Statistical Methods for Clustering and High Dimensional Time Series Analysis

Kunhui Zhang

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:

Ali Shojaie, Chair

Yen-Chi Chen

Abolfazl Safikhani

Program Authorized to Offer Degree:
Department of Statistics

University of Washington

Abstract

Statistical Methods for Clustering
and High Dimensional Time Series Analysis

Kunhui Zhang

Chair of the Supervisory Committee:
Ali Shojaie
Department of Biostatistics

This dissertation mainly explores two statistical tasks, namely clustering and analysis of high-dimensional time series.

Clustering, a very important unsupervised learning problem, studies the structure of unlabeled datasets. The goal of clustering is to partition the data points into subsets such that data points in the same subset are similar and different from those in other subsets. Mode-clustering is a clustering analysis method that partitions the data into groups by the local modes of the underlying density function. Sometimes, finding clusters is not the ultimate goal. The connectivity among clusters may yield valuable information for scientists. This dissertation presents a new clustering method inspired by mode-clustering that not only finds clusters but also assigns each cluster with an attribute label. Clusters obtained from our method show connectivity of the underlying distribution. We also design a local two-sample test based on the clustering result that has more power than a conventional method. We apply our method to the Astronomy and GvHD data and show that our method finds meaningful clusters. In addition, we derive the statistical and computational theory of our method.

Motivated by the challenges of modeling time series data sets that exhibit non-linear patterns, especially in high dimensions, this dissertation also considers the threshold Auto-Regressive (TAR) process. The TAR process provides a family of non-linear auto-regressive

time series models in which the process dynamics are specific step functions of a thresholding variable. While estimation and inference for low-dimensional TAR models have been investigated, high-dimensional TAR models have received less attention. In this dissertation, we develop a new framework for estimating high-dimensional TAR models and propose two different sparsity-inducing penalties. The first penalty corresponds to a natural extension of the classical TAR model to high-dimensional settings, where the same threshold is enforced for all model parameters. Our second penalty develops a more flexible TAR model, where different thresholds are allowed for different auto-regressive coefficients. We show that both penalized estimation strategies can be utilized in a three-step procedure that consistently learns both the thresholds and the corresponding auto-regressive coefficients. However, our theoretical and empirical investigations show that the direct extension of the TAR model is not appropriate for high-dimensional settings and is better suited for moderate dimensions. In contrast, the more flexible extension of the TAR model leads to consistent estimation and superior empirical performance in high dimensions. In addition to the three-step procedure, the dynamic programming approach can successfully handle high dimensions with diverging number of thresholds as well. In particular, extensive numerical analysis and theoretical results demonstrate the advantages of the dynamic programming approach. Finally, we also discuss a method to select the optimal thresholding variable automatically.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

BIC: The Bayesian information criterion.

EBIC: The extended Bayesian information criterion.

HBIC: The high dimensional Bayesian information criterion.

FPR: The false positive rate.

TPR: The true positive rate.

TAR: The threshold Auto-Regressive.

KDE: The kernel density estimator.

IQR: The interquartile range.

ACKNOWLEDGMENTS

I cannot achieve this stage without an overwhelming amount of help from my family, friends, and mentors.

First, I would like to thank my PhD advisor, Ali Shojaie, who offered me a huge amount of support and guidance over the course of my entire graduate education. He is very considerate and patient. When I meet any difficulties, he is always there to support me. It is my pleasure to work with him. I would also like to thank Yen-Chi Chen, Abolfazl Safikhani, and Zaid Harchaoui for their help and valuable insight about research and for serving on my supervisory committee.

Next, I would like to thank the faculty, the department staff, and my fellow PhDs in the Department of Statistics at the University of Washington. They form a very supportive environment and give me lots of help and assistance.

Furthermore, I would like to thank John Dobelman at Rice University for his help and guidance during my graduate school study and PhD application process. I owe my success to support and encouragement from him.

Last but not least, I would like to thank my parents for bringing me up and for their unconditional love and support throughout my entire life. I would also like to thank my friends and relatives for their support and encouragement.

DEDICATION

to my family

Chapter 1

INTRODUCTION

This dissertation mainly explores two statistical tasks, namely clustering and analysis of high-dimensional time series. This first task is discussed in Chapter 2, and is motivated by the Sloan Digital Sky Survey data. Our clustering method allows us to better identify the structures of galaxies and has more extended applications in biology. The second task explores the threshold autoregressive (TAR) models and their estimations. TAR models are among the most widely used non-linear time series models in practice due to their simple and interpretable structure. Chapter 3 and Chapter 4 investigate the TAR models by using different approaches. Chapter 3 establishes a three-step procedure, while Chapter 4 introduces a dynamic programming approach. Both approaches generalize the TAR models in high-dimensions and provide a theoretical consistency of the estimator. In this chapter, we briefly introduce the motivations and problems arising in each chapter and summarize the methods we proposed to solve these problems.

In Chapter 2, we propose a new clustering method inspired by mode-clustering that not only finds clusters but also assigns each cluster with an attribute label. Clusters obtained from our method show connectivity of the underlying distribution. We also design a local two-sample test based on the clustering result that has more power than conventional methods. In addition, we show that our approach provides additional insight into astronomical data and biological flow cytometry data. We also introduce a visualization method using the detected clusters and derive both statistical and computational guarantees of the proposed method.

In Chapter 3, we develop a penalized estimation procedure for learning TAR models. The TAR model is a family of non-linear auto-regressive time series models in which the pro-

cess dynamics are specific step functions of a thresholding variable. It has been extensively studied in univariate and fixed-dimensional settings. However, to the best of our knowledge, methods and theory for high-dimensional TAR models are currently lacking. Given the paucity of the literature on high-dimensional TAR models, we propose a new framework with two estimators for detecting the (unknown) number and values of thresholds and estimating regime-specific auto-regressive parameters in multivariate TAR models. The first approach is a natural extension of the classical TAR model and enforces all auto-regressive parameters to change at the same thresholds. Our theoretical and empirical investigations show that the first approach is only suited for low and moderate dimensions. In contrast, a more flexible TAR model utilizing l_1 penalty allows different thresholds for different auto-regressive coefficients. Applied in a three-step procedure, both estimators can consistently determine both the thresholds and the corresponding auto-regressive coefficients under certain mixing conditions. We also develop efficient algorithms for both methods.

In Chapter 4, we continue discussing the TAR model introduced in Chapter 3 and develop a dynamic programming approach to better estimate the number of thresholds and their corresponding values. In addition, we compare this method to the three-step procedure. We empirically compare the performance of our method with the existing approaches in the simulation section, demonstrating that the dynamic programming approach offers clear advantages in certain cases. Moreover, we establish theoretical results that give a sharper convergence rate of the estimators. The three-step procedure assumes that the minimal jump size v (defined in Assumption B4 in Chapter 3) is independent of the sample size, while the dynamic programming approach in this chapter allows the minimal jump size to decrease with the sample size, and the simulation results corroborate our claims about the advantages of the dynamic programming approach. Finally, we apply both the dynamic programming approach and the three-step procedure to model the stock market data, showing that the dynamic programming approach gives better insight than the three-step procedure and suggests the optimal thresholding variable among other thresholding variables.

In the last chapter, we summarize our previous work and discuss the advantages and limitations of the methods discussed in Chapter 2 to Chapter 4.

Chapter 2

REFINED MODE-CLUSTERING VIA THE GRADIENT OF SLOPE

2.1 Introduction

Mode-clustering is a clustering analysis method that partitions the data into groups by the local modes of the underlying density function Li et al. [2007], Chacón [2012], Arias-Castro et al. [2016], Chen et al. [2016]. A density local mode is often a signature of a cluster, so mode-clustering leads to clusters that are easy to interpret. In practice, we estimate the density function from the data and perform mode-clustering via the density estimator. When we use a kernel density estimator (KDE), there exists a simple and elegant algorithm called the mean-shift algorithm Fukunaga and Hostetler [1975a], Cheng [1995b], Carreira-Perpiñán [2015] that allows us to compute clusters easily. The mean-shift algorithm has made the mode-clustering a numerically friendly problem.

When applied to a scientific problem, we often use a clustering method to gain insight from the data Hastie et al. [2001], Hennig et al. [2015]. Sometimes, finding clusters is not the ultimate goal. The connectivity among clusters may yield valuable information for scientists. To see this, consider the galaxy sample from the Sloan Digital Sky Survey York et al. [2000] in Figure 2.1. While the original data is 3D, here we use a 2D slice of the original data to illustrate the idea. Each black dot indicates the location of a galaxy at a particular location in the sky. Astronomers seek to find clusters of galaxies and their connectivity, since these quantities (clusters and their connections) are associated with the large-scale structures in the universe. Our method finds the underlying connectivity structures without assuming any parametric form of the underlying distribution. In the middle panel, we display the results by the usual mode-clustering method, which only shows clusters, but not how they connect with each other. On the other hand, our proposed method is given in the right panel, which finds a set of dense clusters (purple regions) along with some regions serving as bridges connecting clusters (green areas) and a set of low-density regions (yellow

regions). Thus, our clustering method allows us to better identify the structures of galaxies.

We improve the usual mode-clustering method by (1) adding additional clusters that can further partition the entire sample space, and (2) assigning an attribute label to each cluster. The attribute label will indicate if this cluster is a 'robust cluster' (a cluster around a local mode; purple regions in Figure 2.1), a 'boundary cluster' (a cluster bridging two or more robust clusters; green regions in Figure 2.1), or an 'outlier cluster' (a cluster representing low-density regions; yellow regions in Figure 2.1). With this refined clustering result, we gain further insights into the underlying density function and are able to infer the intricate structure behind the data. Furthermore, we can apply our improved clustering method to the two sample tests. In this case, we can identify the local differences between the two populations and provide a more sensitive result. Note that in the usual case of cluster analysis, adding more clusters is not a preferred idea. However, if our goal is to detect the underlying structures (such as finding the connectivity of high-density regions in the galaxy data in Figure 2.1), using more clusters as an intermediate step to find connectivity could be a plausible approach.

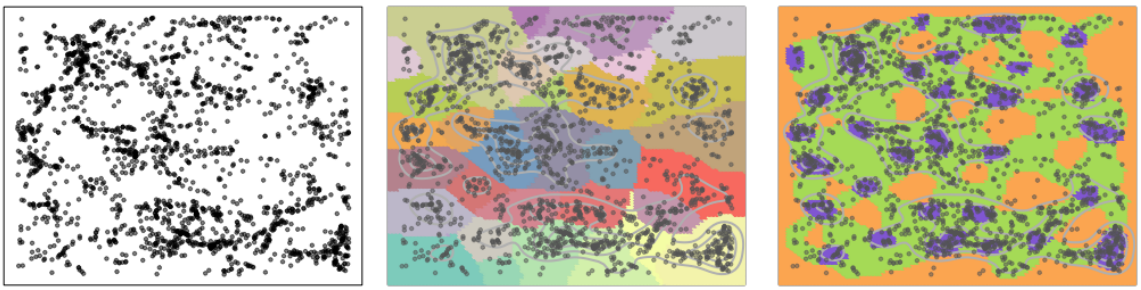


Fig 2.1: Using two clustering methods to learn the cosmic webs. **Left:** the raw galaxy data from the Sloan Digital Sky Survey. **Middle:** the clustering result using the conventional mode/mean-shift clustering. This conventional mode-clustering method fails to detect the connectivity among clusters. **Right:** the clustering result based on our method, where the color indicates different types of clusters.

To summarize, our main contributions are as follows:

- We propose a new clustering method by the slope function that has an additional attribute label of each cluster (Section 2.3).

- We propose new two-sample tests using the clustering result (Section 2.4).
- We introduce a visualization method using the detected clusters (Algorithm 3).
- We derive both statistical and computational guarantees of the proposed method (Section 2.7).

The idea of using local modes to cluster observations can be dated back to Fukunaga and Hostetler [1975a], where the authors used local modes of the KDE to cluster observations and propose the mean-shift algorithm for this purpose Fukunaga and Hostetler [1975a], Comaniciu and Meer [1999]. mode-clustering has been widely studied in statistics and the machine-learning community Chacón and Duong [2013], Chacón et al. [2015b], Carreira-Perpiñán [2015], Arias-Castro et al. [2016], Chen et al. [2016], Chen [2017a]. However, the KDE is not the only option for mode-clustering Li et al. [2007], Scrucca [2016] proposed a Gaussian mixture model method, and Bonis and Oudot [2018] used a fuzzy clustering algorithm, and Jiang and Kpotufe [2017] introduced a nearest-neighbor density method.

This Chapter is organized as follows. We start with a brief review on mode-clustering in Section 2.2 and formally introduce our method in Section 2.3. In Section 2.4, we combine the two-sample test and our approach to create a local two-sample test. We use simulations to illustrate our method on simple examples in Section 2.5. We show the applicability of our approach to three real datasets in Section 2.6. Finally, we study both statistical and computational theories of our method in Section 2.7.

2.2 Review of Mode-Clustering

We start with a review of mode-clustering Chacón [2012], Chacón and Duong [2013], Chen et al. [2016], Menardi [2015]. The concept of mode-clustering is based on the rationale of associated clusters to the regions around the modes of the density. When the density function is estimated by the kernel density estimator, there is an elegant algorithm called the mean-shift algorithm Fukunaga and Hostetler [1975a] that can easily perform the clustering.

In more detail, let p be a probability density function with a compact support $\mathbb{K} \subset \mathbb{R}^d$.

Starting at any point x, mode-clustering creates a gradient ascent flow $\gamma_x(t)$ such that

$$\gamma_x(0) = x, \quad \gamma_x'(t) = \nabla p(\gamma_x(t)).$$

Namely, the flow $\gamma_x(t)$ starts at point x and moves according to the gradient at the present location. Let $\gamma_x(\infty) = \lim_{t\to\infty} \gamma_x(t)$ be the destination of the flow $\gamma_x(t)$. According to the Morse theory Morse [1925], Milnor et al. [1963], when the function is smooth (being a Morse function), such a flow converges to a local maximum of p except for starting points in a set of the Lebesgue measure 0. The mode-clustering partitions the space according to the destination of the gradient flow, that is, for two points x, y, they will be assigned to the same cluster if $\gamma_x(\infty) = \gamma_y(\infty)$. For a local mode η, we define its basin of attraction as $D(\eta) = \{x : \gamma_x(\infty) = \eta\}$. The basin of attraction describes the set of points that belongs to the same cluster.

In practice, we do not know p, so we replace it by a density estimator, $\hat{p}_n$. A common approach to estimate p as the kernel density estimator, in which $\hat{p}_n$ is

$$\hat{p}_n(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

where K is a smooth function (also known, according to the Morse theory, as the kernel function), such as a Gaussian kernel, and $h > 0$ is the smoothing bandwidth that determines the amount of smoothness. Since we used a nonparametric density estimator, we did not need to assume any parametric assumptions on the shape of the distribution.) With this choice, we the define a sample analogue to the flow $\gamma_x(t)$ as

$$\hat{\gamma}_x(0) = x, \quad \hat{\gamma}_x'(t) = \nabla \hat{p}(\hat{\gamma}_x(t))$$

and partition the space according to the destination of $\hat{\gamma}_x$.

2.3 Clustering via the Gradient of Slope

2.3.1 Refining the Clusters by the Gradient of Slope

As is mentioned previously, the mode-clustering has some limitations that the resulting clusters do not provide enough information on the finer structure of the density. To resolve this problem, we introduce a new clustering method by considering gradient descent flows of the 'slope' function. Let $\nabla p(x)$ be the gradient of p. Define the slope function of p as $s(x) = \|\nabla p(x)\|^2$. Namely, the slope function is the squared amplitude of the density gradient.

An interesting property of the slope function is that the minimal points $\{x : s(x) = 0\} = \{x : \nabla p(x) = 0\} = \mathcal{C}$ form the collection of critical points of p, so it contains local modes of p as well as other critical points, such as saddle points and local minima. According to the Morse theory Banyaga and Hurtubise [2013], Matsumoto [2002], there is a saddle point between two nearby local modes when the function is a Morse function. A Morse function is a smooth function f, such that all eigenvalues of Hessian Matrix of f at every critical point are away from 0. This implies that saddle points may be used to bridge connecting regions around two local modes.

With this insight, we propose to create clusters using the gradient 'descent' flow of $s(x)$. Let $\nabla s(x)$ be the gradient of the slope function. Given a starting point $x \in \mathbb{R}^d$, we construct a gradient descent flow as follows:

$$\pi_x(0) = x, \quad \pi_x'(t) = -\nabla s(\pi_x(t)). \tag{2.1}$$

That is, π_x is a flow starting from x and moving along the direction of ∇s. Similar to mode-clustering, we use the destination of gradient flows to cluster the entire sample space.

Note that if the slope function s is a Morse function, the corresponding PDF p will also be a Morse function, as described in the following Lemma.

Lemma 1. *If $s(x)$ is a Morse function, then $p(x)$ is a Morse function.*

Throughout this chapter, we will assume that the slope function is Morse. Thus, the corresponding PDF will also be a Morse function and all critical points of the PDF will be

well-separated.

2.3.2 Type of Clusters

Recall that $\mathcal{C}$ is the collection of critical points of density p. Let $\mathcal{S}$ be the collection of local minima of the slope function $s(x)$. It is easy to see $\mathcal{C} \subset \mathcal{S}$, since any critical point of p has gradient 0, so it is also a local minimum of s.

Thus, the gradient flow in Equation (2.1) leads to a partition of the sample space. Specifically, let $\pi_x(\infty)$ be the destination of the gradient flow $\pi_x(t)$. For an element $m \in \mathcal{C}$, let $\mathbb{S}(m) = \{x : \pi_x(\infty) = m\}$ be its basin of attraction.

We use the sign of eigenvalues of $\nabla^2 p(x)$ to assign an additional attribute to each basin, so the set $\{\mathbb{S}(m) : m \in \mathcal{C}\}$ forms a collection of meaningful disjoint regions. In more detail, for a critical point $m \in \mathcal{C}$ such that $p(m) > \delta$ for a small threshold δ, its $\mathbb{S}(m)$ is classified according to

$$\mathbb{S}(m) \text{ is a } \begin{cases} \text{robust cluster} & \text{if } s(m) = 0, \lambda_1(m) < 0; \\ \text{outlier cluster} & \text{if } s(m) = 0, \lambda_d(m) > 0; \\ \text{boundary cluster} & \text{otherwise,} \end{cases} \tag{2.2}$$

where $\lambda_l(x)$ is the l-th ordered eigenvalue of $\nabla^2 p(x)$ ($\lambda_1(x) \geq \ldots \geq \lambda_d(x)$). In the case of $p(m) \leq \delta$, we always assign it as an outlier cluster. Note that the threshold δ was added to stabilize the numerical calculation. In other words, we refer to a basin of attraction in $\mathbb{S}(m)$ as a robust cluster if $m \in \mathcal{C}$ is a local mode of p. If m is a local minimum of p, then we call its basin of attraction an outlier cluster. The remaining clusters, which are regions connecting robust clusters, are denoted as boundary cluster. Note that the regions outside the support are, by definition, a set of local minima. We assign the same cluster label to those x whose destination $\pi_x(\infty)$ is outside the support, which is an outlier cluster.

