Change point localization in dependent
dynamic nonparametric random dot product graphs

Oscar Hernan Madrid Padilla\textsuperscript{1}, Yi Yu\textsuperscript{2}, and Carey E. Priebe\textsuperscript{3}

\textsuperscript{1}Department of Statistics, University of California, Los Angeles
\textsuperscript{2}Department of Statistics, University of Warwick
\textsuperscript{3}Department of Applied Mathematics and Statistics, Johns Hopkins University

November 19, 2019

Abstract

In this paper, we study the change point localization problem in a sequence of dependent
nonparametric random dot product graphs (e.g. Young and Scheinerman, 2007). To be spe-
cific, assume that at every time point, a network is generated from a nonparametric random
dot product graph model, where the latent positions are generated from unknown underly-
ing distributions. The underlying distributions are piecewise constant in time and change at
unknown locations, called change points. Most importantly, we allow for dependence among
networks generated between two consecutive change points. This setting incorporates the edge-
dependence within networks and across-time dependence between networks, which is the most
flexible setting in the published literature.

To fulfill the task of consistently localizing change points, we propose a novel change point
detection algorithm, consisting of two steps. First, we estimate the latent positions of the
random dot product model, the theoretical result thereof is a refined version of the state-of-
the-art results, allowing the dimension of the latent positions to grow unbounded. Then, we
construct a nonparametric version of CUSUM statistic (e.g. Page, 1954; Padilla et al., 2019a)
that can handle across-time dependence. The consistency is proved theoretically and supported
by extensive numerical experiments, which outperform existing methods.

Keywords: Dependent dynamic networks; Nonparametric random dot product graph mod-
els; Change point localization.

1 Introduction

Computationally-efficient and theoretically-justified change point localization methods that can
handle new data types are in high demand, due to technological advances in a broad range of
application areas including finance, biology, social sciences, to name only a few. The literature
on change point detection is extensive, including the univariate mean case (e.g. Frick et al., 2014;
Fryzlewicz, 2014; Wang et al., 2018b), the high-dimensional mean case (e.g. Wang and Samworth,
2016; Cho, 2016), the robust mean case (e.g. Fearnhead and Rigaill, 2018; Pein et al., 2017), the
covariance case (e.g. Aue et al., 2009; Wang et al., 2017; Avanesov and Buzun, 2018), the univariate
nonparametric case (e.g. Zou et al., 2014; Padilla et al., 2019a), and the multivariate nonparametric
In this paper we are concerned with change point localization in dynamic networks. Let \( \{A(t)\}_{t=1}^{T} \subset \{0,1\}^{n \times n} \) be a sequence of adjacency matrices generated from a sequence of distributions \( \{\mathcal{L}_t\}_{t=1}^{T} \), such that for an unknown sequence of change points \( \{\eta_k\}_{k=1}^{K} \subset \{2,\ldots,T\} \) with \( 1 = \eta_0 < \eta_1 < \ldots < \eta_K \leq T < \eta_{K+1} = T + 1 \), we have that

\[
\mathcal{L}_{t-1} \neq \mathcal{L}_t, \quad \text{if and only if} \quad t \in \{\eta_1,\ldots,\eta_K\}.
\]

The goal is to estimate the change point collection \( \{\eta_k\}_{k=1}^{K} \) accurately.

There has been recently an increasing interest in the literature studying the model described above. Wang et al. (2018a) considered an independent sequence of inhomogeneous Bernoulli networks and presented a nearly optimal change point localization algorithm, accompanied with a phase transition phenomenon. Zhao et al. (2019) assumed an independent sequence of graphon models with independent edges and proposed consistent yet optimal localization result. Other network change point papers include Wang et al. (2014), Cribben and Yu (2017), Liu et al. (2018), Chu and Chen (2017), Mukherjee (2018), among others. More in-depth comparisons with Wang et al. (2018a) will be conducted later in the paper.

1.1 Random dot product graph models

Different from the aforementioned papers, in order to allow for dependence among edges, we assume that at every time point, the network is generated from a random dot product graph (e.g. Young and Scheinerman, 2007; Athreya et al., 2017). We formally define the model in Definitions 1 and 2, which are both from Athreya et al. (2017).

**Definition 1** (Inner product distribution). Let \( F \) be a probability distribution whose support is given by \( X_F \subset \mathbb{R}^d \). We say that \( F \) is a \( d \)-dimensional inner product distribution on \( \mathbb{R}^d \) if for all \( x,y \in X_F \), it holds that \( x^\top y \in [0,1] \).

**Definition 2** (Random dot product graph with distribution \( F \)). Let \( F \) be a \( d \)-dimensional inner product distribution with \( \{X_i\}_{i=1}^{n} \overset{i.i.d.}{\sim} F \). Let \( X = (X_1,\ldots,X_n)^\top \in \mathbb{R}^{n \times d} \). Suppose \( A \) is a random adjacency matrix given by

\[
\mathbb{P} \{ A \mid X \} = \prod_{1 \leq i < j \leq n} (X_i^\top X_j)^{A_{ij}}(1 - X_i^\top X_j)^{1-A_{ij}}.
\]

We write \( A \sim \text{RDPG}(F,n) \).

We would like to make a few comments regarding random dot product graph models. For first time reading, one can safely skip this and jump to Section 1.2.

**Equivalence of distributions**

It can be seen from Definition 2 that the latent positions come into play only through their inner products, i.e. we have

\[
A_{ij} \sim \text{Ber}(X_i^\top X_j), \quad 1 \leq i,j \leq n.
\]

This means that one can apply any orthonormal rotations to all the latent positions and retain the same distribution of \( A \). In light of this rotational invariance, we define the equivalence of inner product distributions below, which is also from Athreya et al. (2017).
Definition 3 (Equivalence of inner product distributions). If both $F(\cdot)$ and $G(\cdot)$ are inner product distributions defined on $\mathbb{R}^d$, and there exists an orthogonal operator $U : \mathbb{R}^d \to \mathbb{R}^d$ such that $F = G \circ U$, then we say $F$ and $G$ are equivalent.

Community structures

The random dot product graph is a generalization of the stochastic block model (Holland et al., 1983), where the latent positions $X$ are assumed to be fixed and satisfy

$$XX^\top = ZQZ^\top,$$

where $Z \in \{0, 1\}^{n \times d}$ is a membership matrix, each row of which consisting one and only one entry being 1 and $Q \in [0, 1]^{d \times d}$ is a connectivity matrix encoding the edge probabilities.

One may be puzzled by the observation that under Definition 2, we have that for any $(i, j) \in \{1, \ldots, n\}^2$, $i \neq j$,

$$E(A_{ij}) = E(X_i^\top X_j) = E(X_1^\top X_2),$$

where the second identity follows from the fact that within a network the latent positions are i.i.d., and therefore one loses the community structure and connections from the stochastic block model.

This observation is due to the randomness of the latent positions. To enforce a version of “communities” under Definition 2, one may introduce a membership vector and treat the distribution $F$ as a mixture distribution. To be specific, we have an alternative to Definition 2 below.

Definition 4. Let $\tau$ be generated from a multinomial distribution with parameter $n, \pi_1, \ldots, \pi_M$, where $M$ is a positive integer. Let $\{F_m\}_{m=1}^M$ be a sequence of $d$-dimensional inner product distributions. Assume that $X_i \mid \tau_i \in \mathbb{R}^d \overset{\text{iid}}{\sim} F_{\tau_i}, \ i = 1, \ldots, n.$

Let $X = (X_1, \ldots, X_n)^\top \in \mathbb{R}^{n \times d}$. Suppose $A$ is a random adjacency matrix given by

$$P \{A \mid X\} = \prod_{1 \leq i < j \leq n} (X_i^\top X_j)^{A_{ij}}(1 - X_i^\top X_j)^{1-A_{ij}}.$$

We write $A \sim \text{RDPG}(F, n)$, where

$$F = \sum_{m=1}^M \pi_m F_m.$$

We remark that Definition 4 is a special case of Definition 2, and therefore the theoretical results based on Definition 2 also hold for Definition 4. The vector $\tau$ prompts the vertex correspondence in a dynamic network. For instance, one may assume a sequence of $\text{RDPG}(F, n)$ using Definition 4, with latent positions drawn independently and the membership vector unchanged. There are also other variants. For instance, one may also assume instead that the membership vector $\tau$ is fixed.

1.2 List of contributions

We highlight the contributions of this paper.

First of all, we propose a novel algorithm for change point localization in dependent dynamic random dot product graph models, see Algorithm 2, which proceeds with first estimating the latent positions $\{\hat{X}_i(t)\}_{i=1,t=1}^{n,T}$, and then translating them to a univariate sequence. Due to the latent
positions’ rotational-invariance properties we discussed in Section 1.1, one pertaining challenge in the RDPG literature is to match the rotations of the latent position estimators of different networks (e.g. Athreya et al., 2017; Cape et al., 2019). We propose a novel way to get around this issue with matching by introducing \( \hat{Y}_{ij}^t = (X_i(t))^\top X_j(t) \), and construct a Kolmogorov–Smirnov CUSUM statistic (Padilla et al., 2019a) based on \( \{\hat{Y}_{ij}^t : (i,j) \in \{(l, n/2 + l) : l = 1, \ldots, n/2\}, t = 1, \ldots, T\} \). One may question the power of the Kolmogorov–Smirnov distance, but it allows for more general distribution functions, which include stochastic block models as special cases.

Secondly, under an appropriate signal-to-noise ratio condition, we prove Algorithm 2 can estimate the number and locations of change points consistently, which will be formally stated in Section 3.2. It is worth mentioning that Theorem 1 handles the situation where there exists dependence across time and among edges. This is not shown in the existing network change point detection literature.

Thirdly, we provide in-depth discussions on the characterization of jumps in Section 3.1. Note that the data we have are a collection of adjacency matrices. However, as stated in Definition 2, the data generating mechanism depends on latent positions’ distributions \( F \)’s. A natural question is whether the changes in \( F \) will lead to the changes in the distributions of the adjacency matrices, and if so, whether we can characterize the changes. The results we developed in Section 3.1 are interesting per se, and can shed light on network testing problems.

Lastly, the numerical experiments provide ample evidence on the strength of our proposed approach. In particular, we highlight the advantage of our method in scenarios with dependent networks.

The rest of the paper is organized as follows. Section 2 provides the formal problem setup and our proposed method in detail. The characterization of the distributional changes and statistical guarantees for our approach are collected in Section 3. We conclude with numerical experiments in Section 4. Technical details are deferred to the Appendix.

2 Methodology

2.1 Setup

We first formally state the full model descriptions.

**Model 1.** Let \( \{A(1), \ldots, A(T)\} \subset \mathbb{R}^{n \times n} \) be a sequence of adjacency matrices of random dot product graphs, satisfying the following.

1. *(Random dot product graphs.)* For any \( t \in \{1, \ldots, T\} \), it holds that

\[
P \{A(t) \mid X(t)\} = \prod_{1 \leq i < j \leq n} (X_i(t))^\top X_j(t)A_{ij}(t)(1 - X_i(t))^\top X_j(t))^{1 - A_{ij}(t)},
\]

where \( X(t) = (X_1(t), \ldots, X_n(t))^\top \in \mathbb{R}^{n \times d} \) satisfies the following.

There exists a sequence \( 1 = \eta_0 < \eta_1 < \ldots < \eta_K \leq T < \eta_{K+1} = T + 1 \) of time points, called change points. For \( k \in \{0, \ldots, K\} \), we have that

\[
X_i(\eta_k) \in \mathbb{R}^d \overset{\text{ind}}{\sim} F_{\eta_k}, \quad i = 1, \ldots, n,
\]
and for $t \in \{\eta_k + 1, \ldots, \eta_{k+1} - 1\}$, we have that

$$X_i(t) = \begin{cases} X_i(t-1), & \text{with probability } \rho, \\ \text{ind } \sim F_{\eta_k}, & \text{with probability } 1 - \rho, \end{cases}$$ (1)

with $F_t$'s satisfying Definition 1. Throughout, we write $P_t = X(t)X(t)^\top$ for the matrix of latent link probabilities at time $t \in \{1, \ldots, T\}$.

2. **(Minimal spacing.)** The minimal spacing between two consecutive change points satisfies

$$\min_{k=1,\ldots,K+1} \{\eta_k - \eta_{k-1}\} = \Delta > 0.$$

3. **(Minimal jump size.)** For each $k \in \{0, \ldots, K\}$ and for any $X,Y \overset{i.i.d.}{\sim} F_{\eta_k}$, denote

$$G_{\eta_k}(z) = \mathbb{P}\left\{X^\top Y \leq z\right\}, \quad z \in [0,1].$$

The magnitudes of the changes in the data generating distribution are such that

$$\min_{k=1,\ldots,K} \sup_{z \in [0,1]} |G_{\eta_k}(z) - G_{\eta_{k-1}}(z)| = \min_{k=1,\ldots,K+1} \kappa_k = \kappa > 0. \quad (2)$$

4. Assume that for every $k \in \{0, \ldots, K\}$,

$$\mathbb{E}\left\{(X(\eta_k))^\top X(\eta_k)\right\} = \Sigma_k \in \mathbb{R}^{d \times d},$$

where $\Sigma_k$ has eigenvalues $\mu_1^k \geq \cdots \geq \mu_d^k > 0$, with $\{\mu_s^k, k = 0, \ldots, K, s = 1, \ldots, d\}$ all being universal constants.

In Model 1, between two consecutive change points, the latent positions are dependent with exponentially decaying correlations; and for latent positions drawn at time points separated by change points, they are independent. If $\rho = 0$ in (1), then all the latent positions are independent. In particular, this implies that the adjacency matrices are independent.

The distributional changes occurring at change points are quantified through cumulative distribution functions $\{G_{\eta_k}\}$ defined in Model 1(3). Intuitively, since the unconditional distributions of $\{A(t)\}$ are completely characterized by the joint distributions of $\{(X_i(t))^\top X_j(t)\}$, it is natural to quantify the changes with respect to $\{G_{\eta_k}\}$. (A more detailed discussion on this can be found in Section 3.1.) In particular, the changes are measured by the Kolmogorov–Smirnov distance in (2), since the Kolmogorov–Smirnov distance does not require assumptions about the moments of the distributions, or about their discrete or continuous nature. With the stochastic block model being a special case of the random dot product graph, the distributions thereof are point-mass distributions, which handicaps the adoption of other (potentially more powerful) distribution distances, including the total variation distance.

Model 1(4) is imposed to guarantee that the latent link probabilities satisfy

$$\mathbb{P}_t\{\text{rank}(P_t) = d\} = 1.$$
2.2 Methods

To arrive at our construction, we start by defining the main statistic, and its population version. Without loss of generality, we assume that the number of nodes $n$ is an even integer. If $n$ is odd, then we randomly ignore a certain but fixed node and all edges connecting to it throughout the whole procedure.

**Definition 5 (CUSUM statistics).** Let $O = \{(i, n/2 + i), i = 1, \ldots, n/2\}$.

- *(Sample version)* With one sample $\{A(t)\}_{t=1}^T \subset \mathbb{R}^{n \times n}$, let
  \[ \hat{X}(t) = U_A(t) \Lambda_A(t)^{1/2}, \]
  where $U_A(t) \in \mathbb{R}^{n \times d}$ is an orthogonal matrix with columns being the leading $d$ eigenvectors of $A(t)$, and $\Lambda_A(t) \in \mathbb{R}^{d \times d}$ is a diagonal matrix with entries being the leading $d$ eigenvalues of $A(t)$.
  For any $t \in \{1, \ldots, T\}$ and $(i, j) \in O$, let
  \[ \hat{Y}_{ij}^t = (\hat{X}_i(t))^\top \hat{X}_j(t), \]
  where $(\hat{X}_i(t))^\top$ is the $i$th row of $\hat{X}(t)$. For any $0 \leq s < t < e \leq T$ and $z \in \mathbb{R}$, we define the CUSUM statistic as
  \[ D_{s,e}^t(z) = \left| \frac{2(e-t)}{n(e-s)(t-s)} \sum_{k=s+1(i,j) \in O}^t 1\{\hat{Y}_{ij}^k \leq z\} - \sqrt{\frac{2(t-s)}{n(e-s)(e-t)}} \sum_{k=t+1(i,j) \in O}^e 1\{\hat{Y}_{ij}^k \leq z\} \right|, \]
  and
  \[ D_{s,e}^t = \sup_{z \in [0,1]} |D_{s,e}^t(z)|. \]

- *(Population version)* With one sample $\{A(t)\}_{t=1}^T \subset \mathbb{R}^{n \times n}$, recall that $P_t = X(t)X(t)^\top$ and write
  \[ X(t) = U_P(t) \Lambda_P(t)^{1/2}, \]
  where $U_P(t) \in \mathbb{R}^{n \times d}$ is an orthogonal matrix with columns being the leading $d$ eigenvectors of $P_t$, and $\Lambda_P(t) \in \mathbb{R}^{d \times d}$ is a diagonal matrix with entries being the leading $d$ eigenvalues of $P_t$.
  For any $t \in \{1, \ldots, T\}$ and $(i, j) \in O$, let
  \[ Y_{ij}^t = (X_i(t))^\top X_j(t), \]
  where $(X_i(t))^\top$ is the $i$th row of $X$. For any $0 \leq s < t < e \leq T$ and $z \in \mathbb{R}$, we define the CUSUM statistic as
  \[ \tilde{D}_{s,e}^t(z) = \left| \sqrt{\frac{2(e-t)}{n(e-s)(t-s)}} \sum_{k=s+1(i,j) \in O}^t \mathbb{E}\left(1\{Y_{ij}^k \leq z\}\right) - \sqrt{\frac{2(t-s)}{n(e-s)(e-t)}} \sum_{k=t+1(i,j) \in O}^e \mathbb{E}\left(1\{Y_{ij}^k \leq z\}\right) \right|, \]
  and
  \[ \tilde{D}_{s,e}^t = \sup_{z \in [0,1]} |\tilde{D}_{s,e}^t(z)|. \]
We remark that in Definition 5, if the $d$th and $(d+1)$th eigenvalues share the same value, then one can randomly pick an eigenvector to construct $\hat{X}, X \in \mathbb{R}^{n \times d}$. In addition, we do not require a specific order of the eigenvectors in constructing $\hat{X}$ and $X$.

Recall that the distributions of the latent positions are equivalent up to a rotation, see Definition 3. To avoid extra efforts in matching the rotations when comparing two latent position distributions, we resort to the inner products of latent positions instead of latent positions itself. We explain this via (3). For any orthogonal matrix $U \in \mathbb{R}^{d \times d}$, it holds that

$$Y_{ij}^t = (X_i(t))^\top X_j(t) = (UX_i(t))^\top UX_j(t).$$

With Definition 5, we arrive at our proposed procedure Algorithm 2 that builds on the wild binary segmentation algorithm (Fryzlewicz, 2014). The method requires first estimating the latent positions, a subroutine shown in Algorithm 1 (adjacency spectral embedding, see e.g. ?). Note that this only needs to be done once regardless of the choice of the tuning parameter $\tau$, and is parallelizable. Since the complexity of the truncated principal component analysis is of order $O(dn^2)$, Algorithm 1 has the computational cost of order $O(Tdn^2)$. Once the latent positions are estimated, we run the remaining steps in Algorithm 2, which amounts to running Algorithm 2 in Padilla et al. (2019a). For a fixed $\tau$ which leads to $\tilde{K}$ change points, we have the computational complexity of order $O(\tilde{K}MTn \log(n))$, which translates to $O(Tdn^2 + \tilde{K}MTn \log(n))$ for the overall cost of Algorithm 2, where $M$ is the number of random intervals drawn in Algorithm 2.

