and we obtain $\hat{q}$ which is the empirical $(1 \quad)$ quantile of the $j\hat{Z}^{(i)}j_1$ over several samples and it can be viewed as an estimate of q in (24).

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

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 $j\hat{V}_t j_1$ for $t = bn + 1; bn + 2; : : : ; n \quad bn;$ where $\hat{V}_t$ is the same as V_t in (9) with Λ therein replaced by $\hat{\Lambda}$: Let $A_1 = f \quad : j\hat{V} j_1 > w^{y}g: w^{y}$ can be set as $\hat{q}$ with to be small (e.g. $= 0:0001$).

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 $k \quad 1$, we let $\hat{}_k = \text{argmax}_{\ 2A_k} j\hat{V} j_1$ and $A_{k+1} = A_k n f \quad : j \quad \hat{}_k j \ 2bng:$ Then the estimate of the number of breaks is $\hat{K}_0 = \max_{k \ 1} fk : A_k \neq ;g:$

Step 7 (**Estimating the sizes of breaks**) We construct $\hat{}_k$ as in Step 3 in Subsection 3.1. We set the estimates of the sizes of the jumps as $\hat{}_k = j\hat{\Lambda}^{\ 1}\hat{}_k j_1$ and their minimum as $\hat{} = \min_{1 \ k \ K_0} \hat{}_k$:

Step 8 (**Constructing confidence intervals for the sizes**) We construct $\tilde{q}$ as in (28). Let $\hat{} = (\hat{}^{1=2}_{1;1}; \hat{}^{1=2}_{2;2}; ...; \hat{}^{1=2}_{p;p})^{>}$: Then the confidence interval for vector $_k$ at level 2 is $(\hat{}_k \quad \tilde{q} \ \hat{}; \hat{}_k + \tilde{q} \ \hat{})$.

Step 9 (**Aggregated jump location estimation and confidence interval construction**) Construct aggregated jump location estimates $\tilde{}_k$ as in (33). The confidence interval for $_k$ is $(\tilde{}_k \quad x; \tilde{}_k + x)$; where x is the $1 \quad =2$ quantile of the distribution $\text{argmax}_r(\ 2 \ ^1\hat{a}_k^2 jrj + \&_k W(r))$ and $\hat{a}_k$ (resp. $\&_k$) is a_k (resp. $\&_k$) with Λ; Σ and $_k$ replaced by their estimations.

Remark 4. We notice that there are a few tuning parameters involved in the procedure. The rigorous study of the selection of tuning parameters deserves further research. We

make some suggestions as follows. For b we can use either a plug-in approach or a cross validation method following section 5.2 of Imbens and Lemieux (2008). We select the maximum bandwidth over all locations for each time. $_{i;j}$ is chosen according to the guidance specified below equation (41). m can be set up as $(bn)^{1=2}$ as in Theorem 5 (ii) initially. And later can be updated by a grid search method which minimizing the out of sample prediction error. We suggest to eliminate the breakpoints and its bn neighborhood when we calculate the prediction error.

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 $p = 20; 50; 100$,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) $f_i(u) = \quad i^2=p^2 + u^2$ and b) $f_i(u) = \sin(2 \ \ u + i=p)$. Let $u_{it} = f_i(t=n)$. $_t$ 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 3), 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 $3=\log(np)$.

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) $_{jit} = 0:05$ for $i = 5; 10$ or ii) $_{jit} = (\sqrt{p}=\log(pn))$.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) $(j\hat{K}_0 \quad K_0 j)$ as in Table

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)
	$p = 100; n = 100$		$p = 150; n = 100$	
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)

1. The averaged distances between the breaks $\sum_{k=1}^{\hat{K}_0} \tilde{j}_k \quad {}_k j_1$ (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.

The estimation accuracy with covariance estimation is included in Tables 4 and 5. And the averaged coverage probabilities of the confidence interval for the breaks (AC) are in Table 6 at the confidence level of 90%.

When the tail of error distribution is light, our long run covariance estimator can still be consistent. Table 3 shows the AD and AM/n in cases: $p = 100$ and $n = 80$; $p = 150$ and $n = 100$. We set $s = 5$, $K_0 = 10$. We find that the method is slightly less accurate with the increasing dimension. To improve the estimation accuracy, a further extension of our estimation to regularized high-dimensional long -run variance estimation can be considered. As the sample sizes increase, the estimation precision is improved. We can see that our method is robust against different data simulation scenarios, and we can achieve good level

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)
	$p = 100; n = 100$		$p = 150; n = 100$	
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)

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 7. The estimation precision of the long-run variance-covariance matrix is maintained across different data-generating processes.

Table 3: AD, AM/n, and the maximum norm of covariance matrix estimation error averaged over 1000 samples in different simulation scenarios, and their standard deviations in bracket.

	$p = 100; n = 80$		$p = 150; n = 100$	
	1)	2)	1)	2)
	AD			
a)	0.185 (0.041)	0.190 (0.044)	0.197 (0.058)	0.186 (0.062)
b)	0.189 (0.042)	0.194 (0.057)	0.192 (0.061)	0.187 (0.063)
	AM/n			
a)	0.057 (0.018)	0.082 (0.019)	0.093 (0.024)	0.099 (0.018)
b)	0.064 (0.021)	0.055 (0.026)	0.067 (0.023)	0.079 (0.027)
	Maximum Norm			
a)	0.080 (0.016)	0.075 (0.019)	0.076 (0.021)	0.083 (0.024)
b)	0.078 (0.034)	0.084 (0.041)	0.081 (0.037)	0.082 (0.022)

Table 4: AD averaged over 1000 samples in different simulation scenarios, and their standard deviations in bracket.

		$p = 20; n = 500$		$p = 30; n = 1000$	
		1)	2)	1)	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)
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)

Table 5: AM/n averaged over 1000 samples in different simulation scenarios, and their standard deviations are in brackets.

		$p = 20; n = 500$		$p = 30; n = 1000$	
		1)	2)	1)	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)
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 6: AC in different simulation scenarios over all the estimated break-points and samples.

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

Table 7: Averaged difference between the variance-covariance and the true matrix. (L_1 norm divided by $p(p \quad 1)=2$).

		$p = 20; T = 500$		$p = 30; T = 1000$	
		1)	2)	1)	2)
a)	i)	0.005	0.008	0.003	0.007
	ii)	0.004	0.006	0.003	0.005
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. Then by Fan and Gijbels (1996), the weights of the local linear estimator take the following form*

$$w_i = \frac{{}_2 \quad {}_1 i/(bn)}{{}_2 \; {}_0 \quad {}_1^2} \frac{K(i/(bn))}{bn} + O((bn)^{\quad 2}):$$

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 $K(:)$; such that

$$bn \max_{0 \; i \; n} |w_i| \quad c_w; \quad \max_{|i \; j| \; m} |w_i \quad w_j| \quad c_w \frac{m}{(bn)^2};$$

$$(bn \sum_{i=0}^{bn} w_i^2)^{1/2} \; ! \; c_w'; \text{ and } \frac{bn}{k} \sum_{i=1}^{k} w_i \quad c_w''; k \quad bn: \tag{43}$$

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

$$bn \sum_{i=1}^{k} w_i/k = F(k/(bn)) + O((bn)^{\quad 1}); \quad \text{where} \quad F(t) = \frac{{}_2 \int_0^t K(x)\mathrm{d}x \quad {}_1 \int_0^t xK(x)\mathrm{d}x}{(\; {}_2 \; {}_0 \quad {}_1^2)t}:$$

Define the numerator function as $g(t) = \; {}_2 \int_0^t K(x)\mathrm{d}x \quad {}_1 \int_0^t xK(x)\mathrm{d}x$: We can see that $g(0) = 0, g(1) > 0$; and the derivative function $g'(t) = (\; {}_2 \quad {}_1 t)K(t)$, which is strictly larger than 0 before ${}_2/{}_1$ and less than 0 afterwards. Therefore we have $F(x) > 0$ on $(0; 1]$: In addition, we note $F(0+) = \; {}_2 K(0)/(\; {}_2 \; {}_0 \quad {}_1^2) > 0$ and $F(1) = 1$: Thus $\inf_{t \in (0;1]} F(t) > 0$ in view of $F(t)$ is a continuous function. □

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

$$\|M_T\|_q^{q'} \quad K_q^{q'} \sum_{t=1}^{T} \| \; {}_t\|_q^{q'}; \text{ where } K_q = \max((q \quad 1)^{\quad 1}; \sqrt{q \quad 1}):$$

B.2 Asymptotic results for Gaussian vector

LEMMA 3 (Comparison). *Let* $X = (X_1; X_2; :::; X_v)^>$ *and* $Y = (Y_1; Y_2; :::; Y_v)^>$ *be two centered Gaussian vectors in* R^v *and let* $d = (d_1; d_2; :::; d_v)^>$ *2* R^v*: We denote* $\Delta = \max_{1 \ i;j \ v} j \ {}^X_{i;j} \quad {}^Y_{i;j} j$*; where we de ne* ${}^X_{i;j} = E(X_i X_j)$ *(resp.* ${}^Y_{i;j} = E(Y_i Y_j)$*). Assume that* Y_i*s have the same variance* ${}^2 > 0$*: Then we have*

$$\sup_{x2R} \left| P\big(jX + dj_1 \quad x\big) \quad P\big(jY + dj_1 \quad x\big) \right| \ . \ \Delta^{1=3} \log(v)^{2=3}; \tag{44}$$

where the constant involved in . *only depends on* :

Proof. It suffices to show for any d *2* R^v*;*

$$\sup_x \left| P(\max_{1 \ i \ v}(X_i + d_i) \quad x) \quad P(\max_{1 \ i \ v}(Y_i + d_i) \quad x) \right| \ . \ \Delta^{1=3} \log(v)^{2=3}:$$

To this end, we define

$$F \ (z) = \quad {}^1 \log\Big(\sum_{j=1}^{v} \exp(\ (z_j + d_j)) \Big):$$

Replace the F ( ) in the proof of Theorem 2 in Chernozhukov et al. (2015) by F (z)*:* Then by the argument in equation (10) in Chernozhukov et al. (2015), we have

$$P\Big(\max_{1 \ i \ v}(X_i + d_i) \quad x \Big) \quad P\Big(\max_{1 \ i \ v}(Y_i + d_i) \quad x + \ + \quad {}^1 \log(v) \Big) + c(\quad {}^2 + \quad {}^1)\Delta;$$