Our classification of $\mathbb{S}(m)$ is based on the following observations. Regions around local modes of p are where we have strong confidence that these points should belong to the cluster represented by their nearby local modes. Regions around local minima of p are the low-density areas where we should treat them as anomaly points/outliers. Figure 2.1 provides a concrete example that our clustering method could lead to more scientific insight–

the connectivity among robust clusters may reveal intricate structure of the underlying distribution.

Defining different types of clusters allows us to partition the whole space into meaningful sub-regions. Given a random sample, to assign the cluster label to each of them, we simply examine which basins of attraction these data points fall in and pass the cluster labels from the regions to the data points. After assigning cluster labels to data points, the cluster categories in Equation (2.2) provide additional information about the characteristics of each data point. Those data points in robust clusters are data points that are highly clustered together; points in the outlier clusters are data points in low-density regions, which could be viewed as anomalies; the rest of points are in the boundary clusters, where these points are not well-clustered and are on the connection regions among different robust clusters.

2.3.3 Estimators

The above procedure is defined when we have access to the true PDF p. In practice, we do not know p, but we have an IID random sample $X_1, \ldots, X_n$ from p with a compact support $\mathbb{K}$. So we estimate p using $X_1, \ldots, X_n$ and then use the estimated PDF to perform the above clustering task.

While there are many choices of density estimators, we consider the kernel density estimator (KDE) in this chapter, since it has a nice form and its derivatives are well-established Wasserman [2006], Chacón et al. [2011], Scott [2015b], Chen [2017a]. In more detail, the KDE is

$$\hat{p}_n(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \qquad \hat{s}_n(x) = \|\nabla \hat{p}_n(x)\|^2,$$

where K is a smooth function (also known as the kernel function) such as a Gaussian kernel, and $h > 0$ is the smoothing bandwidth that determines the amount of smoothness. Note that the bandwidth h in the KDE could be replaced by h_i that depends on each observation. This is called the variable bandwidth KDE in Breiman et al. [1977]. However, since the choice of how h_i depends on each observation is a non-trivial problem, so to simplify the problem, we set all bandwidths to be the same.

Based on $\hat{s}_n(x)$, we first construct a corresponding estimated flow using $\nabla \hat{s}_n(x)$:

$$\hat{\pi}_x(0) = x; \quad \hat{\pi}'_x(t) = -\nabla \hat{s}_n(\hat{\pi}_x(t)). \tag{2.3}$$

An appealing feature is that $\nabla \hat{s}_n(x)$ has an explicit form:

$$\nabla \hat{s}_n(x) = \nabla^2 \hat{p}_n(x) \nabla \hat{p}_n(x), \tag{2.4}$$

where $\nabla \hat{p}_n(x)$ and $\nabla^2 \hat{p}_n(x)$ are the estimated density gradient and Hessian matrix of p. Thus, to numerically construct the gradient flow $\hat{\pi}_x(t)$, we update x by

$$x \leftarrow x - \gamma \cdot \nabla^2 \hat{p}_n(x) \nabla \hat{p}_n(x), , \tag{2.5}$$

where $\gamma > 0$ is the learning rate parameter. Algorithm 1 summarizes the gradient descent approach.

Algorithm 1: Slope minimization via gradient descent.

1. Input: $\hat{p}_n(x)$ and a point x.
2. Initialize $x_0 = x$ and iterate the following equation until convergence: (γ is a step size that could be set to a constant)

$$x_t = x_{t-1} - \gamma \cdot \nabla^2 \hat{p}_n(x_{t-1}) \nabla \hat{p}_n(x_{t-1}).$$

3. Output: x_∞ .

With an output from Algorithm 1, we can group observations into different clusters, with each cluster labeled by a local minimum of $\hat{s}_n$. We assign an attribute to each cluster via the rule in Equation (2.2). Note that the smoothing bias could cause some biases around the boundary of clusters. However, when $h \to 0$, this bias will asymptotically be negligible.

2.4 Enhancements in Two-Sample Tests

Our clustering method can be used as a localized two-sample test. An overview of the idea is as follows. Given two random samples, we first merge them and use clustering method to form partitions of the sample space. Under the null hypothesis, the two samples are

from the same distribution, so the proportion of each sample within each cluster should be similar. By comparing the difference in proportion, we obtain a localized two-sample test. Algorithm 2 summarizes the procedure.

In more detail, suppose we want to compare two samples $G_1 = \{X_1, X_2, \ldots, X_N\}$ and $G_2 = \{Y_1, Y_2, \ldots, Y_M\}$. Let $X_1, \ldots X_N \sim P_X$ and $Y_1, \ldots, Y_M \sim P_Y$. The null hypothesis we want to test is $H_0 : P_X = P_Y$ against $H_1 : P_X \neq P_Y$.

Under H_0, the two samples are from the same distribution, so they have the same PDF q. We first pull both samples together to form a joint dataset

$$G_{\mathsf{all}} = \{X_1, \ldots, X_N, Y_1, \ldots, Y_M\}.$$

We then compute the KDE $\hat{p}_n$ using G_{all} and compute the corresponding estimated slope function $\hat{s}_n$ and apply Algorithm 1 to form clusters. Thus, we obtain a partition of G_{all}. Under H_0, the proportion of Sample 1 in each cluster should be roughly the same as the global proportion $\frac{N}{N+M}$. Therefore, we can apply a simple test of the proportion within each cluster to obtain a p-value. In practice, we often only focus on the robust and boundary clusters and ignore the outlier clusters because of sample size consideration. Let $D_1, \ldots, D_J \subset G_{\mathsf{all}}$ be the robust and boundary clusters, and

$$r_0 = N/(N+M); \tag{2.6}$$

be the global proportion, and

$$r_j = \frac{|D_j \cap G_1|}{|D_j|}. \tag{2.7}$$

be the observed proportion of cluster D_j. We use the test statistic

$$Z_j = \frac{r_j - r_0}{\sqrt{r_0(1-r_0)/n_j}},$$

where $n_j = |D_j|$ is the total number of the pulled sample within cluster D_j, when H_0 is true and the test statistic Z_j follows from a standard normal distribution asymptotically. Note that since we are conducting multiple tests, we reject the null hypothesis after applying the

Bonferroni correction.

Algorithm 2: Local two-sample test.

1. Combine two samples (G_1 and G_2) into one, called G_{all} and compute $r_0 = \frac{N}{N+M}$ from Equation (2.6).
2. Construct a kernel density estimator using G_{all} and its slope function and apply Algorithm 1 to form clusters based on the convergent point.
3. Assign an attribute to each cluster according to Equation (2.2).
4. Let robust clusters and boundary clusters be $D_1, D_2, \ldots, D_J$, where $D_j \subset G_{\mathsf{all}}$ for each j.
5. For each cluster D_j, compute r_j from Equation (2.7) and construct Z statistic:

$$Z_j = \frac{r_j - r_0}{\sqrt{r_0(1-r_0)/n_j}}.$$

   Find the corresponding p-value p_j.
6. Reject H_0 if $p_j < \alpha/J$ for some j under the significance level α.

We can apply this idea to other clustering algorithms. However, we need to be very careful when implementing it because we are using data twice–first to form clusters, then again to do two-sample tests. This could inflate the Type 1 error. Our approach is asymptotically valid because the clusters from the estimated slope converge to the clusters of the population slope (see Section 2.7). Note that our method may not control the Type 1 error in the finite sample situation, but our simulation results in Section 2.5.2 show that this procedure still controls the Type 1 error. This might be due to the conservative result of the Bonferroni correction.

The advantage of this new two-sample test is that we are using the local information, so if the two distributions only differ in a small region, this method will be more powerful than a conventional two-sample test. In particular, the robust clusters are often the ones with more power because they have a higher sample size, and the bumps in the pulled sample's density could be created by a density bump of one sample but not the other, leading to a region with high testing power. In Section 2.5, we demonstrate this through some numerical simulations.

2.4.1 An Approximation Method

The major computational burden of Algorithm 2 comes from Step 2, where we apply Algorithm 1 to 'every observation'. This may be computationally heavy if the sample size is large. Here we propose a quick approximation to the clustering result.

Instead of applying Algorithm 1 to every observation, we randomly subsample the original data (large dimension) or create a grid (low dimension) of points and only apply Algorithm 1 to this smaller set of points. This gives us an approximated set of local minima of the slope function. We then assign a cluster label of each observation according to the 'nearest' local minima.

2.5 Simulations

In this section, we demonstrate the applicability of our method by applying it to some simulation setups. Note that in practice, we need to choose the smoothing bandwidth h in the KDE. Silverman's rule Silverman [1986] is one of the most popular methods for bandwidth selection. The idea is to find the optimal bandwidth by minimizing the mean integrated squared error of the estimated density. Silverman [1986] proposed to use the normal density to approximate the second derivative of the true density, and use the interquartile range providing a robust estimation of the sample standard deviation. For the univariate case, it is defined as follows:

$$h_s = 1.06 \min\{\frac{\text{IQR}}{1.34}, \hat{\sigma}\} n^{-1/5},$$

where $\hat{\sigma}$ is the sample standard deviation and IQR is the interquartile range. As discussed earlier, we choose $h = C' \left(\frac{\log n}{n}\right)^{\frac{1}{d+8}}$, where C' is a constant. This choice is motivated by theoretical analysis in Section 2.7 (Theorem 2). In practice, we do not know C', so we applied a modification of Silverman's rule Silverman [1986]:

$$h = \min\left(\frac{1}{d}\sum_{k=1}^{d}\hat{\sigma}_k, \frac{1}{d}\sum_{k=1}^{d}\frac{\text{IQR}_\text{k}}{1.34}\right) n^{-1/(8+d)}, \tag{2.8}$$

where $\hat{\sigma}_k$ is the standard deviation of the samples on kth dimension, IQR_k is the interquartile range on kth dimension, and $k = 1, 2, \ldots, d$. Note that our procedure involves estimating

both the gradient and Hessian of the PDF. The optimal bandwidth of the two quantities are different, so one may apply two separated bandwidths for gradient and Hessian estimation. However, our empirical studies show that a single bandwidth (optimal for Hessian estimation) still leads to reliable results. Note that this bandwidth selector tends to oversmooth the data in the sense that some density peaks in Figure 2.6b were not detected (not in purple color).

2.5.1 Clustering

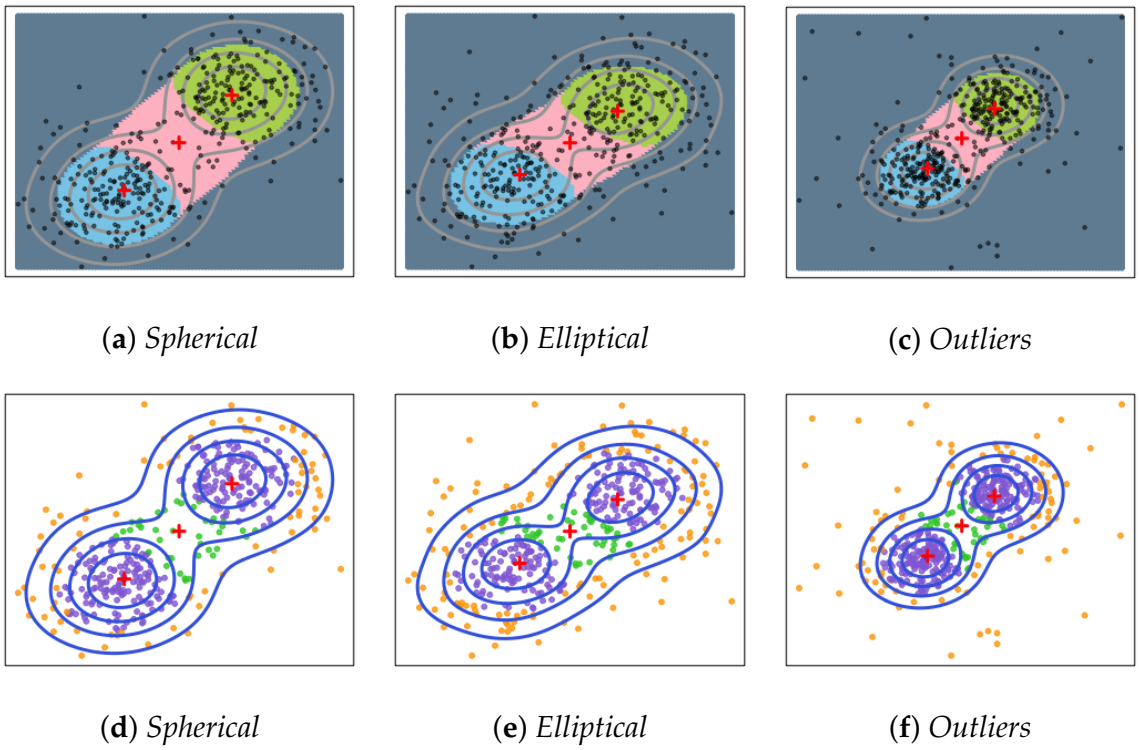
Two-Gaussian mixture. We sample $n = 400$ points from a mixture of two-dimensional normals $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ with equal proportions under the following three scenarios:

- *Spherical*: $\boldsymbol{\mu}_1 = \mathbf{0}$, $\boldsymbol{\mu}_2 = 3e_1 + 3e_2$, and $\Sigma = I_2$.
- *Elliptical*: $\boldsymbol{\mu}_1 = \mathbf{0}$, $\boldsymbol{\mu}_2 = 3e_1 + 3e_2$, and $\Sigma = \mathrm{diag}(1, 3)$. (Note that these clusters are elongated in noise directions.)
- *Outliers*: Same construction as *Spherical*, but with 60 random points (noise) from a uniform distribution over $(-5, 8) \times (-5, 8)$. By design, the outliers differ in such a way that they can only add a little ambiguity.

Note that e_i is the ith standard basis vector, and I_2 is the 2×2 identity matrix. For each scenario, we apply the gradient flow method and draw the contour. If points are outliers, their destinations go to infinity. Thus, we set a threshold to stop them from moving and assign them to outlier clusters.

Figure 2.2 demonstrates that we identify both two clusters and the boundary of these two clusters. Each colored region is the basin of attraction of a local minimum of $s(x)$ in the picture (a–c). Picture (d–f) provide examples of data points clustering. Given the setting of two equal-sized Gaussian mixture, it is straightforward to verify that the gradient flow algorithm can successfully distinguish points according to their destinations. The purple points represent points that belong to corresponding clusters with strong confidence, while green points represent points in low-density areas that belong to the connection regions

among clusters. The yellow points represent points that are not important to any of the clusters. In summary, our proposed method performs well and is not affected by the changes of covariance and outliers.



(**a**) *Spherical* (**b**) *Elliptical* (**c**) *Outliers*

(**d**) *Spherical* (**e**) *Elliptical* (**f**) *Outliers*

Fig 2.2: Simulations with different data settings. Picture (**a**,**d**), picture (**b**,**e**), and picture (**c**,**f**) display, respectively, the three different simulation scenarios: *Spherical*, *Elliptical*, and *Outliers*. In picture (**a**–**c**), each colored region is the basin of attraction of a local minimum of $s(x)$, while the grey regions are the regions that belong to outlier clusters. Picture (**d**–**f**) provides an example of clustering of data points. Points that labeled purple, green, and orange are assigned to robust, boundary, and outlier clusters, respectively.

Four-Gaussian mixture. To show how boundary clusters can serve as bridges among robust clusters, we consider a four-Gaussian mixture. We sample $n = 800$ from a mixture of four two-dimensional normals $N(0, 0.1I_2)$, $N(0.5e_1, 0.1I_2)$, $N(0.5e_2, 0.1I_2)$ and $N(0.5e_1 + 0.5e_2, 0.1I_2)$ with equal proportion. Then we apply our method and display the result in Figure 2.3. Each colored region is the basin of attraction of a local minimum of $s(x)$. The red '+'s are the corresponding local minima to each of the basin of attraction. Clearly, we

see how robust clusters are connected by the boundary clusters so the additional attributes provide useful information on the connectivity among density modes.

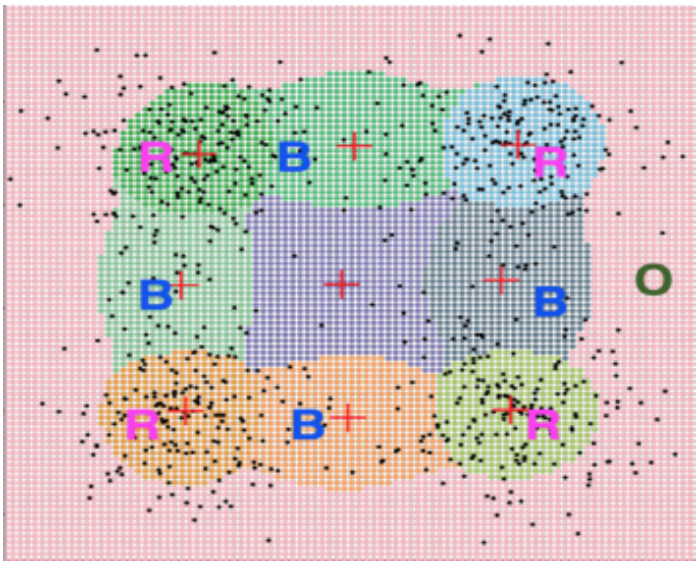


Fig 2.3: Example of the basins of attraction of a Gaussian mixture. Four groups of data are separated into three types of clusters. We partition the space into 10 parts. 'R' represents the region of the robust cluster, 'B' represents the region of the boundary cluster, and 'O' represents the region of the outlier cluster.

Comparison. To better illustrate the strength of our proposed method, we generate an unbalanced four-Gaussian mixture. We sample $n = 2400$ from a mixture of four two-dimensional normals $N(0, 0.5I_2)$, $N(2e_1, 0.5I_2)$, $N(5e_2, 0.5I_2)$ and $N(2e_1 + 5e_2, 0.5I_2)$ with proportion $\frac{5}{12}, \frac{5}{12}, \frac{1}{12}, \frac{1}{12}$, respectively. Then we apply our method and compare it with the density-based spatial clustering of applications with noise (DBSCAN) Ester et al. [1996] in Figure 2.4. DBSCAN is a classical non-parametric, density-based clustering method that estimates the density around each data point by counting the number of points in a certain neighborhood and applies a threshold minPts to identify core, border and noise points. DBSCAN requires two parameters: the minimum number of nearby points required to form a core point (minPts) and the radius of a neighborhood with respect to a certain point (eps). Two points are connected if they are within the distance of eps. Clusters are the connected components of connected core points. Border points are points connected to a

core point, but which do not have enough neighbors to be a core point. Here, we investigate the feasibility of using border points to detect the connectivity of clusters. These two parameters, minPts and eps, are very hard to choose. In the top two rows of Figure 2.4, we set minPts equal to 5 and 10 and change the value of eps to see if we can find the connectivity of core points using border points (gray points). Our results show that it is not possible to use border points to find the connectivity of the top two clusters and the bottom two clusters at the same time. When we are able to detect the connectivity of bottom two clusters (panel (f)), we are not able to find the top two clusters. On the other hand, when we can find the connectivity of the top two clusters (panel (c,h)), the bottom two clusters have already merged into a single cluster. The limitation of DBSCAN is that it is based on the density level set, so when the structures involve different density values, DBSCAN will not be applicable. In contrast, our method only requires one parameter, bandwidth, and it has good performance in this case. From Figure 2.4i–l, our method detects four robust clusters and their boundaries correctly. In addition, this result also shows that our method is robust to the bandwidth selection.