### Algorithm 1 ScaledPCA $(A,d)$

**INPUT:** Matrix $A \in \mathbb{R}^{n \times n}$ and tuning parameter $d \in \mathbb{Z}_+$

$A = (v_1, \ldots, v_n) \text{diag}(\lambda_1, \ldots, \lambda_n)(v_1, \ldots, v_n)^\top$, where $|\lambda_1| \geq \cdots \geq |\lambda_n|.$

$X \leftarrow (v_1, \ldots, v_d) \text{diag}(|\lambda_1|^{1/2}, \ldots, |\lambda_d|^{1/2})$

**OUTPUT:** $X$

In every network, there are $n(n-1)/2$ observations, but note that in Definition 5, we in fact only use $n/2$ of them. This is for technical convenience, since due to the choice of $O$, we obtain independent observations within one network. We acknowledge that there are other variants of this treatment. For instance, instead using a fixed choice of $O$, one can do multiple random sub-samplings and combine the results; one can also gather all the observations and create a $U$-statistic instead. We will show later in Section 3 that using the seemingly most naive choice and in fact most computationally-cheap choice, we are able to achieve consistent estimators, therefore we refrain pursuit on this direction.

### 3 Theory

In this section, we provide the statistical guarantees for Algorithm 2 in Theorem 1. In order to enhance the theoretical understanding, we take a step back and understand how the jump defined in (2) through the cumulative distribution functions of the inner products can be related to the jumps in terms of the distributions of the adjacency matrices.

#### 3.1 Characterizations of the changes

We summarize the notation below and consider two different sets of models.
Algorithm 2 NonPar-RDPG-CPD \( ((s, e), \{(\alpha_m, \beta_m)\}_{m=1}^M, \tau) \)

**INPUT:** A sample \( \{A(t)\}_{t=s+1}^{e} \subseteq \mathbb{R}^{n \times n} \), collection of intervals \( \{(\alpha_m, \beta_m)\}_{m=1}^M \), tuning parameters \( d \in \mathbb{Z}_+ \) and \( \tau > 0 \).

for \( t = s + 1, \ldots, e \) do
    \( X(t) \leftarrow \text{ScaledPCA}(A(t), d) \)
end for

for \( m = 1, \ldots, M \) do
    \( (s_m, e_m) \leftarrow [s, e] \cap [\alpha_m, \beta_m] \)
    if \( e_m - s_m > 1 \) then
        \( b_m \leftarrow \text{argmax}_{s_m \leq t \leq e_m - 1} D_{t}^{s_m, e_m} \)
        \( a_m \leftarrow D_{b_m}^{s_m, e_m} \)
    else
        \( a_m \leftarrow -1 \)
    end if
end for

\( m^* \leftarrow \text{argmax}_{m=1,\ldots,M} a_m \)

if \( a_{m^*} > \tau \) then
    add \( b_{m^*} \) to the set of estimated change points
    \( \text{NonPar-RDPG-CPD}((s, b_{m^*}), \{(\alpha_m, \beta_m)\}_{m=1}^M, \tau) \)
    \( \text{NonPar-RDPG-CPD}((b_{m^*} + 1, e), \{(\alpha_m, \beta_m)\}_{m=1}^M, \tau) \)
end if

**OUTPUT:** The set of estimated change points.

Model 2. We assume the following two independent models:

\[
\{A_{ij}, 1 \leq i < j \leq n\}\{X_i\}_{i=1}^n \xrightarrow{\text{ind}} \text{Ber}(X_i^\top X_j), \quad X_i \sim F \in \mathbb{R}^d;
\]

and

\[
\{
\tilde{A}_{ij}, 1 \leq i < j \leq n\}\{\tilde{X}_i\}_{i=1}^n \xrightarrow{\text{ind}} \text{Ber}(\tilde{X}_i^\top \tilde{X}_j), \quad \tilde{X}_i \sim \tilde{F} \in \mathbb{R}^d.
\]

For \( i \neq j \), the cumulative distribution functions of \( X_i^\top X_j \) and \( \tilde{X}_i^\top \tilde{X}_j \) are denoted by \( G(\cdot) \) and \( \tilde{G}(\cdot) \), respectively. We further write \( \mathcal{L} \) and \( \tilde{\mathcal{L}} \) for the joint unconditional distributions of \( \{A_{ij}, 1 \leq i < j \leq n\} \) and \( \{
\tilde{A}_{ij}, 1 \leq i < j \leq n\} \), respectively.

The rest of this subsection is summarized in Figure 1. The notation \( A \Rightarrow B \) means \( A \) implies \( B \).

**Lemma 1.** With the notation in Model 2, if \( F = \tilde{F} \), then \( G = \tilde{G} \).

This follows automatically from the definitions, and is equivalent to the claim that if \( G \neq \tilde{G} \) then \( F \neq \tilde{F} \), which implies that (2) is equivalent to

\[
F_{n_k} \neq F_{n_{k-1}}, \quad k \in \{1, \ldots, K\}.
\]

However, \( F \neq \tilde{F} \) does not imply \( \mathcal{L} \neq \tilde{\mathcal{L}} \). As a simple toy example, consider \( F \) and \( \tilde{F} \) to be defined in Definition 1, with the same mean but different variances, and \( n = 2 \). Then \( F \neq \tilde{F} \) but \( \mathcal{L} = \tilde{\mathcal{L}} \). Lemma 2 below shows that \( \mathcal{L} \) is determined by the first \( n - 1 \) moments of \( F \).
Lemma 2. Under Model 2, we have that $\mathcal{L} = \tilde{\mathcal{L}}$ if and only if there exists an orthogonal operator $U \in \mathbb{R}^{d \times d}$, such that if $d = 1$,

$$
E_F(X_1^k) = E_{\tilde{F}}((U \tilde{X}_1)^k), \quad k = 1, \ldots, n - 1,
$$

if $d > 1$

$$
E_F(\prod_{l=1}^d X_{1,l}^{k_l}) = E_{\tilde{F}}(\prod_{l=1}^d (U \tilde{X}_1)^{k_l}), \quad k_l \in \mathbb{Z}, \quad k_l \geq 0, \quad \sum_{l=1}^d k_l = k, \quad k = 1, \ldots, n - 1,
$$

where $X_{1,l}$ and $(U \tilde{X}_1)_l$ are the $l$th coordinates of the $X_1$ and $U \tilde{X}_1$.

It can be seen from Lemma 2 that the unconditional distribution of the data matrix is determined by the first $n - 1$ moments of the underlying distribution $F$. Unfortunately, without additional assumptions, the first $n - 1$ moments do not determine the distribution (e.g. Heyde, 1963). This means that only assuming (2) can not guarantee that the data matrices $A$ and $\tilde{A}$ have different distributions.

The final claim we make in this subsection is that under some additional but weak conditions, we will be able to guarantee that $\mathcal{L} \neq \tilde{\mathcal{L}}$.

Assumption 1. Under Model 2, let

$$
\kappa_0 = \sup_{z \in [0,1]} |G(z) - \tilde{G}(z)|.
$$

It holds that

$$
\kappa_0 \sqrt{n} > 3 \sqrt{\log(n)}.
$$

Lemma 3. Assume that Model 2 and Assumption 1 hold. Then we have that

$$
\mathcal{L} \neq \tilde{\mathcal{L}}.
$$

Lemma 3 suggests that under Assumption 1, $G \neq \tilde{G}$ implies $\mathcal{L} \neq \tilde{\mathcal{L}}$. This enhances the rationale of imposing the distributional changes occurring at the change points on the differences on $G$, as detailed in Model 1(4). Assumption 1 is a weak assumption, which will be further elaborated in Section 3.2. The proofs of Lemmas 2 and 3 are collected in Appendix A.

---

1 We are grateful to Richard J. Samworth for pointing out this reference and constructive discussions on this.
3.2 Consistent estimation of change points

We first state a signal-to-noise ratio condition below.

**Assumption 2** (Signal-to-noise ratio). There exists a universal constant $C_{\text{SNR}} > 0$, such that there exists a diverging sequence $a_T \rightarrow \infty$, as $T \rightarrow \infty$, satisfying

$$\kappa \sqrt{\Delta n (1 - \rho)} > C_{\text{SNR}} \sqrt{T} \max\{ \sqrt{\log(n \vee T)}, d^{3/2} \} + a_T.$$ 

To better understand Assumption 2, we would like to use Assumptions 2 and 3 in Wang et al. (2018a) as benchmarks, since Wang et al. (2018a) studied a simpler problem assuming independence within and across networks, and showed a phase transition phenomenon in the minimax sense. However, we would like to emphasize that comparing Assumption 2 and Assumptions 2 and 3 in Wang et al. (2018a) is comparing apples and oranges, to some extent. Even though the jump size $\kappa$ are defined differently in these two papers, both take values in $(0, 1]$. The parameter $\rho$ in this paper indicates the correlation between networks, while the parameter $\rho$ in Wang et al. (2018a) represents the entrywise sparsity. For simplicity, we let $\rho = 1$ in Wang et al. (2018a) for this discussion.

One key difference is that in Assumption 2, the required signal-to-noise ratio is inflated by $\sqrt{T - \rho}$. We might view this as the effective sample size being shrunk from $\Delta$ to $(1 - \rho)\Delta$, due to the dependence across time. In Model 1, we do not allow $\rho = 1$, but allow $\rho \rightarrow 1$, as long as Assumption 2 holds. In the extreme case that $\rho = 1$, between two consecutive change points, there is essentially only one observation. As long as Assumption 1 holds, Lemma 3 shows that the distributions of the adjacency matrices before and after change points are different, which implies that one can identify the change points with probability 1.

Another difference is that in our paper, the signal-to-noise ratio is inflated by $\sqrt{T}$ compared to Wang et al. (2018a). This is due to the fact that we estimate the latent positions separately for every single network, while the graphons were estimated based on a version of sample average of the adjacency matrices in Wang et al. (2018a). The reason we estimate the positions separately roots in the difficulty of deriving theoretical properties of eigenvectors of a sample average matrices. A possible alternative and potentially improved method is to utilize the omnibus embedding (Levin et al., 2017), which however only handles the same latent positions in multiple networks.

We allow the dimensionality $d$ to grow unbounded, provided that Assumption 2 holds. The dimensionality $d$ is essentially the low rank condition imposed in Wang et al. (2018a). The upper bound on the rank $r$ in Wang et al. (2018a) comes into play with the term $\sqrt{r}$, while we have $d^{3/2}$ here. The difference again is rooted in the estimation of the latent positions, although we do not claim optimality here.

The sequence $a_T$ can diverge at any arbitrarily slow rate. We will explain the role of $a_T$ after we state Theorem 1.

Finally, we make connections between Assumptions 1 and 2. Recall that we use Assumption 1 in Lemma 3, where only one observation is available for each distribution, i.e. $\Delta = 1$, $\rho = 0$ and $T = 2$. Ignoring the universal constants, the only difference left between Assumptions 1 and 2 is the term $d^{3/2}$. Of course, if $d = O(1)$, then this is also a universal constant, and there is no difference left. The interesting thing happens when $d$ is allowed to diverge faster than the poly-logarithm term. Assumption 1 is required to differentiate two different distributions, which roughly speaking is related to a testing task; while Assumption 2 is used below in Theorem 1 with the purpose of consistent localization, which is an estimation problem. To this end, the extra $d^{3/2}$ in Assumption 2 is a piece of evidence that estimation is a harder problem than a testing one.
Theorem 1. Let data be from Model 1 and satisfy Assumption 2. Assume the following.

- The tuning parameter $\tau$ in Algorithm 2 satisfies
  \[ c_{r,1}T^{1/2} \max\{\sqrt{\log(n \lor T)}, d^{3/2}\} < \tau < c_{r,2}\kappa\sqrt{\Delta n(1-\rho)}, \]  \hspace{1cm} (4)

  where $c_{r,1}, c_{r,2} > 0$ are universal constants depending on all the universal constants in Model 1 and Assumption 2.

- The tuning parameter $d$ in Algorithms 1 and 2 are the true dimension $d$ of the latent positions.

- The intervals satisfy
  \[ \max_{m=1, \ldots, M} (\alpha_m - \beta_m) \leq C_R \Delta, \]  \hspace{1cm} (5)

  where $C_R > 3/2$ is a universal constant.

Let $\{\hat{\eta}_k\}_{k=1}^K$ be the output of Algorithm 2. We have that

\[ \mathbb{P} \left\{ \hat{K} = K, \quad |\hat{\eta}_k - \eta_k| \leq C \frac{T \max\{\log(n \lor T), d^3\}}{\kappa^2 n(1-\rho)}, \forall k \right\} \]

\[ \geq 1 - C(n \lor T)^{-c} - CTe^{-n} - \exp \left( \log(T/\Delta) - (4C_R)^{-1}T^{-1}M \Delta \right), \]

where $C, c > 0$ are universal constants depending only on the other universal constants.

The proof of Theorem 1 can be found in Appendix D, following two sets of lemmas – technical details on estimating the latent positions and on change point analysis, collected in Appendices B and C, respectively.

It can be seen from Theorem 1 that with probability tending to 1, as $T$ diverges, we have that $\hat{K} = K$ and

\[ \max_{k=1, \ldots, K} \frac{|\hat{\eta}_k - \eta_k|}{\Delta} \leq \max_{k=1, \ldots, K} C \frac{T \max\{\log(n \lor T), d^3\}}{\Delta \kappa^2 n(1-\rho)} \leq C \frac{T \max\{\log(n \lor T), d^3\}}{\Delta \kappa^2 n(1-\rho)} \to 0, \]

where the second inequality follows from the definition of $\kappa$ and the convergence follows from Assumption 2, with the aid of an arbitrarily diverging sequence $a_T$. This implies that the change point estimators we obtain are consistent, with a vanishing localization rate.

Algorithm 2 in fact can handle networks of varying size. For instance, if we do not allow for the dependence across time, then Theorem 1 holds provided that all network sizes are of the same order, which amounts to $c_1 n \leq n_t \leq c_2 n$, $t = 1, \ldots, T$, for universal constants $c_1, c_2 > 0$.

In Theorem 1, we assume that the input $d$ should be the true dimension. This is a seemingly strong condition. We would like to comment on this from a few different angles.

- In the context of stochastic block models, which are simpler than the RDPG models, the parameter $d$ is a lower bound on the number of communities. To estimate the number of communities in a stochastic block model is yet open, despite a tremendous amount of efforts (e.g. Bickel and Sarkar, 2016; Lei, 2016; Chen and Lei, 2018; Li et al., 2016; Franco Saldaña et al., 2017). We do not intend to propose a method to estimate the dimensionality here, but in practice, one could resort to the aforementioned papers.
Without a theoretically-justified method to estimate $d$, we need to discuss on the potential misspecification. If one overestimates $d$, i.e. with an input $d_1 > d$, then our method can still consistently estimate the change points under Assumption 2, with a sufficiently large constant $C_{SNR}$. This is due to the fact our statistic is a function of inner products of latent position estimators. Overestimating $d$ will only add extra noise which is in fact of the same order of the noise introduced when estimating the latent positions with true dimension $d$.

Another possible misspecification is underestimating the dimension $d$, i.e. the input of the algorithms is $d_2 < d$. This is a more damning issue than overestimating $d$, however it does not necessarily lead to inconsistent change point estimators. In order to illustrate this, we further discuss the conditions on $\tau$ in (4). The upper and lower bounds in (4) are the lower bound on the signals and the upper bound on the noise, on a large probability event, respectively. Now we assume a toy example where the true dimension $d = 3$. Recall the definition on the jump size $\kappa$ that

$$\kappa = \min_{k=1, \ldots, K} \sup_{z \in [0, 1]} \left| G_{\eta_k}(z) - G_{\eta_{k-1}}(z) \right|$$

$$= \min_{k=1, \ldots, K} \sup_{z \in [0, 1]} \left| \mathbb{P}_{\eta_k} \left\{ X^TY \leq z \right\} - \mathbb{P}_{\eta_{k-1}} \left\{ X^TY \leq z \right\} \right|$$

$$= \min_{k=1, \ldots, K} \sup_{z \in [0, 1]} \left| \mathbb{P}_{\eta_k} \left\{ \sum_{i=1}^3 X_iY_i \leq z \right\} - \mathbb{P}_{\eta_{k-1}} \left\{ \sum_{i=1}^3 X_iY_i \leq z \right\} \right|.$$ 

If we underestimate $d$ and we miss out the third dimension, our de facto jump size becomes

$$\kappa_1 = \min_{k=1, \ldots, K} \sup_{z \in [0, 1]} \left| \mathbb{P}_{\eta_k} \left\{ \sum_{i=1}^2 X_iY_i \leq z \right\} - \mathbb{P}_{\eta_{k-1}} \left\{ \sum_{i=1}^2 X_iY_i \leq z \right\} \right|.$$ 

Provided that the signal-to-noise ratio condition holds for $\kappa_1$, i.e.

$$\kappa_1 \sqrt{\Delta n(1 - \rho)} > C_{SNR} \sqrt{T} \max\{ \sqrt{\log(n \vee T)}, \, d^{3/2} \} + a_T,$$

with the notation defined in Assumption 2, Theorem 1 still holds.

We conclude this section by commenting on the random intervals. Without assuming (5), and using the trivial bound $C_R \leq T/\Delta$, it can be shown that we will achieve a larger localization error under a stronger scaling, both of which inflate by a factor of polynomials of $T/\Delta$. Finally, in order to guarantee that the probability tends to 1, one needs that $M \gtrsim T \Delta \log (\frac{T}{\Delta})$.

4 Numerical Experiments

4.1 Simulations

We now assess the performance of our proposed estimator NonPar-RDPG-CPD (Algorithm 2) in different scenarios, and compare our results with those produced by the network binary segmentation (NBS) algorithm (Wang et al., 2018a) and the modified neighbourhood smoothing (MNBS) algorithm (Zhao et al., 2019). The measurements we adopt are the absolute error $|\hat{K} - K|$, where

\[\text{Code implementing our method can be found in } \text{https://github.com/hernanmp/RDPG}.\]
\( \hat{K} \) and \( K \) are the numbers of the change point estimators and the true change points, respectively, and the one-sided Hausdorff distance defined as

\[
d(\hat{C}|C) = \max_{\eta \in \hat{C}} \min_{x \in C} |x - \eta|,
\]

where \( C \) is the set of true change points, and \( \hat{C} \) is the set of estimated change points. We also consider the metric \( d(C|\hat{C}) \). For Hausdorff distances, we report the medians over 100 Monte Carlo simulations, and for \( |\hat{K} - K| \), we report the means over 100 Monte Carlos trials. By convention, if \( \hat{C} = \emptyset \), we define \( d(\hat{C}|C) = \infty \) and \( d(C|\hat{C}) = -\infty \).

As for the choice of the tuning parameters, recall that NonPar-RDPG-CPD requires specifying the number of random intervals \( M \), the threshold \( \tau \) for declaring change points, and the dimension of the embedding \( d \). We choose \( \tau \) based on the model selection criteria from Zou et al. (2014). Specifically, we stack all the \( \hat{Y}_{ij}^t \) into one matrix \( \hat{Y} \in \mathbb{R}^{T \times n/2} \). Then for every potential model returned by \( \tau \), we calculate the BIC-type scores defined in Equation (2.4) in Zou et al. (2014), with \( \xi = \log^2(1/n)/5 \) along each column of \( \hat{Y} \). We aggregate all the scores along all the columns of \( \hat{Y} \) producing a single score for each model, e.g. each \( \tau \). We select the model with the smallest score.