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

$$P\Big(\max_{1 \ i \ v}(X_i + d_i) \quad x \Big) \quad P\Big(\max_{1 \ i \ v}(Y_i + d_i) \quad x \Big) \ . \ (\ + \quad {}^1 \log(v)) \sqrt{\log(v)} + (\quad {}^2 + \quad {}^1)\Delta;$$

where the constant in . only depending on *:* Take $= \ {}^1 \log(v)$ and $= \log(v)^{1=6} \Delta^{1=3}$*:* Same argument can be applied in the other direction, and the desired result follows. □

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

$$\sup_{x2\mathsf{R}} \mathsf{P}\Big(\big| jX + dj_1 \quad x \big| \quad e \Big) \quad ce\sqrt{\log(v)}; \tag{45}$$

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

$$I \ := \max_{bn+1 \ i \ n \ bn} \Big| \sum_{t=i \ bn}^{i \ 1} w_{i \ t} \Lambda^{\ 1}{}_{t} \quad \sum_{t=i+1}^{i+bn} w_{t \ i} \Lambda^{\ 1}{}_{t} + d_i \Big|_1 : \tag{46}$$

By (13) and (14) we have

$$jT_n \quad I \ j \quad \max_{bn+1 \ i \ n \ bn} j\mathsf{E}V_i \quad d_i j_1 = O(b^2): \tag{47}$$

Thus we only need to work on I *:* For some $m > 0$*;* let a truncated version of the error term be defined as

$$_{t;m} = \sum_{k=0}^{m \ 1} A_{k \ t \ k}:$$

Consider the m-dependent approximation $I_{;m}$ of I *;* where $I_{;m}$ is I with $_t$ replaced by $_{t;m}$*:* Then we have I $I_{;m}$ for large m*:* Let $I_{z;m}$ be $I_{;m}$ with $_t$ therein replaced by z_t*;* where $(z_t; t \ 2 \ Z)$ are i.i.d. Gaussian vectors with zero mean and identity covariance matrix in R^p*:* Since $I_{;m}$ 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 $I_{;m}$ and $I_{z;m}$ are close. We complete the proof by showing that the distributions of $I_{z;m}$ and $jZ + \underline{d}j_1$ 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). (We note that we use the same order of m in both case i) and ii).). Let $m = (bn)^{1=(1+\ )}$; for any > 0;

$$\mathsf{P}\big((bn)^{1=2}T_n \quad u\big) \quad \mathsf{P}\big((bn)^{1=2}jT_n \quad I_{;m}j \quad \big) + \mathsf{P}\big((bn)^{1=2}I_{;m} \quad u + \ \big)$$

and

$$\mathsf{P}\big((bn)^{1=2}jZ + \underline{d}j_1 \quad u\big) = \mathsf{P}\big((bn)^{1=2}jZ + \underline{d}j_1 \quad u + \ \big) \quad \mathsf{P}\big(u < (bn)^{1=2}jZ + \underline{d}j_1 \quad u + \ \big):$$

Hence

$$\begin{aligned}
&\sup_{u2\mathsf{R}}\Big[\mathsf{P}\big((bn)^{1=2}T_n \quad u\big) \quad \mathsf{P}((bn)^{1=2}jZ + \underline{d}j_1 \quad u\big)\Big] \\
&\mathsf{P}\Big((bn)^{1=2}jT_n \quad I_{;m}j \quad \Big) + \sup_{u2\mathsf{R}}\Big|\mathsf{P}(I_{;m} \quad u) \quad \mathsf{P}(jZ + \underline{d}j_1 \quad u)\Big| \\
&+ \sup_{u2\mathsf{R}}\mathsf{P}\Big(\big|(bn)^{1=2}jZ + \underline{d}j_1 \quad u\big| \quad \Big) =: \mathrm{I}_1 + \mathrm{I}_2 + \mathrm{I}_3:
\end{aligned}$$

For the I_1 part, $jT_n \quad I_{;m}j \quad jT_n \quad Ij + jI \quad I_{;m}j$: Recall (47), then $jT_n \quad Ij \quad c_0b^2$ for some constant $c_0 > 0$: We define $^{0} = 2c_1(np)^{1=q}m \quad$; where the constant c_1 is the one to be defined in Lemma 5. Then by Lemma 5, we have

$$\mathsf{P}((bn)^{1=2}jI \quad I_{;m}j \quad ^{0}) \ . \ (bn)^{\ q=2}:$$

Hence for $\quad = \ ^{0} + c_0(bn)^{1=2}b^2$; $\mathrm{I}_1 = O((bn)^{\ q=2})$:

For the I_2 part, note that

$$\mathrm{I}_2 \quad \sup_{u2\mathsf{R}} j\mathsf{P}(I_{;m} \quad u) \quad \mathsf{P}(I_{z;m} \quad u)j + \sup_{u2\mathsf{R}} j\mathsf{P}(I_{z;m} \quad u) \quad \mathsf{P}(jZ + \underline{d}j_1 \quad u)j =: \mathrm{I}_{21} + \mathrm{I}_{22}:$$

By Lemma 7 (1) for part $\mathrm{I}_{21} = o(1)$ and Lemma 8 for I_{22}, we have

$$\mathrm{I}_2 = O\Big((bn)^{\ 1=6}\log^{7=6}(pn) + ((np)^{2=q}=(bn))^{1=3}\log(pn) + \big(m=(bn) + m \quad\big)^{1=3}\log(np)^{2=3}\Big)$$

For the I_3 part, the diagonal entities in bnQ take the same value i.e. $^2 = 2bn\sum_{i=1}^{bn} w_i^2$; which by (43), converges to $2c_w^{02} > 0$, where c_w^0 is a finite constant. By Lemma 4

$$I_3 \ . \quad \log(np)^{1=2}:$$

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 with $m = (bn)^{1=(1+\ )}$ For the I_1 part, we define $\ = c_1\log(np)^{1=2}m \ + c_0(bn)^{1=2}b^2$; for some constant $c_1 > 0$: Then by Lemma 6, $I_1 = O((np)^{\ q})$: For I_2 part, by Lemma 7 (2) and Lemma 8, we have

$$I_2 = O\Big((bn)^{\ 1=6}\log^{7=6}(pn) + \big(m=(bn) + m \ \big)^{1=3}\log(np)^{2=3}\Big)$$

For I_3; same argument can be applied. Combining the rates of I_1-I_3, we obtain the desired result. □

Lemma 5 and 6 give us concentration inequalities for the m-dependent approximation of I .

LEMMA 5 (m-dependent approximation for polynomial case). *Assume conditions in Theorem 1 (i). For some* $0 < m \ n=2$ *and* $u > 0$*; we have*

$$P\big((bn)^{1=2}jI \ \ I_{;m}j \ u + c_1(np)^{1=q}m \ \big) \ c_2(e^{\ c_3u^2m^2} + npm^{\ q}(bn)^{\ q=2}u^{\ q});$$

where $c_1; c_2; c_3$ *are some positive constants only depending on* $q; c_p; c_w; c_s; \ _q$:

Proof. We note that $I \ I_{;m}$ can be bounded by

$$jI \ I_{;m}j \max_{bn+1 \ i \ n \ bn}\Big(\Big|\sum_{t=i \ bn}^{i \ 1} w_i \ _t\Lambda^{\ 1}(\ _t \ \ _{t;m})\Big|_1 + \Big|\sum_{t=i+1}^{i+bn} w_t \ _i\Lambda^{\ 1}(\ _t \ \ _{t;m})\Big|_1\Big) =: I_1 + I_2:$$

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

$$\mathrm{I}_1 = \max_{\substack{bn+1\ \ i\ \ n\ \ bn \\ 1\ \ j_1\ \ p}} \Big| \sum_{\substack{l\ \ i\ \ m\ \ 1 \\ 1\ \ j_2\ \ p}} E_{i;l;j_1;j_2\ \ l;j_2} \Big|; \tag{48}$$

where $E_{i;l;j_1;j_2}$ is the $(j_1; j_2)$th entity of matrix $E_{i;l}$ and $_{l;j_2}$ is the j_2th entity of $_{l}$. Since $_{l;j_2}$s are independent for different $(l; j_2)$, by Lemma A.2 in Chernozhukov et al. (2013b), for $u > 0$;

$$\mathrm{P}(\sqrt{bn}\mathrm{I}_1 \quad 2\sqrt{bn}\mathrm{E}\mathrm{I}_1 + u) \quad e^{\ u^2=(3\ \ ^2)} + K_q u^{\ q} H_q; \tag{49}$$

where K_q is some constant only depending on q;

$$^2 = bn \max_{\substack{bn+1\ \ i\ \ n\ \ bn \\ 1\ \ j_1\ \ p}} \sum_{\substack{l\ \ i\ \ m\ \ 1 \\ 1\ \ j_2\ \ p}} \mathrm{E}(E_{i;l;j_1;j_2\ \ l;j_2})^2;$$

and

$$H_q = (bn)^{q=2} \sum_{\substack{l\ \ n\ \ bn\ \ m\ \ 1 \\ 1\ \ j_2\ \ p}} \mathrm{E}\left(\max_{\substack{(bn+1)\_(l+m+1)\ \ i\ \ n\ \ bn \\ 1\ \ j_1\ \ p}} jE_{i;l;j_1;j_2\ \ l;j_2}j^q \right):$$

Then we start to analyze the rates of the objects involved in (49). We define $E_{i;l;j_1;}$ to be the j_1th row of $E_{i;l}$: For the ^{2} part, by Assumption 2.5 and (43),

$$jE_{i;l;j_1;}j_2 \quad \sum_{t\ \ l+m} w_{i\ \ t}\ \ ^{1=2}_{j_1;j_1} jA_{t\ \ l;j_1;}j_2 \quad c_w c_s m\ \ =(bn); \tag{50}$$

and

$$\sum_{l\ \ n\ \ bn\ \ m\ \ 1} jE_{i;l;j_1;}j_2 \quad \sum_{t=i\ \ bn}^{i\ \ 1} \sum_{l\ \ t\ \ m} w_{i\ \ t}\ \ ^{1=2}_{j_1;j_1} jA_{t\ \ l;j_1;}j_2 \quad c_w c_s m\ \ : \tag{51}$$

Combining the above arguments and recall that $\mathrm{E}\ ^2_{i;j} = 1$; we have

$$^2 \quad bn \max_{\substack{bn+1\ \ i\ \ n\ \ bn \\ 1\ \ j_1\ \ p}} \Big(\sum_{l\ \ i\ \ m\ \ 1} jE_{i;l;j_1;}j_2 \, _{l\ \ n} \max_{bn\ \ m\ \ 1} jE_{i;l;j_1;}j_2 \Big) \quad (c_w c_s)^2 m^{\ 2}: \tag{52}$$