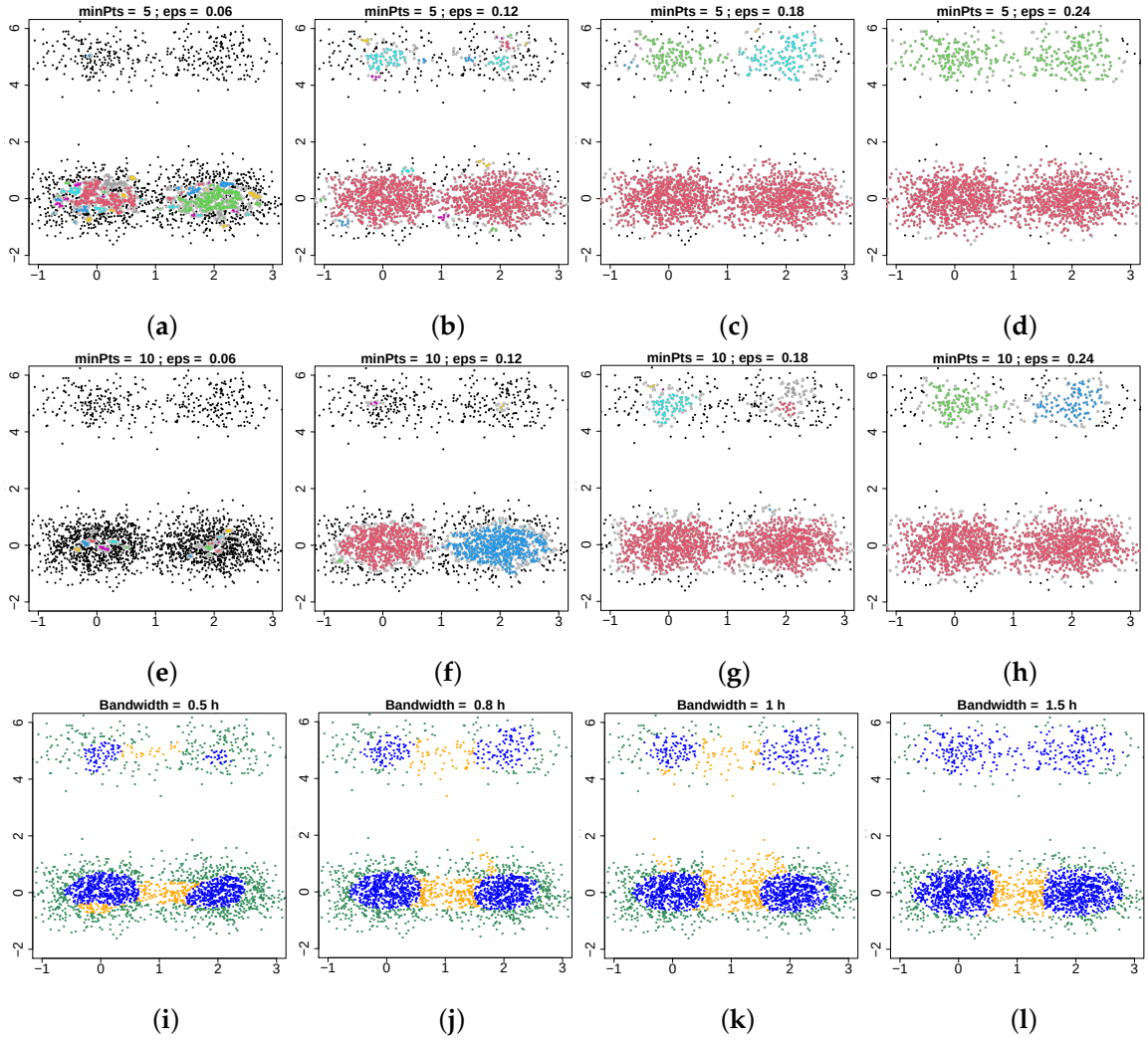


Fig 2.4: Picture (**a**)–(**f**) displays the simulations using DBSCAN with different parameters settings, where minPts represents the the minimum number of points required to form a dense region and eps represents the radius of a neighborhood with respect to certain point. Picture (**i**)–(**l**) displays the simulations using our proposed method with different bandwidth, where h represents the bandwidth selected according to Equation (2.8). In Picture (**a**)–(**h**), each colored region is the cluster detected by DBSCAN, while the gray and black points are points that are border points and outliers, respectively. In Picture (**i**)–(**l**), points that are labeled blue, orange, and green are assigned to robust, boundary, and outlier clusters, respectively.

2.5.2 Two-Sample Test

In this section, we carry out simulation studies to evaluate the performance of the two-sample test in Section 2.4. We compare our method to three other popular approaches: the energy test Székely and Rizzo [2004], the kernel test Gretton et al. [2012], and KS Massey [1951] tests based on each of the two variables.

Our simulation is designed as follows. We draw random samples from a two-Gaussian mixture model in Equation (2.9):

$$p(x) = a\phi(\mu_1, \Sigma_1) + (1-a)\phi(\mu_2, \Sigma_2), \tag{2.9}$$

where $\phi(\cdot)$ is a cumulative distribution function of normal distribution. For the first group, we choose the parameters as $a = 0.7$, $\mu_1 = (-1, 0)$, $\mu_2 = (0, 1)$, $\Sigma_1 = \text{diag}\,(0.3, 0.3)$, and $\Sigma_2 = \text{diag}\,(0.3, 0.3)$.

In our first experiment (left panel of Figure 2.5), we generate the second sample from a Gaussian mixture with identical setup, except that the second covariance matrix $\Sigma_2 = \text{diag}(\sigma_2, 0.3)$, and we gradually increase σ_2 from 0.3 (H_0 is correct) to 0.8 to see how the power of the test changes. We generate $n_1 = n_2 = 500$ observations in both samples and repeat the process 500 times to compute the power of the test. This experiment investigates the power as a function of signal strength.

In the second experiment (right panel of Figure 2.5), we consider a similar setup except that we fix $\Sigma_2 = \text{diag}\,(0.35, 0.3)$ and vary the sample size from $n_1 = n_2 = 500$ to $n_1 = n_2 = 4000$ and examine how the power changes under different sample size. This experiment examines the power as a function of sample size.

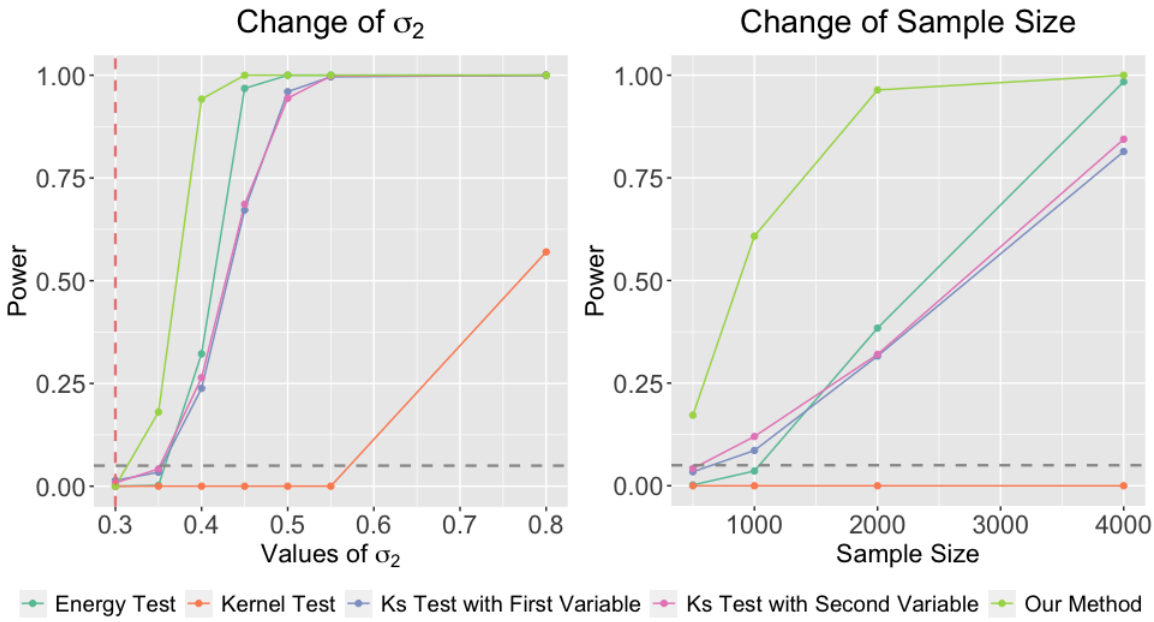


Fig 2.5: Power analysis of the proposed method. We compare the power of our two-sample test with three other approaches: the energy test, the kernel test, the KS test with only the first variable, and the KS test with only the second variable. In the left panel, we vary the variance of the second Gaussian. In the right panel, we fix the two distributions and increase the sample size. In both cases, our method has a higher power than the other three naive approaches.

In both experiments, all methods control the Type 1 errors. However, our method has better power in both experiments compared to the other alternatives. Our method is more powerful because we utilize the local information from clustering. In this simulation setup, the difference between the two distributions is the width of second Gaussian component. Our method is capable of capturing this local difference and using it as evidence in the hypothesis test.

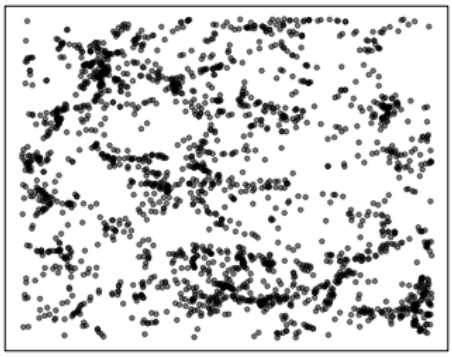
Finally, we would like to emphasize again that two-sample test after clustering has to be used with caution; we are using data twice, so we may not be able to control the Type 1 error. One needs to theoretically justify that the resulting clusters converge to a population limit and apply numerical analysis to investigate the finite-sample coverage.

2.6 Real Data Application

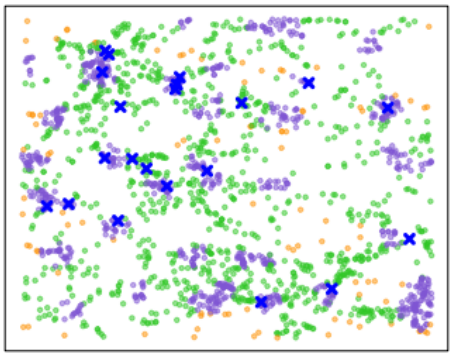
2.6.1 Applications to Astronomy

We apply our method to detect the Cosmic Webs Bond et al. [1996] from the galaxy sample of the Sloan Digital Sky Survey York et al. [2000]. It is known that galaxies inside our universe are not uniformly distributed. There are low-dimensional structures where matters are aggregated together. Roughly speaking, there are four types of structures in the Cosmic Webs: galaxy clusters, filaments, sheets, and voids Bond et al. [1996]. Galaxy clusters are small regions with lots of matter. Filaments are regions with moderate matter density which connect galaxy clusters. Sheets are weakly dense regions where clusters and filaments are distributed. Voids are vast regions with very low matter density. Because of their properties, galaxy clusters are like zero-dimensional objects (points), filaments are one-dimensional curve-like structures, sheets are two-dimensional surface-like structures, and voids are three-dimensional regions.

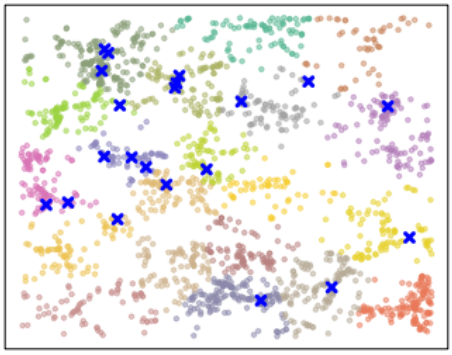
Figure 2.6 displays our result. Note that it is the same data as Section 2.1. Panel (a) of Figure 2.6 shows the scatter plot of galaxies in the thin slice of the universe. In Panel (b), we color galaxies according to the types of clusters they belong to; purple, green and orange regions are the robust boundary and outlier clusters, respectively. We mark the locations of known galaxy clusters as blue "×"s Koester et al. [2007a]. These galaxy clusters are obtained using imaging analysis Koester et al. [2007b], which is a completely different approach. As can easily be seen, there is a strong agreement between galaxy clusters and the robust regions. Out of the 21 galaxy clusters, 85.71% fall into the robust clusters, and 14.29% fall into the boundary clusters. Moreover, the boundary clusters (green), connecting the robust clusters (purple), behave like the filaments in the Cosmic Webs, and the low-density outlier clusters are similar to the void structures. Figure 2.6 As for comparison, we display the results from k-means (Figure 2.6c), traditional mode-clustering (Figure 2.6d)), and Gaussian mixture model (Figure 2.6e), which are not structurally correlated with the locations of blue "×"s.



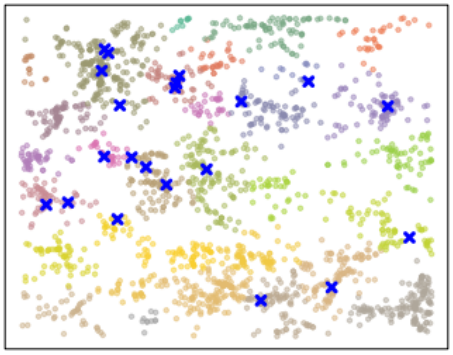
(**a**) Original Cosmic Web data



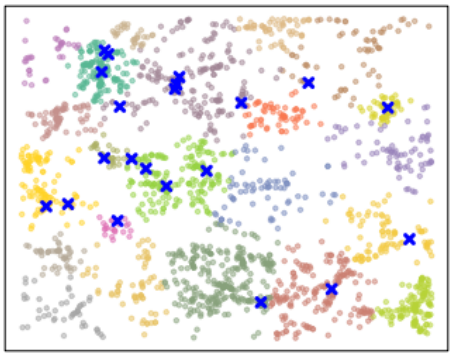
(**b**) Three types of clusters with data



(**c**) k-means for Cosmic Web data



(**d**) Mode clustering for Cosmic Web data



(**e**) Gaussian mixture model for Cosmic Web data

Fig 2.6: We show that the gradient flow method is better in detecting the 'Cosmic Web' Bond et al. [1996] in our universe. For comparison, we perform the k-means clustering method with 20 centers and traditional mode-clustering to show that our proposed method is better to detect the 'Cosmic Web' in our universe. The blue "×"s are the points from image analysis. The results do not structurally correlate with the locations of blue "×"s.

Thus, this analysis reveals the potential of our approach as a good method for detecting the Cosmic Webs with less information. Note that, since our dataset is two-dimensional, we cannot define the cosmic sheet structures.

2.6.2 Application to GvHD Data

We also apply our method to the GvHD (Graft-versus-Host Disease) data from Brinkman et al. [2007]. The GvHD is a famous example for two-sample test problem. It contains a positive/disease sample and a control/normal sample. There are 9083 observations in the positive sample and 6809 observations in the control sample. Each observation consists of four biomarkers: CD4, CD8b, CD3, and CD8. Our goal is to test whether the positive sample and control sample are from the same distribution or not.

Since the sample size is non-trivial and the dimension is 4, naively applying Algorithm 2 will be computationally heavy, so we apply the approximation method in Section 2.4.1. We first random select 5% of the whole dataset, including both positive and control samples, as initial points in Algorithm 2. Then, the algorithm to find the local minima and add the attribute label is based on Equation (2.2). Finally, we assign a cluster label and attribute it to each observation according to an observation's nearest detected local minima of the slope.

Having identified clusters, we perform the two-sample test, and the result is summarized in Table 2.1. According to Table 2.1, all groups are significantly different. Thus, we can conclude that the positive sample is from a different distribution than the control sample.

Table 2.1: Summary of estimated proportion in each group. Note that "Proportion" in the table is referred to as the proportion of the positive group.

Cluster	**Proportion**	**5% CI**	**95% CI**	**Z Score**	**Cluster Type**
1	0.910	0.900	0.920	46.980	Robust Cluster
2	0.010	0.010	0.020	−69.620	Robust Cluster
3	0.680	0.650	0.720	5.550	Robust Cluster
4	0.370	0.350	0.390	−17.570	Boundary Cluster
5	0.800	0.770	0.830	11.470	Boundary Cluster
6	0.410	0.380	0.440	−9.920	Boundary Cluster
7	0.920	0.900	0.940	19.170	Robust Cluster
8	0.420	0.370	0.470	−5.930	Boundary Cluster
Overall Proportion	0.570				

The clustering result can be used to visualize the data, since the robust and boundary clusters characterize regions with non-trivial probability mass and each cluster is represented by a minimum of the slope function. The slope minimum within each cluster is the center of that cluster. Algorithm 3 provides a summary of the visualization algorithm. In more detail, we first compute the minimal distance of two different clusters to decide whether two clusters (robust or boundary) are connected. If the value is less than $4 \times \sqrt{h^2 \times d}$, two clusters are connected (neighboring to each other), where d is the number of dimensions. Then we apply multi-dimensional scaling to the centers of robust and boundary clusters to reduce the dimension to 2. Each of these points represents a particular cluster. If two clusters are connected, we add an edge to them on the graph. Finally, we add a pie chart at each cluster's center with a radius corresponding to the total number of observations in that cluster, and partition the pie chart according to the composition from the two samples. Figure 2.7 shows the 2D visualization of the GvHD data, along with the composition of the two samples in each cluster.

Algorithm 3: Visualization based on slope function.

1-4. The same steps as Algorithm 2.

5. Let robust clusters be $\{R_1, R_2, \ldots, R_{J_1}\}$ and boundary clusters be $\{B_1, B_2, \ldots, B_{J_2}\}$.

6. For each pair of R_{j_1} and B_{j_2}, compute their Hausdorff distance (minimal distance of all pairs):

$$\text{edge}_{\text{j}_1,\text{j}_2} = \text{Haus}\left(\text{R}_{\text{j}_1}, \text{B}_{\text{j}_2}\right).$$

7. Apply multidimensional scaling to local minima corresponding to robust and boundary clusters. Let their 2 dimensional representation point be $s_1^*, \cdots s_{J_1+J_2}^*$.

8. For each cluster D_j in $\{R_1, R_2, \ldots, R_{J_1}, B_1, B_2, \ldots, B_{J_2}\}$, plot a pie chart centered at corresponding s_j^* with radius proportional to $\sqrt{|D_j|}$. The pie chart contains two groups, each with ratio $\left(\frac{|D_j \cap G_1|}{|D_j|}, \frac{|D_j \cap G_2|}{|D_j|}\right)$.

9. Label the robust clusters and boundary clusters, and add an edge between a pair of robust cluster R_{j_1} and boundary cluster B_{j_2} if $\text{edge}_{\text{j}_1,\text{j}_2} \leq 4 \times \sqrt{h^2 \times d}$, where d is the number of dimensions.

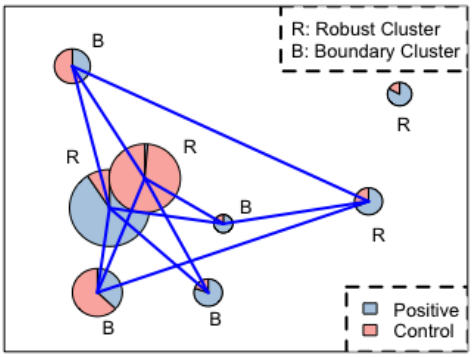


Fig 2.7: Visualization of GvHD dataset. We apply Algorithm 3 for visulization. Blue lines represent the connections among clusters. Each pie chart describes the total amount of corresponding clusters that is divided between the positive group and the control group.