As for the dimension of the latent positions \( d \), we set it as 10. We find the procedure very robust with the choice of \( d \), which supports our discussions on the misspecification after Theorem 1. We also set \( M = 120 \). As for NBS, we follow the proposal by the authors in Wang et al. (2018a) setting \( \tau \) to be of order \( n \log^2(T) \). For the MNBS, we use the default choice of its tuning parameters with code generously provided by the authors of Zhao et al. (2019).

**Disclaimer**: We would like to emphasize that the comparisons to the competitors might not be fair, due to the fact that the tuning parameter choosing schemes in Zhao et al. (2019) and Wang et al. (2018a) are not meant for dependent networks.

We construct four different models, in each of which, \( T = 150 \) and \( K = 2 \). The locations of the change points are evenly spaced, giving rise to three disjoint intervals \( A_1 = [1, 50] \), \( A_2 = [51, 100] \) and \( A_3 = [101, 150] \). As for the sizes of networks, we consider \( n \in \{100, 200, 300\} \).

**Scenario 1. Stochastic block models.** We construct two matrices of probabilities, \( P, Q \in \mathbb{R}^{n \times n} \). The matrix \( P \) satisfies

\[
P_{i,j} = \begin{cases} 
0.5, & i, j \in B_l, l \in \{1, \ldots, 4\}, \\
0.3, & \text{otherwise}, 
\end{cases}
\]

where \( B_1, \ldots, B_4 \) are evenly sized communities of nodes that form a partition of \( \{1, \ldots, n\} \). The matrix \( Q \) satisfies

\[
Q_{i,j} = \begin{cases} 
0.45, & i, j \in B_l, l \in \{1, \ldots, 4\}, \\
0.2, & \text{otherwise}. 
\end{cases}
\]

We then construct a sequence of matrices \( \{E(t)\}_{t=1}^T \subset \mathbb{R}^{n \times n} \) such that

\[
E_{i,j}(t) = \begin{cases} 
P_{i,j}, & t \in A_1 \cup A_3, \\
Q_{i,j}, & \text{otherwise}, 
\end{cases}
\]

for every \( i, j \in \{1, \ldots, n\} \).
The data are then generated with a correlation parameter \( \rho \in \{0, 0.5, 0.9\} \). Specifically, for any \( \rho \), we have \( A_{i,j}(1) \sim \text{Ber}(P_{i,j}(1)) \), and between two consecutive change points,

\[
A_{i,j}(t + 1) \sim \begin{cases} 
\text{Ber}((1 - E_{i,j}(t + 1))\rho + E_{i,j}(t + 1)), & A_{i,j}(t) = 1, \\
\text{Ber}((E_{i,j}(t + 1))(1 - \rho)), & A_{i,j}(t) = 0,
\end{cases}
\]

for \( 1 \leq i < j \leq n \).

**Scenario 2.** We first generate

\[
X_i(t) \sim \text{Uniform}[0.2, 0.8], \quad i = 1, \ldots, n, \ t \in A_1 \cup A_3.
\]

Then for any \( \varepsilon \in \{0.05, 0.15, 0.3\} \), we generate

\[
X_i(t) = \begin{cases} 
Z_i(t) + 0.2, & i \in \{1, \ldots, \lfloor n\varepsilon \rfloor\}, \\
Z_i(t), & \text{otherwise},
\end{cases}
\]

where \( Z_i(t) \sim \text{Uniform}[0.2, 0.8] \) for \( i \in \{1, \ldots, n\} \) and \( t \in A_2 \). Then we generate \( A_{i,j}(t) \sim \text{Ber}(X_i(t)X_j(t)) \).

**Scenario 3.** For \( t \in \{1, 101\} \), we generate \( Z_i(t) \sim \mathcal{N}(0, I_3) \), and for \( t \in A_1 \cup A_3 \setminus \{1, 101\} \), we generate

\[
Z_i(t) \sim \begin{cases} 
\mathcal{N}(0, I_3), & \text{with probability 0.9,} \\
Z_i(t - 1), & \text{with probability 0.1.}
\end{cases}
\]

We then set

\[
P_{i,j}(t) = \frac{\exp\{Z_i(t)^\top Z_j(t)\}}{1 + \exp\{Z_i(t)^\top Z_j(t)\}}.
\]

Furthermore, we generate \( P_{i,j}(51) \sim \text{Beta}(100, 100) \), and for \( t \in \{52, \ldots, 100\} \) we generate

\[
P(t) = \begin{cases} 
P(t - 1), & \text{with probability 0.9,} \\
\sim \text{Beta}(100, 100), & \text{with probability 0.1.}
\end{cases}
\]

Once the mean matrices \( \{P(t)\}_{t=1}^{T} \sim \mathbb{R}^{n \times n} \) have been constructed, we independently draw \( A_{i,j}(t) \sim \text{Ber}(P_{i,j}(t)) \), for all \( i, j \in \{1, \ldots, n\} \) and \( t \in \{1, \ldots, T\} \).

**Scenario 4.** For \( t \in \{1, 101\} \) we generate \( X_i \in \mathbb{R}^5 \) as

\[
X_i(t) \sim \text{Dirichlet}(1, 1, 1, 1, 1),
\]

for all \( i \in \{1, \ldots, n\} \). Then for \( t \in A_1 \cup A_3 \setminus \{1, 101\} \),

\[
X_i(t) = \begin{cases} 
X_i(t - 1), & \text{with probability 0.9,} \\
\sim \text{Dirichlet}(1, 1, 1, 1, 1), & \text{otherwise},
\end{cases}
\]

for all \( i \in \{1, \ldots, n\} \).
Table 1: Scenario 1

| Method          | $n$ | $\rho$ | $|K - \hat{K}|$ | $d(\hat{C}|C)$ | $d(C|\hat{C})$ |
|-----------------|-----|--------|----------------|----------------|----------------|
| NonPar-RDPG-CPD | 300 | 0      | 0.1            | 1.0            | 1.0            |
| NBS             | 300 | 0      | 0.0            | 1.0            | 1.0            |
| MNBS            | 300 | 0      | 1.16           | 50.0           | 0.0            |
| NonPar-RDPG-CPD | 200 | 0      | 0.0            | 1.0            | 1.0            |
| NBS             | 200 | 0      | 0.0            | 1.0            | 1.0            |
| MNBS            | 200 | 0      | 1.92           | $\inf$        | $-\inf$        |
| NonPar-RDPG-CPD | 100 | 0      | 0.2            | 1.0            | 1.0            |
| NBS             | 100 | 0      | 0.0            | 1.0            | 1.0            |
| MNBS            | 100 | 0      | 0.84           | 50.0           | 0.0            |
| NonPar-RDPG-CPD | 300 | 0.5    | 0.0            | 0.0            | 0.0            |
| NBS             | 300 | 0.5    | 21.2           | 1.0            | 43.0           |
| MNBS            | 300 | 0.5    | 0.0            | 0.0            | 0.0            |
| NonPar-RDPG-CPD | 200 | 0.5    | 0.04           | 0.0            | 0.0            |
| NBS             | 200 | 0.5    | 21.3           | 1.0            | 4.30           |
| MNBS            | 200 | 0.5    | 0.0            | 0.0            | 0.0            |
| NonPar-RDPG-CPD | 100 | 0.5    | 0.16           | 0.0            | 0.0            |
| NBS             | 100 | 0.5    | 21.3           | 1.0            | 42.0           |
| MNBS            | 100 | 0.5    | 0.12           | 0.0            | 0.0            |
| NonPar-RDPG-CPD | 300 | 0.9    | 0.0            | 0.0            | 0.0            |
| NBS             | 300 | 0.9    | 21.0           | 1.0            | 43.0           |
| MNBS            | 300 | 0.9    | 3.12           | 0.0            | 36.0           |
| NonPar-RDPG-CPD | 200 | 0.9    | 0.0            | 0.0            | 0.0            |
| NBS             | 200 | 0.9    | 21.0           | 1.0            | 43.0           |
| MNBS            | 200 | 0.9    | 2.88           | 0.0            | 35.0           |
| NonPar-RDPG-CPD | 100 | 0.9    | 0.0            | 1.0            | 1.0            |
| NBS             | 100 | 0.9    | 21.04          | 1.0            | 43.0           |
| MNBS            | 100 | 0.9    | 3.28           | 0.0            | 35.0           |

for all $i \in \{1, \ldots, n\}$. We also have

$$X_i(51) \sim \begin{cases} \text{Dirichlet}(500, 500, 500, 500, 500), & i \in \{1, \ldots, \lfloor n\varepsilon \rfloor\}, \\ \text{Dirichlet}(1, 1, 1, 1, 1), & i \in \{\lfloor n\varepsilon \rfloor + 1, \ldots, n\}, \end{cases}$$

and for $t \in A_2 \backslash \{51\}$,

$$X_i(t) \sim \begin{cases} X_i(t - 1), & \text{with probability 0.9}, \\ \text{Dirichlet}(500, 500, 500, 500, 500), & \text{with probability 0.1 if } i \in \{1, \ldots, \lfloor n\varepsilon \rfloor\}, \\ \text{Dirichlet}(1, 1, 1, 1, 1), & \text{with probability 0.1, if } i \in \{\lfloor n\varepsilon \rfloor + 1, \ldots, n\}, \end{cases}$$

for all $i \in \{1, \ldots, n\}$, where $\varepsilon \in \{0.05, 0.15, 0.3\}$.

Examples of matrices $A(t)$ generated in each scenario are depicted in Figures 2-3. We can see qualitative differences among Scenarios 1-4. In particular, Scenario 1 produces adjacency matrices with block structure. Interpretation is less clear for the other models, but we see that Scenario 3 seems to generate more dense graphs than Scenarios 2 and 4.

Results comparing NonPar-RDPG-CPD with NBS are provided in Tables 1-4. We observe that, overall, NonPar-RDPG-CPD provides generally reliable estimation of the number of change points and their locations.
Figure 2: The top row shows two instances of data generated in Scenario 1. The left panel corresponds to $A(t)$ for $t$ before the first change point, and the right panel to $A(t)$ between the first and second change points. The bottom row shows the corresponding plots for Scenario 2 with $\varepsilon = 0.05$.

In Scenario 1 with $\rho = 0$, a model where the marginal distributions of $A(t)$ only change in mean, we see from Table 1 that NBS outperforms our proposed approach. This does not come as a surprise since NBS is designed to detect change points in mean. However, as $\rho$ increases and the number of samples decreases, the most robust method seems to be NonPar-RDPG-CPD.

Scenario 2 poses an interesting example where the behaviour of only a fraction of nodes in the network changes at the change points. Furthermore, the data are generated under an RDPG model. As shown in Table 2, NonPar-RDPG-CPD seems to be the best method for estimating the number of change points. A possible explanation is that the underlying changes in the distributions of $A(t)$ not only occur at the level of the means, and hence the NBS might not be the ideal for this
Figure 3: The top row shows two instances of data generated in Scenario 3. The left panel corresponds to $A(t)$ for $t$ before the first change point, and the right panel to $A(t)$ between the first and second change points. The bottom row shows the corresponding plots for Scenario 4 with $\varepsilon = 0.05$.

scenario even though it outperforms MNBS in this framework. Our method was constructed under the assumption of the RDPG model.

To assess the robustness of our method to misspecification, we can look at the performance of our method in the context of Scenario 3 which is not an RDPG. Interestingly, Table 3 shows that NonPar-RDPG-CPD is the best in this model with MNBS coming in second. In contrast, NBS suffers greatly, overestimating the number of change points. This makes sense since between change points, the latent positions $X(t)$ remain constant with probability 0.9 and change with probability 0.1. Hence, some of these changes in $X(t)$ could be confused as change points by NBS.

Finally, Scenario 4 consists of an example of Model 1. However, similarly as Scenario 2, the
Table 2: Scenario 2

| Method          | n   | $\varepsilon$ | $|K - \hat{K}|$ | $d(\hat{C}|C)$ | $d(C|\hat{C})$ |
|-----------------|-----|---------------|-----------------|----------------|----------------|
| NonPar-RDPG-CPD | 300 | 0.3           | 0.04            | 0.0            | 0.0            |
| NBS             | 300 | 0.3           | 0.28            | 1.0            | 1.0            |
| MNBS            | 300 | 0.3           | 0.76            | 0.0            | 21.0           |
| NonPar-RDPG-CPD | 200 | 0.3           | 0.0             | 0.0            | 0.0            |
| NBS             | 200 | 0.3           | 0.32            | 1.0            | 1.0            |
| MNBS            | 200 | 0.3           | 0.48            | 0.0            | 1.0            |
| NonPar-RDPG-CPD | 100 | 0.3           | 0.08            | 3.0            | 0.0            |
| NBS             | 100 | 0.3           | 0.08            | 1.0            | 1.0            |
| MNBS            | 100 | 0.3           | 0.64            | 0.0            | 18.0           |
| NonPar-RDPG-CPD | 300 | 0.15          | 0.0             | 2.0            | 2.0            |
| NBS             | 300 | 0.15          | 0.4             | 1.0            | 1.0            |
| MNBS            | 300 | 0.15          | 0.76            | 0.0            | 21.0           |
| NonPar-RDPG-CPD | 200 | 0.15          | 0.04            | 3.0            | 3.0            |
| NBS             | 200 | 0.15          | 0.28            | 1.0            | 1.0            |
| MNBS            | 200 | 0.15          | 0.76            | 0.0            | 20.0           |
| NonPar-RDPG-CPD | 100 | 0.15          | 0.28            | 4.0            | 10.0           |
| NBS             | 100 | 0.15          | 0.32            | 1.0            | 1.0            |
| MNBS            | 100 | 0.15          | 0.48            | 1.0            | 5.0            |
| NonPar-RDPG-CPD | 300 | 0.05          | 0.72            | 36.0           | 5.0            |
| NBS             | 300 | 0.05          | 0.84            | 1.0            | 9.0            |
| MNBS            | 300 | 0.05          | 1.24            | 1.0            | 21.0           |
| NonPar-RDPG-CPD | 200 | 0.05          | 0.64            | 37.0           | 6.0            |
| NBS             | 200 | 0.05          | 0.76            | 3.0            | 11.0           |
| MNBS            | 200 | 0.05          | 0.6             | 4.0            | 8.0            |
| NonPar-RDPG-CPD | 100 | 0.05          | 0.72            | 19.0           | 15.0           |
| NBS             | 100 | 0.05          | 1.4             | inf            | inf            |
| MNBS            | 100 | 0.05          | 1.88            | inf            | inf            |

Table 3: Scenario 3

| Method          | n   | $|K - \hat{K}|$ | $d(\hat{C}|C)$ | $d(C|\hat{C})$ |
|-----------------|-----|-----------------|----------------|----------------|
| NonPar-RDPG-CPD | 300 | 0.24            | 0.0            | 0.0            |
| NBS             | 300 | 15.04           | 1.0            | 43.0           |
| MNBS            | 300 | 0.84            | 25             | 36             |
| NonPar-RDPG-CPD | 200 | 0.08            | 0.0            | 0.0            |
| NBS             | 200 | 14.4            | 43.0           | 1.0            |
| MNBS            | 200 | 0.84            | 23             | 36             |
| NonPar-RDPG-CPD | 100 | 0.52            | 3.0            | 5.0            |
| NBS             | 100 | 13.96           | 1.0            | 43.0           |
| MNBS            | 100 | 1.16            | 23             | 35             |

change points correspond to shifts in the behaviour of only some of the nodes in the network. In particular, Table 4 suggests that our method performs reasonably well, improving its performance when the signal-to-noise ratio increases. This is different from the NBS which once again tends to overestimate the number of change points. As for the MNBS, we see that this method is unable to detect the change points in this example.
Figure 4: Examples of adjacency matrices, down-sampled to a 100 × 100, between the change points estimated by NonPar-RDPG-CPD in the zebrafish example. From left to right and from top to bottom, the panels correspond to \( t = 3, 7, 15, 32, 40, 45, 52, 60, 65, 75, 80 \) and 87.
Table 4: Scenario 4

| Method            | $n$  | $\varepsilon$ | $|\hat{K} - K|$ | $d(\hat{C}|C)$ | $d(C|\hat{C})$ |
|-------------------|------|---------------|-----------------|----------------|----------------|
| NonPar-RDPG-CPD   | 300  | 0.3           | 0.72            | 35.0           | 12.0           |
| NBS               | 300  | 0.3           | 19.4            | 1.0            | 43.0           |
| MNBS              | 300  | 0.3           | 2.0             | inf            | inf            |
| NonPar-RDPG-CPD   | 200  | 0.3           | 0.84            | 40.0           | 10.0           |
| NBS               | 200  | 0.3           | 19.4            | 1.0            | 43.0           |
| MNBS              | 200  | 0.3           | 2.0             | inf            | inf            |
| NonPar-RDPG-CPD   | 100  | 0.3           | 1.0             | 30.0           | 20.0           |
| NBS               | 100  | 0.3           | 9.44            | 3.0            | 41.0           |
| MNBS              | 100  | 0.3           | 2.0             | inf            | inf            |
| NonPar-RDPG-CPD   | 300  | 0.15          | 0.8             | 34.0           | 17.0           |
| NBS               | 300  | 0.15          | 20.24           | 1.0            | 43.0           |
| MNBS              | 300  | 0.15          | 2.0             | inf            | inf            |
| NonPar-RDPG-CPD   | 200  | 0.15          | 0.96            | 40.0           | 11.0           |
| NBS               | 200  | 0.15          | 17.0            | 1.0            | 43.0           |
| MNBS              | 200  | 0.15          | 2.0             | inf            | inf            |
| NonPar-RDPG-CPD   | 100  | 0.15          | 0.84            | 34.0           | 18.0           |
| NBS               | 100  | 0.15          | 10.64           | 1.0            | 41.0           |
| MNBS              | 100  | 0.15          | 2.0             | inf            | inf            |
| NonPar-RDPG-CPD   | 300  | 0.05          | 0.80            | 33.0           | 17.0           |
| NBS               | 300  | 0.05          | 20.48           | 1.0            | 43.0           |
| MNBS              | 300  | 0.05          | 2.0             | inf            | inf            |
| NonPar-RDPG-CPD   | 200  | 0.05          | 0.88            | 38.0           | 19.0           |
| NBS               | 200  | 0.05          | 17.56           | 1.0            | 43.0           |
| MNBS              | 200  | 0.05          | 2.0             | inf            | inf            |
| NonPar-RDPG-CPD   | 100  | 0.05          | 1.04            | 32.0           | 16.0           |
| NBS               | 100  | 0.05          | 11.48           | 3.0            | 41.0           |
| MNBS              | 100  | 0.05          | 2.0             | inf            | inf            |

4.2 Real data

Our goal is to estimate change points in the context of the neuronal activity in larval zebrafish. The data consist of simultaneous whole-brain neuronal activity data at near single cell resolution (Prevedel et al., 2014). The original data format is a matrix of size 5379 $\times$ 5000. This corresponds to the neural activity of 5379 neurons over 5000 frames, where one second in time corresponds to 20 frames.