For the H_q part, by Assumption 2.5 and (43), $\max_{i;j_1;j_2} jE_{i;l;j_1;j_2}j \quad c_w c_s((1 \quad l) \_ m) \quad =(bn)$: Recall that $\tilde{p} \quad c_p p$: Then we have

$$H_q \quad (bn)^{q=2} \sum_{l \;\; n \;\; bn \;\; m \;\; 1} [c_w c_s((1 \quad l) \_ m) \quad =(bn)]^q \;\; {}^q_q$$
$$(c_w c_s)^q \;\; {}^q_q (bn) \;\; {}^{q=2} \tilde{p} \Big(\sum_{m \;\; l \;\; n \;\; bn \;\; m} m \quad {}^q + \sum_{l< \;\; m} (1 \quad l) \quad {}^q \Big) \quad c_0 (bn) \;\; {}^{q=2} npm \quad {}^q; \tag{53}$$

where $c_0 = 3c_p(c_w c_s)^q \;\; {}^q_q$:

For EI_1 part, note that $\mathrm{EI}_1 \quad k\mathrm{I}_1 k_q$: By Lemma 2, we have

$$\mathrm{EI}_1 \quad \Big(\sum_{i;j_1} \mathrm{E}(j \sum_{l;j_2} E_{i;l;j_1;j_2} \;\; {}_{l;j_2} j^q) \Big)^{1=q} \quad \Big(\sum_{i;j_1} \big((q \quad 1) \sum_l jE_{i;l;j_1;} j_2^2 \;\; {}^2_q \big)^{q=2} \Big)^{1=q}:$$

Thus by (50) and (51) we have

$$\mathrm{EI}_1 \;\; . \;\; (bn) \;\; {}^{1=2} m \quad (np)^{1=q}; \tag{54}$$

where the constant in . only depends on $c_w; c_s; \;\; {}_q; q$: Our conclusions follows by applying (52), (53) and (54) into (49) and a similar argument for I_2. □

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

$$\mathrm{P}\big((bn)^{1=2} jI \quad I \;\; {}_{;m} j \quad u\big) \quad \begin{cases} 2npe \;\; {}^{a_1 m^2 \;\; u^2}; & \text{if} \quad u < a_2 (bn)^{1=2} m \quad ; \\ 2npe \;\; {}^{a_3 m \;\; (bn)^{1=2} u}; & \text{if} \quad u \quad a_2 (bn)^{1=2} m \quad ; \end{cases}$$

where $a_1; a_2; a_3$ are some positive constants only depending on $a_0; c_w; c_s; \;\; {}_e$:

Proof. Recall the definition of I_1 and I_2 in the proof of Lemma 5. Let $e = c_w c_s m =(bn)$ and $c = a_0=e$: Then by (50), $Ee^{cE_{i;l;j_1;j_2} \ l;j_2} < 1$; for any $0 < c \quad c$; and we have

$$E(e^{cI_1}) \quad \sum_{\substack{bn+1 \ i \ n \ bn \\ 1 \ j_1 \ p}} E\left(\exp\Big\{ c \sum_{\substack{l \ n \ bn \ m \ 1 \\ 1 \ j_2 \ p}} E_{i;l;j_1;j_2 \ l;j_2} \Big\} + \exp\Big\{ \quad c \sum_{\substack{l \ n \ bn \ m \ 1 \\ 1 \ j_2 \ p}} E_{i;l;j_1;j_2 \ l;j_2} \Big\} \right)$$

$$=: I_{11} + I_{12}:$$

Since $E_{i;j} = 0$; for $E_{i;l;j_1;j_2} \neq 0$; we have

$$\begin{aligned} E(e^{cE_{i;l;j_1;j_2} \ l;j_2}) &= 1 + \frac{E(e^{cE_{i;l;j_1;j_2} \ l;j_2} \quad 1 \quad cE_{i;l;j_1;j_2} \ l;j_2)}{c^2 E^2_{i;l;j_1;j_2}} c^2 E^2_{i;l;j_1;j_2} \\ &\quad 1 + \frac{E(e^{c \ e \ j \ l;j_2 j} \quad 1 \quad c \ e \ j \ l;j_2 j)}{(c \ e \ )^2} c^2 E^2_{i;l;j_1;j_2} \\ &\quad 1 + \frac{e}{a_0^2} c^2 E^2_{i;l;j_1;j_2}; \end{aligned}$$

where the first inequality is because that for any $x > 0$, the function $g_x(t) = (e^{tx} \quad 1 \quad tx)=t^2$ increases on $t \ 2 \ (0; \ 1 \ )$; and $e^t \quad t \quad e^{jtj} \quad jtj$: We define $c^{\emptyset} = \ _e=a_0^2$; and the rate of I_{11} is derived as follows,

$$I_{11} \quad \sum_{i;j_1} \prod_{l;j_2} (1 + c^{\emptyset} c^2 E^2_{i;l;j_1;j_2}) \quad \sum_{\substack{bn+1 \ i \ n \ bn \\ 1 \ j_1 \ p}} \exp\Big\{ c^{\emptyset} c^2 \sum_{\substack{l \ n \ bn \ m \ 1 \\ 1 \ j_2 \ p}} E^2_{i;l;j_1;j_2} \Big\};$$

$$np \exp\big\{ c^2 c_1 m \ ^2 =(bn) \big\};$$

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

$$P(I_1 \quad u) \quad e^{\ cu} E(e^{cI_1}) \quad e^{\ cu} (I_{11} + I_{12}):$$

We define $c = bnm^2 \ u=(2c_1)$: Hence if $c < c$; then $P(I_1 \quad u) \quad 2npe^{\ u^2 m^2 \ bn=(4c_1)}$; if $c \quad c$; then $P(I_1 \quad u) \quad 2npe^{\ e^{bn} \ a_0=(c_w c_s) bnm \ u}$: The proof for I_2 is similar and therefore omitted. □

LEMMA 7. *Let $m \geq 1$ and $m/(bn) \to 0$:*

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

$$\sup_{u \in \mathbb{R}} |P(I_{*;m} \leq u) - P(I_{z;m} \leq u)| \cdot (bn)^{-1/6} \log^{7/6}(pn) + ((np)^{2/q}/(bn))^{1/3} \log(pn);$$

where the constant in $\cdot$ only depends on c_w, c_w^0, c_s and ε_q.

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

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

where the constant in $\cdot$ only depends on c_w, c_w^0, c_s, a_0 and ε_e.

Proof. First we consider the case of (1). Denote

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

Then $I_{*;m}$ can be rewritten into

$$I_{*;m} = \max_{bn+1 \leq i \leq n-bn} \left| \sum_{i-m+1 \vee bn \vee l \leq i \leq 1} D_{i;l} \, {}_{l} + \sum_{i-m+2 \vee l \leq i \leq i+bn} D_{i;l-l} + d_i \right|_1 :$$

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

$$G_{i;l} = \begin{cases} D_{i;l} & \text{if } i - m + 1 \vee bn \vee l \leq i \leq m + 1; \\ D_{i;l} - D_{i;l} & \text{if } i - m + 2 \vee l \leq i \leq 1; \\ -D_{i;l} & \text{if } i \vee l \leq i + bn; \end{cases} \tag{56}$$

and elsewhere zero. We define $d_{i;j_1}$ to be the j_1th 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

$$N_2^{1=2} I_{;m} = \max_{\substack{bn+1 \\ 1}} \Big| \sum_{\substack{2 \\ 1}} g_{i;l;j_1;j_2} + N_2^{1=2} d_{i;j_1} \Big|; \text{ where } g_{i;l;j_1;j_2} = N_2^{1=2} G_{i;l;j_1;j_2} \quad {}_{l;j_2}: \tag{57}$$

For any r q; we denote

$$M_r := \max_{\substack{bn+1 \\ 1}} \quad {}_{ij_1;r}; \quad \text{where} \quad {}^{r}_{ij_1;r} := \sum_{\substack{2 \\ 1}} \mathsf{E} j g_{i;l;j_1;j_2} j^r = N_1 = \sum_{l=2}^{n} j G_{i;l;j_1;} j_r^r \quad {}^{r}_{r} N_2^{r=2} = N_1:$$

By Assumption 2.5 and (43), for any r 2;

$$j D_{i;l;j_1;} j_r \quad j D_{i;l;j_1;} j_2 \quad c_w c_s = (bn); \text{ and similarly } j D_{i;l;j_1;} j_r \quad c_w c_s = (bn): \tag{58}$$

Then by (56), $\max_{i;l;j_1} j G_{i;l;j_1;} j_r$ $2c_w c_s = (bn)$: Since $G_{i;l}$ is zero for $l < i$ $m+1$ bn or $l > i + bn$;

$$M_r \quad (4c_w c_s \quad {}_r)(N_1 = bn)^{1=2 \quad 1=r}: \tag{59}$$

Especially, for $r = 2$; M_r c_1 where $c_1 = 4c_w c_s \quad {}_2$: By (55), for i bn l i m; and $S_m = \sum_{k=0}^{m} {}^{1} A_k$;

$$G_{i;l} = D_{i;l} = \sum_{t=l}^{m+l} {}^{1} w_i \quad {}_t \Lambda \quad {}^{1} A_t \quad {}_l = w_i \quad {}_l \Lambda \quad {}^{1} S + w_i \quad {}_l \Lambda \quad {}^{1} (S_m \quad S) + \sum_{t=l}^{l+m} {}^{1} (w_i \quad {}_t \quad w_i \quad {}_l) \Lambda \quad {}^{1} A_t \quad {}_l:$$

Therefore by Assumption 2.5 and (43), we have

$$\begin{aligned} \big| j G_{i;l;j_1;} j_2 \quad w_i \quad {}_l \big| = \quad & {}^{1=2}_{j_1;j_1} w_i \quad {}_l \sum_{k \quad m} j A_{k;j_1;} j_2 + \quad {}^{1=2}_{j_1;j_1} \sum_{t=l}^{l+m} {}^{1} j w_i \quad {}_t \quad w_i \quad {}_l j j A_t \quad {}_{l;j_1;} j_2 \\ & c_w c_s m \quad = (bn) + c_w c_s m = (bn)^2: \end{aligned}$$