2.7 Theory

In this section, we study both statistical and algorithmic convergence of our method. We start with the convergence of estimated minima $\hat{\mathcal{S}}$ to the population minima $\mathcal{S}$ along with the convergence of the gradient flow. Then we discuss the algorithmic convergence of Algorithm 1.

For a set $\mathcal{D}$, we denote its cardinality by $|\mathcal{D}|$. For a function f, we define $\|f\|_\infty = \sup_x |f(x)|$ to be the $\mathcal{L}_\infty$-norm. Let ∇f and $\nabla^2 f$ be the gradient and Hessian matrix of f, respectively. We define $\|f\|_{l,\infty}$ as the element-wise $\mathcal{L}_\infty$-norm for l-th order derivatives of f. Specifically, $\|f\|_{0,\max} = \|f\|_\infty$,

$$\|f\|_{1,\max} = \max_k \|[\nabla f(x)]_k\|_\infty \ , \ \|f\|_{2,\max} = \max_{kk'} \|[\nabla^2 f(x)]_{kk'}\|_\infty,$$

for $k = 1, 2, \ldots, d$ and $k' = 1, 2, \ldots, d$. A twice-differentiable function f is called Morse Morse [1925], Milnor et al. [1963], Banyaga and Hurtubise [2013] if all eigenvalues of the Hessian matrix of f at critical points are away from 0.

Recall that our data are random sample $X_1, \ldots, X_n$ from a PDF $p(x)$ and $s(x) = \|\nabla p(x)\|_2^2$. Additionally, $\hat{p}_n$, $\nabla \hat{p}_n$ and $\nabla^2 \hat{p}_n$ are the estimated PDF, gradient, and Hessian matrix, respectively. In our analysis, we consider the following assumptions.

Assumptions.

(P) The density function $p(x)$ is four-times bounded and continuously differentiable.

(L) $s(x)$ is a Morse function.

(K) The kernel K is four-times bounded and continuously differentiable. Moreover, the collection of kernel functions and their partial derivatives up to the third order satisfy the VC-type conditions in Giné and Guillou [2002]. See Appendix A for more details.

Assumption (P) is slightly stronger than the conventional assumptions for density estimation that we need to be four-times differentiable. This is because we are working with gradient of 'slope', which already involves second derivatives. To control the bias, we need additionally two derivatives, leading to a requirement on the fourth-order derivatives. Assumption (L) is slightly stronger than the conventional Morse function assumption on $p(x)$. We need the slope function to be Morse so that the gradient system is well-behaved. In fact, Assumption (L) implies that $p(x)$ is Morse function due to Lemma 1. Assumption (K) is a common assumption to ensure uniform convergence of a kernel-type estimator; see, for example Genovese et al. [2012, 2014].

2.7.1 Estimation Consistency

With the above assumption, we can show that the local minima of $\hat{s}_n$ converge to the local minima of s.

Theorem 2 (Consistency of local minima of s)**.** *Assume (K), (P) and (L). Let c_1 be the bound for the partial derivatives of s up to the third order and denote the l-th largest eigenvalues of $\nabla^2 s(x)$ by $\lambda_{(s,l)}(x)$ ($l = 1, 2, \ldots, d$, where d is the dimension). Assume:*

(A1) There exists $\eta_1 > 0$ such that for any point x with $\|\nabla s(x)\| \leq \eta_1$ and $0 > -\lambda_0'/2 \geq \lambda_{(s,d)}(x)$, we have $\min_{m \in \mathcal{S}} \|m - x\| \leq \frac{\lambda_0'}{2dc_1}$, where $0 < \lambda_0' \leq |\lambda_{(s,l)}(m)|$ for $l = 1, 2, \ldots, d$ and $m \in \mathcal{S}$.

When $\|\hat{p}_n - p\|_{4,\max}$ is sufficiently small, we have

- *$|\mathcal{S}| = |\hat{\mathcal{S}}|$, and*
- *for every point $m \in \mathcal{S}$, there exists a unique element $\hat{m} \in \hat{\mathcal{S}}$ such that*

$$\|\hat{m} - m\| = O(h^2) + O_P\left(\sqrt{\frac{1}{nh^{d+4}}}\right).$$

Theorem 2 shows two results. First, asymptotically, there will be a one–one corresponding relationship between a population's local minimum and an estimated local minimum. The second result shows the rate of convergence, which is the rate of estimating second derivatives. This is reasonable, since the local minima of s is defined through the gradient of $s(x) = \|\nabla p(x)\|^2$, which requires second derivatives of p.

Note that the fourth-order derivative assumption (P) can be relaxed to a smoothed third-order derivative conditions. We use this stronger condition to simplify the derivation. , since the global minima of s are the critical points of p, the consistency of estimating a global minimum only requires a third-order derivative (or a smooth second-order derivative) assumption; see, for example, Vieu [1996], Chazal et al. [2017].

Theorem 2 also implies the rate of the set estimator $\hat{\mathcal{S}}$ in terms of the Hausdorff distance.

For given two sets A, B, their Hausdorff distance is

$$\text{Haus}(A, B) = \max\left\{\sup_{x\in A} d(x, B), \sup_{x\in B} d(x, A)\right\},$$

where $d(x, A) = \inf_{y\in A} \|x - y\|$ is the projection distance from point x to the set A.

Corollary 3. *Assume (K),(P), (L), and (A1). When $\|\hat{p}_n - p\|_{4,\max}$ is sufficiently small,*

$$Haus(\hat{\mathcal{S}}, \mathcal{S}) = O(h^2) + O_P\left(\sqrt{\frac{1}{nh^{d+4}}}\right).$$

The above results describe the statistical consistency of the convergent points (local minima) of a gradient flow system. In what follows, we show that the gradient flows will also converge under the same set of assumptions.

Theorem 4 (Consistency of gradient flows)**.** *Assume (K),(P) and (L). Then for a fixed point x, when $\frac{nh^{d+8}}{\log n} \to \infty$, $h \to 0$,*

$$\sup_{t\geq 0} \|\hat{\pi}_x(t) - \pi_x(t)\| = \left\{O(h^{2\alpha}) + O_P\left(\left(\frac{\log n}{nh^{d+4}}\right)^{\frac{\alpha}{2}}\right)\right\} \wedge \left\{O(h) + O_P\left(\sqrt[4]{\frac{\log n}{nh^d}}\right)\right\},$$

where $\mu_{\min}(x)$ and $\mu_{\max}(x)$ are the minimal and maximal eigenvalues of the Hessian matrix of s evaluated at the destination $\pi_x(\infty)$, and $\alpha = \mu_{\min}(x)/\left(\mu_{\min}(x) + \mu_{\max}(x)\right)$.

Theorem 4 is mainly inspired by Theorem 2 in Arias-Castro et al. [2016]. It shows that starting at a given point x, the estimated gradient flow $\hat{\pi}_x(t)$ is a consistent estimator to the population gradient flow $\pi_x(t)$. One may notice that this result shows that the convergence rate is slowed down by the factor α, which comes from the curvature of s around the local minimum. This is due to the fact that when a flow is close to its convergent point (a local minimum), the speed of flow is decreasing until 0 (when it arrives at a minimum), so the eigenvalues determine the rate of how fast the speed of a flow decreases along a particular direction. When the eigengap (difference between $\mu_{\min}(x)$ and $\mu_{\max}(x)$) is large, even a small perturbation could change the orientation of the flow drastically, leading to a slower convergence rate.

Remark 1. *It is possible to obtain the clustering consistency in the sense that the clustering based on s and $\hat{s}_n$ are asymptotically the same Chen et al. [2017b]. In Chen et al. [2017b], the authors placed conditions on the density function and showed that the mode-clustering of $\hat{p}$ leads to a consistent partition of the data compared to the mode-clustering of p. If we generalize their conditions to the slope s, we will obtain a similar clustering consistency result.*

2.7.2 Algorithmic Consistency

In this section, we study the algorithmic convergence of Algorithm 1. For simplicity, we consider the case where the gradient descent algorithm is applied to s. The convergence analysis of gradient descent has been well studied in the literature Nesterov [2014], Ruder [2016] under convex/concave setups. Our algorithm is a gradient descent algorithm but is applied to a non-convex scenario. Fortunately, if we consider a small ball around each local minimum, the function s will still be a convex function, so the conventional techniques apply.

Specifically, we need an additional assumption that is slightly stronger than (L).

(A2) There are positive numbers $R_0, \eta_1, \lambda_0 > 0$ such that for all $x \in B(m, R_0)$, where $m \in \mathcal{S}$, and $B(m, R_0)$ is a ball with center m and radius R_0, all eigenvalues of Hessian matrix $\nabla^2 s(x)$ are above λ_0 and $\|\nabla s(x)\| \leq \eta_1$.

The assumption (A2) is a local strongly convex condition.

Theorem 5 (Convergence of Algorithm 1)**.** *Assume conditions (P), (K), (A1) and (A2). Let the step size in Algorithm 1 be γ. Recall that x_t is the point at iteration time t and x_0 is the initial point. Assume that the step size $\gamma < 1/L$, where $L = \sup_x \|\nabla s(x)\|$. For any initial point x_0 within the ball $B(m, R_0)$, there exists a constant $C_0 < 1$ such that:*

$$\|x_t - m\| \leq (1 - \gamma L)^t \|x_0 - m\|,$$
$$\|s(x_t) - s(m)\| \leq C_0^t \|s(x_0) - s(m)\|.$$

Note that λ_0 is the constant in assumption (A2) and satisfies $\lambda_0 \leq L$; see the proof of this theorem.

Theorem 5 shows that when the initial point is sufficiently close to a local minimum, the algorithm converges linearly Nesterov [2014], Ruder [2016] to the local minimum. Additionally, this implies that the ball $B(m, R_0)$ is always in the basin of attraction of m. However, note that the actual basin could be much larger than $B(m, R_0)$.

2.8 Conclusions

In this chapter, we introduced a novel clustering approach based on the gradient of the slope function. The resulting clusters are associated with an attribute label, which provides additional information on each cluster. With this new clustering method, we propose a two-sample test using local information within each cluster, which improves the testing power. Finally, we developed an informative visualization tool that gives the structure of multi-dimensional data.

We studied our improved method's performance empirically and theoretically. Simulation studies show that our refined clustering method is capable of capturing fine structures within the data. Furthermore, as a two-sample test procedure, our clustering method has better power than conventional approaches. The analysis on Astronomy and GvHD data shows that our method finds meaningful clusters. Finally, we studied both statistical and computational theory of our proposed method. Our proposed method demonstrated good empirical performance and statistical and numerical properties. Finally, we would like to note that while our method works well for the GvHD data (d = 4), it may not be applicable for any higher dimensional data, since our method is a nonparametric procedure involving derivative estimation. The curse of dimensionality prevents us from applying it to data with more dimensions.

Chapter 3

PENALIZED ESTIMATION OF THRESHOLD AUTO-REGRESSIVE MODELS WITH MANY COMPONENTS AND THRESHOLDS

3.1 Introduction

The threshold Auto-Regressive (TAR) model [Tong, 1978, Tong and Lim, 1980] allows regime-specific auto-regressive parameters, where the regimes are governed by a thresholding random variable, typically some previous lag of the time series (see formal definition in Section 3.2). Thanks to its flexibility, the TAR model has become a popular framework for analyzing non-linear time series from diverse application domains, from economics [Lee et al., 2002] and finance [Chen et al., 2011] to genomics [Jiang et al., 2014] and epidemiology [Watier and Richardson, 1995]. Applications in macroeconomics have been particularly diverse: Enders et al. [2007] modeled the U.S. GDP growth, and constructed confidence intervals for the parameters; Juvenal and Taylor [2008] explored the validity of the law of one price in nine European countries; and Aslan et al. [2018] applied a TAR model to commodity prices, and used it to represent abrupt changes, time-irreversibility, and regime-shifting behavior. See Hansen [2011] for a selective review of threshold autoregression in economics.

TAR models have been extensively studied in univariate and fixed-dimensional settings. For example, Chan [1993] investigated the asymptotic properties of the least squares estimation for TAR models with two regimes, Chen [1995] proposed an estimation procedure when the thresholding variable is unknown, Bruce [1997] derived the asymptotic distribution of general TAR models, and Li et al. [2012] developed the asymptotic theory of the least squares estimator for a moving average TAR model. In other related work, Chan and Kutoyants [2012] proved the consistency of a Bayesian estimator of the TAR model, while Chan et al. [2015] proposed a novel modified LASSO approach for threshold estimation and established its consistency in multiple threshold models. Tsay [1998b] first extended univariate TAR models to multivariate settings, and proposed to use grid search based on

the Akaike information criterion (AIC) to select the thresholds. Later, Lo and Zivot [2001], Hansen and Seo [2002], Dueker et al. [2011], Li and Tong [2016] used grid search based methods to study the multivariate TAR models assuming either a known number of thresholds or an upper bound on the number of thresholds. However, these approaches may not work in practice, as the number of thresholds is often unknown. More recently, Calderón V and Nieto [2017], Orjuela and Villanueva [2021] introduced Bayesian methodologies for the estimation of thresholds in multivariate TAR models with an unknown number of thresholds. These methods bypass the assumptions on the number of thresholds, but do not establish the consistency of the number of the estimated thresholds. Another limitation of existing approaches is that they are not applicable in high dimensions. The advantages and limitations of existing approaches are summarized in Table B.1 in Appendix B.0.3. See also Tong [2011] for a review of threshold models in time series analysis.

High-dimensional time series models have received considerable attention in recent years [Basu and Michailidis, 2015b, Lam and Yao, 2012, Han and Liu, 2013]. In this setting, the ambient dimension is of the same order or larger than the sample size. This poses numerous practical and theoretical challenges. While a number of theoretical results have been established for linear time series models in high dimensions, with few exceptions [e.g., Chen et al., 2017a, Tank et al., 2017], their non-linear counterparts have received less attention. In the context of threshold models, the recent work by Liu and Chen [2020] investigates the estimation of threshold factor models with growing number of variables. However, this work assumes a single threshold, which limits the flexibility of the model. Moreover, while the number of time series components is allowed to grow, it is assumed to be smaller than the sample size (see Theorem 1 in Liu and Chen [2020]). In fact, to the best of our knowledge, methods and theory for high-dimensional TAR models are currently lacking.

Given the paucity of the literature on high-dimensional TAR models, in this chapter, we propose two estimators for detecting the (unknown) number and values of thresholds and estimating regime-specific auto-regressive parameters in multivariate TAR models with many components. Both approaches are based on a three-step estimation framework and utilize similar penalized estimation strategies, but they differ in one key aspect. The first approach is a natural extension of the classical TAR model and enforces all auto-regressive

parameters to change at the same thresholds. As we discuss in Section 3.3, this assumption may be too restrictive in high-dimensional settings with many components. In fact, our theoretical and empirical investigations indicate that the extension of the classical TAR is not appropriate for high-dimensional settings and is better suited for moderate dimensions. As such, we refer to this first version as the multivariate TAR (mvTAR) model. To mitigate the limitation of the mvTAR model, we then propose a more flexible high-dimensional TAR model (hdTAR) where different auto-regressive parameters are allowed to change at different thresholds. This flexibility seems to introduce a new challenge, as the model may have many thresholds. However, our theoretical and empirical investigations show that this flexibility is indeed necessary in high dimensions and leads to improved theoretical guarantees and empirical performances. We develop efficient algorithms for both methods and establish the consistency of the thresholds and auto-regressive parameters under certain mixing conditions.

To establish our theoretical results, we address two key challenges that arise in penalized estimation of high-dimensional TAR models. The first challenge involves verifying appropriate concentration inequalities, including two main ingredients in high-dimensional statistics: (1) a restricted eigenvalue condition and (2) a deviation bound condition [Loh and Wainwright, 2011]. These conditions are crucial in deriving consistency results in high-dimensional settings, as hinted in Bickel et al. [2009]. The conditions have been previously verified in the setting of i.i.d. observations and, more recently, studied in certain linear time series models [Basu and Michailidis, 2015b, Safikhani and Shojaie, 2020]. However, extending these results to non-linear TAR models is challenging. This is primarily due to the random ordering of the design matrix based on the threshold (switching) variable (see e.g. Equation (3.3)). To address this challenge, we develop a bracketing argument [Van der Vaart, 2000, Chan et al., 2015] specifically designed to handle the threshold-type structure (see Lemmas 16 and 18 in the Appendix). These results are verified under certain mixing conditions (see Assumption B2 in Section 3.4) and are of independent interest in the context of non-linear high-dimensional time series models. The second challenge concerns our screening step to consistently estimate the number of thresholds. Many theoretical results in the context of TAR models assume that the number of thresholds is known [Lo and Zivot,

2001, Liu and Chen, 2020]. This assumption may not be realistic in practice; in fact, it is appealing to infer the number of thresholds from data. To that end, the second step of our proposed algorithms utilizes an information criterion that screens candidate thresholds identified in the first step and removes redundant ones. This step successfully resolves the challenge by consistently estimating the number of thresholds with high probability (see Theorem 7).

The rest of the chapter is organized as follows. After formally defining the multivariate TAR model in Section 3.2, we describe our algorithms in Section 3.3 and establish their theoretical properties in Section 3.4. In Section 3.5, we propose data-driven methods to select the hyper-parameters. While the required hyper-parameters are characterized in our asymptotic results, these rates involve unknown constants and cannot be used in practice. The empirical performance of the proposed methods is investigated using both simulated and real data sets, in Section 3.6 and Section 3.7, respectively. We conclude with a brief summary in Section 3.8.

3.2 Multivariate TAR Formulations

The classical TAR model, proposed by Tong and Lim [1980], is defined as

$$x_t = a^0_{(j)} + \sum_{k=1}^{K} a^k_{(j)} x_{t-k} + \sigma_{(j)}\epsilon_t, \quad \text{if } r_{j-1} < z_t \leq r_j, \tag{3.1}$$

where m_0 denotes the number of thresholds, r_js are the threshold parameters which partition the time series into m_0+1 regimes, K is the number of lags to be considered in the model, z_t is a switching variable (maybe functions of some components of x_t), σ_js are segment-specific error variances, and $a^0_{(j)}$ and $a^k_{(j)}$ are coefficients in regime j, for $j = 1, \ldots, m_0+1$ (they are allowed to be different in each regime). The noise or innovation, ϵ_t, is an i.i.d. sequence of random variables with zero mean and unit variance.