To construct the final sequence of networks, we proceed as in Lyzinski et al. (2017). Specifically, we first remove artificial neurons leaving us with a 5105 $\times$ 5000 matrix. Then we bin the data into 100 non-overlapping periods. Each period corresponds to 2.5 seconds of the original data. The resulting time series is then $Z(t) \in \mathbb{R}^{5105 \times 50}$ for $t \in \{1, \ldots, 100\}$. Following Lyzinski et al. (2017), we finally construct the adjacency matrices $A(t) \in \mathbb{R}^{5105 \times 5105}$ as

$$A_{i,j}(t) = 1\{\text{corr}(Z_i(t), Z_j(t)) > 0.7\}, \quad t = 1, \ldots, T,$$

where $T = 100$.

With the time series $\{A(t)\}_{t=1}^{T}$ in hand, we proceed to run change point detection with Algorithm 2. The implementation details are the same as those in Section 4.1. However, to facilitate
computations at every instance of time we randomly sample 800 nodes in the network and work with a down-sampled version of $A(t)$. After running our method, we estimate change points at locations 5, 10, 29, 36, 42, 50, 57, 62, 71, 79, 85, and 89. In the original 250 seconds time stamp, the changes correspond to 12.5, 25.0, 72.5, 90.0, 105.0, 125.0, 142.5, 155.0, 177.5, 197.5, 212.5, and 222.5 seconds. Simple inspection suggests that our estimated change points are in agreement with the extracted intensity signal of Ca2+ fluorescence using spatial filters in Figure 3 (c) in Prevedel et al. (2014). As remarked in Park et al. (2015), a lab scientist induced a change-point at the 16th second, by giving an olfactory stimulus to the zebrafish. In the scale of our time series $\{A(t)\}_{t=1}^T$, this change corresponds to $t=6$ which seems to be captured by our algorithm that detected a change point at $t=5$.

We also considered change point detection with the algorithm NBS (Wang et al., 2018a). The set of estimated change points is roughly the same to that estimated by NonPar-RDPG-CPD: 10, 14, 22, 26, 32, 36, 42, 50, 58, 62, 66, 72, 80, and 90. One important difference, however, is that NBS did not detect a change point near $t=6$, the change point created by the lab scientist. We also tried the MNBS method (Zhao et al., 2019), but this only detected changes at 14, 45, 66, 80. Finally, we have included Figure 4 which shows down-sampled versions of $A(t)$ for values of $t$ between estimated change points. This reinforces our intuition that the structural breaks estimated with NonPar-RDPG-CPD are meaningful.

A Technical details of Section 3.1

Proof of Lemma 2. For any $i, j \in \{1, \ldots, n\}$, $i \neq j$, it holds that

$$P\{A_{ij}|X_i, X_j\} = X_i^T X_j = X_i^T U^T U X_j,$$

for any orthogonal operator $U \in R^{d \times d}$. In this proof, by the equivalence in terms of the distributions $F$ and $\tilde{F}$, we mean the equivalence up to a rotation, which is detailed in Definition 3. Without loss of generality, if a rotation is needed, we omit it in the notation.

We divide this proof into two cases: (a) $d = 1$ and (b) $d > 1$.

(a) $p = 1$.

Since the entries of $A$ and $\tilde{A}$ are Bernoulli random variables, they only take values in $\{0, 1\}^{n \times n}$. For any symmetric matrix $v \in \{0, 1\}^{n \times n}$, we have

$$P\{A = v\} = E \left\{ E \left( \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} 1 \{ A_{ij} = v_{ij} \} \left| \{X_i\}_{i=1}^{n} \right. \right) \right\}$$

$$= E \left( \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} \{ (X_i X_j) v_{ij} + (1 - X_i X_j)(1 - v_{ij}) \} \right).$$

(6)

If $\mathcal{L} = \tilde{\mathcal{L}}$, then we have the following.

- If $v_{ij} \equiv 1$, then

$$E \left( \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} (X_i X_j) \right) = \left\{ E(X_1^{n-1}) \right\}^n.$$
which implies that $E_F(X_i^{n-1}) = E_F(\tilde{X}_i^{n-1})$. Note that in order to have an edge, $n \geq 2$, which implies that $n - 1 \geq 1$.

- If there is one and only one pair $(i, j)$, $i < j$, such that $v_{ij} = v_{ji} = 0$, and $v_{kl} = 1$, $(k, l) \notin \{(i, j), (j, i)\}$, then without loss of generality, we let $(i, j) = (1, 2)$. If $n = 2$, then

\[(6) = 1 - \{E(X_1)\}^2,\]

which implies that $E_F(X_1^{n-1}) = E_F(\tilde{X}_1^{n-1})$.

If $n \geq 3$, then

\[(6) = E \left[ \prod_{i=3}^{n} \prod_{j=i+1}^{n} (X_i X_j) \cdot \prod_{r=1}^{2} \prod_{l=3}^{n} (X_r X_l) \cdot (1 - X_1 X_2) \right] \]

\[= \{E(X_1^{n-2})\}^2 \{E(X_1^{n-1})\}^{n-2} - \{E(X_1^{n-1})\}^n,\]

which implies $E_F(X_1^{n-2}) = E_F(\tilde{X}_1^{n-2})$.

- If $n \geq 3$, then for $k \in \{2, \ldots, n - 1\}$, without loss of generality, let $v_{1j} = v_{j1} = 0$, $j \in \{2, \ldots, k + 1\}$, and $v_{rs} = v_{sr} = 1$ otherwise. We have that

\[(6) = E \left[ \prod_{i=k+2}^{n} \prod_{j=i+1}^{n} (X_i X_j) \cdot \prod_{l=k+2}^{n} \prod_{i=l+1}^{k+1} (X_l X_i) \cdot \prod_{r=2}^{k+1} (1 - X_1 X_r) \right] \]

\[= \{E(X_1^{n-1})\}^{n-k-1} E \left[ \prod_{i=1}^{k+1} X_i^{n-k-1} \cdot \prod_{r=2}^{k+1} (1 - X_1 X_r) \right] \]

\[= \{E(X_1^{n-1})\}^{n-k-1} \sum_{r=0}^{k} \binom{k}{r} (-1)^r E(X_1^{n-k+r}) \left[E(X_1^{n-k})\right]^r. \tag{7} \]

Note that, if $k = 2$, then the summands in (7) include moments $n - 1$, $n - 2$ and $n - 3$. We have already shown that $E_F(X_1^{n-1}) = E_F(\tilde{X}_1^{n-1})$ and $E_F(X_1^{n-2}) = E_F(\tilde{X}_1^{n-2})$, therefore (7) implies that $E_F(X_1^{n-3}) = E_F(\tilde{X}_1^{n-3})$.

- By induction, for $n > k_0$ and $k_0 \geq 3$, if it holds that $E_F(X_1^{n-s}) = E_F(\tilde{X}_1^{n-s})$, $s = 1, \ldots, k_0$, then we have $E_F(X_1^{n-k_0-1}) = E_F(\tilde{X}_1^{n-k_0-1})$, due to the fact that the summands in (7) include moment $n - s$, $s = 1, \ldots, k_0 + 1$.

We conclude that if $L = \tilde{L}$, then $E_F(X_1^k) = E_F(\tilde{X}_1^k)$, $k = 1, \ldots, n - 1$.

If $E_F(X_1^k) = E_F(\tilde{X}_1^k)$, $k = 1, \ldots, n - 1$, then it follows from that for any $v$,

\[(6) = \sum_{l=0}^{\sum_{i<j} 1\{v_{ij} = 0\}} \left(\sum_{i<j} 1\{v_{ij} = 0\}\right)^l (-1)^{\sum_{i<j} 1\{v_{ij} = 0\} - l} \]
For any symmetric matrix $v$ which is a function solely of $E_{1}$, $k = 1, \ldots, n - 1$. We, therefore, have that $L = \tilde{L}$.

(b) $d > 1$.

Since the entries of $A$ and $\tilde{A}$ are Bernoulli random variables, they only take values in $\{0, 1\}^{n \times n}$. For any $v \in \{0, 1\}^{n \times n}$, we have

$$P\{A = v\} = \mathbb{E}\left\{ \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} \mathbb{1}\{A_{ij} = v_{ij}\} \middle| \{X_l\}_{l=1}^{n}\right\}$$

$$= \mathbb{E}\left[ \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} \left( \sum_{k=1}^{d} X_{i,k}X_{j,k} \right) v_{ij} + \left( 1 - \sum_{k=1}^{d} X_{i,k}X_{j,k} \right) (1 - v_{ij}) \right].$$

If $L = \tilde{L}$, then we have the following.

- If $v_{ij} \equiv 1$, then

$$= \mathbb{E}\left[ \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} \left( \sum_{k=1}^{d} X_{i,k}X_{j,k} \right) \right] = \mathbb{E}\left[ \prod_{i=2}^{n} \left( \sum_{k=1}^{d} X_{i,k}X_{j,k} \right) \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} \left( \sum_{k=1}^{d} X_{i,k}X_{j,k} \right) \right]$$

$$= \mathbb{E}\left[ \prod_{k_2, \ldots, k_i=1}^{d} \left( \prod_{i=2}^{n} X_{1,k_i} \right) \prod_{i=2}^{n} \prod_{j=i+1}^{n} \left( \sum_{k=1}^{d} X_{i,k}X_{j,k} \right) \right],$$

where the third identity follows from the independence assumption. Note that for any $(k_2, \ldots, k_n) \in \{1, \ldots, p\}^{\otimes(n-1)}$, the term

$$\mathbb{E}\left[ \prod_{i=2}^{n} X_{1,k_i} \right] \prod_{i=2}^{n-1} \prod_{j=i+1}^{n} \left( \sum_{k=1}^{d} X_{i,k}X_{j,k} \right)$$

in (9) does not involve $X_1$, and the term

$$\mathbb{E}\left( \prod_{i=2}^{n} X_{1,k_i} \right)$$

includes all possible terms of the form

$$\mathbb{E}\left( \prod_{l=1}^{d} X_{1,l}^{k_l} \right), \quad \sum_{l=1}^{d} k_l = n - 1, \quad k_l \geq 0, \ l \in \{1, \ldots, d\}. \quad (10)$$
Due to the exchangeability, we conclude that (8) is solely a function of polynomials of (10).

If \( n = 2 \), then due to Definition 1, we have that \( L = \tilde{L} \) implies that

\[
E \left( \prod_{l=1}^{d} X_{1,l}^{k_l} \right) = E \left( \prod_{l=1}^{d} \tilde{X}_{1,l}^{k_l} \right), \quad \sum_{l=1}^{d} k_l = n - 1, \quad k_l \geq 0, \ l \in \{1, \ldots, d\}.
\]

• If \( n \geq 3 \), then we prove by induction. Assume that

\[
E \left( \prod_{l=1}^{p} X_{1,l}^{k_l} \right) = E \left( \prod_{l=1}^{p} \tilde{X}_{1,l}^{k_l} \right), \quad \sum_{l=1}^{p} k_l = n - k, \ldots, n - 1, \quad k_l \geq 0, \ l \in \{1, \ldots, p\},
\]

where \( n - 1 \geq n - k \geq 2 \). We now proceed to prove that

\[
E \left( \prod_{l=1}^{p} X_{1,l}^{k_l} \right) = E \left( \prod_{l=1}^{p} \tilde{X}_{1,l}^{k_l} \right), \quad \sum_{l=1}^{p} k_l = n - k - 1, \ldots, n - 1, \quad k_l \geq 0, \ l \in \{1, \ldots, d\}. \tag{11}
\]

To show this, we assume that \( v_{1j} = v_{j1} = 0, \ j \in \{2, \ldots, k + 1\} \), and \( v_{rs} = 1 \) otherwise. We have that

\[
(8) = E \left[ \prod_{j=2}^{k+1} \left( 1 - \sum_{s=1}^{n} X_{1,s} \cdot X_{j,s} \right) \cdot \prod_{l=k+2}^{n} \left( \sum_{s=1}^{d} X_{1,s} \cdot X_{l,s} \right) \cdot \prod_{i=2}^{n-1} \prod_{r=i+1}^{n} \left( \sum_{s=1}^{d} X_{i,s} \cdot X_{r,s} \right) \right]
\]

\[
= (-1)^k E \left\{ \prod_{l=k+2}^{n} \left( \sum_{s=1}^{d} X_{1,s} \cdot X_{l,s} \right) \cdot \prod_{i=2}^{n-1} \prod_{r=i+1}^{n} \left( \sum_{s=1}^{d} X_{i,s} \cdot X_{r,s} \right) \right\} + f(X)
\]

\[
= (-1)^k \sum_{s_{k+2}, \ldots, s_n=1}^{d} E \left( \prod_{l=k+2}^{n} X_{1,s_l} \right) E \left\{ \prod_{l=k+2}^{n} \left( \sum_{s=1}^{d} X_{l,s} \right) \cdot \prod_{i=2}^{n-1} \prod_{r=i+1}^{n} \left( \sum_{s=1}^{d} X_{i,s} \cdot X_{r,s} \right) \right\} + f(X),
\]

where \( f(X) \) is solely a function of

\[
E \left( \prod_{l=1}^{d} X_{1,l}^{k_l} \right), \quad \sum_{l=1}^{d} k_l = n - k, \ldots, n - 1, \quad k_l \geq 0, \ l \in \{1, \ldots, d\}.
\]

Note that

\[
\sum_{s_{k+2}, \ldots, s_n=1}^{d} E \left( \prod_{l=k+2}^{n} X_{1,s_l} \right)
\]

is a function of

\[
E \left( \prod_{l=1}^{d} X_{1,l}^{k_l} \right), \quad \sum_{l=1}^{d} k_l = n - k - 1, \quad k_l \geq 0, \ l \in \{1, \ldots, d\}.
\]

Therefore we have shown (11).
To this end, we have that $\mathcal{L} = \tilde{\mathcal{L}}$ implies that

$$
\mathbb{E} \left( \prod_{l=1}^{d} X_{k_l}^{i_l, l} \right) = \mathbb{E} \left( \prod_{l=1}^{d} \tilde{X}_{k_l}^{i_l, l} \right), \quad \sum_{l=1}^{d} k_l = 1, \ldots, n - 1, \quad k_l \geq 0, \ l \in \{1, \ldots, d\}. \tag{12}
$$

To show that (12) implies that $\mathcal{L} = \tilde{\mathcal{L}}$, we notice that for any $v$,

$$
(8) = \sum_{l=0}^{\sum_{i<j} 1\{v_{ij}=0\}} \left( -1 \right)^{l} \left[ \sum_{\left\{(i_1,j_1),\ldots,(i_l,j_l)\right\} \in \{(i,j): v_{ij}=0, i<j\}} \left( \prod_{l=1}^{d} \sum_{k=1}^{\sum_{i\geq 1, j\leq n} 1\{k_{ij} \geq 0\} \quad l \in \{1, \ldots, d\}}, \right) \right)
$$

which is solely a function of

$$
\mathbb{E} \left( \prod_{l=1}^{d} X_{k_l}^{i_l, l} \right), \quad \sum_{l=1}^{d} k_l = 1, \ldots, n - 1, \quad k_l \geq 0, \ l \in \{1, \ldots, d\}.
$$

The final claim holds.

\(\square\)

**Proof of Lemma 3.** For simplicity, we assume $n$ is an even number. Let $\mathcal{O} = \{(i, n/2 + i), \ i = 1, \ldots, n/2\}$. Let

$$z_{*} \in \arg\sup_{z \in [0,1]} |G(z) - \tilde{G}(z)|.
$$

Note that

$$
\left| \sqrt{\frac{2}{n}} \sum_{(i,j) \in \mathcal{O}} \left( 1\{Y_{ij} \leq z_{*}\} - 1\{\tilde{Y}_{ij} \leq z_{*}\} \right) \right|
$$

$$
= \sqrt{\frac{2}{n}} \sum_{(i,j) \in \mathcal{O}} \left( \left[ 1\{Y_{ij} \leq z_{*}\} - \mathbb{E} \left[ 1\{Y_{ij} \leq z_{*}\} \right] \right) - \left[ 1\{\tilde{Y}_{ij} \leq z_{*}\} - \mathbb{E} \left[ 1\{\tilde{Y}_{ij} \leq z_{*}\} \right] \right) \right.
$$

$$
+ \left. \sqrt{\frac{n}{2}} \left[ \mathbb{E} \left[ 1\{Y_{ij} \leq z_{*}\} \right] - \mathbb{E} \left[ 1\{Y_{ij} \leq z_{*}\} \right] \right) \right]
$$

$$
\geq \sqrt{\frac{n}{2}} \left[ \mathbb{E} \left[ 1\{Y_{ij} \leq z_{*}\} \right] - \mathbb{E} \left[ 1\{Y_{ij} \leq z_{*}\} \right] \right) - \left| \sqrt{\frac{2}{n}} \sum_{(i,j) \in \mathcal{O}} \left( 1\{Y_{ij} \leq z_{*}\} - \mathbb{E} \left[ 1\{Y_{ij} \leq z_{*}\} \right] \right) \right|
$$

$$
- \left| \sqrt{\frac{2}{n}} \sum_{(i,j) \in \mathcal{O}} \left( 1\{\tilde{Y}_{ij} \leq z_{*}\} - \mathbb{E} \left[ 1\{\tilde{Y}_{ij} \leq z_{*}\} \right] \right) \right|
$$

$$
= \kappa_{0} \sqrt{n/2} - \sqrt{\frac{2}{n}} \sum_{(i,j) \in \mathcal{O}} \left( 1\{Y_{ij} \leq z_{*}\} - \mathbb{E} \left[ 1\{Y_{ij} \leq z_{*}\} \right] \right)
$$
- $\sqrt{\frac{2}{n}} \sum_{(i,j) \in O} \left( 1 \{ \tilde{Y}_{ij} \leq z_s \} - E \left[ 1 \{ \tilde{Y}_{ij} \leq z_s \} \right] \right)$.