Note $m=(bn) = o(1)$, thus by (43),

$$\min_{\substack{bn \\ 1}} \min_{\substack{i \\ j_1}} \min_{\substack{n \\ p}} \min_{\substack{bn \\ ij_1;2}} \Big(\sum_{l=i}^{\substack{i \\ bn}} \sum_{\substack{m \\ }} jG_{i;l;j_1;} j_2^2 \; {}_2^2 N_2 = N_1 \Big)^{1=2} \quad c_w^{\emptyset} \; {}_2 \quad o(1) \quad c_1; \tag{60}$$

some constant $c_1 > 0$: Since $\max_{i;l;j_1;j_2} jG_{i;l;j_1;j_2}j \quad 2c_w c_s=(bn);$

$$\max_{l;j_2} \mathsf{E}(\max_{i;j_1} jg_{i;l;j_1;j_2} j^q) = \max_{i;l;j_1;j_2} jG_{i;l;j_1;j_2} N_2^{1=2} j^q \quad (2c_w c_s)(N_1=(bn))^{q=2}:$$

Note

$$B_n := \max \Big\{ M_3^3; M_4^2; \big(\max_{l;j_2} \mathsf{E}(\max_{i;j_1} jg_{i;l;j_1;j_2} j^q) \big)^{1=q} \Big\} \; . \; (N_1=(bn))^{1=2};$$

where the constant in . only depends on $c_w; c_s; \; {}_q$: By Proposition 2.1 in Chernozhukov et al. (2017) we have

$$\begin{aligned} &\sup_{u2\mathsf{R}} j\mathsf{P}(N_2^{1=2} I_{\;;m} \quad u) \quad \mathsf{P}(N_2^{1=2} I_{z;m} \quad u) j \\ &. \big(B_n^2 \log^7(pn)=N_1 \big)^{1=6} + \big(B_n^2 \log^3(pn)=N_1^{1 \;\; 2=q} \big)^{1=3} \\ &. (bn)^{\;\; 1=6} \log^{7=6}(pn) + ((np)^{2=q}=(bn))^{1=3} \log(pn): \end{aligned}$$

For part (2), let $M = \log_2(\;\; {}_e)\_1$; and $B_n^{\emptyset} = (2c_w c_s M=a_0)(N_1=(bn))^{1=2}$: Since $\max_{i;l;j_1;j_2} jG_{i;l;j_1;j_2}j$ $2c_w c_s=(bn)$; we have

$$\max_{i;l;j_1;j_2} \mathsf{E}(e^{g_{i;l;j_1;j_2}=B_n^{\emptyset}}) \quad 2:$$

Note

$$B_n := \max f M_3^3; M_4^2; B_n^{\emptyset} g \; . \; (N_1=(bn))^{1=2}:$$

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

$$\sup_{u2\mathsf{R}} \big| \mathsf{P}(N_2^{1=2} I_{\;;m} \quad u) \quad \mathsf{P}(N_2^{1=2} I_{z;m} \quad u) \big| \; . \; \big(B_n^2 \log^7(pn)=N_1 \big)^{1=6} \; . \; (bn)^{\;\; 1=6} \log^{7=6}(pn):$$

□

LEMMA 8. *Assume conditions in Theorem 1 (i) or (ii), for* $m \; ! \; 1$; $m=(bn) \; ! \; 0$;

$$\sup_{u2R} jP(I_{z;m} \quad u) \quad P(jZ + \underline{d}j_1 \quad u)j \; . \; \left(m=(bn) + m \quad \right)^{1=3} \log(np)^{2=3};$$

where the constant in . *only depends on* $c_w; c_w^0$ *and* c_s:

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

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

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

$$\begin{aligned} jGG^> \quad G \; G^>j_{\max} \quad & jH^0H^{0>}j_{\max} + 2j(H^1 \quad G \;)G^>j_{\max} + j(H^1 \quad G \;)(H^1 \quad G \;)^>j_{\max} \\ & =: \mathrm{I}_1 + \mathrm{I}_2 + \mathrm{I}_3: \end{aligned}$$

By (58), $\max_{i;l;j} jG_{i;l;j;}j_2 \quad 2c_w c_s=(bn)$: Therefore

$$(bn)\mathrm{I}_1 \quad (bn) \max_{i_1;i_2;j_1;j_2} \sum_{l=2 \quad m}^{0} jG_{i_1;l;j_1;}j_2 jG_{i_2;l;j_2;}j_2 \quad (2c_w c_s)^2 m=(bn):$$

Denote $\Delta_{i;l} = G_{i;l} \quad G_{i;l}$: For $i \quad m+1 \quad bn \quad l < i \quad bn$; $\Delta_{i;l} = D_{i;l}$; and thus $j\Delta_{i;l;j;}j_2$ $c_w c_s=(bn)$: For $i \quad bn \quad l \quad i \quad m+1$; we have

$$\Delta_{i;l} = D_{i;l} \quad w_{i \; l}\Lambda^{\; 1}S = \sum_{t=l}^{m+l \; 1} (w_{i \; t} \quad w_{i \; l})\Lambda^{\; 1}A_{t \; l} \quad w_{i \; l}\Lambda^{\; 1} \sum_{t \; m} A_t: \tag{61}$$

Hence $j\Delta_{i;l;j;}j_2 \quad c_w c_s m=(bn)^2 + c_w c_s m \quad =(bn)$: For $i \quad m+1 \quad l \quad i \quad 1$; $\Delta_{i;l} = D_{i;l} \quad D_{i;l}$ $w_{i \; l}\Lambda^{\; 1}S$: Then $j\Delta_{i;l;j;}j_2 \quad 3c_w c_s=(bn)$: Similarly we can bound $j\Delta_{i;l;j;}j_2$ for $i \quad l \quad i + bn$: For the rest l; $\Delta_{i;l} = 0$: We note that $jG_{i;l;j;}j \quad c_w c_s=(bn)$: Consequently,

$$(bn)\mathrm{I}_2 \quad (bn) \max_{i_1;i_2;j_1;j_2} \sum_{l=1}^{n} j\Delta_{i_1;l;j_1;}j_2 jG_{i_2;l;j_2;}j_2 \; . \; m=(bn) + m \quad ;$$

where the constant in . only depends on $c_w; c_s$: Similarly we have $(bn)\mathrm{I}_3$. $m=(bn)+m$: Combining I_1-I_3;

$$(bn)jGG^{>} \quad G\ G^{>}j_{\max} \ . \ m=(bn)+m \quad :$$

By (43), for any j we have $bnQ_{j;j} = 2bn\sum_{i=1}^{n} w_i^2 \ ! \ 2c_w^{02}$: Then the desired result follows from Lemma 3. □

B.4 Proof of Theorem 2

Proof of (i). Note 1 $\Phi(x)$ (2) $^{1=2}x\ ^{1}e\ ^{x^2=2}$; where $\Phi(\ )$ is the cumulative distribution function of a standard normal distribution. Recall $G_{i;l}$ in (15). Let $G_{i;} = (G_{i;1}; G_{i;2}; : : : ; G_{i;n})$ and $\underline{z}$ be a Gaussian vector in R^{np} with zero mean and identity covariance matrix. Let $G_{i;;j;}$ be the jth row of $G_{i;}$; then

$$\mathsf{P}((bn)^{1=2}jG\ \underline{z}j_1 \quad u) \quad \sum_{i=bn}^{n}\sum_{j=1}^{bn}\sum_{}^{p}\mathsf{P}((bn)^{1=2}jG_{i;;j;}^{>}\underline{z}j \quad u) \quad np(2\ )\ ^{1=2}(\ =u)e\ ^{u^2=(2\ ^2)}; \tag{62}$$

where $= (bn)^{1=2}jG_{i;;j;}j_2 = (2bn\sum_{t=1}^{bn} w_t^2)^{1=2}$; which by (43) converges to $2\ ^{1=2}c_w^{0} > 0$: Thus

$$\mathsf{P}\big(jG\ \underline{z}j_1 \quad 2c_w^{0}\log(np)^{1=2}(bn)\ ^{1=2}\big) \ ! \ 0: \tag{63}$$

Let $S := f1 \quad i \quad n : ji \quad {}_kj > bn;$ for all $1 \quad k \quad K_0g$: For any $i \ 2 \ S; d_i = 0$: Hence by Theorem 1,

$$\sup_{u2\mathsf{R}}\Big|\mathsf{P}\big(\max_{i2S}jV_ij_1 \quad u\big) \quad \mathsf{P}\big(\max_{i2S}jG_{i;}\underline{z}j_1 \quad u\big)\Big| \ ! \ 0: \tag{64}$$

Since $\max_{i2S}jG_{i;}\underline{z}j_1 \quad jG\ \underline{z}j_1$; by (63) and (64) we have $\mathsf{P}(\max_{i2S}jV_ij_1 \quad !^{y}) \ ! \ 0$: Thus we obtain

$$\lim_{n\ !\ 1}\mathsf{P}\big(8t \ 2 \ A_1; 91 \quad k \quad K_0; jt \quad {}_kj \quad bn\big) = 1: \tag{65}$$

Recall that $d_k = \Lambda^{-1}\ _k$: Since $|d_k + G_{k;}\underline{z}_1| \quad |d_k|_1 \quad |G_{k;}\underline{z}_1|$; we have

$$\mathsf{P}\Big(\min_{1\ k\ K_0}|d_k + G_{k;}\underline{z}_1| \quad !^y\Big) \quad \mathsf{P}\Big(\max_{1\ k\ K_0}|G_{k;}\underline{z}_1| \quad \min_{1\ k\ K_0}|d_k|_1 \quad !^y\Big)$$
$$\mathsf{P}\big(|G\ \underline{z}|_1 \quad !^y\big):$$

Therefore $\mathsf{P}(\min_{1\ k\ K_0}|d_k + G_{k;}\underline{z}|_1 \quad !^y) \ ! \ 0$: Subsequently the break statistics will be bigger than the threshold at the points of break with probability approach 1,

$$\mathsf{P}(\min_{1\ k\ K_0}|V_k|_1 \quad !^y) \ ! \ 0;$$

in view of

$$\sup_{u2\mathsf{R}}\Big|\mathsf{P}(|V_k|_1 \quad u) \quad \mathsf{P}(|d_k + G_{k;}\underline{z}|_1 \quad u)\Big| \ ! \ 0:$$

Therefore we have

$$\mathsf{P}(\ _k \ 2\ A_1; 1 \quad k \quad K_0) \ ! \ 1: \tag{66}$$

Let $B(\ ; r) = \{t : |t| \quad j \quad rg$: By (65) and (66), we have

$$\lim_n \mathsf{P}\Big(\{\ _1;\ _2; : : : ;\ _{K_0}\} \quad A_1 \quad [_{1\ k\ K_0} B(\ _k; bn)\Big) = 1:$$