The original TAR model was restricted to univariate time series, but can be extended to multivariate settings, as described in Tsay [1998a]. Formally, a multivariate time series $\{\boldsymbol{x}_t\}$ follows TAR model with one switching variable if

$$\boldsymbol{x}_t = \sum_{k=1}^{K} \boldsymbol{A}^{(k,j)} \boldsymbol{x}_{t-k} + \boldsymbol{\Sigma}_j^{1/2} \boldsymbol{\epsilon}_t, \quad \text{if } r_{j-1} < z_t \leq r_j, \tag{3.2}$$

where $\boldsymbol{x}_t = (x_{(t,1)}, x_{(t,2)}, \ldots, x_{(t,p)})'$ is the observed process in $\mathbb{R}^p$ at time t, p is the number of time series components, and K is the number of lags considered in the model. Here $\boldsymbol{\epsilon}_t = (\epsilon_{(t,1)}, \epsilon_{(t,2)}, \ldots, \epsilon_{(t,p)})' \in \mathbb{R}^p$ is a multivariate i.i.d. sequence with zero mean in all components. The covariance matrix $\boldsymbol{\Sigma}_j$ for the j-th regime, $\boldsymbol{\Sigma}_j$, is allowed to be different in each regime. To simplify the notations, when there is no ambiguity, we simply denote the error term by $\boldsymbol{\epsilon}_t$ instead of $\boldsymbol{\Sigma}_j^{1/2}\boldsymbol{\epsilon}_t$. The transition matrices $\boldsymbol{A}^{(k,j)} \in \mathbb{R}^{p\times p}$ is the coefficient matrix corresponding to the k-th lag of a TAR process in regime j. More specifically, similar to the modeling framework of Chan et al. [2015], we assume there exist m_0 threshold values $-\infty < r_1 < r_2 < \ldots < r_{m_0} < +\infty$ with $r_0 = -\infty$ and $r_{m_0+1} = +\infty$ which partition the process into $m_0 + 1$ regimes. For each regime, the total transition matrices $\boldsymbol{A}^{(\cdot,j)} = \left(\boldsymbol{A}^{(1,j)}, \boldsymbol{A}^{(2,j)}, \ldots, \boldsymbol{A}^{(K,j)}\right) \in \mathbb{R}^{p\times pK}$ are fixed where $r_{j-1} < z_t \leq r_j$ for $j = 1, \ldots, m_0 + 1$.

Our goal is to estimate the number of thresholds, i.e. m_0, together with the threshold values, r_j, and the auto-regressive parameters in each regime.

Next, we introduce some additional notations. For a symmetric matrix $\mathbf{X}$, let $\lambda_{\min}(\mathbf{X})$ and $\lambda_{\max}(\mathbf{X})$ denote its minimum and maximum eigenvalues. Let the h-th row of $\boldsymbol{A}^{(\cdot,j)}$ be $\boldsymbol{A}_h^{(\cdot,j)}$, and set the number of non-zero elements in $\boldsymbol{A}_h^{(\cdot,j)}$ to $d_{h,j}$ for $h = 1, 2, \ldots, p$ and $j = 1, 2, \ldots, m_0 + 1$. Denote the total sparsity of the model by $d_n^* = \sum_{j=1}^{m_0+1} \sum_{h=1}^{p} d_{h,j}$. Further, let $\mathcal{I}_{h,j}$ represent the set of all column indexes of $\boldsymbol{A}_h^{(\cdot,j)}$, $\mathcal{I} = \cup_{h,j}\mathcal{I}_{h,j}$ and define $d_n = \max_{1\leq h\leq p, 1\leq j\leq 1+m_0} |\mathcal{I}_{h,j}|$. Note that p, m_0 and the sparsity may increase with the number of time points, T, specifically, $p \equiv p(n)$ and $m_0 \equiv m_0(n)$ and $d_{h,j} \equiv d_{h,j}(n)$, where $n = T - K$. For simplicity, we suppress the n-index. Finally, let $\epsilon_{t,l}$ be error term of l-th time series, and recall that $\boldsymbol{\epsilon}_t = \left(\epsilon_{(t,1)}, \epsilon_{(t,2)}, \ldots, \epsilon_{(t,p)}\right)'$. Throughout the chapter, positive constants $C, C_1, C_2, \ldots$ are used to denote universal constant, $\boldsymbol{A}'$ denotes the transpose of a matrix $\boldsymbol{A}$, and $\|\boldsymbol{A}\|_1$ and $\|\boldsymbol{A}\|_2$ denotes its ℓ_1 and Frobenius norms, respectively. We denote the ℓ_1 and ℓ_2 norms of a vector v by $\|v\|_1$ and $\|v\|$, respectively.

3.3 Regularized Estimation of High-Dimensional TARs

The number of parameters in the TAR model Equation (3.2), $(m_0+1)(Kp^2)$, increases with the number of time series p and the number of thresholds m_0. Estimating these parameters becomes especially challenging when the model has more than one threshold, i.e. $m_0 > 1$, and the number of thresholds is unknown. This is because identifying the thresholds would require a search over all possible values of threshold levels z_t, which is infeasible.

To overcome the above challenges, in Section 3.3.1 we first reformulate the TAR estimation problem via a non-parametric model with $(T-K)p^2K$ parameters. This over-parameterization allows us to use regularized estimation strategies to efficiently obtain an initial estimate of the thresholds by solving a penalized least squares estimation problem. In particular, we use a total variation penalty [Tibshirani et al., 2005] to obtain piecewise constant estimates of $\boldsymbol{A}^{(k,j)}$ for regime j with respect to the threshold variable z_t.

The classical multivariate TAR model Equation (3.2) requires the parameters of transition matrices $\boldsymbol{A}^{(k,j)}$ to change at the same threshold values z_t. To obtain such an estimate, we consider a grouped fused lasso penalty in Section 3.3.2. The resulting estimate, referred to as mvTAR, is suitable for low-to-moderate-dimensional problems, where p is fixed or small compared to the number of observations T. However, for problems with large p, especially when $p \gg T$, requiring that all transition matrix parameters change at the same threshold value becomes restrictive. Moreover, the theoretical advantages of the group lasso penalty dissipate when grouped parameters do not follow the same sparsity pattern [Huang and Zhang, 2010]. These limitations are reflected in our theoretical and numerical analyses in Sections 3.4 and 3.6. To achieve efficient estimation in high-dimensions, in Section 3.3.2 we propose a more flexible high-dimensional TAR model, named hdTAR, in which transition matrix parameters are allowed to change at different thresholds. As we show, this flexibility results in theoretical and empirical advantages. The difference between the flexible TAR model and the original version is illustrated in Figure 3.1.

Both our group and regular fused lasso penalties overestimate the number of thresholds. This is because a key requirement for consistency of ℓ_1-regularized estimation strategies, namely the restricted eigenvalue property [Bickel et al., 2009] is not guaranteed to hold in

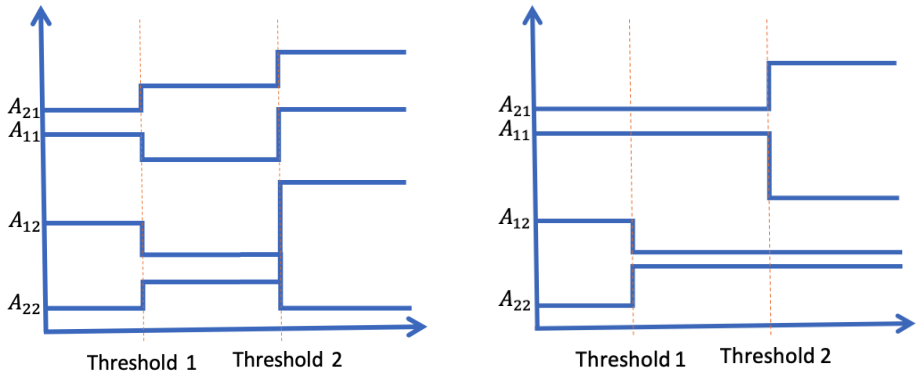


Fig 3.1: Example of changes of transition matrices. The left panel depicts the situation in which the classical TAR multivariate TAR model (mvTAR) in which all elements of the transition matrices change together at all threshold values. The right panel illustrates the proposed flexible TAR model for high dimensions (hdTAR) in which different elements of the transition matrices would not change at some threshold values.

our setting (see Section 3.4). To remove the redundant selected thresholds, we introduce a screening criterion in Section 3.3.3 that consistently estimates the (many) unknown thresholds. In Section 3.3.4, we obtain consistent estimates of high-dimensional auto-regressive parameters within each estimated regime.

3.3.1 Reparametrization of the TAR Model

In this section, we reparametrize the TAR model Equation (3.2) by considering n transition matrices for each value of the ordered switching variable z_t (assuming, without loss of generality, that z_t assumes unique values).

Let $n = T - K$ and let $\pi(i)$ be the time index of the i-th smallest element of z_t for

$i = 1, \ldots, n$. Then the TAR model Equation (3.2) with lag K can be written as

$$\begin{pmatrix} \boldsymbol{x}'_{\pi(1)} \\ \boldsymbol{x}'_{\pi(2)} \\ \vdots \\ \boldsymbol{x}'_{\pi(n)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}'_{\pi(1)-1} & \cdots & \boldsymbol{x}'_{\pi(1)-K} & & \mathbf{0} & & \cdots \\ \boldsymbol{x}'_{\pi(2)-1} & \cdots & \boldsymbol{x}'_{\pi(2)-K} & \boldsymbol{x}'_{\pi(2)-1} & \cdots & \boldsymbol{x}'_{\pi(2)-K} & \cdots \\ & \vdots & & & & & \ddots \\ \boldsymbol{x}'_{\pi(n)-1} & \cdots & \boldsymbol{x}'_{\pi(n)-K} & \boldsymbol{x}'_{\pi(n)-1} & \cdots & \boldsymbol{x}'_{\pi(n)-K} & \cdots \end{pmatrix} \times \begin{pmatrix} \boldsymbol{\theta}'_1 \\ \boldsymbol{\theta}'_2 \\ \vdots \\ \boldsymbol{\theta}'_n \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon}'_{\pi(1)} \\ \boldsymbol{\epsilon}'_{\pi(2)} \\ \vdots \\ \boldsymbol{\epsilon}'_{\pi(n)} \end{pmatrix}. \tag{3.3}$$

$$\begin{pmatrix} \mathbf{0} & \\ \mathbf{0} & \\ & \\ \boldsymbol{x}'_{\pi(n)-1} \cdots & \boldsymbol{x}'_{\pi(n)-K} \end{pmatrix}$$

Let $\boldsymbol{\theta}_1 = \left(\boldsymbol{A}^1_{\pi(1)}, \ldots, \boldsymbol{A}^K_{\pi(1)}\right) \in \mathbb{R}^{p \times pK}$, and $\boldsymbol{\theta}_i = \left(\boldsymbol{A}^1_{\pi(i+1)} - \boldsymbol{A}^1_{\pi(i)}, \ldots, \ \boldsymbol{A}^K_{\pi(i+1)} - \boldsymbol{A}^K_{\pi(i)}\right)$, where $\boldsymbol{A}^k_{\pi(i)}$ is the transition matrix for i-th ordered observation at lag k. Denote the response matrix, the design matrix, the model parameters and the error term in Equation (3.3) by $\mathcal{Y}, \mathcal{X}, \Theta$ and E, respectively. Then, Equation (3.3) can be written as $\mathcal{Y} = \mathcal{X}\Theta + E$. Moreover, letting $\boldsymbol{Y} = \text{vec}(\mathcal{Y}) \in \mathbb{R}^{np \times 1}$, $\boldsymbol{\Theta} = \text{vec}(\Theta) \in \mathbb{R}^{np^2K \times 1}$, $\boldsymbol{E} = vec(E) \in \mathbb{R}^{np \times 1}$, and $\boldsymbol{Z} = I_p \otimes \mathcal{X} \in \mathbb{R}^{np \times np^2K}$ with $\otimes$ denoting the tensor product, Equation (3.3) can be written in vector form as

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\Theta} + \boldsymbol{E}. \tag{3.4}$$

While redundant, the over-parametrization in Equation (3.3) has an important benefit: $\boldsymbol{\theta}'_i \neq 0$ if and only if the auto-regressive coefficients change in TAR process at time $\pi(i)$. Thus, finding the thresholds is equivalent to finding non-zero $\boldsymbol{\theta}_i$s for $i > 1$. In other words, the problem of threshold estimation can be translated to a high-dimensional variable selection problem in Equation (3.4).

3.3.2 Penalties for Moderate- and High-Dimensional TARs

Sparsity-inducing penalties, such as lasso, are particularly suitable for estimating $\boldsymbol{\Theta}$ in Equation (3.4): A sparse estimate $\hat{\boldsymbol{\theta}}_1$ gives an interpretable estimate of the transition matrices for the smallest value of z_t, while sparsity in $\hat{\boldsymbol{\theta}}_i$ for $i > 1$ would imply no changes

in the transition matrices over z_t. Such a strategy corresponds to a fused lasso, or total variation, penalty [Tibshirani et al., 2005, Rinaldo, 2009]. In this chapter, we consider a similar strategy and obtain an estimate of $\boldsymbol{\Theta}$ by solving

$$\hat{\boldsymbol{\Theta}} = \arg\min_{\boldsymbol{\Theta}} \|\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\Theta}\|_2^2 + \lambda_1 \|\boldsymbol{\Theta}\|_{\diamond} + \lambda_2 \sum_{i=1}^{n} \left\| \sum_{i'=1}^{i} \boldsymbol{\theta}_{i'} \right\|_1 , \tag{3.5}$$

The first penalty in Equation (3.5), $\|\cdot\|_{\diamond}$, encodes either an ℓ_2, or grouped fused lasso penalty, $\|\cdot\|_2$, or an ℓ_1, or fused lasso penalty, $\|\cdot\|_1$. The group fused lasso penalty encourages all entries of the transition matrices to change at the same threshold values. In contrast, the fused lasso penalty provides a more flexible TAR model in which different transition matrix parameters are allowed to change at different thresholds. As discussed earlier, the group fused lasso penalty is only suitable for low to moderate-dimensional problems (where p is allowed to grow, but $p < T$), whereas the more flexible fused lasso penalty is appropriate for both moderate- and high-dimensional problems (where $p \gg T$); see also Figure 3.1. In both cases, the magnitude of the penalty is controlled by the tuning parameter λ_1, which is chosen data-adaptively via cross validation; see Section 3.5 for more details.

The second penalty in Equation (3.5), controlled by tuning parameter λ_2, further encourages the overall sparsity of the estimated transition matrices by penalizing changes in transition matrices after each potential threshold index i. While often not needed in practice, this additional sparsity results in improved estimation and allows us to obtain better rates of convergence for the proposed estimator in Section 3.4.

With either ℓ_2 or ℓ_1 penalties, the optimization problem in Equation (3.5) is convex and can be solved efficiently. With the ℓ_2 penalty, the problem can be solved using a sub-gradient descent algorithm. However, the problem further simplifies when $\lambda_2 = 0$ and we can instead use a more efficient proximal gradient descent algorithm; see Algorithm 6 in the Appendix. With the ℓ_1 penalty, the problem is easy to solve efficiently using a path-wise coordinate descent algorithm [Friedman et al., 2007] regardless of the value of λ_2. This is because, by Proposition 1 in Friedman et al. [2007], it suffices to first find the solution for $\lambda_2 = 0$, and then apply an element-wise soft thresholding operator; see Algorithm 5 in the Appendix.

3.3.3 Threshold Selection

Using Equation (3.5), we can define a set of candidates threshold estimates as

$$\hat{\mathcal{A}}_n = \left\{ z_{\pi(i-1)} : \|\hat{\boldsymbol{\theta}}_i\|_2 \neq 0 \, , i \geq 2 \right\}. \tag{3.6}$$

Let $\hat{r}_j$ be the j-th sorted (from the lowest to the highest) estimated threshold in the set $\hat{\mathcal{A}}_n$, and let $\hat{m}$ be the cardinality of the set $\hat{\mathcal{A}}_n$. As we show in Section 3.4, it is likely for the fused lasso to over-estimate the number of thresholds [Harchaoui and Lévy-Leduc, 2010]. Thus, we need to remove the redundant thresholds. In our screening step, we aim to keep exactly m_0 points in $\hat{\mathcal{A}}_n$ which are close enough to the true threshold values. To that end, we develop an *information criterion* by modifying the screening procedure of Safikhani and Shojaie [2020] to make it more suitable for the threshold structure of model Equation (3.2). Essentially, this step consists of estimating the transition parameters within each estimated regime $\{t : \hat{r}_j < z_t \leq \hat{r}_{j+1}\}$ for $j = 0, 1, \ldots, \hat{m}$ with $\hat{r}_0 = -\infty$ and $\hat{r}_{\hat{m}+1} = +\infty$ and comparing the total sum of squared error (SSE) before and after excluding a certain estimated threshold $\hat{r}_j$. The basic idea is to keep the estimated thresholds for which the value of SSE increases significantly if we remove them. More specifically, for a given set of estimated thresholds $\{-\infty, s_1, s_2, \ldots, s_m, +\infty\}$ with $1 \leq m \leq \hat{m}$, and for j-th estimated threshold s_j, denote by $\mathcal{T}_{(s_{j-1}, s_j)} = \left\{ i : s_{j-1} < z_{\pi(i)} \leq s_j \right\}$ the set of orders of z_ts for which their corresponding ordered switching variable $z_{\pi(i)}$s fall into the interval $[s_{j-1}, s_j]$. Now, given a fixed number of thresholds m, we obtain the estimator $\hat{\boldsymbol{\theta}}_{s_1, s_2, \ldots, s_m}$ of $\boldsymbol{\theta}_{s_1, s_2, \ldots, s_m}$ by minimizing the following penalized regression problem

$$\sum_{j=1}^{m+1} \frac{1}{\left| \mathcal{T}_{(s_{j-1}, s_j)} \right|} \sum_{i \in \mathcal{T}_{(s_{j-1}, s_j)}} \left\| \boldsymbol{x}_{\pi(i)} - \boldsymbol{\theta}_{(s_{j-1}, s_j)} \boldsymbol{Y}_{\pi(i)} \right\|_2^2 + \eta_{(s_{j-1}, s_j)} \|\boldsymbol{\theta}_{(s_{j-1}, s_j)}\|_1, \tag{3.7}$$

where $\boldsymbol{Y}_{\pi(i)} = \begin{pmatrix} \boldsymbol{x}'_{\pi(i)-1} & \ldots & \boldsymbol{x}'_{\pi(i)-K} \end{pmatrix}'$, $\boldsymbol{\theta}_{s_1, s_2, \ldots, s_m} = \begin{pmatrix} \boldsymbol{\theta}'_{(s_0, s_1)}, \boldsymbol{\theta}'_{(s_1, s_2)}, \ldots, & \boldsymbol{\theta}'_{(s_{m-1}, s_m)} \end{pmatrix}$, and tuning parameters $\boldsymbol{\eta}_n = \left(\eta_{(-\infty, s_1)}, \ldots, \eta_{(s_m, +\infty)} \right)$. The `glmnet` package [Friedman et al., 2010] readily solves the problem.