(13)

Next, it follows from Hoeffding’s inequality that

$$\mathbb{P} \left\{ \max \left\{ \left| \sqrt{\frac{2}{n}} \sum_{(i,j) \in O} \left( 1 \{ Y_{ij} \leq z_s \} - E \left[ 1 \{ Y_{ij} \leq z_s \} \right] \right) \right| , \left| \sqrt{\frac{2}{n}} \sum_{(i,j) \in O} \left( 1 \{ \tilde{Y}_{ij} \leq z_s \} - E \left[ 1 \{ \tilde{Y}_{ij} \leq z_s \} \right] \right) \right| \right\} > \sqrt{\log(n)} \right\} \leq 2n^{-4}. \quad (14)$$

Combining (13) and (14), we have that with probability at least $1 - 2n^{-4}$,

$$\left| \sqrt{\frac{2}{n}} \sum_{(i,j) \in O} \left( 1 \{ Y_{ij} \leq z_s \} - E \left[ 1 \{ Y_{ij} \leq z_s \} \right] \right) \right| \geq \kappa_0 \sqrt{n/2} - 2 \sqrt{\log(n)}. \quad (15)$$

We then prove by contradiction. If $\mathcal{L} = \tilde{\mathcal{L}}$, then it follows from Hoeffding’s inequality that

$$\mathbb{P} \left\{ \left| \sqrt{\frac{2}{n}} \sum_{(i,j) \in O} \left( 1 \{ Y_{ij} \leq z_s \} - 1 \{ \tilde{Y}_{ij} \leq z_s \} \right) \right| \leq \sqrt{\log(n)} \right\} \geq 1 - 2n^{-4}. \quad (16)$$

Due to Assumption 1, (15) and (16) contradict with each other, which implies that $\mathcal{L} \neq \tilde{\mathcal{L}}$.

\[\square\]

**B Large probability events**

Define

$$\Delta_{s,e}^{t}(z) = \sum_{k=s+1}^{e} w_k \sum_{(i,j) \in O} \left( 1 \{ \hat{Y}_{ij}^{k} \leq z \} - E \left[ 1 \{ Y_{ij}^{k} \leq z \} \right] \right),$$

where

$$w_k = \begin{cases} \sqrt{\frac{2}{n} \frac{e-t}{(e-s)(t-s)}}, & k = s+1, \ldots, t, \\ -\sqrt{\frac{2}{n} \frac{t-s}{(e-s)(e-t)}}, & k = t+1, \ldots, e. \end{cases}$$

In this section, we are to show the following two events hold with probability tending to 1, as $(n \vee T) \to \infty$,

$$\mathcal{B}_1 = \left\{ \max_{0 \leq s \leq t \leq e \leq T} \Delta_{s,e}^{t} \leq C_0 \sqrt{\frac{T}{1 - \rho}} \max \{ \log(n \vee T), d^{3/2} \sqrt{\log(n \vee T)} \} \right\}$$

and

$$\mathcal{B}_2 = \left\{ \max_{0 \leq s < t < e \leq T} \sup_{z \in [0,1]} \left| \sqrt{\frac{2}{n(e-s)}} \sum_{k=s+1}^{e} \sum_{(i,j) \in O} \left( 1 \{ \hat{Y}_{ij}^{k} \leq z \} - E \left[ 1 \{ Y_{ij}^{k} \leq z \} \right] \right) \right| \right\}$$

26
Under Model 1, we have that
\[ \epsilon \geq \frac{T}{1 - \rho} \max \{ \log(n \vee T), d^{3/2} \sqrt{\log(n \vee T)} \}. \]

This is formally stated in Lemma 10. To reach there, we denote
\[ \mathcal{E}_1 = \left\{ \max_{t=1,\ldots,T} \|U_{P_t}^\top (A(t) - P_t) U_{P_t}\|_F \leq C_1 \sqrt{n \log(n \vee T)} \right\}, \]
\[ \mathcal{E}_2 = \left\{ \max_{t=1,\ldots,T} \| (A(t) - P_t) U_{P_t} \|_F \leq C_2 \sqrt{\log(n \vee T)} \right\}, \]
\[ \mathcal{E}_3 = \left\{ \max_{t=1,\ldots,T} \| A(t) - P_t \|_{op} \leq C_3 \sqrt{n} \right\} \]
and
\[ \mathcal{E}_4 = \left\{ 2^{-n} \min_{k=1,\ldots,K} \mu_d^k \leq \min_{t=1,\ldots,T} \lambda_d(P_t) \leq \max_{t=1,\ldots,T} \lambda_1(P_t) \leq (3/2) n \max_{k=1,\ldots,K} \mu_1^k \right\}, \]
where \( C_1 > 4\sqrt{6}, C_2 > 4\sqrt{6}, C_3 > 0 \) are universal constants.

**Lemma 4.** Under Model 1, for any \( t \in \{1, \ldots, T\} \), it holds that
\[ \mathbb{P}\{\lambda_{d+1}(P_t) = 0\} = 1. \]

**Proof.** For any \( t \in \{1, \ldots, T\} \), we have that
\[ P_t = X(t)(X(t))^\top. \]

For any realisation of \( X(t) \in \mathbb{R}^{n \times d} \), \( \lambda_{d+1}(P_t) = 0 \). Thus the final claim holds. \( \square \)

**Lemma 5.** Under Model 1, we have that
\[ \max \left\{ \mathbb{P}\{ \mathcal{E}_1 \mid \{X(t)\}_{t=1}^T \}, \mathbb{P}\{ \mathcal{E}_1 \} \right\} \geq 1 - (n \vee T)^{-c_1}, \]
\[ \max \left\{ \mathbb{P}\{ \mathcal{E}_2 \mid \{X(t)\}_{t=1}^T \}, \mathbb{P}\{ \mathcal{E}_2 \} \right\} \geq 1 - (n \vee T)^{-c_2} \]
and
\[ \max \left\{ \mathbb{P}\{ \mathcal{E}_3 \mid \{X(t)\}_{t=1}^T \}, \mathbb{P}\{ \mathcal{E}_3 \} \right\} \geq 1 - 4Te^{-n}, \]
where \( c_1, c_2 > 0 \) are universal constants depending on \( C_1 \) and \( C_2 \), respectively.

**Proof.** We start with \( \mathbb{P}\{ \mathcal{E}_1 \mid \{X(t)\}_{t=1}^T \} \). For any \((i, j) \in \{1, \ldots, d\}^2\) and any \( t \in \{1, \ldots, T\} \), it satisfies that
\[ [U_{P_t}^\top (A(t) - P_t) U_{P_t}]_{ij} = 2 \sum_{k=1}^{n-1} \sum_{l=k+1}^{n} (U_{P_t})_{ii} (A(t) - P_t)_{lk} (U_{P_t})_{kj} + \sum_{k=1}^{n} (U_{P_t})_{ki} (A(t) - P_t)_{kk} (U_{P_t})_{kj}. \]

For any \( \epsilon > 0 \), it holds that
\[ \mathbb{P}\left\{ \left| 2 \sum_{k=1}^{n-1} \sum_{l=k+1}^{n} (U_{P_t})_{ii} (A(t) - P_t)_{lk} (U_{P_t})_{kj} \right| > \epsilon \{X(t)\}_{t=1}^T \right\} \]
Lemma 6. Under Model 1, it holds that

\[
\frac{\epsilon^2}{8} \sum_{k=1}^{n-1} \sum_{l=k+1}^{n} (U_{P_l})^2_{kl} (U_{P_l})^2_{kj} \geq 2 \exp\left\{-\frac{\epsilon^2}{8}\right\},
\]

where the first inequality follows from Theorem 2.2.5 in Vershynin (2018), and the identity follows from the definitions of \( U_P \). Moreover,

\[
\left| \sum_{k=1}^{n} (U_{P_k})_{kl}(A(t) - P_t)_{kl}(U_{P_t})_{kj} \right| \leq \sum_{k=1}^{p} |(U_{P_k})_{kl}(U_{P_t})_{kj}| \leq \sqrt{\sum_{k=1}^{P} (U_{P_k})^2_{kl} \sum_{k=1}^{P} (U_{P_k})^2_{kj}} = 1.
\]

Combining (20), (21) and (22), and taking \( \epsilon \) to be \((C/2)\sqrt{\log(n \lor T)}\), we have that

\[
P\{\mathcal{E}_t^c | \{X(t)\}_{t=1}^T \leq 2T \exp\left\{-\frac{C_2^2}{32} \log(n \lor T)\right\} \leq (n \lor T)^{-c_1},
\]

where \( c_1 > 0 \) depends on \( C_1 \).

In addition, it holds that

\[
P\{\mathcal{E}_t \} = E \left\{ P\{\mathcal{E}_t | \{X(t)\}_{t=1}^T \} \right\} \geq 1 - (n \lor T)^{-c_1},
\]

therefore, (17) holds and (18) holds following almost identical arguments.

Lastly, it follows from Eq.(4.18) in Vershynin (2018) that there exists a universal constant \( C_3 > 0 \), such that

\[
P\{\|A(t) - P_t\|_{op} > C\sqrt{n} | \{X(t)\}_{t=1}^T \} \leq 4e^{-n},
\]

which leads to (19).

\[ \square \]

**Lemma 6.** Under Model 1, it holds that

\[
P\left\{ \frac{2^{-1}n}{k=1,...,K} \min_{t=1,...,T} \mu_d^k \leq \min_{t=1,...,T} \lambda_d(P_t) \leq \max_{t=1,...,T} \lambda_1(P_t) \leq (3/2)n \max_{k=1,...,K} \mu^k_1 \right\} > 1 - (n \lor T)^{-c_5},
\]

**Proof.** We first fix \( t \in \{1, \ldots, T\} \) and for simplicity drop the dependence on \( t \) notationally. For \( i \in \{1, \ldots, n\} \), let \( Y_i = X_i \Sigma^{-1/2} \) and \( Y = (Y_1, \ldots, Y_n)^T = X \Sigma^{-1/2} \), satisfying \( E\{n^{-1}Y^TY\} = I_d \).

It follows from Lemma 4.1.5 in Vershynin (2018) that for any \( \epsilon > 0 \), if

\[
\|n^{-1}Y^TY - I\|_{op} \leq \max\{\epsilon, \epsilon^2\},
\]

then the eigenvalues of \( n^{-1}Y^TY \) satisfy

\[
(1 - \max\{\epsilon, \epsilon^2\})^2 \leq \lambda_{\min}(n^{-1}Y^TY) \leq \lambda_{\max}(n^{-1}Y^TY) \leq (1 + \max\{\epsilon, \epsilon^2\})^2,
\]

which implies that

\[
n(1 - \max\{\epsilon, \epsilon^2\})^2 \leq \lambda_{\min}(\Sigma^{-1/2}X^T X \Sigma^{-1/2}) \leq \lambda_{\max}(\Sigma^{-1/2}X^T X \Sigma^{-1/2}) \leq n(1 + \max\{\epsilon, \epsilon^2\})^2.
\]

28
Denote $S = \Sigma^{-1/2}X^\top X \Sigma^{-1/2}$. We then have

$$
\lambda_1(P) = \lambda_{\max}(X^\top X) = \lambda_{\max}(\Sigma^{1/2} S \Sigma^{1/2}) \leq n(1 + \max\{\varepsilon, \varepsilon^2\})^2 \max_{k=1,\ldots,K} \mu_k^k
$$

and

$$
\lambda_d(P) = \lambda_{\min}(X^\top X) = \lambda_{\min}(\Sigma^{1/2} S \Sigma^{1/2}) = \max_{\dim(E) = d} \min_{v \in S_E} \langle S \Sigma^{1/2} v, \Sigma^{1/2} v \rangle = \max_{\dim(E) = d} \min_{v \in S_E} \|\Sigma^{1/2} v\|^2 \left\langle S \frac{\Sigma^{1/2}}{\|\Sigma^{1/2} v\|}, \frac{\Sigma^{1/2}}{\|\Sigma^{1/2} v\|} \right\rangle
$$

$$
\geq \max_{\dim(E) = d} \min_{v \in S_E} \left\langle S \frac{\Sigma^{1/2}}{\|\Sigma^{1/2} v\|}, \frac{\Sigma^{1/2}}{\|\Sigma^{1/2} v\|} \right\rangle \min_{k=1,\ldots,K} \mu_d^k \geq \max_{\dim(E) = d} \min_{v \in S_E} \langle S v, v \rangle \min_{k=1,\ldots,K} \mu_d^k
$$

$$
\geq n(1 - \max\{\varepsilon, \varepsilon^2\})^2 \min_{k=1,\ldots,K} \mu_d^k.
$$

Now it suffices to investigate (23). Since

$$
\|n^{-1}Y^\top Y - I\|_{op} = \sup_{v \in S^{d-1}} \left| \frac{1}{n} \sum_{i=1}^n \left( (Y_i^\top v)^2 - 1 \right) \right|
$$

taking $N$ to be a $1/4$-net on $S^{d-1}$, it holds that

$$
P \left\{ \|n^{-1}Y^\top Y - I\|_{op} > C \sqrt{\frac{\log(n \vee T)}{n}} \right\} \leq 9^d \max_{v \in N} P \left\{ \left| \frac{1}{n} \sum_{i=1}^n \left( (Y_i^\top v)^2 - 1 \right) \right| > C \sqrt{\frac{\log(n \vee T)}{n}} \right\}
$$

$$
\leq 2 \times 9^d \exp\{-c \log(n \vee T)\},
$$

where $C, c > 0$ are universal constants.

Thus we have that

$$
P \left\{ \frac{2^{-1}n}{K} \min_{k=1,\ldots,K} \mu_d^k \leq \min_{t=1,\ldots,T} \lambda_d(P_t) \leq \max_{t=1,\ldots,T} \lambda_1(P_t) \leq (3/2)n \max_{k=1,\ldots,K} \mu_1^k \right\} > 1 - (n \vee T)^{-c_4},
$$

where $c_4 > 0$ is a universal constant.

Lemma 7 is adapted from Theorem 8 in Athreya et al. (2017).

**Lemma 7.** It holds that

$$
P \left\{ \max_{t=1,\ldots,T} \min_{W \in \mathcal{O}_d} \|\tilde{X}(t) - X(t)W\|_F > C_W \sqrt{\frac{\log(n \vee T)}{n^{1/2}}} \right\}
$$

$$
\leq 1 - (n \vee T)^{-c_1} - (n \vee T)^{-c_2} - 4Te^{-n} - (n \vee T)^{-c_4}.
$$

**Proof of Lemma 7.** We first work on a fixed $t \in \{1, \ldots, T\}$, and then use union bounds arguments to reach the final conclusion. For simplicity, we drop the dependence on $t$ for now. Recall that

$$
\tilde{X} = U_A S_A^{1/2} \quad \text{and} \quad X = U_P S_P^{1/2}.
$$
Define \( W^* = W_1W_2^T \), where \( W_1 \) and \( W_2 \) are the left and right singular vectors of \( U_P^T U_A \), that \( U_P^T U_A = W_1A_1W_2^T \). Since \( W^* \in \mathbb{O}_d \), we have that
\[
\min_{W \in \mathbb{O}_d} \| \hat{X} - XW \|_F \leq \| \hat{X} - XW^* \|_F.
\]

In the rest of this proof, denote by \( \lambda_1, \ldots, \lambda_n \) as the eigenvalues of \( P \), with \( |\lambda_1| \geq \cdots \geq |\lambda_n| \); denote by \( \hat{\lambda}_1, \ldots, \hat{\lambda}_n \) as the eigenvalues of \( P \), with \( |\hat{\lambda}_1| \geq \cdots \geq |\hat{\lambda}_n| \).

**Step 1.** We first provide a deterministic upper bound for \( \| W^*S_A^{1/2} - S_P^{1/2}W^* \|_F \).

We have,
\[
W^*S_A = (W^* - U_P^T U_A)S_A + U_P^T U_A S_A = (W^* - U_P^T U_A)S_A + U_P^T A U_A
\]
\[
= (W^* - U_P^T U_A)S_A + U_P^T (A - P) U_A + U_P^T P U_A
\]
\[
= (W^* - U_P^T U_A)S_A + U_P^T (A - P) (U_A - U_P U_P^T U_A) + U_P^T (A - P) U_P + S_P U_P^T U_A
\]
\[
= (W^* - U_P^T U_A)S_A + U_P^T (A - P) (U_A - U_P U_P^T U_A) + U_P^T (A - P) U_P + S_P (U_P^T U_A - W^*) + S_P W^*,
\]
where the second and the fourth inequalities are due to
\[
A U_A = U_A S_A U_A^T U_A = U_A S_A \quad \text{and} \quad U_P^T P = U_P^T U_P U_P^T U_A = S_P U_P^T,
\]
respectively. Therefore,
\[
\| W^* S_A - S_P W^* \|_F \leq \| W^* - U_P^T U_A \|_F (\| S_A \|_{\text{op}} + \| S_P \|_{\text{op}}) + \| U_P^T (A - P) (U_A - U_P U_P^T U_A) \|_F
\]
\[
+ \| U_P^T (A - P) U_P \|_F
\]
\[
\leq \| I_n - \Lambda_1 \|_F \| W_1 \|_{\text{op}} \| W_2 \|_{\text{op}} (\| S_A \|_{\text{op}} + \| S_P \|_{\text{op}})
\]
\[
+ \| A - P \|_{\text{op}} \| U_A - U_P U_P^T U_A \|_F + \| U_P^T (A - P) U_P \|_F
\]
\[
\leq \| I_n - \Lambda_1 \|_F (2\lambda_1 + \| A - P \|_{\text{op}}) + \| A - P \|_{\text{op}} \| U_A - U_P U_P^T U_A \|_F
\]
\[
+ \| U_P^T (A - P) U_P \|_F = (I) + (II) + (III),
\]
where \( \lambda_1 \) is the largest singular value of \( P \) and the last inequality is due to Weyl’s inequality.

In addition, let \( \{ \theta_1, \ldots, \theta_d \} \) be the principal angles between the column spaces spanned by \( U_A \) and \( U_P \). We thus have
\[
\| I_n - \Lambda_1 \|_F = \sqrt{\sum_{i=1}^d (1 - \cos \theta_i)^2} \leq \sqrt{d} (1 - \cos^2 \theta_1) = \sqrt{d} \sin^2 \theta_1 = \sqrt{d} \min_{W \in \mathbb{O}_d} \| U_A - U_P W \|_{\text{op}}^2
\]
\[
\leq \sqrt{d} \min_{W \in \mathbb{O}_d} \| U_A - U_P W \|_F^2 = \frac{4d^{3/2} \| A - P \|_{\text{op}}^2}{\lambda_d^2}
\]
(25)
where the first and second inequalities are due to \( \cos \theta_i, \sin \theta_i \in [0, 1] \), and the last inequality is due to Theorem 2 in Yu et al. (2014) and the fact that \( \lambda_{d+1} = 0 \).

As for term \((II)\), there exists \( W \in \mathbb{O}_d \) such that
\[
\| U_A - U_P U_P^T U_A \|_F = \sqrt{\text{tr}(U_A U_A^T - U_A U_A^T U_P U_P^T U_A^T U_A)} = \sqrt{d - \text{tr}(U_A^T U_P W W^T U_P^T U_A)}
\]
30
\[
\sum_{i=1}^{d} (1 - \cos^2 \theta_i) = \sum_{i=1}^{d} \sin^2 \theta_i \leq \frac{2\sqrt{d} \|A - P\|_{op}}{\lambda_d}. 
\] (26)

Term (III) is dealt in Lemma 5.