Since for $k_1 \neq k_2$; $|\ _{k_1} \quad _{k_2}| \quad bn$; for any $k_1 \neq k_2$ and $t\ 2\ B(\ _{k_1}; bn)$; for all large n; $B(t; 2bn) \setminus B(\ _{k_2}; 2bn) = ;$: Thus we complete the proof. □

Proof of (ii). First consider the case (1). Let $^{(l)}_i$ (resp. $U^{(l)}_i$) be $\hat{}^{(l)}_i$ with Y_i therein replaced by $(i=n)$ (resp. $_i$). Similarly we can define $^{(r)}_i$ and $U^{(r)}_i$: Let $\Delta\ _i = {}^{(l)}_i \quad {}^{(r)}_i$ and $\Delta U_i = U^{(l)}_i \quad U^{(r)}_i$: Let Δf_i be $\Delta\ _i$ with replaced by f:

For any $1 \quad k \quad K_0$, and any t such that $jt \quad {}_k j \quad bn;$ we have $\Delta \ {}_t = (1 \quad \sum_{i=1}^{jt} {}^{kj} w_i) \ {}_k + \Delta f_t:$ Hence

$$\begin{aligned} V_t &= \Lambda \ {}^1\Delta \ {}_t + \Lambda \ {}^1\Delta U_t \\ &= (1 \quad \sum_{i=1}^{jt \quad kj} w_i)\Lambda \ {}^1 \ {}_k + \Lambda \ {}^1\Delta f_t + \Lambda \ {}^1\Delta U \ {}_k + (\Lambda \ {}^1\Delta U_t \quad \Lambda \ {}^1\Delta U \ {}_k): \end{aligned} \tag{67}$$

Note $\hat{}_k = \operatorname{argmax}_{ft:jt \quad {}_k j \ bng} jV_t j_1:$ The proceeding proof contains three steps.

Step 1. Let $j \ {}_k = \operatorname{argmax}_j jV_{\hat{}_k;j} j;$ where $V_{\hat{}_k;j}$ is the jth entity of $V_{\hat{}_k}:$ This step shows

$$\liminf_n \ {}_1 \min_{k \quad K_0} j(\Lambda \ {}^1 \ )_{j \ {}_k} j= \qquad 1:$$

We shall show by contradiction. By (47), $j\Delta f_t j_1 = O(b^2):$ If there exists $1 \quad k \quad K_0$, such that $j(\Lambda \ {}^1 \ )_{j \ {}_k} j \quad c \quad ;$ for some $c < 1;$ then by (67), $jV_{\hat{}_k} j_1 \quad c \quad + O(b^2) + j\Lambda \ {}^1\Delta U_{\hat{}_k} j_1:$ Let $\tilde{U}_t$ be U_t with ${}_{i;j}$ replaced by $z_{i;j}$ where $z_{i;j}$ are i.i.d standard normal random variables. Then $\max_t j\Lambda \ {}^1\Delta \tilde{U}_t j_1 = O_{\mathsf{P}}((bn) \ {}^{1=2}\log(np)^{1=2}):$ Then by Gaussian approximation Theorem 1,

$$\max_{1 \quad k \quad K_0;jt \quad {}_k j \quad bn} j\Lambda \ {}^1\Delta U_t j_1 = O_{\mathsf{P}}((bn) \ {}^{1=2}\log(np)^{1=2}):$$

Since $(bn) \ {}^{1=2}\log(np)^{1=2};$ we have $jV_{\hat{}_k} j_1 \quad c \quad (1 + o_{\mathsf{P}}(1)):$ On the other hand, by (67), $jV \ {}_k j_1 \quad + O(b^2) \quad j\Lambda \ {}^1\Delta U \ {}_k j_1 = \quad (1 + o_{\mathsf{P}}(1)):$ These imply $\mathsf{P}(V_{\hat{}_k} < V \ {}_k) \ ! \ 1;$ which is a contradiction.

Step 2. This step shows

$$\max_{1 \quad k \quad K_0;jt \quad {}_k j \quad bn} j\Lambda \ {}^1\Delta U \ {}_k \quad \Lambda \ {}^1\Delta U_t j_1 = jt \quad {}_k j^{1=2} = O_{\mathsf{P}}((np)^{1=q} = (bn)):$$

Let t $_k$; the other direction can be similarly dealt with. Note that

$$\Delta U_t \quad \Delta U_{\ k} = \sum_{i=t \ \ bn}^{k \ \ bn \ \ 1} w_{t \ \ i \ \ i} + \sum_{i= \ k \ \ bn}^{t \ \ 1} (w_{t \ \ i} \quad w_{\ k \ \ 1}) \ _i \quad \sum_{i=t+1}^{k \ \ 1} (w_{i \ \ t} + w_{\ k \ \ i}) \ _i$$

$$\sum_{i= \ k+1}^{t+bn} (w_{i \ \ t} \quad w_{i \ \ k}) \ _i \quad \sum_{i=t+bn+1}^{k+bn} w_{i \ \ k \ \ i} + w_{\ k \ \ t \ \ t} \quad w_{\ k \ \ t \ \ k} =: \sum_{k=1}^{7} r_k:$$

For r_1 we have

$$j\Lambda \ \ ^1 r_1 j_{\ 1} = \max_{1 \ \ j_1 \ \ p} \Big| \sum_{l \ \ k \ \ bn \ \ 1;1 \ \ j_2 \ \ p} E_{l;j_1;j_2 \ \ l;j_2} \Big|; \text{ where } E_l = \sum_{i=(t \ \ bn)\_l}^{k \ \ bn \ \ 1} w_{t \ \ i} \Lambda \ \ ^1 A_{i \ \ l};$$

and $E_{l;j_1;j_2}$ is $(j_1; j_2)$th entity of matrix E_l: Then

$$\max_{1 \ \ k \ \ K_0;jt \quad _kj \ \ bn} j\Lambda \ \ ^1 r_1 j_{\ 1} = jt \quad _k j^{1=2} = O_{\mathrm{P}} f(np)^{1=q} = (bn) g;$$

uniformly over $t; k$: A similar argument leads to the same bound for r_3 and r_5: For r_2; we can rewrite

$$j\Lambda \ \ ^1 r_2 j_{\ 1} = \max_{1 \ \ j_2 \ \ p} \Big| \sum_{l \ \ t \ \ 1} E^{0}_{l;j_1;j_2 \ \ l;j_2} \Big|; \text{ where } E^{0}_{l} = \sum_{i=(\ k \ \ bn)\_l}^{t \ \ 1} (w_{t \ \ i} \quad w_{\ k \ \ i}) \Lambda \ \ ^1 A_{i \ \ l}:$$

Then similarly we have $\max_{1 \ \ k \ \ K_0;jt \quad _kj \ \ bn} j\Lambda \ \ ^1 r_2 j_{\ 1} = j_{\ k} \quad tj^{3=2} = O_{\mathrm{P}} f(np)^{1=q} = (bn)^2 g$: We obtain the desired result by summing up all the above bounds.

Step 3. Without loss of generality, assume $_{j \ k} > 0$: Then by the argument in step 1, with probability tending to 1, $V_{\hat{}_k;j \ \ k} > 0$: By Assumption 2.1, we have $j\Delta f_t \quad \Delta f \ j_{\ 1} = O(jt \quad j=n)$; uniformly over $t; \ :$ With probability tending to 1, by (67),

$$jV_{\ _k} j_{\ 1} \quad jV_{\hat{}_k} j_{\ 1} \quad V_{\ _k;j \ \ k} \quad V_{\hat{}_k;j \ \ k}$$

$$\sum_{i=1}^{j\hat{}_k \quad _k j} w_i (\Lambda \ \ ^1 \ \ )_{j \ \ k} \quad O(j\hat{} _k \quad _k j=n) \quad j\Lambda \ \ ^1 \Delta U_{\ _k} \quad \Lambda \ \ ^1 \Delta U_{\hat{}_k} j_{\ 1}:$$

By (43), we have $\sum_{i=1}^{jt} {}^{j} w_i \quad c_w^{00} jt \quad j=(bn)$: For finite moment case, by Step 1 and Step 2 we further derive

$$jV_{\ k}j_1 \quad jV_{\wedge_k}j_1 \quad c_w^{00}j\hat{}_k \quad {}_kj \quad =(bn) \quad O(j\hat{}_k \quad {}_kj=n) \quad O_{\mathsf{P}}\big(j_{\ k} \quad \hat{}_k j^{1=2}(np)^{1=q}=(bn)\big);$$

uniformly over k: Since $jV_{\ k}j_1 < jV_{\wedge_k}j_1$; we have

$$\max_{1 \ k \ K_0} j_{\ k} \quad \hat{}_k j = O_{\mathsf{P}} f(np)^{2=q} = \ {}^2 g:$$

Similar argument can be applied for sub exponential case. □

Proof of (iii). Recall the definition of ${}_t^{(l)}$; ${}_t^{(r)}$; $U_t^{(l)}$ and $U_t^{(r)}$ in the proof of (ii) and $M = bn$: Since $M \quad \log(np)= \ {}^2$;

$$j \ {}_{\wedge_k \ M}^{(l)} \quad ((\hat{}_k \quad M)=n)j_1 = jf_{\wedge_k \ M}^{(l)} \quad f((\hat{}_k \quad M)=n)j_1 = O(b^2):$$

Similarly $j \ {}_{\wedge_k + M}^{(r)} \quad ((\hat{}_k + M)=n)j_1 = O(b^2)$: Since $\max_{1 \ j \ p} jf_j^{0}j$ is bounded,

$$j \ ((\hat{}_k + M)=n) \quad ((\hat{}_k \quad M)=n) \quad {}_k \ j_1 = jf((\hat{}_k + M)=n) \quad f((\hat{}_k \quad M)=n)j_1 = O(M=n):$$

Hence

$$\begin{aligned} j\Lambda \ {}^1(\hat{}_k \quad {}_k \ )j_1 = \big|\Lambda \ {}^1(\ {}_{\wedge_k + M}^{(r)} \quad {}_{\wedge_k \ M}^{(l)} \quad {}_k \ ) + \Lambda \ {}^1U_{\wedge_k + M}^{(r)} \quad \Lambda \ {}^1U_{\wedge_k \ M}^{(l)}\big|_1 \\ O(b + M=n) + j\Lambda \ {}^1U_{\wedge_k \ M}^{(l)} \quad \Lambda \ {}^1U_{\wedge_k + M}^{(r)} j_1: \end{aligned} \tag{68}$$