Denoting

$$L_n(s_1, s_2, \ldots, s_m\,; \boldsymbol{\eta}_n) = \sum_{j=1}^{m+1} \sum_{i \in \mathcal{T}_{(s_{j-1}, s_j)}} \left\| \boldsymbol{x}_{\pi(i)} - \boldsymbol{\theta}_{(s_{j-1}, s_j)} \boldsymbol{Y}_{\pi(i)} \right\|_2^2 + \sum_{j=1}^{m+1} \eta_{(s_{j-1}, s_j)} \left\| \boldsymbol{\theta}_{(s_{j-1}, s_j)} \right\|_1, \tag{3.8}$$

we construct our information criterion as

$$\mathrm{IC}\,(s_1, s_2, \ldots, s_m\,; \boldsymbol{\eta}_n) = L_n(s_1, s_2, \ldots, s_m\,; \boldsymbol{\eta}_n) + m\omega_n, \tag{3.9}$$

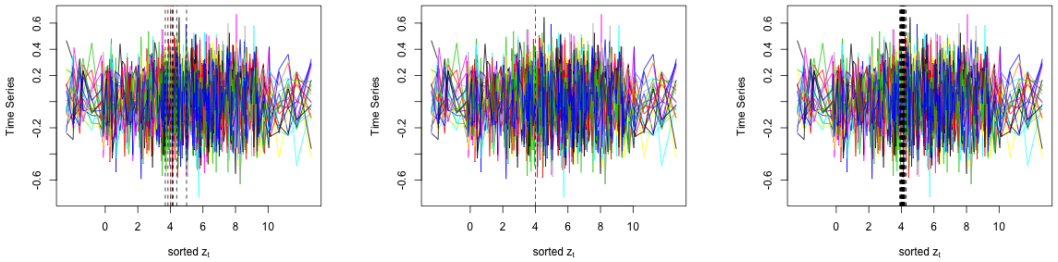
where ω_n is a carefully chosen sequence defined in Section 3.4. We then select a subset of the initial $\hat{m}$ candidate threshold values by solving

$$(\tilde{m}, \tilde{r}_1, \tilde{r}_1, \ldots, \tilde{r}_{\tilde{m}}) = \mathrm{argmin}_{0 \le m \le \hat{m}, \boldsymbol{s}=(s_1, s_2, \ldots, s_m) \in \hat{\mathcal{A}}_n} \mathrm{IC}\,(\boldsymbol{s}; \boldsymbol{\eta}_n)\,. \tag{3.10}$$

Practical choices for tuning parameters $\boldsymbol{\eta}_n$ and ω_n are discussed in Section 3.5.

The over-estimation of the thresholds and the effect of the screening step are illustrated in Figure 3.2. The left panel of Figure 3.2 — which is obtained for one replicate of simulation Scenario 1 in Section 3.6 — clearly shows that the first step of our procedure detects more threshold values. The middle panel shows that second step successfully screens out the extra threshold estimates and keep a single value which is very close to the true threshold (here, the true threshold value is 4). The right panel of Figure 3.2 confirms that the final estimated thresholds across all 200 replicates are indeed close to the true thresholds.

When the number of estimated thresholds selected in Step 1 is large, it might be computationally demanding to find the minimizer of the IC. In such cases, we propose to approximate the optimal thresholds using the backward elimination algorithm (BEA) proposed in Safikhani and Shojaie [2020]. Starting with the set of initial thresholds $\hat{\mathcal{A}}_n$, the algorithm reduces the computational cost by removing one threshold at a time until IC does not reduce any further.



(a) First Step. (b) Second Step. (c) Second Step for 200 replicates.

Fig 3.2: Estimated thresholds in Simulation Scenario 1 with hdTAR. On average around 8 points are selected in the first step, and Figure 3.2a shows the result of one single run in first step. Figure 3.2b shows the results of final selected threshold estimates for single simulation in Figure 3.2a, and Figure 3.2c shows the final selected threshold estimates all 200 simulation runs.

3.3.4 Estimation of Auto-Regressive Parameters

Given the estimated thresholds, we simply take each estimated regime $\mathcal{T}_{(\tilde{r}_{j-1},\tilde{r}_j)} = \left\{i : \tilde{r}_{j-1} < z_{\pi(i)} \leq \tilde{r}_j\right\}$ with $\tilde{r}_0 = -\infty$ and $\tilde{r}_{\tilde{m}+1} = -\infty$ for $j = 1, \ldots, \tilde{m}+1$, and estimate the transition matrices in each regime separately. More specifically, for a fixed $j = 1, \ldots, \tilde{m}+1$ we solve

$$\hat{\boldsymbol{\beta}}^{(\cdot,j)} = \arg\min_{\boldsymbol{\beta}} \left(\sum_{i \in \mathcal{T}_{(\tilde{r}_{j-1},\tilde{r}_j)}} \left\| \boldsymbol{x}_{\pi(i)} - \boldsymbol{\beta} \boldsymbol{Y}_{\pi(i)} \right\|_2^2 + \alpha_j \left\| \boldsymbol{\beta} \right\|_1 \right), \tag{3.11}$$

where α_j is the tuning parameter for the j-th regime for $j = 1, 2, \ldots, \tilde{m}$. It can be solved efficiently using existing software and HBIC can be used to select α_j. As an alternative to the separate estimation in Equation (3.11), if the distances between consecutive threshold values are of the same order, the auto-regressive parameters can also be jointly estimated [Saegusa and Shojaie, 2016].

3.4 Theoretical Properties

In this section, we establish the consistency of our procedure proposed in Section 3.3.2. Recall that in the first step of our procedure we use either the ℓ_2 or the ℓ_1 penalty in Equation (3.5), corresponding to classical (mvTAR) and flexible (hdTAR) TAR models. More specifically, in the following, $\hat{\boldsymbol{\Theta}}$ is the estimator defined in Equation (3.5) with either the ℓ_1 penalty or the ℓ_2 penalty, $\hat{\boldsymbol{\theta}}_{s_1,s_2,\ldots,s_m}$ is the estimator defined in Equation (3.7), and finally, $\hat{\boldsymbol{\beta}}^{(.,j)}$ is the estimator defined in Equation (3.11). We make the following assumptions.

Assumption B1. *$\{\boldsymbol{\epsilon}_t\}$ is a sequence of i.i.d. sub-Weibull random variables with bounded continuous and positive density and sub-Weibull constant K_ϵ and sub-Weibull parameter $\varkappa_c > 0$; specifically, there exist constants K_ϵ and $\varkappa_c > 0$ such that $\|\boldsymbol{\epsilon}_t\|_\psi \leq K_\epsilon$ where $\|\boldsymbol{\epsilon}_t\|_\psi := \sup_{c \geq 1 \wedge \varkappa_c} c^{-\frac{1}{\varkappa_c}} \left(\mathbb{E}\left|\boldsymbol{\epsilon}_t\right|^c\right)^{1/c}$.*

Assumption B2. *For each $j = 1, 2, \ldots, m_0 + 1$, the process*

$$\boldsymbol{x}_t = \sum_{k=1}^{K} \boldsymbol{A}^{(k,j)} \boldsymbol{x}_{t-k} + \boldsymbol{\epsilon}_t$$

is sub-Weibull with sub-Weibull parameter $\varkappa_1 > 0$ and β-mixing stationary with a geometrically decaying mixing coefficient b_n; specifically, there exist constants $c_b > 0$ and $\varkappa_2 > 0$ such that for all $n \in \mathbb{N}$, $b(n) \leq \exp(-c_b n^{\varkappa_2})$ and for all $t, \tau > 0$, $(\boldsymbol{x}_t, \ldots, \boldsymbol{x}_{t+n}) \overset{d}{=} (\boldsymbol{x}_{t+\tau}, \ldots \boldsymbol{x}_{t+\tau+n})$, where $\overset{d}{=}$ denotes equality in distribution. Moreover, $\mathbb{E}[\boldsymbol{x}_t] = \boldsymbol{0}_{p\times 1}$. In addition, assume $2/3 \leq \varkappa_0 < 1$, where $\varkappa_0 := \left(\frac{2}{\varkappa_1} + \frac{1}{\varkappa_2}\right)^{-1}$.

Assumption B3. *The matrices $\boldsymbol{A}^{(.,j)}$ are sparse for $j = 1, \ldots, m_0 + 1$. More specifically, for all $h = 1, 2, \ldots p$ and $j = 1, 2, \ldots m_0 + 1, d_{hj} \ll p$, i.e., $d_{kj}/p = o(1)$. Moreover, there exists a positive constant $M_A > 0$ such that*

$$\max_{1 \leq j \leq m_0+1} \left\| \boldsymbol{A}^{(.,j)} \right\|_\infty \leq M_A.$$

Assumption B4. *There exists a positive constant ν such that*

$$\min_{1 \leq j \leq m_0} \left\| \boldsymbol{A}^{(.,j+1)} - \boldsymbol{A}^{(.,j)} \right\|_2 \geq \nu > 0.$$

Moreover, there exist constants l and u such that $r_j \in [l, u]$ for $1 \leq j \leq m_0$. In addition, there exists a vanishing positive sequence γ_n such that as $n \to \infty$, $\min_{1 \leq j \leq m_0+1} |r_j - r_{j-1}| / \gamma_n \to +\infty$. For hdTAR, we assume $d_n^* \frac{\log(p^2 K)}{\sqrt{n\gamma_n}} \to 0$, whereas for mvTAR we assume $\sqrt{p^2 K} d_n^* \frac{\log(p^2 K)}{\sqrt{n\gamma_n}} \to 0$.

Assumption B5. *$\{z_t\}$ is a β-mixing stationary process with a geometric decaying mixing coefficient and positive density. In addition, $\mathbb{E}|z_t|^{2+\iota} < \infty$ for $\iota > 0$.*

The above assumptions are natural in high-dimensional settings and commonly used in the literature. Assumptions B1 and B2 are utilized to derive appropriate concentration inequalities needed to verify the asymptotic properties of the proposed methodology and have been used in the literature [Li et al., 2012, Wong et al., 2020]. The sub-Weibull distribution of error terms controls the tail effects, while the β-mixing condition ensures the dependence structure can be controlled appropriately. The latter is specifically needed due to the temporal correlation among observations. We can relax the β-mixing assumption to α-mixing if we restrict to Gaussian distributions, rather than sub-Weibull processes. However, to keep the distributional assumption more general, we consider here the β-mixing assumption. In Appendix 4, we also develop a sufficient condition for β-mixing processes by imposing constrains on the operator norm of transition matrices; this implies that the β-mixing condition is less restrictive. The assumption $\varkappa_0 \geq 2/3$ is to ensure a sharp consistency rate for estimating the thresholds and can be removed at the cost of worsening the consistency rate (see additional details in Remark 2). Assumption B3 ensures the sparsity of the model and is needed to quantify the effect of model misspecification, since exact recovery of threshold values is not possible. A similar assumption has been used in Safikhani and Shojaie [2020] in the context of change point detection. Further, Assumption B4 puts a minimum jump size on the transition matrices ensuring a detectable change occurred at threshold r_j; it also puts certain conditions on the detection rate, which is related to γ_n. Assumption B4 can be seen as an extension of Assumption H4 in Chan et al. [2015] to high-dimensions. It can be seen that the assumption is more stringent for mvTAR, rendering this procedure not suitable for high dimensions. Finally, Assumption B5 is used to build the relationship between the length of each regime and the number of observations in that regime.

Our first theoretical result concerns the first step, i.e., the initial estimation of thresholds using group or regular fused lasso penalties. The penalized estimation Equation (3.5) in this step does not guarantee parameter estimation consistency since the design matrix $\boldsymbol{Z}$ in Equation (3.4) may not satisfy the restricted eigenvalue condition [Basu and Michailidis, 2015b], which is critical for establishing the parameter estimation consistency in high-dimensions [Bickel et al., 2009]. However, with either penalty, the estimator over-estimates the true number of thresholds, as established next.

Let $\mathcal{A}_n = \{r_1, r_2, \ldots, r_{m_0}\}$ be the set of the sorted true thresholds. Define the Hausdorff distance between two countable sets as:

$$d_H(A, B) = \max_{b \in B} \min_{a \in A} |b - a|.$$

Though not a distance, $d_H(A, B)$ proves useful in Theorem 6.

Theorem 6. *Under assumptions B1 to B5, there exist large constants $C_1, C_2 > 0$ such that $\tilde{\lambda}_{1,n} = C_1 \frac{\log(p^2 K)}{\sqrt{n}}$, and $\tilde{\lambda}_{2,n} = \frac{C_2}{n} \frac{\log(p^2 K)}{\sqrt{n\gamma_n}}$, where for hdTAR $\lambda_{1,n} = \tilde{\lambda}_{1,n}$ and $\lambda_{2,n} = \tilde{\lambda}_{2,n}$, whereas for mvTAR, $\lambda_{1,n} = \sqrt{p^2 K}\tilde{\lambda}_{1,n}$ and $\lambda_{2,n} = \sqrt{p^2 K}\tilde{\lambda}_{2,n}$. Then,*

$$\min \left\{ \mathbb{P}\left(|\hat{\mathcal{A}}_n| \geq m_0 \right), \mathbb{P}\left(d_H\left(\mathcal{A}_n, \hat{\mathcal{A}}_n \right) \leq \gamma_n \right) \right\} \to 1.$$

Theorem 6 shows that the number of estimated thresholds $\hat{m}$ in Step 1 is no less than the true number of thresholds m_0 with high probability. In addition, there exists at least one estimated threshold in the γ_n-radius neighborhood of the true thresholds. The rate of consistency for threshold detection, γ_n, depends on the number of time series p, the maximum considered lag K, and the minimum distance between consecutive true thresholds in the model. In addition, the convergence rate for using $\hat{r}_j$ to estimate r_j could be as low as $\log \log n \left(\log\left(p^2 K\right)\right)^2 / n$ when m_0 is finite.

The rate of consistency for thresholds detection, γ_n, for mvTAR also depends on the number of time series p, the maximum considered lag K, and the minimum distance between consecutive true thresholds in the model. However, the assumptions on γ_n for hdTAR and mvTAR are different, so the consistency rate for thresholds detection is different for these

two methods. In addition, when using the ℓ_2 penalty, the convergence rate for using $\hat{r}_j$ to estimate r_j could be as low as $\log\log n \left(\log\left(p^2K\right)\right)^2 p^2K/n$ when m_0 is finite. Thus, convergence of mvTAR is only guaranteed in low to moderate dimensions and not in high dimensions. Finally, the minimum sample size requirement depends on the sub-Weibull parameter $\varkappa_1$ and β-mixing parameter $\varkappa_2$. For example, as mentioned in Lemma 16, we need $n \geq c_0 \left(\log(p^2K)\right)^{2/\varkappa_0 - 1}$ where $\varkappa_0 := \left(\frac{2}{\varkappa_1} + \frac{1}{\varkappa_2}\right)^{-1}$. This indicates that if the sub-Weibull parameter $\varkappa_1$ increases (i.e., the tail probability decays faster), the minimum sample size will decrease; similarly, the minimum sample size decreases as the β-mixing parameter $\varkappa_2$ increases.

Next, we state Theorem 7 which shows the screening procedure Equation (3.10) consistently estimates the number and values of thresholds. For that, we need two additional assumptions.

Assumption B6. *Let* $\Delta_n = \min_{1 \leq j \leq m_0+1} |r_j - r_{j-1}|$. *Then,*

$$m_0 \left(n\gamma_n\right)^{3/2} d_n^{*2}/\omega_n \to 0, \text{ and } n\Delta_n/(m_0\omega_n) \to +\infty.$$

Assumption B7. *There exist positive constants* c, c_1, c_2 *and* c_3 *such that for indexes* j' *and* $j'-1$ *and corresponding estimated thresholds* $s_{j'}$ *and* $s_{j'-1}$,
(a) if $|s_{j'} - s_{j'-1}| \leq \gamma_n$, *then* $\eta_{(s_{j'-1}, s_{j'})} = c\sqrt{n\gamma_n} \log\left(p^2K\right)$;
(b) if there exist r_j *and* r_{j+1} *such that* $|s_{j'-1} - r_j| \leq \gamma_n$ *and* $|s_{j'} - r_{j+1}| \leq \gamma_n$, *then,*

$$\eta_{(s_{j'-1}, s_{j'})} = \frac{2}{c_3}\left(c_1 \frac{\log(p^2K)}{\sqrt{n(s_{j'} - s_{j'-1})}} + c_2 M_A d_n^* \frac{\gamma_n}{s_{j'} - s_{j'-1}}\right);$$

(c) otherwise $\eta_{(s_{j'-1}, s_{j'})} = \frac{2}{c_3}\left(c_1 \frac{\log(p^2K)}{\sqrt{n(s_{j'} - s_{j'-1})}} + c_2 M_A d_n^*\right)$.

Assumption B6 makes a unique connection between three important quantities: (1) minimum spacing between consecutive thresholds, Δ_n; (2) the consistency rate for estimating the threshold values, γ_n; (3) the penalty term in the definition of the information criterion, ω_n. This connection helps with quantifying the consistency rate for estimating the threshold values as discussed after Theorem 7.

Assumption B7 specifies three different tuning parameter rates for the screening step. Although this assumption may seem technical, but it is needed to get the sharpest consistency rate. It is possible to define a fixed tuning parameter for all cases in Assumption B7, but the consistency results will be worsened. Remark 5 in [Safikhani and Shojaie, 2020] shed some light into this issue.

Theorem 7. *Under Assumptions B1 to B7, if $n \to +\infty$, the minimizer*

$$(\tilde{m}, \tilde{r}_j, j = 1, 2, \ldots, \tilde{m})$$

of Equation (3.10) satisfies:

$$\mathbb{P}(\tilde{m} = m_0) \to 1. \tag{3.12}$$

In addition, there exists a constant $B > 0$ such that:

$$\mathbb{P}\left(\max_{1 \le j \le m_0} |\tilde{r}_j - r_j| \le B m_0 \left(\gamma_n\right)^{3/2} d_n^{*2} \sqrt{n}\right) \to 1. \tag{3.13}$$

When $p = cn^{\kappa}$, where $c > 0$ and $\kappa \in (0, 1)$, the proposed procedure for both hdTAR and mvTAR can also be applied to low-dimensional time series. The consistency results would be similar to those in Theorem 7. It is challenging to select ηs in practice, since the distance between estimated thresholds to the true thresholds is unknown. Instead, we set ηs to be the same and apply BIC/HBIC to select them.

Although the consistency rates for mvTAR and hdTAR are both functions of γ_n, the assumptions on γ_n for the two methods are different, leading to different rates of consistency. To illustrate this point, consider the case when m_0 is finite. Then, when using the ℓ_1 penalty in the first step, we can set $\gamma_n = (\log n)^{\rho} \left(\log\left(p^2 K\right)\right)^{2+2\rho} / n$ for some $\rho > 0$. With this rate, the hdTAR model can have total sparsity $d_n^* = o\left(\left(\log n \left(\log\left(p^2 K\right)\right)^2\right)^{\rho/2}\right)$. The consistency rate then becomes of order $\left((\log n)^{\frac{5}{2}\rho} \left(\log\left(p^2 K\right)\right)^{3+5\rho}\right) / n$. In comparison, when using the ℓ_2 penalty, we can set $\gamma_n = (\log n)^{\rho'} \left(\log\left(p^2 K\right)\right)^{2+2\rho'} \left(p^2 K\right)^{1+\rho'} / n$ for some $0 < \rho' < 1$ to ensure that Assumption B3 is satisfied. With this rate, the mvTAR model can have total sparsity $d_n^* = o\left(\left(p^2 K \log n \left(\log\left(p^2 K\right)\right)^2\right)^{\rho'/2}\right)$. Using a similar calculation, the

consistency rate for mvTAR becomes of order $\left((\log n)^{\frac{5}{2}\rho'}\left(\log\left(p^2K\right)\right)^{3+5\rho'}\left(p^2K\right)^{\frac{3}{2}+\frac{5}{2}\rho'}\right)/n$, further highlighting that mvTAR is not suitable in high dimensions, when $p = cn^{\kappa}$, where $c > 0$ and $\kappa \geq 1$.