As for \( \|W^* S_A^{1/2} - S_P^{1/2} W^*\|_F \), we note that the \( ij \)-th entry of \( W^* S_A^{1/2} - S_P^{1/2} W^* \) satisfies that

\[
|W^*_{ij}(\hat{\lambda}_j^1 - \lambda_1^{1/2})| = \frac{|W^*_{ij}(\hat{\lambda}_j - \lambda_i)|}{\lambda_j^{1/2} + \lambda_1^{1/2}} \leq \frac{|(W^* S_A - S_P W^*)_{ij}|}{\lambda_d^{1/2}},
\]

which means

\[
\|W^* S_A^{1/2} - S_P^{1/2} W^*\|_F \leq \frac{\|W^* S_A - S_P W^*\|_F}{\lambda_d^{1/2}} \leq \frac{8\delta^2 \|A - P\|_{op}^2 \lambda_d}{\lambda_d^{5/2}} + \frac{4d\delta^2 \|A - P\|_{op}^3}{\lambda_d^{3/2}} + \frac{2d\delta^2 \|A - P\|_{op}^2}{\lambda_d^{1/2}} + \frac{\|U_P^T(A - P)U_P\|_F}{\lambda_d^{1/2}}. 
\] (27)

**Step 2.** We then provide an upper bound for \( \min_{W \in \mathcal{O}_d} \|\hat{X} - XW\|_F \). Since

\[
\min_{W \in \mathcal{O}_d} \|\hat{X} - XW\|_F \leq \|\hat{X} - XW^*\|_F,
\]
in the rest of this step, we work on \( \hat{X} - XW^* \). We have that

\[
\|\hat{X} - XW^*\|_F = \|U_A S_A^{1/2} - U_P S_P^{1/2} W^*\|_F = \|U_A S_A^{1/2} - U_P W^* S_A^{1/2} + U_P (W^* S_A^{1/2} - S_P^{1/2} W^*)\|_F
\]

\[
\leq \|(U_A - U_P U_P^T U_A) S_A^{1/2}\|_F + \|U_P (U_P^T U_A - W^*) S_A^{1/2}\|_F + \|U_P (W^* S_A^{1/2} - S_P^{1/2} W^*)\|_F
\]

\[
= (I) + (II) + (III).
\] (28)

As for term (I), it holds that

\[
(U_A - U_P U_P^T U_A) S_A^{1/2} = (A - P) U_A S_A^{1/2} - U_P U_P^T (A - P) U_A S_A^{1/2}
\]

\[
= (A - P) U_P W^* S_A^{1/2} - U_P U_P^T (A - P) U_P W^* S_A^{1/2} + (I - U_P U_P^T)(A - P) (U_A - U_P W^*) S_A^{1/2}
\]

\[
= (I.1) + (I.2) + (I.3),
\]

which satisfies

\[
\| (A - P) U_P W^* S_A^{1/2}\|_F \leq \|(A - P) U_P\|_F (\hat{\lambda}_d)^{-1/2},
\]

\[
\|U_P U_P^T (A - P) U_P W^* S_A^{1/2}\|_F \leq \|U_P^T (A - P) U_P\|_F (\hat{\lambda}_d)^{-1/2}
\]

and

\[
\| (I - U_P U_P^T)(A - P)(U_A - U_P W^*) S_A^{1/2}\|_F \leq \|A - P\|_{op}\|U_A - U_P W^*\|_F (\hat{\lambda}_d)^{-1/2}
\]

\[
\leq \frac{\sqrt{d} \|A - P\|_{op} (\hat{\lambda}_d)^{-1/2}}{\lambda_d}.
\]
As for term (II), it holds that
\[
\|U_P(U_P U_A - W^*)S_A^{1/2}\|_F \leq \|I - \Lambda_1\|_F (\lambda_1 + \|A - P\|_{op})^{1/2} \leq \frac{4d^{3/2}\|A - P\|_{op}^2}{\lambda_d^2} \sqrt{\frac{3\lambda_1}{2}}.
\]

As for term (III), it holds that \(\|U_P(W^* S_A^{1/2} - S_P^{1/2} W^*)\|_F \leq \|(W^* S_A^{1/2} - S_P^{1/2} W^*)\|_F\).

Combining (27), (28) and all above, we have that
\[
\min_{\Theta_d} \|\hat{X} - XW\|_F \leq \|U_P^T(A - P)U_P\|_F (\hat{\lambda}_d)^{-1/2} + \|U_P^T(A - P)U_P\|_F (\hat{\lambda}_d)^{-1/2}
\]
\[
+ \frac{\sqrt{d}\|A - P\|_{op}(\hat{\lambda}_d)^{-1/2}}{\lambda_d} + \frac{4d^{3/2}\|A - P\|_{op}^2}{\lambda_d^2} \sqrt{\frac{3\lambda_1}{2}}
\]
\[
+ \frac{8d^{3/2}\|A - P\|_{op}^2}{\lambda_d^2} + \frac{4d^{3/2}\|A - P\|_{op}^3}{\lambda_d^2} + \frac{2d^{1/2}\|A - P\|_{op}^2}{\lambda_d^{3/2}} + \|U_P^T(A - P)U_P\|_F.
\]

Noting that \(\hat{\lambda}_d \geq \lambda_d - \|A - P\|_{op}\), it holds on the event \(E_1 \cap E_2 \cap E_3 \cap E_4\), that
\[
\mathbb{P}\left\{ \max_{t=1,\ldots,T \in \Theta_d} \min_{W \in \Theta_d} \|\hat{X}(t) - X(t)W\|_F > C_W \sqrt{\frac{\log(n \vee T)}{n}} \right\}
\]
\[
\leq \mathbb{P}\left\{ \max_{t=1,\ldots,T \in \Theta_d} \min_{W \in \Theta_d} \|\hat{X}(t) - X(t)W\|_F > \sqrt{2} (3C_1 + 2C_2) \sqrt{\frac{\log(n \vee T)}{n}} + \frac{2^{5/2}C_3}{(\min_{k=1,\ldots,K} \mu_d^k)^{3/2}} \frac{\sqrt{d}}{n}
\]
\[
+ \frac{2^{5/2}C_3^2}{(\min_{k=1,\ldots,K} \mu_d^k)^{3/2}} \frac{\sqrt{d}}{n} + \frac{24C_3^2}{(\min_{k=1,\ldots,K} \mu_d^k)^2} \frac{d^{3/2}}{\sqrt{n}} + \frac{2^{9/2}C_3^2}{(\min_{k=1,\ldots,K} \mu_d^k)^{5/2}} \frac{d^{3/2}}{\sqrt{n}} \right\}
\]
\[
\leq 1 - (n \vee T)^{-c_1} - (n \vee T)^{-c_2} - (n \vee T)^{-c_4} - 4Te^{-n},
\]
where \(C_W > 0\) is a universal constant depending only on \(C_1, C_2, C_3, \max_{k=1,\ldots,K} \mu_1^k\) and \(\min_{k=1,\ldots,K} \mu_d^k\).

We first state a weakly dependent version of Bernstein inequality. This is in fact Theorem 4 in Delyon (2009). The notation in Lemma 8 only applies within Lemma 8.

**Lemma 8.** Let \(\{X_1, \ldots, X_T\}\) be centred random variables. Define
\[
g = \sum_{t=2}^{T} \sum_{s=1}^{t-1} \|X_s\|_\infty \|\mathbb{E}(X_t \mid F_s)\|_\infty, \quad v = \sum_{t=1}^{T} \|\mathbb{E}(X_t^2 \mid F_{t-1})\|_\infty
\]
and
\[
m = \max_{t=1,\ldots,T} \|X_t\|_\infty,
\]
where \(F_s = \sigma\{X_1, \ldots, X_s\}, s \geq 1\), is the natural \(\sigma\)-field generated by \(\{X_i\}_{i=1}^s\). For any \(\varepsilon > 0\), it holds that
\[
\mathbb{P}\left\{ \sum_{t=1}^{T} X_t > \varepsilon \right\} \leq 2 \exp\left( -\frac{\varepsilon^2}{2(v + 2g) + 2\varepsilon m/3} \right).
\]
Lemma 9. Under Model 1, it holds that for any $z \in \mathbb{R}$,
\[
\mathbb{P}\left\{ \max_{0 \leq s < t \leq T} \left| \Delta_{s, t}^k(z) \right| \geq C_8 \sqrt{T} \max\{\sqrt{\log(n \vee T)}, d^{3/2}\} \right\} \leq 4(n \vee T)^{-c} + 4Te^{-n},
\]
where $c = \min\{c_1, c_2, c_4, c_5\} - 1 > 0$ is a universal constant.

In addition,
\[
\mathbb{P}\left\{ \max_{0 \leq s < t \leq T} \left| \sqrt{\frac{2}{n(e-s)}} \sum_{i=s+1}^{e} \sum_{(i,j) \in O} \left( 1\{Y_{ij}^k \leq z\} - \mathbb{E}\left\{ 1\{Y_{ij}^k \leq z\} \right\} \right) \right| \geq C_8 \sqrt{\frac{T}{1 - \rho}} \max\{\sqrt{\log(n \vee T)}, d^{3/2}\} \right\} \leq 4(n \vee T)^{-c} + 4Te^{-n},
\]
where $c = \min\{c_1, c_2, c_4, c_5\} - 1 > 0$ is a universal constant.

Proof. For any $(i, j) \in O$ and $t \in \{1, \ldots, T\}$, it holds that
\[
\left| \hat{Y}_{ij}^t - Y_{ij}^t \right| = \left| (\hat{X}_i(t))^\top \hat{X}_j(t) - (X_i(t))^\top X_j(t) \right| \leq \left| (\hat{X}_i(t) - W_tX_i(t))^\top W_tX_j(t) \right| + \left| (\hat{X}_i(t) - W_tX_i(t))^\top (W_tX_j(t) - \hat{X}_j(t)) \right|
\]
\[
\leq 2 \max_{t=1, \ldots, T} \min_{W \in \mathbb{O}_d} \| \hat{X}(t) - X(t)W^\top \|_F \max_{t=1, \ldots, T} \|X_i(t)\| + \left( \max_{t=1, \ldots, T} \min_{W \in \mathbb{O}_d} \| \hat{X}(t) - X(t)W^\top \|_F \right)^2
\]
\[
\leq 2 \max_{t=1, \ldots, T} \min_{W \in \mathbb{O}_d} \| \hat{X}(t) - X(t)W^\top \|_F + \left( \max_{t=1, \ldots, T} \min_{W \in \mathbb{O}_d} \| \hat{X}(t) - X(t)W^\top \|_F \right)^2,
\]
where $W_t \in \mathbb{O}_d$ satisfies
\[
\| \hat{X}(t) - X(t)W_t^\top \|_F = \min_{W \in \mathbb{O}_d} \| \hat{X}(t) - X(t)W^\top \|_F.
\]

We fix the chosen pairs $O \subset \{1, \ldots, n\}^2$ with $|O| = n/2$, which is assumed to be an integer. As for the sequence $\{w_k\}$, it holds that
\[
\sum_{k=s+1}^{e} \sum_{(i,j) \in O} w_k^2 = 1.
\] (30)

We have for any $z \in \mathbb{R}$, it holds that
\[
|\Delta_{s, t}^k(z)| \leq \left| \sum_{k=s+1}^{e} \sum_{(i,j) \in O} w_k \left( 1\{Y_{ij}^k \leq z\} - 1\{Y_{ij}^k \leq z\} \right) \right| + \left| \sum_{k=s+1}^{e} \sum_{(i,j) \in O} w_k \left( 1\{Y_{ij}^k \leq z\} - \mathbb{E}\left\{ 1\{Y_{ij}^k \leq z\} \right\} \right) \right| \leq (I) + (II).
Term \((II)\). As for \((II)\), notice that
\[
\mathbb{E} \left( \mathbb{1}\{Y^k_{i,j} \leq z\} - \mathbb{E} \left\{ \mathbb{1}\{Y^k_{i,j} \leq z\} \right\} \right) = 0.
\]

In order to apply Lemma 8, we let
\[
V_i(k) = w_k \mathbb{1}\{Y^k_{i,j} \leq z\} - w_k \mathbb{E} \left\{ \mathbb{1}\{Y^k_{i,j} \leq z\} \right\},
\]
with \(i = 1, \ldots, n/2, \ k = 1, \ldots, T\). We order \(\{V_i(k)\}\) as
\[
V_1(1), \ldots, V_1(T), V_2(1), \ldots, V_2(T), \ldots, V_{n/2}(1), \ldots, V_{n/2}(T). \tag{31}
\]
Denote \(F_{i,t}\) as the natural \(\sigma\)-field generated by \(V_i(k)\) and all the random variables before it in the order of (31), and denote \(F_{i,t,−}\) as the natural \(\sigma\)-field generated by all the random variables before \(Y_i(t)\) in the order of (31) excluding \(Y_i(t)\). If \((i,t) = (1,1)\), then \(F_{i,t,−}\) is the \(\sigma\)-field generated by constants.

In addition, for the notation in Lemma 8, we have that
\[
v = \frac{n}{2} \sum_{i=1}^{n/2} \sum_{t=s+1}^{e} \|\mathbb{E}(V_i(t)^2 | F_{i,t,−})\|_{\infty}
\]
\[
= \frac{n}{2} \sum_{i=1}^{n/2} \sum_{k: \eta_k \in (s,e)} (w_{\eta_k+1})^2 \mathbb{E} \left\{ \mathbb{1}\{Y^\eta_k+1_{i,j} \leq z\} \right\} (1 - \mathbb{E} \left\{ \mathbb{1}\{Y^\eta_k+1_{i,j} \leq z\} \right\})
\]
\[
+ \sum_{i=1}^{n/2} \sum_{t \in (s,e)} (1 - \rho)(w_t)^2 \mathbb{E} \left\{ \mathbb{1}\{Y^t_{i,j} \leq z\} \right\} (1 - \mathbb{E} \left\{ \mathbb{1}\{Y^t_{i,j} \leq z\} \right\})
\]
\[
+ \sum_{i=1}^{n/2} \sum_{t \in (s,e)} \rho(w_t)^2 \|\mathbb{1}\{Y^{t-1}_{i,j} \leq z\} - \mathbb{E} \left\{ \mathbb{1}\{Y^{t-1}_{i,j} \leq z\} \right\}\|_{\infty}^2 \leq 1 + \rho, \tag{32}
\]
where the last inequality is due to (30),
\[
m \leq \max_{t=1,\ldots,T} |w_t|, \tag{33}
\]
and
\[
g = (n/2) \sum_{k: \eta_k \in (s,e)} \left( \sum_{t=\eta_k+2}^{\eta_k+1} \sum_{u=\eta_k+1}^{t-1} + \sum_{t=s+1}^{\eta_k+1} \sum_{u=s+2}^{t-1} \right) |w_t w_u| \rho^{t-u}. \tag{34}
\]
Combining (32), (33), (34) and Lemma 8, we have for any \(\varepsilon > 0\), it holds that
\[
\mathbb{P} ((II) \geq \varepsilon) \leq 2 \exp \left\{ -C \varepsilon^2 / ((1 - \rho)^{-1} + \varepsilon) \right\}.\]
We thus denote
\[
E_5 = \left\{ \max_{1 \leq s < t \leq T} \left| \sum_{k=s+1}^{t} w_k \sum_{(i,j) \in O} \left( \mathbb{1}\{Y_{ij}^k \leq z\} - \mathbb{E}\{Y_{ij}^k \leq z\} \right) \right| \geq C_5 \sqrt{\frac{\log(n \lor T)}{1 - \rho}} \right\},
\]
where \( C_5 > 0 \) is a universal constant, and therefore it holds that
\[
P\{E_5\} \leq (n \lor T)^{-c_5},
\]
where \( c_5 > 0 \) is a universal constant.

**Term (I).** As for (I), we have that
\[
E \left\{ \left| \mathbb{1}\{\hat{Y}_{ij}^k \leq z\} - \mathbb{1}\{Y_{ij}^k \leq z\} \right| \right\}
\leq \max \left\{ P\left\{ \left( \hat{Y}_{ij}^k \leq z \right) \cap \left( Y_{ij}^k > z \right) \right\}, P\left\{ \left( \hat{Y}_{ij}^k > z \right) \cap \left( Y_{ij}^k \leq z \right) \right\} \right\} = \max\{(I.1), (I.2)\}.
\]
Let
\[
E_6 = \left\{ \max_{t=1, \ldots, T} \min_{W \in O}_{\delta} \| \hat{X}_t - X_t W \| \leq C_W \sqrt{\frac{\log(n \lor T) \lor d^{3/2}}{n^{1/2}}} \right\}.
\]
On the event \( E_6 \), it holds that
\[
\max_{t=1, \ldots, T} \left| \hat{Y}_{ij}^t - Y_{ij}^t \right| \leq 3C_W \sqrt{\frac{\log(n \lor T) \lor d^{3/2}}{n^{1/2}}} = \delta
\]
and
\[
P\left\{ \max_{t=1, \ldots, T} \min_{W \in O}_{\delta} \| \hat{X}_t - X_t W \| \leq \delta \right\}
\geq 1 - (n \lor T)^{-c_1} - (n \lor T)^{-c_2} - (n \lor T)^{-c_4} - 4T e^{-n} = 1 - p_\delta.
\]
Therefore,
\[
(I.1) = P\left\{ \left( \hat{Y}_{ij}^k \leq z \right) \cap \left( Y_{ij}^k > z \right) \mid Y_{ij}^k > z + \delta \right\} P\{Y_{ij}^k > z + \delta\} + P\left\{ \left( \hat{Y}_{ij}^k \leq z \right) \cap \left( Y_{ij}^k > z \right) \mid Y_{ij}^k < z + \delta \right\} P\{Y_{ij}^k < z + \delta\}
\leq p_\delta (1 - F_k(z + \delta)) + F_k(z + \delta) - F_k(z) \leq p_\delta + \delta C_F
\]
and
\[
(I.2) = P\left\{ \left( \hat{Y}_{ij}^k > z \right) \cap \left( Y_{ij}^k \leq z \right) \mid Y_{ij}^k \leq z - \delta \right\} P\{Y_{ij}^k \leq z - \delta\} + P\left\{ \left( \hat{Y}_{ij}^k > z \right) \cap \left( Y_{ij}^k \leq z \right) \mid Y_{ij}^k > z - \delta \right\} P\{Y_{ij}^k > z - \delta\}
\leq p_\delta F_k(z - \delta) + F_k(z) - F_k(z - \delta) \leq p_\delta + \delta C_F.
\]
Then we have,

\[
\mathbb{E} \left| \sum_{k=s+1}^{e} w_k \sum_{(i,j) \in \mathcal{O}} \left( 1 \{ \hat{Y}_{ij}^k \leq z \} - 1 \{ Y_{ij}^k \leq z \} \right) \right| \leq 2 \sqrt{\frac{n}{2}} \frac{(e-t)(t-s)}{e-s} (p_\delta + \delta C_F) \\
\leq 2 \sqrt{\frac{n}{2}} \min\{\sqrt{e-t}, \sqrt{t-s}\} (p_\delta + \delta C_F).
\]

Therefore, following from similar arguments as those used in bounding (II), we have that for any \( \varepsilon > 0 \), it holds that

\[
P\left\{ \left| \sum_{k=s+1}^{e} w_k \sum_{(i,j) \in \mathcal{O}} \left( 1 \{ \hat{Y}_{ij}^k \leq z \} - 1 \{ Y_{ij}^k \leq z \} \right) \right| > \varepsilon \right\} \\
\leq 2 \exp \left\{ -C \varepsilon^2 / ((1 - \rho)^{-1} + \varepsilon) \right\},
\]

which implies that

\[
P\left\{ \left| \sum_{k=s+1}^{e} w_k \sum_{(i,j) \in \mathcal{O}} \left( 1 \{ \hat{Y}_{ij}^k \leq z \} - 1 \{ Y_{ij}^k \leq z \} \right) \right| > \mathbb{E} \left| \sum_{k=s+1}^{e} w_k \sum_{(i,j) \in \mathcal{O}} \left( 1 \{ \hat{Y}_{ij}^k \leq z \} - 1 \{ Y_{ij}^k \leq z \} \right) \right| + \varepsilon/2 \right\} \\
\leq \mathbb{P} \left\{ \left| \sum_{k=s+1}^{e} w_k \sum_{(i,j) \in \mathcal{O}} \left( 1 \{ \hat{Y}_{ij}^k \leq z \} - 1 \{ Y_{ij}^k \leq z \} \right) \right| > 2 \sqrt{\frac{n}{2}} \min\{\sqrt{e-t}, \sqrt{t-s}\} (p_\delta + \delta C_F) + \varepsilon/2 \right\} \\
\leq 2 \exp \left\{ -C \varepsilon^2 / ((1 - \rho)^{-1} + \varepsilon) \right\} + p_\delta.
\]

Lastly, we have that

\[
P\left\{ \max_{0 \leq s \leq t \leq T} |\Delta_{s,e}(z)| \geq C_8 \sqrt{\frac{T}{1 - \rho}} \max\{\sqrt{\log(n \vee T)}, d^{3/2}\} \right\} \\
\leq \mathbb{P} \left\{ |\Delta_{s,e}(z)| > C_5 \sqrt{\frac{\log(n \vee T)}{1 - \rho}} + \sqrt{2n} \min\{\sqrt{e-t}, \sqrt{t-s}\}(p_\delta + \delta C_F) \right\} \\
\leq 4(n \vee T)^{-c} + 4T e^{-n},
\]

where \( c = \min\{c_1, c_2, c_4, c_5\} - 1 > 0 \) is a universal constant.