By Gaussian approximation and (63) we have

$$\mathsf{P}\big(j\Lambda \ {}^1U_{\wedge_k \ M}^{(l)} \quad \Lambda \ {}^1U_{\wedge_k + M}^{(r)} j_1 \quad 2c_w^{0}\log(np)^{1=2}=(bn)^{1=2}\big) \ ! \ 0:$$

Inserting the above equation into (68) and we obtain the desired result. □

B.5 Proof of Theorem 3

Proof. We recall $M = bn$: Similar to (68), we have $j\Lambda^{-1}(\hat{}_k \quad _k) \quad \Lambda^{-1}(U^{(r)}_{\hat{}_k+M} \quad U^{(l)}_{\hat{}_k \quad M})j_1$ $cM=n$: Therefore

$$\sup_{u2R}\Big|P\big((bn)^{1=2}j\Lambda^{-1}(\hat{}_k \quad _k)j_1 \quad u\big) \quad P\big((bn)^{1=2}j\tilde{Z}j_1 \quad u\big)\Big|$$

$$\sup_{u2R}P\big(j(bn)^{1=2}j\tilde{Z}j_1 \quad uj \quad c(bn)^{1=2}M=n\big)$$

$$+\sup_{u2R}\big|P(j\Lambda^{-1}(U^{(l)}_{\hat{}_k \quad M} \quad U^{(r)}_{\hat{}_k+M})j_1 \quad u) \quad P(j\tilde{Z}j_1 \quad u)\big| = \mathrm{I}_1 + \mathrm{I}_2:$$

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 $2c_w^{02} > 0$. Therefore by Lemma 4,

$$\mathrm{I}_1 = Of(bn)^{1=2}(M=n)\log(np)^{1=2}g = o(1):$$

Let $\tilde{G} = (\tilde{G}_1; \tilde{G}_2; : : : ; \tilde{G}_n)$; where $\tilde{G}_l = w_{\hat{}_k \quad M \quad l}\Lambda^{-1}S$; if $\hat{}_k \quad M \quad bn \quad l \quad \hat{}_k \quad M \quad 1$; and $\tilde{G}_l = w_{l \quad (\hat{}_k+M)}\Lambda^{-1}S$; if $\hat{}_k + M + 1 \quad l \quad \hat{}_k + M + bn$ and elsewhere zero. Let $\underline{z}$ be Gaussian vector in R^{np} 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

$$\sup_{u2R}\Big|P(j\Lambda^{-1}(U^{(l)}_{\hat{}_k \quad M} \quad U^{(r)}_{\hat{}_k+M})j_1 \quad u) \quad P(j\tilde{G}\underline{z}j_1 \quad u)\Big| = o(1):$$

Thus $\mathrm{I}_2 = o(1)$ and we complete the proof. □

B.6 Proof of Theorem 4

Proof of (i). We shall condition on the event where $\hat{S}_k = S_k$ and $j\hat{}_k \quad _k j \quad bn$: By Theorem 2 and Corollary 2, the event would take place with probability tending to 1. Denote $"_t = \sum_{j2S_k}(\Lambda^{-1}\hat{}_k)_j(\Lambda^{-1} \quad _t)_j$; and $\hat{a}_k = \sum_{j2S_k}(\Lambda^{-1} \quad _k)_j(\Lambda^{-1}\hat{}_k)_j$: Then we have

$$X_t = \hat{a}_k 1_{t \quad _k} + \sum_{j2S_k} f_j(t=n)(\Lambda^{-1}\hat{}_k)_j + "_t:$$

Let $I(t) = \sqrt{(4bn+1)=((t \quad t_0)(4bn+1 \quad t+t_0))}$, $t_0 = \hat{}_k \quad 2bn$, $t_1 = \hat{}_k + 2bn$ and

$$D_t = \Big(\sum_{s=t_0}^{t_1} X_s \frac{t \quad t_0}{4bn+1} \quad \sum_{s=t_0}^{t \ 1} X_s \Big) I(t):$$

For any $r > 0$; we have

$$\begin{aligned} D_{\ k+r} \quad D_{\ k} =& \Big(\sum_{s=t_0}^{t_1} X_s \frac{k \quad t_0}{4bn+1} \quad \sum_{s=t_0}^{k \ 1} X_s \Big) \big(I(\ _k + r) \quad I(\ _k) \big) \\ &+ \Big(\sum_{s=t_0}^{t_1} X_s \frac{r}{4bn+1} \quad \sum_{s=\ _k}^{k+r \ 1} X_s \Big) I(\ _k + r) = \mathrm{I}_1 + \mathrm{I}_2: \end{aligned}$$

Denote $\mathrm{I}_i(f)$ (resp. $\mathrm{I}_i('')$, $\mathrm{I}_i(a)$) to be I_i with X_t therein replaced by $f(t=n)$ (resp. $''_t$, $\hat{a}_k \mathbf{1}_{t \quad _k}$), $i = 1; 2$.

Firstly, consider the f part. Note $jI(\ _k + r) \quad I(\ _k)j \ . \ (bn)^{\ 3=2} r:$ Thus by the continuity of f_j; for I_1 part, we have

$$\begin{aligned} \max_j \Big| \sum_{s=t_0}^{t_1} f_j(s=n) \frac{k \quad t_0}{4bn+1} \quad \sum_{s=t_0}^{k \ 1} f_j(s=n) \Big| jI(\ _k + r) \quad I(\ _k)j &= O((b^2 n) jI(\ _k + r) \quad I(\ _k)j) \\ &= O(br(bn)^{\ 1=2}): \end{aligned}$$

Similarly we can handle the f part in I_2 and therefore

$$j\mathrm{I}_1(f) + \mathrm{I}_2(f)j = O(br(bn)^{\ 1=2} j\tilde{}_k j_1):$$

Secondly, let us consider the drift part, for $r \quad bn;$

$$\begin{aligned} &\mathrm{I}_1(a) + \mathrm{I}_2(a) \\ =& (t_1 \quad _k + 1) \hat{a}_k \frac{k \quad t_0}{4bn+1} \big(I(\ _k + r) \quad I(\ _k) \big) + \Big((t_1 \quad _k + 1) \hat{a}_k \frac{r}{4bn+1} \quad r\hat{a}_k \Big) I(\ _k + r) \\ =& \quad (bn)^{\ 1=2} \hat{a}_k r=2(1 + o(1)): \end{aligned}$$

Thirdly, let us focus on the ε part. By Theorem 2 (iii), $j(\Lambda^{-1}\hat{\Sigma}_k)_{j2}S_k j_2 = j\tilde{\Sigma}_k j_2(1+o_P(1))$ and thus $\hat{a}_k = j\tilde{\Sigma}_k j_2(1+o_P(1))$. Then together with (35), we obtain that the long run variance for ε_k is $\sigma_k^2(1+o_P(1))$. Hence by Theorem 1 in El Machkouri et al. (2013),

$$\Big|\sum_{s=\hat{\tau}_k-2bn}^{\hat{\tau}_k+2bn}\varepsilon_s\Big| = O_P((bn)^{1/2}\sigma_k\log(bn)^{1/2}) \quad \text{and} \quad \Big|\sum_{s=\tau_k}^{\tau_k+r}\varepsilon_s\Big| = O_P(r^{1/2}\sigma_k).$$

Therefore

$$\mathrm{I}_1(\varepsilon) = O_P\big((bn)^{1/2}\sigma_k\log(bn)^{1/2}|\beta(\tau_k+r) \quad \beta(\tau_k)|\big) = O_P\big((bn)^{-1}r\sigma_k\log(bn)^{1/2}\big).$$

For I_2 part, we have

$$\begin{aligned}\mathrm{I}_2(\varepsilon) &= \Big(\sum_{s=t_0}^{t_1}\varepsilon_s\frac{r}{4bn+1} \quad \sum_{s=\tau_k}^{\tau_k+r}\Big)^{-1}\varepsilon_s\Big)\beta(\tau_k+r)\\ &= O_P\Big((bn)^{-1}r\sigma_k\log(bn)^{1/2} + r^{1/2}\sigma_k(bn)^{-1/2}\Big).\end{aligned}$$

Combining all the previous parts we have

$$\begin{aligned}&D_{\tau_k+r} \quad D_{\tau_k}\\ =\ & (bn)^{-1/2}a_k r(1+o_P(1))/2 + O(br(bn)^{-1/2}|j\tilde{\Sigma}_k j_1|) + O_P\Big((bn)^{-1}r\sigma_k\log(bn)^{1/2} + r^{1/2}\sigma_k(bn)^{-1/2}\Big)\end{aligned} \tag{69}$$

Note $|j\tilde{\Sigma}_k j_1| \quad |jS_k j^{1/2}|j\tilde{\Sigma}_k j_2|$; thus

$$|j\tilde{\Sigma}_k j_1|b \quad |jS_k j^{1/2}|j\tilde{\Sigma}_k j_2|b \quad |jS_k j^{1/2}|j\tilde{\Sigma}_k j_2|^{y} \quad |j\tilde{\Sigma}_k j_2^2| = a_k.$$

Therefore in (69), we have $O(br(bn)^{-1/2}|j\tilde{\Sigma}_k j_1|) = o((bn)^{-1/2}a_k r)$.