Remark 2. *If we remove the assumption $\varkappa_0 \geq 2/3$ and only keep $\varkappa_0 < 1$, then, according to Lemma 18, the choice of γ_n would also depend on $\varkappa_0$. For the hdTAR model, we can set $\gamma_n = (\log n)^{\rho}\left(\log\left(p^2K\right)\right)^{2/\varkappa_0-1+2\rho}/n$ for some $\rho > 0$, and keep the total sparsity the same as above. The consistency rate then becomes of order $\left((\log n)^{\frac{5}{2}\rho}\left(\log\left(p^2K\right)\right)^{3/\varkappa_0-3/2+5\rho}\right)/n$. Similarly, for mvTAR model, we can set $\gamma_n = (\log n)^{\rho'}\left(\log\left(p^2K\right)\right)^{2/\varkappa_0-1+2\rho'}\left(p^2K\right)^{1+\rho'}/n$ for some $0 < \rho' < 1$, and keep the total sparsity the same as above. The consistency rate for mvTAR becomes of order $\left((\log n)^{\frac{5}{2}\rho'}\left(\log\left(p^2K\right)\right)^{3/\varkappa_0-3/2+5\rho'}\left(p^2K\right)^{\frac{3}{2}+\frac{5}{2}\rho'}\right)/n$.*

Our last theorem establishes the consistent estimation of regime-specific transition matrices in the third step.

Theorem 8. *Under Assumptions B1 to B7, and selecting*

$$\alpha_j = C\sqrt{\log(p^2K)/\left(n\gamma_n\right)}$$

for some large enough $C > 0$, with high probability approach to 1, there exists a positive constant C' such that we have for each fixed regime j:

$$\left\|\hat{\boldsymbol{\beta}}^{(.,j)} - \boldsymbol{A}^{(.,j)}\right\|_2 \leq C'\sqrt{d_n^* \log(p^2K)/\left(n\gamma_n\right)}. \tag{3.14}$$

The consistency rate derived in Theorem 8 is similar to that of Wong et al. [2020], Basu and Michailidis [2015b] for high-dimensional vector auto-regressive models.

3.5 Tuning Parameter Selection

We next provide guidance on selecting the tuning parameters for our three-step procedure.

$\lambda_{1,n}$ We choose $\lambda_{1,n}$ by cross-validation. We first randomly choose the order of switching variable z_t with equal space. Let $\mathbb{T}$ be a set of time points corresponding to selected switching variable. We use the rest of observations to estimate $\boldsymbol{\Theta}$ in the first step for a

range of $\lambda_{1,n}$. To choose the optimal value of $\lambda_{1,n}$, we use the estimated $\boldsymbol{\Theta}$ to predict the series at time points in $\mathbb{T}$. The optimal $\lambda_{1,n}$ is selected as the value corresponding to the minimum mean squared prediction error over $\mathbb{T}$.

$\lambda_{2,n}$ The rate for $\lambda_{2,n}$ vanishes fast as n increases. Thus, to lower the computational cost, we set $\lambda_{2,n}$ to zero. It is possible to select $\lambda_{2,n}$ using cross-validation as well at the cost of increasing computation time. However, the sensitivity analysis reported in Safikhani and Shojaie [2020] indicates that setting $\lambda_{2,n} = 0$ is a reasonable choice.

$\boldsymbol{\eta_n}$ Selecting η_n is in general difficult. For $0 \leq m \leq \hat{m}$ ($\hat{m}$ is the number of estimated thresholds in step 1), we choose different ηs for different regimes, and use HBIC and eBIC [Wang and Zhu, 2011] across all regimes. For each time series l, $l \in 1, 2, \ldots, p$, and $j = 1, 2, \ldots, m+1$, set η_j^l as the tuning parameter for l-th time series at j-th regime. Then, the HBIC for interval $[s_{j-1}, s_j]$ is defined as

$$\text{HBIC}\left(j, \eta_j^l\right) = \log\left(\text{SSE}_{l,j} / \left|\mathcal{T}_{(s_j - s_{j-1})}\right|\right) + \frac{\gamma_1 \left\|\hat{\boldsymbol{\theta}}^l_{s_{j-1}, s_j}\right\|_0}{\left|\mathcal{T}_{(s_j - s_{j-1})}\right|} \log\left(pK\right),$$

where $\gamma_1 = 2.8$ that is within the recommended range in Wang and Zhu [2011]. Similarly, the eBIC for interval $[s_{j-1}, s_j]$ is defined as

$$\text{eBIC}\left(j, \eta_j^l\right) = \log\left(\text{SSE}_{l,j} / \left|\mathcal{T}_{(s_j - s_{j-1})}\right|\right) + \frac{\gamma_2 \left\|\hat{\boldsymbol{\theta}}^l_{s_{j-1}, s_j}\right\|_0}{\left|\mathcal{T}_{(s_j - s_{j-1})}\right|} \left(\log\left(pK\right) + \log\left(\left|\mathcal{T}_{(s_j - s_{j-1})}\right|\right)\right),$$

where $\gamma_2 = 1.4$ that is within the recommended range in Wang and Zhu [2011] as well. If $\left|\mathcal{T}_{(s_j - s_{j-1})}\right| \geq pK$, η_j^l is selected as:

$$\hat{\eta}_j^l = \arg\min_{\eta_j^l} \text{eBIC}\left(j, \eta_j^l\right). \tag{3.15}$$

If $\left|\mathcal{T}_{(s_j - s_{j-1})}\right| < pK$, η_j^l is selected as:

$$\hat{\eta}_j^l = \arg\min_{\eta_j^l} \text{HBIC}\left(j, \eta_j^l\right). \tag{3.16}$$

ω_n We first perform the backward elimination algorithm (BEA) until no break points are left. Then, we cluster the differences in the objective function L_n into two subgroups, small and large. If removing a threshold only leads to a small decrease in L_n, then the removed threshold is likely redundant. In contrast, true thresholds lead to larger decrease. We choose the smallest decrease in the second group as the optimal value of ω_n. To this end, we first calculate the minimum sum of squared error for removing all thresholds in $\hat{\mathcal{A}}_n$ one by one, denoted as $L_0', L_1', \ldots, L_{\hat{m}}'$. Then, ω_n is selected as the maximum values among $L_{j+1}' - L_j'$ for $j = 0, 1, \ldots, \hat{m} - 1$.

α_i For simplicity, we let all time series share the same α_i, denoted by α_n. For low to moderate dimensions, the tuning parameter α_n for parameter estimation is selected as the minimizer of the combined HBIC over all regimes. For $j = 1, 2, \ldots, \tilde{m} + 1$, we define the HBIC on interval $[\tilde{r}_{j-1}, \tilde{r}_j]$ as:

$$\text{HBIC}\left(j, \alpha_n\right) = \log\left(\det \hat{\boldsymbol{\Sigma}}_{\epsilon,j}\right) + \frac{\gamma \left\|\hat{\boldsymbol{\beta}}^{(\cdot,j)}\right\|_0}{\left|\mathcal{T}_{(\tilde{r}_{j-1}, \tilde{r}_j)}\right|} \log\left(p^2 K\right),$$

where $\hat{\boldsymbol{\Sigma}}_{\epsilon,j}$ is the residual sample covariance matrix with $\hat{\boldsymbol{\beta}}$ estimated in Equation (3.11) and $\gamma = 2.8$. For high dimensions, we choose α_n by 10-fold cross validation.

3.6 Empirical Evaluations

In this section, we present simulation results evaluating the performance of the proposed procedure in both moderate dimensions and high dimensions; the first four simulations scenarios presented are moderate-dimensional, while the last one is high-dimensional. Details of simulation settings are presented in Appendix 7. All results are averaged over 200 replicates.

We compare our method with Tsay [1998b], Li and Tong [2016], and the threshold vectorized auto-regressive method [Lo and Zivot, 2001]. These methods, which are denoted as Tsay (1998), Li (2016) and TVAR, respectively, assume a known number of thresholds or at least a known upper bound on the number of thresholds when establishing the asymptotic properties of their estimators. In practice, Tsay [1998b] proposes to perform a grid search to select the number of thresholds, when unknown, by minimizing AIC. They are also restricted to low dimensions. For instance, TVAR estimates the number of thresholds and the values of thresholds using two separate steps and assumes the number of thresholds is at most 2. This is in contrast to our developed mvTAR and hdTAR methods, which do not make any assumptions about the number of thresholds. Though Calderón V and Nieto [2017] and Orjuela and Villanueva [2021] do not require a known number of thresholds, we did not include a comparison with these methods, as they cannot handle larger dimensions, e.g., $p = 20$.

We compare the estimated thresholds and the percentage of simulations where thresholds are correctly estimated; this is defined as cases where the selected thresholds are close to the true thresholds. More specifically, a selected threshold is considered as close to the first true threshold, z_1, if it is in the interval $[-\infty, z_1 + 0.5(z_2 - z_1))$; similarly, a selected threshold is considered as close to the second true threshold, z_2, if it falls in the interval $[z_1 + 0.5(z_2 - z_1), \infty]$. Note that the number of thresholds is set to be known for Tsay (1998), Li (2016) and TVAR, since the first two require a known number of thresholds and TVAR does not perform well in selecting the number of threshold in its first step (Note that when the number of thresholds is not provided, TVAR's rates of correctly identifying the correct number of thresholds are 84%, 87%, 65%, and 16% (11.5% for $T = 300$) in the first four scenarios and the method is not applicable in the last scenario (scenario 5) due to high-dimensionality of model.)

3.6.1 Simulation Results

We next compare the performances of the proposed hdTAR and mvTAR methods with Tsay (1998), Li (2016) and TVAR. Here, the selection rate of Tsay (1998), Li (2016) and

TVAR is based on whether the estimated thresholds are within one standard deviation of true threshold.

Table 3.1 summarizes the results of threshold estimation. In all simulations, if any of the methods does not select a thresholds, we set the minimum value of the threshold variable as the selected threshold. The results indicate that Tsay (1998) and Li (2011) do not work well even for the first three scenarios, while hdTAR, mvTAR, and TVAR perform well in the first three scenarios; however, the estimation error and standard deviation of TVAR are larger than those of hdTAR and mvTAR. In Scenario 4, in which only a portion of time series components change at threshold values, the detection rate for both mvTAR and TVAR drops significantly, while hdTAR still achieves 100% threshold detection rate. This is expected since hdTAR is more flexible and mvTAR only works well for scenarios in which auto-regressive components change at the same threshold values. In Scenario 4, mvTAR tends to choose a large λ_1 which leads to selecting smaller number of threshold values in the first step than needed. Nonetheless, when the changes in the transition matrices are large enough, the threshold values can still be detected using the ℓ_2 penalty. Finally, hdTAR continues to offer excellent threshold detection in the high-dimensional setting of Scenario 5; in contrast, the other methods are not well suited for this scenario and are not included.

	Threshold(s)	Methods	Mean	Std	Selection Rate
Scenario 1		hdTAR	4.05	0.05	1.00
		mvTAR	4.04	0.05	1.00
	4	TVAR	4.23	1.13	0.95
		Tsay (1998)	5.36	2.01	0.57
		Li (2016)	7.66	0.34	0.00
Scenario 2		hdTAR	4.05	0.06	1.00
		mvTAR	4.04	0.05	1.00
	4	TVAR	4.15	1.17	0.93
		Tsay (1998)	5.46	2.00	0.56
		Li (2016)	7.66	0.34	0.00
Scenario 3		hdTAR	4.04	0.06	1.00
		mvTAR	4.04	0.05	1.00
	4	TVAR	4.15	1.17	0.93
		Tsay (1998)	7.63	0.76	0.03
		Li (2016)	7.66	0.34	0.00
Scenario 4		hdTAR	4.00	0.15	1.00
(T = 600)	4	mvTAR	2.44	1.24	0.93
		TVAR	3.82	1.34	0.85
		hdTAR	6.02	0.09	1.00
	6	mvTAR	5.30	1.37	0.82
		TVAR	6.11	1.28	0.88
Scenario 4		hdTAR	4.03	0.47	1.00
(T = 300)	4	mvTAR	2.52	0.91	0.67
		TVAR	3.92	1.48	0.81
		hdTAR	6.00	0.31	1.00
	6	mvTAR	4.78	1.16	0.42
		TVAR	6.19	1.42	0.84
Scenario 5		hdTAR	5.06	0.29	1.00
	5	mvTAR	–	–	–
		TVAR	–	–	–

Table 3.1: Mean and standard deviation of estimated thresholds, the percentage of simulation runs where thresholds are correctly detected (selection rate) in different simulation

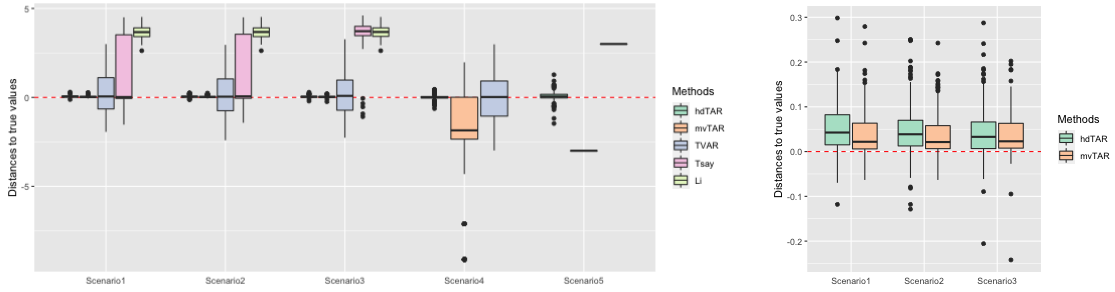


Fig 3.3: Box plot of distances between the estimated final points and true values. The left panel shows the results for all the five scenarios with all the five models. The right panel zooms in the results in the first three scenarios using hdTAR and mvTAR.

Table 3.2 summarizes the performance of the five methods in terms of auto-regressive parameter estimation. Since Tsay (1998) does not provide coefficients estimates, so we use the method in our Step 3 to estimate the parameters given the thresholds obtained by Tsay (1998). The results indicate that both hdTAR and mvTAR perform well in the first three scenarios, as measured by their high true positive rates and low false positive rates. Since TVAR does not perform variable selection, all estimated values of transition matrices using this method are non-zero. This leads to true positive and false positive rates that are both equal to 1, which are not meaningful and are hence excluded from the table.

The results also indicate that in Scenario 4 with $T = 600$ and Scenario 5 hdTAR performs satisfactorily, while in Scenario 4 with $T = 300$, its FPR increases to around 20%. This is primarily due to the smaller sample size in this scenario for each of the three regimes. Recall from Table 3.1 that the selection rate of mvTAR in both of these scenarios was very low; as a result, in many simulation replicates there were fewer number estimated regimes than needed to obtain estimates of auto-regressive parameters. As a result, mvTAR is not included in the comparisons for Scenarios 4 and 5. These findings underscore the advantages of hdTAR in settings with complex patterns of changes in auto-regressive parameters as well as in high dimensions.

Box plots summarizing the results in Table 3.1 are presented in Figure 3.3.

	Method	REE	SD(REE)	FPR	TPR
Scenario 1	hdTAR	0.31	0.04	0.03	0.95
	mvTAR	0.32	0.04	0.03	0.95
	TVAR	0.85	0.19	–	–
	Tsay (1998)	0.70	0.30	0.04	0.57
	Li (2016)	1.50	0.44	1.00	1.00
Scenario 2	hdTAR	0.31	0.04	0.03	0.95
	mvTAR	0.31	0.04	0.03	0.94
	TVAR	0.89	0.43	–	–
	Tsay (1998)	0.69	0.31	0.04	0.55
	Li (2016)	1.49	0.55	1.00	1.00
Scenario 3	hdTAR	0.34	0.04	0.04	0.89
	mvTAR	0.34	0.04	0.04	0.89
	TVAR	0.69	0.65	–	–
	Tsay (1998)	0.88	0.05	0.03	0.36
	Li (2016)	1.33	0.64	1.00	1.00
Scenario 4 (T = 600)	hdTAR	0.5	0.05	0.02	0.77
	mvTAR	–	–	–	–
	TVAR	0.67	0.07	–	–
Scenario 4 (T = 300)	hdTAR	0.77	0.09	0.19	0.71
	mvTAR	–	–	–	–
	TVAR	0.87	0.15	–	–
Scenario 5	hdTAR	0.80	0.04	0.51	0.86
	mvTAR	–	–	–	–
	TVAR	–	–	–	–

Table 3.2: Results of parameter estimation for simulation scenarios. The table shows mean and standard deviation of relative estimation error (REE), true positive rate (TPR), and false positive rate (FPR) for estimated coefficients.

3.7 Real Data Application

We demonstrate the utility of our penalized estimation framework in financial econometric applications by analyzing a bank balance sheet data. The data consists of total balances of the top 10 largest US banks over time, each measured in thousands of dollars (available from `www.fdic.gov`).

To assess the relationship between the state of the banking sector and the overall economic conditions, we fit a multivariate TAR model of the quarterly bank balance sheet data over the period of 1995 to 2018 with the growth rate of the US GDP as the switching

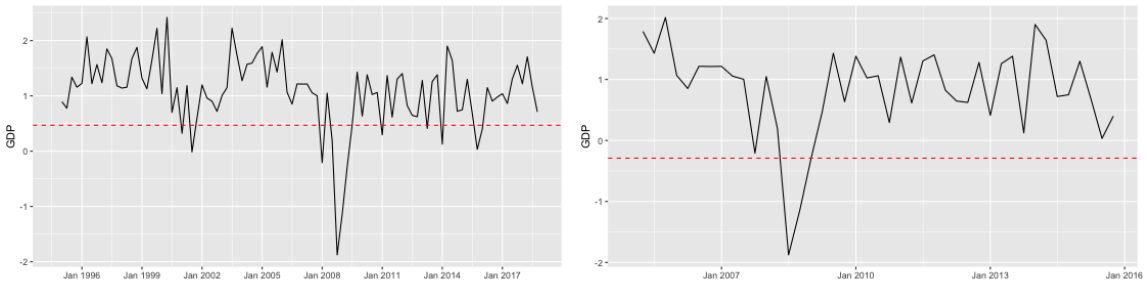


Fig 3.4: The GDP growth rate and detected thresholds using data from ten top banks. The red dash line shows the estimated threshold. The left panel shows the GDP growth rate and detected thresholds based on data from 1995 to 2018, while the right panel shows the GDP growth rate and detected thresholds based on data from 2005 to 2015. In both cases, the proposed method divides economic patterns into only two conditions — recession and non-recession.

variable. For the quarterly GDP data $y_t; t = 1, 2, \ldots, T$ over T observations, the growth rate is defined as

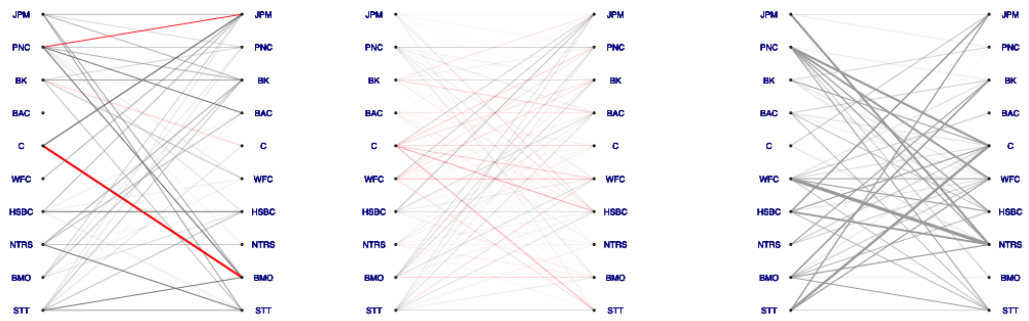
$$z_t = 100(\log y_t - \log y_{t-1}), \quad t = 2, 3, \ldots, T.$$

To reduce the non-stationarity, the bank balance sheet data $v_t; t = 1, 2, \ldots, T$, is also transformed as

$$x_t = \log v_t - \log v_{t-1}, \quad t = 2, 3, \ldots, T.$$

We applied the hdTAR on the entire time series consisting of $T = 98$ quarterly observations from 1995 to 2018. To examine how results change with smaller sample sizes, we also analyze the shorter time period of quarterly observations from 2005 to 2015. The detected threshold for both time periods are shown in Figure 3.4. Although hdTAR does not enforce the coefficients to change at the same threshold value, irrespective of the sample size it identifies a single threshold corresponding to the great recession of 2008. This further highlights the flexibility and adaptability of hdTAR for both moderate- and high-dimensional TAR models. As a comparison, we also applied the mvTAR to the same two data sets, but exclude the results due to the inconsistency in the estimated thresholds using mvTAR when applied to the same two data sets.