The result (29) follows from the identical arguments. \qed

**Lemma 10.** Let

\[ \Delta_{s,e}^t = \sup_{z \in \mathbb{R}} |\Delta_{s,e}(z)|. \]
It holds that
\[
P\left\{ \max_{0 \leq s < t < e \leq T} \Delta^{t}_{s,e} > C_9 T^{1/2} (1 - \rho)^{-1/2} \max\{ \sqrt{\log(n \lor T)}, d^{3/2} \} \right\} \leq 11(n \lor T)^{-c} + 8Te^{-n}.
\]

In addition,
\[
P\left\{ \max_{0 \leq s < t < e \leq T} \sup_{z \in \mathbb{R}} \left| \sqrt{\frac{2}{n(e - s)}} \sum_{k=s+1}^{e} \sum_{(i,j) \in \mathcal{O}} \left( 1 \{ \hat{Y}^{k}_{ij} \leq z \} - \mathbb{E} \{ 1 \{ Y^{k}_{ij} \leq z \} \} \right) \right| \right\} \leq C_9 T^{1/2} (1 - \rho)^{-1/2} \max\{ \sqrt{\log(n \lor T)}, d^{3/2} \} \leq 11(n \lor T)^{-c} + 8Te^{-n}. \tag{35}
\]

Proof. Let
\[
\delta = 3C_W \sqrt{\frac{\log(n \lor T)}{n^{1/2}}} \tag{36}
\]
Let \(z_{m} = m\delta\), \(m = 1, \ldots, \lceil 1/\delta \rceil\). Let \(I_m = [z_m - \delta, z_m + \delta]\), for \(m = 1, \ldots, \lceil 1/\delta \rceil - 1\), and \(I_{\lceil 1/\delta \rceil} = [z_{\lceil 1/\delta \rceil} - 1, 1]\). Let \(M = \lceil 1/\delta \rceil\). Then
\[
\sup_{z \in \mathbb{R}} |\Delta^{t}_{s,e}(z)| \leq \max_{j = 1, \ldots, M} \left\{ |\Delta^{t}_{s,e}(z_j)| + \sup_{z \in I_j} |\Delta^{t}_{s,e}(z_j) - \Delta^{t}_{s,e}(z)| \right\}. \tag{37}
\]

It follows from Lemma 9 that
\[
P\left\{ \max_{j = 1, \ldots, M} |\Delta^{t}_{s,e}(z_j)| \geq C_8 \sqrt{T}(1 - \rho)^{-1/2} \max\{ \sqrt{\log(n \lor T)}, d^{3/2} \} \right\} \leq 4(n \lor T)^{-c} + 4Te^{-n}. \tag{38}
\]

For every \(z \in \mathbb{R}\), on the event
\[
\left\{ \max_{k = 1, \ldots, e} \max_{(i,j) \in \mathcal{O}} \left| \hat{Y}^{k}_{ij} - Y^{k}_{ij} \right| \leq \delta \right\},
\]
it holds that
\[
\left| 1 \{ \hat{Y}^{k}_{ij} \leq z \} - 1 \{ Y^{k}_{ij} \leq z \} \right| \leq 1 \{ Y^{k}_{ij} \in [z - \delta, z + \delta] \}.
\]

For any \(z \in \mathbb{R}\), there exist \(z_m\) and \(z_{m+1}\), \(m \in \{1, \ldots, M = 1\}\), such that
\[
[z - \delta, z + \delta] \subset [z_m - \delta, z_m + \delta] \cup [z_{m+1} - \delta, z_{m+1} + \delta].
\]

Let
\[
B_m = \sum_{k=s+1}^{e} \sum_{(i,j) \in \mathcal{O}} 1 \{ Y^{k}_{ij} \in I_m \}, \quad m = 1, \ldots, M.
\]

Therefore
\[
|\Delta^{t}_{s,e}(z_m) - \Delta^{t}_{s,e}(z)|
\]

37
\[
\begin{aligned}
&\leq \sum_{k=s+1}^e \sum_{(i,j) \in \Omega} w_k \{ \mathbb{1}[\hat{Y}_{i,j}^k \leq z_m] - \mathbb{1}[Y_{i,j}^k \leq z_m] \} + \sum_{k=s+1}^e \sum_{(i,j) \in \Omega} w_k \{ \mathbb{1}[\hat{Y}_{i,j}^k \leq z] - \mathbb{1}[Y_{i,j}^k \leq z] \}
\end{aligned}
\]
\[
+ \sum_{k=s+1}^e \sum_{(i,j) \in \Omega} w_k \{ \mathbb{1}[\hat{Y}_{i,j}^k \leq z_m] - \mathbb{1}[Y_{i,j}^k \leq z_m] \} + \sum_{k=s+1}^e \sum_{(i,j) \in \Omega} w_k \{ G_k(z_m) - G_k(z) \}
\]
\[
\leq \left( \sqrt{\frac{2(e-t)}{n(e-s)(t-s)}} + \sqrt{\frac{2(t-s)}{n(e-s)(e-t)}} \right) \left( \sum_{k=s+1}^e \sum_{(i,j) \in \Omega} |1[Y_{i,j}^k \leq z_m] - 1[Y_{i,j}^k \leq z_m]| \right)
\]
\[
+ \sum_{k=s+1}^e \sum_{(i,j) \in \Omega} |1[Y_{i,j}^k \leq z] - 1[Y_{i,j}^k \leq z]| + \sum_{k=s+1}^e \sum_{(i,j) \in \Omega} 1[Y_{i,j}^k \leq I_m]
\]
\[
+ \left( \sum_{k=s+1}^e \sum_{(i,j) \in \Omega} \left| w_k \right| \max_{k=s+1,...,e} |G_k(z) - G_k(z_m)| \right)
\]
\[
\leq 4 \left( \sqrt{\frac{2(e-t)}{n(e-s)(t-s)}} + \sqrt{\frac{2(t-s)}{n(e-s)(e-t)}} \right) \max_{m=1,...,M} B_m + \frac{2n(e-t)(t-s)}{e-s} \delta C_G. \tag{39}
\]

Since
\[
\max_{m=1,...,M} B_m \leq \sum_{k=s+1}^e \sum_{(i,j) \in \Omega} (\mathbb{1}[Y_{i,j}^k \in I_m] - \mathbb{P}\{Y_{i,j}^k \in I_m\}) + \sum_{k=s+1}^e \sum_{(i,j) \in \Omega} \mathbb{P}\{Y_{i,j}^k \in I_m\}
\]
\[
\leq \sum_{k=s+1}^e \sum_{(i,j) \in \Omega} (\mathbb{1}[Y_{i,j}^k \in I_m] - \mathbb{P}\{Y_{i,j}^k \in I_m\}) + (e-s)n\delta C_G,
\]
and
\[
\mathbb{P}\left\{ \max_{m=1,...,M} \sum_{k=s+1}^e \sum_{(i,j) \in \Omega} (\mathbb{1}[Y_{i,j}^k \in I_m] - \mathbb{P}\{Y_{i,j}^k \in I_m\}) \leq C_9 \sqrt{\frac{n(e-s) \log(n \vee T)}{1-\rho}} \right\} \geq 1-(n \vee T)^{-c_9},
\]
where $C_9, c_9 > 0$ are universal constants, we have that
\[
\mathbb{P}\left\{ \left( \sqrt{\frac{2(e-t)}{n(e-s)(t-s)}} + \sqrt{\frac{2(t-s)}{n(e-s)(e-t)}} \right) \max_{m=1,...,M} B_m \geq C_{10} T^{1/2}(1-\rho)^{-1/2}(\log(n \vee T) \vee d^{3/2}) \right\}
\]

38
\[ (n \lor T)^{-c_{10}}, \quad (40) \]

where \( C_{10}, c_{10} > 0 \) are universal constants.

Combining (36), (37), (38), (39) and (40), the proof is complete.

C Change point analysis lemmas

Lemma 11. Under Model 1, for any pair \((s, e) \subset (0, T)\) satisfying

\[ \eta_{k-1} \leq s \leq \eta_{k} \leq \ldots \leq \eta_{k+q} \leq e \leq \eta_{k+q+1}, \quad q \geq 0, \]

let

\[ b_1 \in \arg\max_{b=s+1, \ldots, e-1} \bar{D}_s^b. \]

Then \( b_1 \in \{\eta_1, \ldots, \eta_K\} \).

Let \( z \in \text{argmax}_{x \in \mathbb{R}} |\bar{D}_s^b (x)| \). If \( \bar{D}_t^b (z) > 0 \) for some \( t \in (s, e) \), then \( \bar{D}_t^b (z) \) is either monotonic or decreases and then increases within each of the interval \((s, \eta_k), (\eta_k, \eta_{k+1}), \ldots, (\eta_{k+q}, e)\).

This is identical to Lemma 7 in Padilla et al. (2019a) and we omit the proof here.

Lemma 12. Under Model 1, let \( 0 \leq s < \eta_k < e \leq T \) be any interval satisfying

\[ \min\{\eta_k - s, e - \eta_k\} \geq c_1 \Delta, \]

with \( c_1 > 0 \). Then we have that

\[ \max_{t=s+1, \ldots, e-1} \bar{D}_t^b \geq \frac{2^{-3/2}c_1 \kappa \Delta \sqrt{n}}{\sqrt{e-s}}. \]

Proof. Recall that

\[ G_{\eta_k} (z) = \mathbb{P} \left\{ (X_1(\eta_k))^\top X_2(\eta_k) \leq z \right\}. \]

Let

\[ z_0 \in \text{argmax}_{z \in [0,1]} |G_{\eta_k} (z) - G_{\eta_k+1} (z)|. \]

Without loss of generality, assume that \( F_{\eta_k} (z_0) > F_{\eta_k+1} (z_0) \). For \( s < t < e \), note that

\[ \bar{D}_s^t (z_0) = \left| \sqrt{\frac{n(e-t)}{2(e-s)(t-s)}} \sum_{k=s+1}^{t} G_k (z_0) - \sqrt{\frac{n(t-s)}{2(e-s)(e-t)}} \sum_{k=t+1}^{e} G_k (z_0) \right| \]

\[ = \left| \sqrt{\frac{n(e-s)}{2(t-s)(e-t)}} \sum_{k=s+1}^{t} \tilde{G}_k (z_0) \right|, \]

where \( \tilde{G}_k (z_0) = G_k (z_0) - (e-s)^{-1} \sum_{k=s+1}^{e} G_k (z_0) \).

Under Model 1, it holds that \( \tilde{G}_{\eta_k} (z_0) > \kappa/2 \). Therefore

\[ \sum_{k=s+1}^{\eta_k} \tilde{G}_k (z_0) \geq (c_1/2) \kappa \Delta, \quad \text{and} \quad \sqrt{\frac{n(e-s)}{2(t-s)(e-t)}} \geq \sqrt{\frac{n}{2(e-s)}}. \]

39
Then
\[
\max_{t=s+1,\ldots,e-1} \bar{D}_{s,e}^t \geq \frac{2^{-3/2}c_1 \Delta \sqrt{n}}{\sqrt{e-s}}.
\]

Lemma 13. Under Model 1, if \( \eta_k \) is the only change point in \((s,e)\), then
\[
\tilde{D}_{s,e}^{\eta_k} \leq \kappa_k \sqrt{n/2 \min\{\sqrt{\eta_k - s}, \sqrt{e - \eta_k}\}};
\]
if \((s,e) \subset (0,T)\) contain two and only two change points \( \eta_k \) and \( \eta_{k+1} \), then we have
\[
\max_{t=s+1,\ldots,e-1} \tilde{D}_{s,e}^{\eta_k} \leq \sqrt{n/2} \sqrt{e - \eta_{k+1}\kappa_{k+1} + \sqrt{n/2} \sqrt{\eta_k - s}\kappa_k};
\]
if \((s,e) \subset (0,T)\) contains two or more change points, including \( \eta_k \) and \( \eta_{k+1} \), which satisfy that \( \eta_k - s \leq c_1 \Delta \), for \( c_1 > 0 \), then
\[
\tilde{D}_{s,e}^{\eta_k} \leq \sqrt{c_1} \tilde{D}_{s,e}^{\eta_{k+1}} + \sqrt{2(\eta_k - s)n\kappa_k}.
\]

Proof. As for (41), it is due to that
\[
\tilde{D}_{s,e}^{\eta_k} = \sqrt{\frac{n(\eta_k - s)(e - \eta_k)}{2(e - s)}} \sup_{z \in \mathbb{R}} |G_{\eta_k}(z) - G_{\eta_{k+1}}(z)| \leq \kappa_k \sqrt{n/2 \min\{\sqrt{\eta_k - s}, \sqrt{e - \eta_k}\}}.
\]
Eq. (42) follows similarly.

As for (43), we consider the distribution sequence \( \{H_t\}_{t=s+1}^e \) be such that
\[
H_t = \begin{cases} 
G_{\eta_{k+1}}, & t = s+1, \ldots, \eta_k, \\
G_t, & t = \eta_k+1, \ldots, e.
\end{cases}
\]
For any \( s < t < e \), define
\[
\mathcal{H}_{s,e}^t = \sup_{z \in \mathbb{R}} |\mathcal{H}_{s,e}^t(z)|,
\]
where
\[
\mathcal{H}_{s,e}^t(z) = \sqrt{\frac{n(t-s)(e-t)}{2(e-s)}} \left\{ \frac{1}{t-s} \sum_{l=s+1}^{t} H_l(z) - \frac{1}{e-t} \sum_{l=t+1}^{e} H_l(z) \right\}.
\]
For any \( t \geq \eta_k \) and \( z \in \mathbb{R} \), it holds that
\[
|\bar{D}_{s,e}^t(z) - \mathcal{H}_{s,e}^t(z)| = \sqrt{\frac{2(e-t)}{n(e-s)(t-s)}} \frac{n(\eta_k - s)}{2} |G_{\eta_{k+1}}(z) - G_{\eta_k}(z)| \leq \sqrt{\frac{n(\eta_k - s)}{2}} \kappa_k.
\]
Thus we have
\[
\tilde{D}_{s,e}^{\eta_k} = \sup_{z \in \mathbb{R}} |\bar{D}_{s,e}^{\eta_k}(z) - \mathcal{H}_{s,e}^{\eta_k}(z)| \leq \sup_{z \in \mathbb{R}} |\bar{D}_{s,e}^{\eta_k}(z) - \mathcal{H}_{s,e}^{\eta_k}(z)| + \mathcal{H}_{s,e}^{\eta_k}
\leq \mathcal{H}_{s,e}^{\eta_k} + \sqrt{\frac{n(\eta_k - s)}{2}} \kappa_k \leq \sqrt{\frac{(\eta_k - s)(e - \eta_{k+1})}{(\eta_{k+1} - s)(e - \eta_{k+1})}} \mathcal{H}_{s,e}^{\eta_{k+1}} + \sqrt{\frac{n(\eta_k - s)}{2}} \kappa_k
\leq \sqrt{c_1} \tilde{D}_{s,e}^{\eta_{k+1}} + \sqrt{2n(\eta_k - s)\kappa_k}.
\]
Lemma 14. For any $z_0 \in \mathbb{R}$ and $(s,e) \subset (0,T)$ satisfying the following: there exits a true change point $\eta_k \in (s,e)$ such that

$$\min\{\eta_k - s, e - \eta_k\} \geq c_{1}\Delta, \quad (44)$$

$$\tilde{D}_{s,e}^{\eta_k}(z_0) \geq (c_{1}/2)\sqrt{n/2}\frac{\kappa_{\Delta}}{\sqrt{e - s}}, \quad (45)$$

where $c_{1} > 0$ is a sufficiently small constant, and that

$$\max_{t = s + 1, \ldots, e} |\tilde{D}_{s,e}^{\eta_k}(z_0)| - \tilde{D}_{s,e}^{\eta_k}(z_0) \leq 2^{-3/2}c_{1}^{3}(e - s)^{-7/2}\Delta^{4}\sqrt{n}, \quad (46)$$

for all $d \in (s,e)$ satisfying

$$|d - \eta_k| \leq c_{1}\Delta/32, \quad (47)$$

it holds that

$$\tilde{D}_{s,e}^{\eta_k}(z_0) - \tilde{D}_{s,e}^{d}(z_0) > c|d - \eta_k|\Delta\tilde{D}_{s,e}^{\eta_k}(z_0)(e - s)^{-2},$$

where $c > 0$ is a sufficiently small constant.

Proof. The proof is identical to the proof of Lemma 11 in Padilla et al. (2019a) after letting $n_{\min} = n_{\max} = n/2$. □

Lemma 15. Under Model 1, consider any generic $(s,e) \subset (0,T)$, satisfying

$$\min_{l=1,\ldots,K} \min\{\eta_l - s, e - \eta_l\} \geq \Delta/16, \quad \eta_k \in (s,e). \quad \text{and} \quad e - s \leq C_{R}\Delta.$$

Let

$$\kappa_{s,e}^{\max} = \max_{l=1,\ldots,K} \kappa_{l}, \quad \text{and} \quad b \in \arg \max_{s<t<e} D_{s,e}^{l}. \quad \text{For some } c_{1} > 0 \text{ and } \gamma > 0, \text{ suppose that}$$

$$D_{s,e}^{b} \geq c_{1}\kappa_{s,e}^{\max}\sqrt{\Delta n}, \quad (48)$$

$$\max_{t = s + 1, \ldots, e - 1} \sup_{z \in \mathbb{R}} |\Delta_{s,e}^{t}(z)| \leq \gamma, \quad (49)$$

and

$$\max_{0 \leq s < e \leq T} \sup_{z \in \mathbb{R}} \left| \sqrt{\frac{2}{n(e - s)}} \sum_{t = s + 1}^{e} \sum_{i,j \in \mathcal{O}} (1\{\hat{Y}_{t,i,j} \leq z\} - G_{t}(z)) \right| \leq \gamma. \quad (50)$$

If there exits a sufficiently small $0 < c_{2} < c_{1}/2$ such that

$$\gamma \leq c_{2}\kappa_{s,e}^{\max}\sqrt{\Delta n}, \quad (51)$$

then there exists a change point $\eta_k \in (s,e)$ such that

$$\min\{e - \eta_k, \eta_k - s\} \geq \Delta/4 \quad \text{and} \quad |\eta_k - b| \leq C_{e}\frac{\gamma^{2}}{\kappa_{s,e}^{2} n},$$

where $C_{e} > 0$ is a sufficiently large constant.
Proof. Without loss of generality, assume that $\tilde{D}_{s,e}^b > 0$ and that $\tilde{D}_{s,e}^t$ is locally decreasing at $b$. Observe that there has to be a change point $\eta_k \in (s, b)$, or otherwise $\tilde{D}_{s,e}^b > 0$ implies that $\tilde{D}_{s,e}^t$ is decreasing, as a consequence of Lemma 11. Thus, if $s \leq \eta_k \leq b \leq e$, then

$$\tilde{D}_{s,e}^{\eta_k} \geq \tilde{D}_{s,e}^b \geq D_{s,e}^b - \gamma \geq (c_1 - c_2)\kappa_{s,e}^{\max} \Delta n/2 \geq 2^{-3/2} c_1 \kappa_{s,e}^{\max} \sqrt{\Delta n},$$

where the second inequality follows from (49), and the third inequality follows from (55), and (48). Observe that $e - s \leq C_R \Delta$ and that $(s, e)$ contains at least one change point.