Since $\tilde{\Sigma}_k$ is a covariance matrix with diagonal entities 1, $|j\tilde{\Sigma}_k j_2| \quad |jS_k j|$. Note $\sigma_k^2 \quad |j\tilde{\Sigma}_k j_2|a_k$; thus

$$\sigma_k \quad |jS_k j^{1/2}a_k^{1/2}.$$

Then we have

$$\&_k(bn)^{1=2}\log(bn)^{1=2} \quad jS_kj^{1=2}a_k^{1=2}(bn)^{1=2}\log(bn)^{1=2} \quad jS_kj^{1=2}a_k^{1=2} \; y \quad a_k;$$

where the last inequality is because $a_k = j\tilde{}_kj_2^2 \quad {}^{y2}jS_kj:$ Therefore $O_P\big((bn)^{1=2}r\&_k\log(bn)^{1=2}\big) = o_P((bn)^{1=2}a_kr)$ in (69). Inserting the above equations into (69) leads to

$$(bn)^{1=2}(D_{k+r} \quad D_k) \quad a_kr(1=2 + o_P(1)) + O_P(\&_kr^{1=2}):$$

Since $D_{\sim k}$ is the maximum, $D_{k+r} \quad D_k > 0:$ Therefore $r = O_P(\&_k^2=a_k^2):$ By a similar argument for the $r < 0$ part, the desired result follows. □

Proof of (ii). Let F_t be the -field generated by $f_{s;j}; s \quad t; 1 \quad j \quad pg:$ Denote the projection operator $P_t = \mathsf{E}(\,jF_t) \quad \mathsf{E}(\,jF_{t\ 1}):$ Let $^{-}_{k;j} = (\Lambda^{\ 1}_{\ k})_j;$ if $j \ 2 \ S_k$, $^{-}_{k;j} = 0$ if $j \not\in S_k$. Let $(\ ^{0}_{t})$ be an i.i.d copy of $(\ _t)$. Then

$$c_t := kP_0{}''_tk_4 \quad k^{-}_{k}{}^{>}\Lambda^{\ 1}A_t(\ _0 \quad ^{0}_{0})k_4 \ . \ j^{-}_{k}{}^{>}\Lambda^{\ 1}A_tj_{2\ 4};$$

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

$$\sum_{s\ m} c_s \ . \ \sum_{s\ m} j^{-}_{k}{}^{>}\Lambda^{\ 1}A_sj_2 \quad \sum_{j=1}^{p}\sum_{s\ m} {}^{-}_{k;j}{}^{\ }_{j;j}{}^{1=2}jA_{s;j;}j_2 \ . \ j^{-}_kj_1m \quad = j\tilde{}_kj_1m \quad :$$

Thus by Corollary 2.1 in Berkes et al. (2014), strong invariance principle holds for $\sum_{s\ t}{}''_s:$ Thus similar to (69), we have

$$(bn)^{1=2}(D_{k+r} \quad D_k) = \quad 2^{\ 1}a_kr(1 + o_P(1)) + I_2(\ )\ ^{P}_{?} \quad 2^{\ 1}a_kr + \&_k\mathsf{B}(r):$$

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

$$S = \{k \mid A_k \text{ or } A_{k-1} \text{ contains break points}\}.$$

Then by assumption $|S| \le 2K_0$. We look at estimators without the break point first.

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

Let

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

Define functions

$$B_{i,j}^+(u,x) = \tilde{\theta}_{i,j} - u + \psi_{i,j}[(\tilde{\theta}_{i,j} - u)^2 + v_{i,j}^2]/2 + x,$$
$$B_{i,j}^-(u,x) = \tilde{\theta}_{i,j} - u - \psi_{i,j}[(\tilde{\theta}_{i,j} - u)^2 + v_{i,j}^2]/2 - x.$$

The proof contains four steps.

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

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

By (40), $\psi(x) \le x + x^2/2$ and thus

$$\mathsf{E}\bar{h}_{i,j}(u) \le \sum_{k \notin S} \big(\mathsf{E}(\hat{\theta}_{i,j,k} - u) + \psi_{i,j}\mathsf{E}(\hat{\theta}_{i,j,k} - u)^2/2\big)/N_2 = B_{i,j}^+(u,0).$$

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 $x \& (N_2 \log(N_2))^{1=2}$;

$$\sum_{i;j=1}^{p} \mathsf{P}\Big(\sup_{ju \quad i;jj \quad C_0} j\bar{h}_{i;j}(u) \quad \mathsf{E}\bar{h}_{i;j}(u)j \quad x(\ {}^{1=2}_{i;i}\ {}^{1=2}_{j;j})=N_2 \Big) \ . \ p^2\big(N_2 \log(n)^{q=4} x^{\ q=2} + e^{\ x^2=(cN_2)}\big); \tag{70}$$

where c and the constant in $.$ are independent of $n; p$:

First introduce some notation. For any random variable X; denote $\mathsf{E}_0 X = X \quad \mathsf{E}X$; the centering operator. Let $F_k = (\ {}_t;\ t\ 2\ [_s\ {}_k A_s)$ and $F_{k;fsg};\ s \quad k;$ be F_k with ${}_t; t\ 2\ A_s$ therein replaced by ${}^{\emptyset}_t$; where ${}^{\emptyset}_t$ are i.i.d copy of ${}_t$: For any random variable $X = h(F_k)$; let $X_{fig} = h(F_{k;fig})$: Denote $\Delta\ {}_k = \ {}_k \quad {}_{k\ 1}$: We now show that the temporal dependence measure decays with polynomial rate. Let ${}_{i;j;k}(u) = \quad {}_{i;j}(\hat{}_{i;j;k} \quad u)$: Since $j\ {}^{\emptyset}j_1 \quad 1$, we have for any $s\ 2\ \mathsf{N}$ and any u;

$$\begin{aligned} & k \sup_u j\ {}_{i;j;k}(u) \quad {}_{i;j;k;fk\ \ sg}(u)jk_{q=2} \quad k\hat{}_{i;j;k} \quad \hat{}_{i;j;k;fk\ \ sg}k_{q=2} \\ & 2\ {}^1m\Big(k\mathsf{E}_0\Delta\ {}_{k;i}(\Delta\ {}_{k;j} \quad \Delta\ {}_{k;j;fk\ \ sg})k_{q=2} + k\mathsf{E}_0(\Delta\ {}_{k;i} \quad \Delta\ {}_{k;i;fk\ \ sg})\Delta\ {}_{k;j;fk\ \ sg}k_{q=2}\Big) \\ & =: 2\ {}^1m(\mathrm{I}_1 + \mathrm{I}_2): \end{aligned}$$

Let $U_{k;i}$ (resp. $f_{k;i}$) be ${}_{k;i}$ with Y_t replaced by ${}_t$ (resp. $f(t=n)$). Then ${}_{k;i} = \ {}_{k;i} + f_{k;i}$ when there is no break. Let $\Delta U_{k;i} = U_{k;i} \quad U_{k\ 1;i}$ and $\Delta f_{k;i} = f_{k;i} \quad f_{k\ 1;i}$: Then we have

$$\mathrm{I}_1 \quad \Big\|\mathsf{E}_0\Delta f_{k;i}(\Delta U_{k;j} \quad \Delta U_{k;j;fk\ \ sg})\Big\|_{q=2} + \Big\|\mathsf{E}_0\Delta U_{k;i}(\Delta U_{k;j} \quad \Delta U_{k;j;fk\ \ sg})\Big\|_{q=2} =: \mathrm{I}_{11} + \mathrm{I}_{12}: \tag{71}$$

Since $\max_j jf_j^{\emptyset}j_1 < f\ ;$

$$\max_{1\ \ j\ \ p} j\Delta f^{\emptyset}_{k;j}j \quad f\ m=n: \tag{72}$$

Let $E_{k;l;i;} = \sum_{t=(km+1)\_l}^{(k+1)m} A_{t\ l;i;}$; where $A_{t\ l;i;}$ is the ith row of matrix $A_{t\ l}$: Then

$$U_{k;i} = \sum_{l\ (k+1)m} E_{k;l;i;\ \ l}=m$$

and

$$\Delta U_{k;i} \quad \Delta U_{k;i;fk\ \ sg} = \begin{cases} \sum_{l2A_{k\ \ s}} (E_{k;l;i;} \quad E_{k\ \ 1;l;i;})(\ _l \quad {}^{0}_{l})=m; & s \quad 1; \\ \sum_{l2A_{k\ \ s}} E_{k;l;i;}(\ _l \quad {}^{0}_{l})=m; & s=0: \end{cases} \tag{73}$$

Since $_l$ are i.i.d, by Lemma 2, (72) and (73),

$$k\Delta f_{k;i}(\Delta U_{k;j} \quad \Delta U_{k;j;fk\ \ sg})k_{q=2} \quad 2f\ c_q\Big(\sum_{l2A_{k\ \ s}} (jE_{k;l;i}\, j_2 + E_{k\ \ 1;l;i}\, j_2)^2\Big)^{1=2}\ _{q=2}=n: \tag{74}$$

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

$$\sum_{l2A_{k\ \ s}} jE_{k;l;i;}\, j_2^2 \ .\ m(m(s \quad 1))^{\ 2}\ _{i;i}; \quad \text{and} \quad \sum_{l\ (k+1)m} jE_{k;l;i;}\, j_2^2 \ .\ m\ _{i;i}; \tag{75}$$

where the constant in $.$ only depending on $;c_s$. Hence by (74) and (75),

$$\mathrm{I}_{11} \ .\ m^{1=2}n^{\ 1}(m(s \quad 1))^{\ 1=2}_{i;i}; \tag{76}$$

where the constant in $.$ only depends on $;c_s;q;\ _q;f$: By Lemma 2 and (75)

$$\begin{aligned} k\mathrm{E}_0 U_{k;i}(U_{k;j} \quad U_{k;j;fk\ \ sg})k_{q=2} &= \Big\|\mathrm{E}_0\Big(\sum_{l_1\ (k+1)m} E_{k;l_1;i;\ \ l_1} \sum_{l_2 2A_{k\ \ s}} E_{k;l_2;j;}(\ _{l_2} \quad {}^{0}_{l_2})\Big)\Big\|_{q=2} m^{\ 2} \\ &\quad .\ m^{\ 2}\Big(\sum_{l_1\ (k+1)m} \sum_{l_2 2A_{k\ \ s}} jE_{k;l_1;i;}\, j_2^2 jE_{k;l_2;j;}\, j_2^2\Big)^{1=2} \\ &\quad .\ m^{\ 1}(m(s \quad 1))^{\ 1=2}_{i;i}\ ^{1=2}_{j;j}; \end{aligned} \tag{77}$$

where the constant in $.$ only depends on $_q;q;;c_s$: Thus $\mathrm{I}_{12} \ .\ m^{\ 1}(m(s \quad 1))^{\ 1=2}_{i;i}\ ^{1=2}_{j;j}$: By combining the bounds for I_{11} and I_{12} and a similar argument for I_2; we have

$$_s := \max_k \big\| \sup_u j\ _{i;j;k}(u) \quad _{i;j;k;fk\ \ sg}(u)j\big\|_{q=2} \ .\ ((ms) \quad \mathbf{1}_{s>1} + \mathbf{1}_{s\ \ 1})^{\ 1=2}_{i;i}\ ^{1=2}_{j;j}; \tag{78}$$

where the constant in . only depends on $_q; c_s; c$; ; $q; f$:

Let := $^{1=2}_{i;i}$ $^{1=2}_{j;j}x=(2N_2)$ and A_n be the net for fu : ju $_{i;j}j$ C_0g: Denote $f(u) = \bar{h}_{i;j}(u)$ $E\bar{h}_{i;j}(u)$: Then by j $^0j_1$ 1;

$$\sup_{jv} \sup_{i;jj} \min_{C_0 u2A_n} jf(u) \quad f(v)j \quad :$$

Therefore $jA_nj = 2C_0= = O(n)$ and

$$P\Big(\sup_{ju} \sup_{i;jj} j\bar{h}_{i;j}(u) \quad E\bar{h}_{i;j}(u)j \quad x(\ ^{1=2}_{i;i}\ ^{1=2}_{j;j})=N_2\Big) \quad P\Big(\max_{u2A_n} j\bar{h}_{i;j}(u) \quad E\bar{h}_{i;j}(u)j \quad x(\ ^{1=2}_{i;i}\ ^{1=2}_{j;j})=(2N_2)\Big):$$

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

Step 3. This step shows for the estimator

$$\max_{1 \ i;j \ p} j\tilde{}_{i;j} \quad _{i;j}j = O(m \ ^{=(\ +1)}\ ^{1=2}_{i;i}\ ^{1=2}_{j;j} + m^3=n^2); \quad \text{and} \quad v^2_{i;j} = O(\ _{i;i}\ _{j;j}); \qquad (79)$$

where the convergence is uniform for 1 $i; j$ p:

Let $\hat{}_{i;j;k}$ be $\hat{}_{i;j;k}$ with Y_t replaced by $_t$ and let $_{i;j} = E\hat{}_{i;j;1}$: Then by (72),

$$j\tilde{}_{i;j} \quad _{i;j}j \quad m\sum_{k2S} j\Delta f_{k;i}jj\Delta f_{k;j}j=(2N_2) = O(m^3=n^2): \qquad (80)$$

Note the convergence in above $O(\ )$ and all the followings are uniform for $i; j$: Let $_{i;j;k} = E(\ _{0;i}\ _{k;j})$: Then for any $L < m$;

$$jmE(U_{1;i}U_{1;j}) \quad _{i;j}j = \Big|m^{\ 1} \sum_{m<k<m} (m \quad jkj)\ _{i;j;k} \quad _{i;j}\Big| = O(\sum_{jkj\ \ L} j\ _{i;j;k}j + Lm^{\ 1}\sum_{k2Z} j\ _{i;j;k}j):$$

By Assumption 2.5, $\sum_{jkj\ \ L} j\ _{i;j;k}j \quad \sum_{t2Z;jkj\ \ L} jA_{t;i;}j_2jA_{t+k;j;}j_2 = O(L \quad ^{1=2}_{i;i}\ ^{1=2}_{j;j})$: Take $L = m^{1=(\ +1)}$; then $jmE(U_{1;i}U_{1;j}) \quad _{i;j}j = O(m \ ^{=(\ +1)}\ ^{1=2}_{i;i}\ ^{1=2}_{j;j})$: And similarly $jmE(U_{1;i}U_{2;j})j = O(m \ ^{=(\ +1)}\ ^{1=2}_{i;i}\ ^{1=2}_{j;j})$: Hence

$$_{i;j} = m\big(E(U_{1;i}U_{1;j}) + E(U_{2;i}U_{2;j}) \quad E(U_{1;i}U_{2;j}) \quad E(U_{2;i}U_{1;j})\big)=2 = \ _{i;j} + O(m \ ^{=(\ +1)}\ ^{1=2}_{i;i}\ ^{1=2}_{j;j}):$$

Together with (80) we obtain the first part in (79).

Since $\hat{}_{i;j;k} = m(\Delta f_{k;i} + \Delta U_{k;i})(\Delta f_{k;j} + \Delta U_{k;j})=2;$ we have $\mathsf{E}\hat{}_{i;j;k} = m\Delta f_{k;i}\Delta f_{k;j}=2 + \ _{i;j}:$ By (72) and (80),

$$v_{i;j}^2 = \sum_{k \neq i,j} \mathsf{E}\hat{}_{i;j;k}^2=N_2 \quad \tilde{}_{i;j}^2 = \sum_{k \neq i,j} \mathrm{Var}(\hat{}_{i;j;k})=N_2 + O(m^3 n^{\ 2\ \ 1=2}_{i;i\ \ j;j}):$$

Note by (73) and (75) we have

$$m^2\mathrm{Var}(\Delta f_{k;i}\Delta U_{k;j}) = O(m^3 n^{\ 2}\ _{j;j}); \quad \text{and} \quad m^2\mathrm{Var}(\Delta U_{k;i}\Delta U_{k;j}) = O(\ _{i;i}\ \ _{j;j}):$$

Thus $\mathrm{Var}(\hat{}_{i;j;k}) = O(\ _{i;i}\ \ _{j;j})$ and the second part in (79) holds.

Step 4. Since $jSj \quad 2K_0;$ for any $i; j;$ and $j \ \ j_1 \quad \log(2);$

$$jN_1 h_{i;j}(u)=N_2 \quad \bar{h}_{i;j}(u)j \quad 2\log(2)K_0=(\ _{i;j}N_2): \tag{81}$$

Combining (81), Step 1 and step 2 with $x = N_2^{1=2}\log^{1=2}(np)$, then with probability tending 1, for all $1 \quad i; j \quad p;$ and $ju \quad _{i;j}j \quad C_0;$

$$B_{i;j}(u;\Delta) \quad N_1N_2\ ^1h_{i;j}(u) \quad B_{i;j}^{+}(u;\Delta); \tag{82}$$

where

$$\Delta = h\ ^{1=2}_{i;i}\ ^{1=2}_{j;j} + 2\log(2)K_0=(\ _{i;j}N_2) \quad \text{and} \quad h = x=N_2: \tag{83}$$

Note if

$$^2_{i;j}v_{i;j}^2 + 2\ _{i;j}\Delta \quad 1; \tag{84}$$

then $B_{i;j}^{+}(u;\Delta)$ exists real roots. Denote the smaller one as u^{+}, which satisfies u^{+} $\tilde{}_{i;j} + \ _{i;j}v_{i;j}^2 + 2\Delta:$ Take $\ _{i;j} = \ _{i;j}\ ^{1=2}_{i;i}\ ^{1=2}_{j;j}:$ By Step 3 and Assumption 2.3, if (84), then

$$_{i;i}^{1=2}\ _{j;j}^{1=2}(u^{+} \quad _{i;j}) = O\Big\{m^{\ =(\ +1)} + m^3=n^2 + \ _{i;j} + h + mK_0=(\ _{i;j}n)\Big\}: \tag{85}$$

Similar bound can be obtained for u as well. When (82) holds, u $\hat{}_{i;j}$ u^{+}. Take $_{i;j} = (m=n)^{1=2}$; then with probability tending to 1,

$$_{i;i}^{1=2}\ _{j;j}^{1=2}\left|\hat{}_{i;j}\right| \quad _{i;j}\left| \ . \ m \ ^{=(\ +1)} + m^3=n^2 + (m=n)^{1=2} + N_2\ ^{1=2}\log(np);\right.$$

where the convergence is uniform for all $1 \quad i; j \quad p$: Since there exists some constant $c_1; c_2 > 0$; such that $c_1 \quad \bar{}_{j;j}= \ _{j;j} \quad c_2$; and any $i; j$ with probability tending to 1. Thus the desired result follows. □

Proof of (ii). Same argument as for (i), except that we need to replace Step 2 by Step 2' with $x \quad N_2=\log(np)^{2:5}$. Then we obtain the desired result.

Step 2'. This step shows

$$\sum_{i;j=1}^{p} \mathrm{P}\Big(\sup_{ju \quad _{i;j}\left| \ \right. C_0} \left|\bar{h}_{i;j}(u) \quad \mathrm{E}\bar{h}_{i;j}(u)\right| \quad x(\ _{i;i}^{1=2}\ _{j;j}^{1=2})=N_2\Big) \ . \ p^2 ne^{\ cxN_2\ ^{1=2}}; \qquad (86)$$

where c and the constant in $.$ 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). x_1 represent t and x_2 represents i.

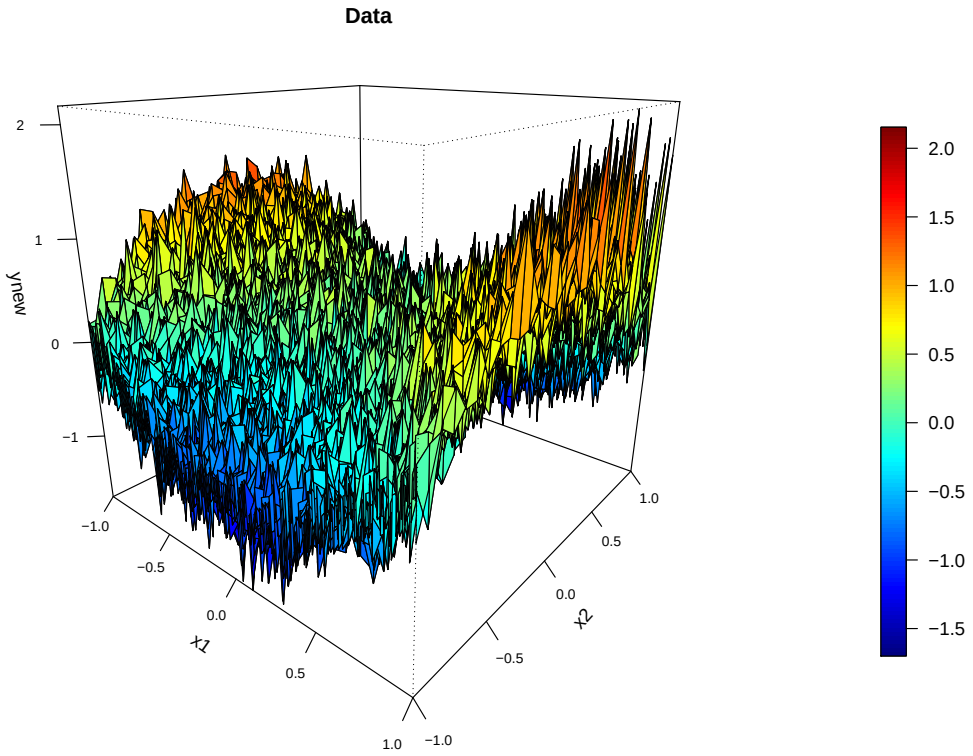


Figure 6: Visualization of the real (left) and estimated correlation matrix (using the robust estimation method).