**Step 1.** In this step, we are to show that

$$\min \{\eta_k - s, e - \eta_k\} \geq \min \{1, c_1^2\} \Delta / 16.$$  \tag{53}

Suppose that $\eta_k$ is the only change point in $(s, e)$. Then (53) must hold or otherwise it follows from (41) in Lemma 13, we have

$$D_{s,e}^{\eta_k} \leq \kappa_k \sqrt{\Delta n} c_1 / 4,$$

which contradicts (52).

Suppose $(s, e)$ contains at least two change points. Then $\eta_k - s < \min \{1, c_1^2\} \Delta / 16$ implies that $\eta_k$ is the most left change point in $(s, e)$. Therefore it follows from (43) that

$$\tilde{D}_{s,e}^{\eta_k} \leq \frac{c_1}{4} \tilde{D}_{s,e}^{\eta_k+1} + \sqrt{2n(\eta_k - s)}\kappa_k \leq \frac{c_1}{4} \max_{t=s+1,\ldots,e} \tilde{D}_{s,e}^t + \frac{c_1\kappa_k \sqrt{n \Delta}}{4\sqrt{2}}$$

$$\leq \frac{c_1}{4} \max_{t=s+1,\ldots,e} D_{s,e}^t + \frac{c_1}{4} \gamma + \frac{c_1\kappa_k \sqrt{n \Delta}}{4\sqrt{2}}$$

$$\leq \frac{\gamma}{4} \max_{t=s+1,\ldots,e} D_{s,e}^t - \gamma,$$ \tag{54}

where the last inequality follows from that

$$\max_{t=s+1,\ldots,e} D_{s,e}^t = D_{s,e}^b \geq 2^{-3/2} c_1 \kappa_{s,e}^{\max} \sqrt{\Delta n},$$

as implied by (52). Therefore, (54) contradicts

$$\tilde{D}_{s,e}^{\eta_k} \geq \tilde{D}_{s,e}^b - \gamma,$$

which is also implied by (52).

**Step 2.** It follows from Lemma 14 that

$$\tilde{D}_{s,e}^{\eta_k} - \tilde{D}_{s,e}^{\eta_k+1} \Delta / 32 \geq \frac{c_1 \Delta}{32} \Delta \tilde{D}_{s,e}^{\eta_k} (e - s) \geq \frac{c_{c1}}{32 C_R^2} (c_1 \kappa \sqrt{\Delta n} - 2 \gamma) \geq 2 \gamma. \tag{55}$$

We claim that $b \in (\eta_k, \eta_k + c_1 \Delta / 32)$. By contradiction, suppose that $b \geq \eta_k + c_1 \Delta / 32$. Then

$$\tilde{D}_{s,e}^b \leq \tilde{D}_{s,e}^{\eta_k+1} \Delta / 32 < \tilde{D}_{s,e}^{\eta_k} - 2 \gamma \leq \max_{t=s+1,\ldots,e} \tilde{D}_{s,e}^t - 2 \gamma \leq \max_{t=s+1,\ldots,e} D_{s,e}^t - \gamma = D_{s,e}^b - \gamma,$$ \tag{56}

where the first inequality follows from Lemma 11, the second follows from (55), and the fourth follows from (49). Note that (56) shows that

$$\tilde{D}_{s,e}^b < D_{s,e}^b - \gamma,$$
which is a contradiction with (52) showing that
\[ \bar{D}_{s,e}^b \geq \bar{D}_{s,e}^b - \gamma. \]
Therefore we have \( b \in (\eta_k, \eta_k + c_1 \Delta/32) \).

**Step 3.** This follows from the identical arguments as those in **Step 3** in the proof of Lemma 15 in Padilla et al. (2019a) by letting \( n_{\min} = n_{\max} = n/2 \) and translating notation appropriately. We have that
\[ |b - \eta_k| \leq C_\epsilon \frac{\gamma^2}{n \kappa_k^2}, \]
where \( C_\epsilon > 0 \) is a universal constant.

**D  Proof of Theorem 1**

**Proof of Theorem 1.** Since \( \epsilon \) is the upper bound of the localisation error, by induction, it suffices to consider any interval \((s, e) \subset (1, T)\) that satisfies
\[ \eta_{k-1} \leq s \leq \eta_k \leq \ldots \leq \eta_{k+q} \leq \epsilon \leq \eta_{k+q+1}, \quad q \geq -1, \]
and
\[ \max\{\min\{\eta_k - s, s - \eta_{k-1}\}, \min\{\eta_{k+q+1} - e, e - \eta_{k+q}\}\} \leq \epsilon, \]
where \( q = -1 \) indicates that there is no change point contained in \((s, e)\).

By Assumption 2, it holds that
\[ \epsilon = C_\epsilon T \max\{\log(n \lor T), d^2\} \leq \Delta/4. \]
It has to be the case that for any change point \( \eta_k \in (0, T) \), either \( |\eta_k - s| \leq \epsilon \) or \( |\eta_k - s| \geq \Delta - \epsilon \geq 3\Delta/4 \). This means that \( \min\{|\eta_k - s|, |\eta_k - e|\} \leq \epsilon \) indicates that \( \eta_k \) is a detected change point in the previous induction step, even if \( \eta_k \in (s, e) \). We refer to \( \eta_k \in (s, e) \) an undetected change point if \( \min\{|\eta_k - s|, |\eta_k - e|\} \geq 3\Delta/4 \).

In order to complete the induction step, it suffices to show that we (i) will not detect any new change point in \((s, e)\) if all the change points in that interval have been previous detected, and (ii) will find a point \( b \in (s, e) \) such that \( |\eta_k - b| \leq \epsilon \) if there exists at least one undetected change point in \((s, e)\).

Recall the definitions \( Y_{ij}^k = (X_i(k))^\top X_j(k) \) and \( \hat{Y}_{ij}^k = (\hat{X}_i(k))^\top \hat{X}_j(k) \). For \( j = 1, 2 \), define the events
\[ B_j(\gamma) = \left\{ \max_{1 \leq s < b < e \leq T} \sup_{z \in [0,1]} \left| \sum_{k=s+1}^e w_k^{(j)} \sum_{(i,j) \in O} \left( 1 \{ \hat{Y}_{ij}^k \leq z \} - E \{ 1 \{ Y_{ij}^k \leq z \} \} \right) \right| \leq \gamma \right\}, \]
where
\[ w_k^{(1)} = \begin{cases} \sqrt{\frac{2}{n}} \sqrt{\frac{(e-b)}{(b-s)(e-s)}}, & k = s + 1, \ldots, b, \\ -\sqrt{\frac{2}{n}} \sqrt{\frac{(b-s)}{(e-b)(e-s)}}, & k = b + 1, \ldots, e, \end{cases} \]
\[ w_k^{(2)} = \sqrt{\frac{2}{n}} \frac{1}{\sqrt{e-s}}, \]
and
\[ \gamma = C_{\gamma} T^{1/2} \max\{ \sqrt{\log(n \vee T)}, \, d^{3/2} \}, \]
with a sufficiently large constant \( C_{\gamma} > 0 \).

Define
\[ S = \bigcap_{k=1}^{K} \{ \alpha_s \in [\eta_k - 3\Delta/4, \eta_k - \Delta/2], \beta_s \in [\eta_k + \Delta/2, \eta_k + 3\Delta/4], \text{ for some } s = 1, \ldots, S \}. \]

It follows from Lemma 10 that for \( j = 1, 2 \), it holds that
\[ \mathbb{P}\{B_j\} \geq 1 - 11(n \vee T)^{-c} - 8 Te^{-n}. \]

The event \( S \) is studied in Lemma 13 in Wang et al. (2018b). The rest of the proof assumes the event \( B_1(\gamma) \cap B_2(\gamma) \cap S \).

**Step 1.** In this step, we will show that we will consistently detect or reject the existence of undetected change points within \((s, e)\). Let \( a_m, b_m \) and \( m^* \) be defined as in Algorithm 2. Suppose there exists a change point \( \eta_k \in (s, e) \) such that \( \min\{\eta_k - s, e - \eta_k\} \geq 3\Delta/4 \). In the event \( S \), there exists an interval \((\alpha_m, \beta_m)\) selected such that \( \alpha_m \in [\eta_k - 3\Delta/4, \eta_k - \Delta/2] \) and \( \beta_m \in [\eta_k + \Delta/2, \eta_k + 3\Delta/4] \).

Following Algorithm 2, \((s_m, e_m) = (\alpha_m, \beta_m) \cap (s, e)\). We have that \( \min\{\eta_k - s_m, e_m - \eta_k\} \geq (1/4)\Delta \) and \((s_m, e_m)\) contains at most one true change point.

It follows from Lemma 12, with \( c_1 \) there chosen to be \( 1/4 \), that
\[ \max_{s_m < t < e_m} \tilde{D}_{s_m, e_m}^t \geq \frac{2^{-7/2} \kappa \Delta \sqrt{n}}{\sqrt{e - s}}, \]

Therefore
\[ a_m = \max_{s_m < t < e_m} D_{s_m, e_m}^t \geq \max_{s_m < t < e_m} \tilde{D}_{s_m, e_m}^t - \gamma \geq 2^{-7/2}C_{R}^{-1/2} \kappa \sqrt{\Delta n} - \gamma. \]

Thus for any undetected change point \( \eta_k \in (s, e) \), it holds that
\[ a_m^* = \sup_{1 \leq m \leq S} a_m \geq 2^{-7/2}C_{R}^{-1/2} \kappa \sqrt{\Delta n} - \gamma \geq c_{\tau, 2} \kappa \sqrt{\Delta n}, \tag{57} \]
where the last inequality is from the choice of \( \gamma \) and \( c_{\tau, 2} > 0 \) is achievable with a sufficiently large \( C_{\text{SNR}} \) in Assumption 2. This means we accept the existence of undetected change points.

Suppose that there are no undetected change points within \((s, e)\), then for any \((s_m, e_m)\), one of the following situations must hold.

(a) There is no change point within \((s_m, e_m)\);
(b) there exists only one change point \( \eta_k \in (s_m, e_m) \) and \( \min\{\eta_k - s_m, e_m - \eta_k\} \leq \epsilon_k \); or
(c) there exist two change points \( \eta_k, \eta_{k+1} \in (s_m, e_m) \) and \( \eta_k - s_m \leq \epsilon_k, e_m - \eta_{k+1} \leq \epsilon_{k+1} \).
Observe that if (a) holds, then we have
\[
\max_{s_m < t < e_m} D^t_{s_m, e_m} \leq \max_{s_m < t < e_m} \tilde{D}^t_{s_m, e_m} + \gamma = \gamma < \tau,
\]
so no change points are detected.

Cases (b) and (c) are similar, and case (b) is simpler than (c), so we will only focus on case (c). It follows from Lemma 13 that
\[
\max_{s_m < t < e_m} \tilde{D}^t_{s_m, e_m} \leq \sqrt{\frac{n}{2}} \sqrt{\frac{e_m - \eta_k + 1}{e_m}} + \sqrt{\frac{n}{2}} \sqrt{\eta_k - s_m} \kappa_k \\
\leq \sqrt{2C \epsilon T^{1/2}} \max\{\sqrt{\log(n \vee T)}, d^{3/2}\},
\]
therefore
\[
\max_{s_m < t < e_m} D^t_{s_m, e_m} \leq \max_{s_m < t < e_m} \tilde{D}^t_{s_m, e_m} + \gamma \leq 2\gamma < \tau.
\]

Under (4), we will always correctly reject the existence of undetected change points.

**Step 2.** Assume that there exists a change point \(\eta_k \in (s, e)\) such that \(\min\{\eta_k - s, \eta_k - e\} \geq 3\Delta/4\). Let \(s_m, e_m\) and \(m^*\) be defined as in Algorithm 2. To complete the proof it suffices to show that, there exists a change point \(\eta_k \in (s_{m^*}, e_{m^*})\) such that \(\min\{\eta_k - s_{m^*}, \eta_k - e_{m^*}\} \geq \Delta/4\) and \(|b_{m^*} - \eta_k| \leq \epsilon\).

To this end, we are to ensure that the assumptions of Lemma 14 are verified. Note that (48) follows from (57), (49) and (50) follow from the definitions of events \(B_1(\gamma)\) and \(B_2(\gamma)\), and (51) follows from Assumption 2.

Thus, all the conditions in Lemma 14 are met. Therefore, we conclude that there exists a change point \(\eta_k\), satisfying
\[
\min\{e_{m^*} - \eta_k, \eta_k - s_{m^*}\} > \Delta/4 \quad (58)
\]
and
\[
|b_{m^*} - \eta_k| \leq C \epsilon \gamma^2 \kappa_k^2 \leq \epsilon,
\]
where the last inequality holds from the choice of \(\gamma\) and Assumption 2.

The proof is completed by noticing that (58) and \((s_{m^*}, e_{m^*}) \subset (s, e)\) imply that
\[
\min\{e - \eta_k, \eta_k - s\} > \Delta/4 > \epsilon.
\]
As discussed in the argument before **Step 1**, this implies that \(\eta_k\) must be an undetected change point.

\[\square\]

**References**

Arlot, S., Celisse, A. and Harchaoui, Z. (2012). A kernel multiple change-point algorithm via model selection. *arXiv preprint arXiv:1202.3878*.

Athreya, A., Fishkind, D. E., Tang, M., Priebe, C. E., Park, Y., Vogelstein, J. T., Levin, K., Lyzinski, V. and Qin, Y. (2017). Statistical inference on random dot product graphs: a survey. *The Journal of Machine Learning Research*, 18 8393–8484.
Aue, A., Hörmann, S., Horváth, L. and Reimherr, M. (2009). Break detection in the covariance structure of multivariate nonlinear time series models. The Annals of Statistics, 37 4046–4087.

Avanesov, V. and Buzun, N. (2018). Change-point detection in high-dimensional covariance structure. Electronic Journal of Statistics, 12 3254–3294.

Bickel, P. J. and Sarkar, P. (2016). Hypothesis testing for automated community detection in networks. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 78 253–273.

Cape, J., Tang, M. and Priebe, C. E. (2019). The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. The Annals of Statistics, 47 2405–2439.

Chen, K. and Lei, J. (2018). Network cross-validation for determining the number of communities in network data. Journal of the American Statistical Association, 113 241–251.

Cho, H. (2016). Change-point detection in panel data via double cusum statistic. Electronic Journal of Statistics, 10 2000–2038.

Chu, L. and Chen, H. (2017). Asymptotic distribution-free change-point detection for modern data. arXiv preprint.

Cribben, I. and Yu, Y. (2017). Estimating whole-brain dynamics by using spectral clustering. Journal of the Royal Statistical Society: Series C (Applied Statistics), 66 607–627.

Delyon, B. (2009). Exponential inequalities for sums of weakly dependent variables. Electronic Journal of Probability, 14 752–779.

Fearnhead, P. and Rigaill, G. (2018). Changepoint detection in the presence of outliers. Journal of the American Statistical Association 1–15.

Franco Saldaña, D., Yu, Y. and Feng, Y. (2017). How many communities are there? Journal of Computational and Graphical Statistics, 26 171–181.

Frick, K., Munk, A. and Sieling, H. (2014). Multiscale change point inference. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76 495–580.

Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. The Annals of Statistics, 42 2243–2281.

Garreau, D. and Arlot, S. (2018). Consistent change-point detection with kernels. Electronic Journal of Statistics, 12 4440–4486.

Heyde, C. C. (1963). On a property of the lognormal distribution. Journal of the Royal Statistical Society: Series B (Methodological), 25 392–393.

Holland, P. W., Laskey, K. B. and Leinhardt, S. (1983). Stochastic blockmodels: First steps. Social Networks 109–137.

Lei, J. (2016). A goodness-of-fit test for stochastic block models. The Annals of Statistics, 44 401–424.
Levin, K., Athreya, A., Tang, M., Lyzinski, V. and Priebe, C. E. (2017). A central limit theorem for an omnibus embedding of multiple random dot product graphs. In 2017 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE, 964–967.

Li, T., Levina, E. and Zhu, J. (2016). Network cross-validation by edge sampling. arXiv preprint arXiv:1612.04717.

Liu, F., Choi, D., Xie, L. and Roeder, K. (2018). Global spectral clustering in dynamic networks. Proceedings of the National Academy of Sciences of the United States of America.

Lyzinski, V., Park, Y., Priebe, C. E. and Trosset, M. (2017). Fast embedding for jofc using the raw stress criterion. Journal of Computational and Graphical Statistics, 26 786–802.

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

Mukherjee, S. S. (2018). On Some Inference Problems for Networks. Ph.D. thesis.

Padilla, O. H. M., Yu, Y., Wang, D. and Rinaldo, A. (2019a). Optimal nonparametric change point detection and localization. arXiv preprint arXiv:1905.10019.

Padilla, O. H. M., Yu, Y., Wang, D. and Rinaldo, A. (2019b). Optimal nonparametric multivariate change point detection and localization. arXiv preprint arXiv:1910.13289.

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

Park, Y., Wang, H., Nöbauer, T., Vaziri, A. and Priebe, C. E. (2015). Anomaly detection on whole-brain functional imaging of neuronal activity using graph scan statistics. Neuron, 2 4–000.

Pein, F., Sieling, H. and Munk, A. (2017). Heterogeneous change point inference. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 79 1207–1227.

Prevedel, R., Yoon, Y.-G., Hoffmann, M., Pak, N., Wetzstein, G., Kato, S., Schrödel, T., Raskar, R., Zimmer, M. and Boyden, E. S. (2014). Simultaneous whole-animal 3d imaging of neuronal activity using light-field microscopy. Nature methods, 11 727.

Vershynin, R. (2018). High-dimensional probability: An introduction with applications in data science, vol. 47. Cambridge University Press.

Wang, D., Yu, Y. and Rinaldo, A. (2017). Optimal covariance change point localization in high dimension. arXiv preprint arXiv:1712.09912.

Wang, D., Yu, Y. and Rinaldo, A. (2018a). Optimal change point detection and localization in sparse dynamic networks. arXiv preprint arXiv:1809.09602.

Wang, D., Yu, Y. and Rinaldo, A. (2018b). Univariate mean change point detection: Penalization, cusum and optimality. arXiv preprint arXiv:1810.09498.

Wang, H., Tang, M., Park, Y. and Priebe, C. E. (2014). Locality statistics for anomaly detection in time series of graphs. IEEE Transactions on Signal Processing, 62 703–717.
WANG, T. and SAMWORTH, R. J. (2016). High-dimensional changepoint estimation via sparse projection. *arXiv preprint arXiv:1606.06246*.

YOUNG, S. J. and SCHEINERMAN, E. R. (2007). Random dot product graph models for social networks. In *International Workshop on Algorithms and Models for the Web-Graph*. Springer, 138–149.

YU, Y., WANG, T. and SAMWORTH, R. J. (2014). A useful variant of the davis–kahan theorem for statisticians. *Biometrika, 102* 315–323.

ZHAO, Z., CHEN, L. and LIN, L. (2019). Change-point detection in dynamic networks via graphon estimation. *arXiv preprint arXiv:1908.01823*.

ZOU, C., YIN, G., FENG, L. and WANG, Z. (2014). Nonparametric maximum likelihood approach to multiple change-point problems. *The Annals of Statistics, 42* 970–1002.