The Granger causal networks [Basu et al., 2015] of interactions among these ten banks in



(a) (b)

Fig 3.5: The Granger causality graph for the top ten banks across time. Each vertex represents a bank, and the links display directed interactions between banks. Panel (a) corresponds to the longer time series (1995–2018) and panel (b) corresponds to the shorter time series (2005–2015). The left figure in panel (a) shows the interactions during the recession; the right figure shows the interactions in non-recession. The red links in each panel represent the interactions that occur in that economic period only. Panel (b) only show the interactions among banks identified in non-recession period from the shorter time series. Given the very small number of observations in the recession period in the shorter time series, the Granger causality graph for this period is not estimated.

both recession and non-recession periods during 1995–2018 are shown in Figure 3.5a. The red links in each panel represent the interactions that occur in that economic period only. The results show strong interactions between Citibank and Harris Bank and a comparable strong interaction between PNC and JPMorgan Chase during the recession. The interactions become weaker during the non-recession period, but more interactions appear among banks. A similar observation was made in Lin and Michailidis [2017].

We only plot the estimated network structures during non-recession period from 2005–2015. This is because the detected threshold is very close to the lower boundary of the sorted values of the switching variable, resulting in very few observations in the recession regime. From Figure 3.5b, the interactions among banks in non-recession period from 2005 to 2015 are similar to the structures detected using full data set. This further confirms the satisfactory performance of hdTAR in both larger data and smaller data sets.

3.8 Discussion

We developed a three-step algorithm to estimate the number and values of thresholds, as well as the auto-regressive parameters in possibly high-dimensional TAR model. The proposed algorithm can utilize either an ℓ_2 or an ℓ_1 penalty, or more specifically, a grouped or regular fused lasso penalty. The ℓ_2 penalty corresponds to the natural extension of the original multivariate TAR model in which all coefficients are forced to change at the same thresholds. The ℓ_1 penalty, in contrast, is more flexible allowing each coefficient to potentially change at different thresholds. Although this flexibility potentially comes at the cost of a larger number of thresholds in the TAR model, our theoretical and empirical results indicate that mvTAR is not appropriate for high-dimensional settings and is better suited for moderate dimensions. In contrast, the more flexible hdTAR leads to consistent estimation and superior empirical performance in both moderate and high dimensions.

We established that both versions of our algorithm, termed mvTAR and hdTAR, consistently estimate the model parameters under natural conditions on the distribution and on the level of temporal correlations in the model. The consistency rates for both models depend explicitly on several model characteristics. Specifically, when the total number of thresholds, m_0, is finite, the rate of consistency for detecting the thresholds is based on: (1) the effective number of time points, n, (2) the number of time series components, p, (3) the number of lags, K, and (4) the total sparsity of the model, d_n^*. For mvTAR, if we set $d_n^* = o\left(\left(\log n \left(\log\left(p^2K\right)\right)^2 p^2K\right)^{\rho'/2}\right)$ for small $0 < \rho' < 1$, then the consistency rate becomes of order $\left((\log n)^{\frac{5}{2}\rho'}\left(\log\left(p^2K\right)\right)^{3+5\rho'}\left(p^2K\right)^{\frac{3}{2}+\frac{5}{2}\rho'}\right)/n$. This confirms that mvTAR is suitable for moderate dimension but may not work in high dimensions. In contrast, for hdTAR, setting $d_n^* = o\left(\left(\log n \left(\log\left(p^2K\right)\right)^2\right)^{\rho/2}\right)$ for some small positive ρ, the consistency rate becomes of order $\left((\log n)^{\frac{5}{2}\rho}\left(\log\left(p^2K\right)\right)^{3+5\rho}\right)/n$. The first component of the rate, i.e. $(\log n)^{\frac{5}{2}\rho}$, is similar to some existing consistency rates for univariate TAR models [Chan et al., 2015] while the additional term $\left(\log\left(p^2K\right)\right)^{3+5\rho}$ quantifies the difficulty in estimating the thresholds in high-dimensions.

A limitation of the proposed procedure is that it requires several hyperparameters, especially in the second step. To lower the computational cost, we chose similar tuning

parameters in the second step according to eBIC/ HBIC. However, regime-specific tuning parameters may improve the estimation performance in finite samples. Fast selection of regime-specific tuning parameters is an interesting future research direction. Identifying the switching variable in the TAR model is another challenge, specifically in applications. For example, in the bank data, we selected the GDP as the switching variable. However, it is not obvious whether this is an optimal choice; for example, the unemployment rate or the inflation rate could also serve as the switching variable. Selecting optimal (data-driven) switching variable is another fruitful future research direction.

Chapter 4

DYNAMIC PROGRAMMING APPROACH FOR HIGH-DIMENSIONAL THRESHOLD AUTO-REGRESSIVE MODELS WITH MANY COMPONENTS AND THRESHOLDS

4.1 Introduction

In this chapter, we continue discussing TAR model introduced in Chapter 3. However, in this chapter, we provide a dynamic programming approach, named dpTAR, to better estimate the number of thresholds and their corresponding values for TAR models. In addition, we have empirically compared the performance of our method with the existing approaches in the simulation section, demonstrating that our method offers clear advantages. In addition, we establish theoretical results that give a sharper convergence rate of the estimators.

Recall that the multivariate TAR model has been well studied in the literature (Tsay [1998b], Lo and Zivot [2001], Hansen and Seo [2002], Dueker et al. [2011], Li and Tong [2016]). Chapter 3 and Appendix B.0.3 have thoroughly reviewed the existing multivariate TAR estimation methods. It is worth noting that without knowing the number of thresholds, Tsay [1998b] (or other existing approaches for multivariate TAR models that test the existence of a threshold) is not straightforward. The few existing approaches that can estimate the number of thresholds can only handle the finite number of thresholds and only work in low dimensions. To our knowledge, Nieto [2005], Calderón V and Nieto [2017], Calderón V and Nieto [2017], and Zhang et al. [2022] are the only methods that do not require a known number of thresholds or a bound on the number of thresholds. Except Zhang et al. [2022], Nieto [2005], Calderón V and Nieto [2017], and Calderón V and Nieto [2017] utilize a Bayesian estimation framework. However, the consistency of the number of estimated thresholds is not investigated for these Bayesian methods, which could be a challenging problem and could be a good direction for future research. Extending these methods to high dimensions can also be challenging. To the best of our knowledge, the approach in Chapter 3 [Zhang et al., 2022] and the dynamic programming approach (dpTAR)

discussed in this chapter are the only two methods that can deal with the diverging number of thresholds m_0 and high-dimensional problems. The method provided in Chapter 3 (Zhang et al. [2022]) assumes that the minimal jump size v (defined in Assumption B4 in Chapter 3) is independent of the sample size, while the dynamic programming approach in this chapter allows the minimal jump size to decrease with the sample size. The simulation results corroborate our claims about the advantages of the dynamic programming approach compared with existing approaches, including the three-step procedure. In addition, the consistency rate derived by the dynamic programming approach is sharper than existing approaches with the combination of β-mixing and sub-Weibull assumption.

Our method is motivated by Wang et al. [2019], which describes a dynamic programming approach in high-dimensional autoregressive processes. The dynamic programming approach is a type of exact search method that is commonly used in the change point detection problems. The change point detection problems have been widely used in diverse application domains, from economics and finance [Andreou and Ghysels, 2002, Modisett and Maboudou-Tchao, 2010] to genomics and biology [Braun et al., 2000, Bleakley and Vert, 2011]. It has been well studied in both univariate and multivariate time series [Killick et al., 2012, Harchaoui and Lévy-Leduc, 2010]. See Brodsky and Darkhovsky [2013], Truong et al. [2020] for the reviews of the current findings of change point detection problems. In addition, change point detection problems in high-dimensional time series models have received considerable attention in recent years [Safikhani and Shojaie, 2020, Wang et al., 2017, Grundy et al., 2020, Wang et al., 2019]. Safikhani and Shojaie [2020] proposed a three-stage procedure for consistent estimation of both structural change points and parameters of high-dimensional piece-wise vector autoregressive models. Wang et al. [2017] extended binary segmentation algorithms Vostrikova [1981] for covariance change point localization in high dimensions. Grundy et al. [2020] developed an approach that takes inspiration from geometry to map a high-dimensional time series to two dimensions. Wang et al. [2019] established a combination of dynamic programming and Lasso-type estimators approach to localizing changes in high-dimensional.

This chapter is organised as follows. Section 4.2 describe the multivariate TAR model; Section 4.3 introduce the dynamic programming approach and the corresponding algorithm;

Section 4.4 discuss theoretical properties and compare them with theoretical results in Chapter 3. In Section 4.5, we propose data-driven methods to select the hyper-parameters. The empirical performance of the proposed methods is investigated using both simulated and real data sets, in Section 4.6 and Section 4.7, respectively. We discuss and summarize our results in Section 4.8.

4.2 TAR Model Recap

Recall the multi-variate TAR model discussed in Chapter 3. Formally, a multivariate time series $\{\boldsymbol{x}_t\}$ follows TAR model with one switching variable, z_t, if

$$\boldsymbol{x}_t = \sum_{k=1}^{K} \boldsymbol{A}^{(k,j)} \boldsymbol{x}_{t-k} + \boldsymbol{\Sigma}_j^{1/2} \boldsymbol{\epsilon}_t, \quad \text{if } r_{j-1} < z_t \leq r_j, \tag{4.1}$$

where $\boldsymbol{x}_t = (x_{(t,1)}, x_{(t,2)}, \ldots, x_{(t,p)})'$ is the observed process in $\mathbb{R}^p$ at time t, p is the number of time series components, and K is the number of lags considered in the model. Here $\boldsymbol{\epsilon}_t = (\epsilon_{(t,1)}, \epsilon_{(t,2)}, \ldots, \epsilon_{(t,p)})' \in \mathbb{R}^p$ is a multivariate i.i.d. sequence with zero mean in all components. The covariance matrix $\boldsymbol{\Sigma}_j$ for the j-th regime, $\boldsymbol{\Sigma}_j$, is allowed to be different in each regime. To simplify the notations, when there is no ambiguity, we simply denote the error term by $\boldsymbol{\epsilon}_t$ instead of $\boldsymbol{\Sigma}_j^{1/2}\boldsymbol{\epsilon}_t$. The transition matrices $\boldsymbol{A}^{(k,j)} \in \mathbb{R}^{p\times p}$ is the coefficient matrix corresponding to the k-th lag of a TAR process in regime j. More specifically, similar to the modeling framework of Chan et al. [2015], we assume there exist m_0 threshold values $-\infty < r_1 < r_2 < \ldots < r_{m_0} < +\infty$ with $r_0 = -\infty$ and $r_{m_0+1} = +\infty$ which partition the process into $m_0 + 1$ regimes. For each regime, the total transition matrices $\boldsymbol{A}^{(\cdot,j)} = \left(\boldsymbol{A}^{(1,j)}, \boldsymbol{A}^{(2,j)}, \ldots, \boldsymbol{A}^{(K,j)}\right) \in \mathbb{R}^{p\times pK}$ are fixed where $r_{j-1} < z_t \leq r_j$ for $j = 1, \ldots, m_0 + 1$.

Now suppose we have multiple switching variables. Let $z_{t,l}$ be the l-th switching variable and $\mathcal{P}$ be an interval partition of $\{z_{1,l}, \ldots, z_{T,l}\}$ into $m + 1$ regimes, that is

$$\mathcal{P} = \{(-\infty, r_1], (r_1, r_2], \ldots, (r_m, \infty)\},$$

where $|\mathcal{P}|$ represents the cardinality of $\mathcal{P}$. Set $\boldsymbol{Y}_{\pi(i),l} = \begin{pmatrix} \boldsymbol{x}'_{\pi(i)-1,l} & \boldsymbol{x}'_{\pi(i)-1,l} & \cdots & \boldsymbol{x}'_{\pi(i)-K,l} \end{pmatrix}$, where $\pi(\cdot)$ is the function which projects order statistics of the observations to the corre-

sponding indexes of the observations, and $\boldsymbol{x}'_{\pi(i)-1,l}$ is the ordered $\boldsymbol{x}_t$ according to $z_{t,l}$.

Our goal is (1) to select the optimal switching variable z_t^* among all $z_{t,l}$s, (2) estimate the number of thresholds, i.e. m_0, (3) estimate the thresholds' values, r_js, and (3) find the auto-regressive parameters in each regime.

4.3 Dynamic Programming Approach for TAR model

Notations: We use the similar definition as in Chapter 3. Denote r_j is the j-th threshold and $\hat{r}_j$ is the estimated j-th threshold. For a given interval $(s, e]$, denote $\mathcal{T}^l_{(s,e)} = \left\{i : s < z_{\pi(i),l} \leq e\right\}$ as the set of orders of $z_{t,l}$s for which their corresponding ordered switching variable $z_{\pi(i),l}$s fall into the interval $(s, e]$. And $\left|\mathcal{T}_{(s,e)}\right|$ represents the number of $z_{\pi(i),l}$s that fall into the interval $(s, e]$. For simplicity, we use $\mathcal{T}$ to represent the regime $\mathcal{T}_{(s,e)}$, use z_t to represent all the $z_{t,l}$s, and use $z_{\pi(i)}$ to represent all the $z_{\pi(i),l}$s. For a symmetric matrix $\mathbf{X}$, let $\lambda_{\min}(\mathbf{X})$ and $\lambda_{\max}(\mathbf{X})$ denote its minimum and maximum eigenvalues. Let the h-th row of $\boldsymbol{A}^{(\cdot,j)}$ be $\boldsymbol{A}_h^{(\cdot,j)}$, and set the number of non-zero elements in $\boldsymbol{A}_h^{(\cdot,j)}$ to $d_{h,j}$ for $h = 1, 2, \ldots, p$ and $j = 1, 2, \ldots, m_0 + 1$. Denote the total sparsity of the model by $d_n^* = \sum_{j=1}^{m_0+1} \sum_{h=1}^{p} d_{h,j}$. Further, let $\mathcal{I}_{h,j}$ be the set of all column indexes of $\boldsymbol{A}_h^{(\cdot,j)}$, $\mathcal{I} = \cup_{h,j} \mathcal{I}_{h,j}$ and define $d_n = \max_{1 \leq h \leq p, 1 \leq j \leq 1+m_0} |\mathcal{I}_{h,j}|$. Note that p, m_0 and the sparsity may increase with the number of time points, T, specifically, $p \equiv p(n)$ and $m_0 \equiv m_0(n)$ and $d_{h,j} \equiv d_{h,j}(n)$, where $n = T - K$. For simplicity, we suppress the n-index. Finally, let $\epsilon_{t,l}$ be error term of l-th time series, and recall that $\boldsymbol{\epsilon}_t = \left(\epsilon_{(t,1)}, \epsilon_{(t,2)}, \ldots, \epsilon_{(t,p)}\right)'$. Throughout the paper, positive constants $C, C_1, C_2, \ldots$ are used to denote universal constant, $\boldsymbol{A}'$ denotes the transpose of a matrix $\boldsymbol{A}$, and $\|\boldsymbol{A}\|_1$ and $\|\boldsymbol{A}\|_2$ denotes its ℓ_1 and Frobenius norms, respectively. We denote the ℓ_1 and ℓ_2 norms of a vector v by $\|v\|_1$ and $\|v\|$, respectively.

To estimate the number of thresholds and their values correspondingly, we first view the TAR model as a minimal partition problem and then apply a dynamic programming

approach to solve it. We first construct the loss function,

$$L(\mathcal{T}) = \begin{cases} \sum_{z_{\pi(i)} \in \mathcal{T}} \|\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}} \boldsymbol{Y}_{\pi(i)}\|_2^2, & |\mathcal{T}| \geq \omega \\ 0, & \text{otherwise,} \end{cases} \tag{4.2}$$

with

$$\hat{\boldsymbol{A}}_{\mathcal{T}} = \arg\min_{\boldsymbol{A}} \left(\sum_{i \in \mathcal{T}} \|\boldsymbol{x}_{\pi(i)} - \boldsymbol{A}\boldsymbol{Y}_{\pi(i)}\|_2^2 + \lambda \sqrt{|\mathcal{T}|} \|\boldsymbol{A}\|_1 \right), \tag{4.3}$$

where the tuning parameter $\omega > 0$. In this step, we utilize the l_1 penalty, the fused lasso penalty, to estimate the transition matrices in Equation (4.3). The fused lasso penalty allows the sparsity of the transition matrix at each regime. In addition, we construct a loss function, $L(\mathcal{T})$, in Equation (4.2) where $L(\mathcal{T})$ can be evaluated on each regime. The total loss of the TAR model is given by the sum of loss on each regime.

Next, we solve:

$$\hat{\mathcal{P}} \in \arg\min_{\mathcal{P}'} \left\{ \sum_{\mathcal{T} \in \mathcal{P}'} L(\mathcal{T}) + \omega |\mathcal{P}'| \right\}. \tag{4.4}$$

$\mathcal{P}' = \{(a, b] \mid a, b \in \{-\infty, z_{\pi(1)}, \ldots, z_{\pi(n)}, \infty\}\}$ is an interval partition of $\{z_1, \ldots, z_T\}$, and $\mathcal{P}'$ could contain all the possible regimes. The minimizer $\hat{\mathcal{P}}$ is obtained by considering all possible regimes. In addition, the transition matrices are jointly estimated based on the final estimated regimes $\hat{\mathcal{P}}$.

In this last step, we solve the minimal partition problem in Equation (4.4). The tuning parameter ω is used to control the number of the estimated thresholds to avoid overestimation, since $\sum_{\mathcal{T} \in \mathcal{P}'} L(\mathcal{T})$ is usually monotonically decreasing in the number of the possible regimes $|\mathcal{P}'|$ since the TAR model becomes more complex and flexible as $|\mathcal{P}'|$ increases. Noting that the estimation accuracy is greatly affected by the choice of the tuning parameter ω, we discuss it both theoretically and practically in Sections 4.4 and 4.5.

4.3.1 Algorithm

In this section, we present the algorithm of the dynamic programming approach.

Algorithm 4: Penalized Dynamic Programming

1. Input: x_t, z_t, ngrid $\in \mathbb{R}^+$, and parameters $\lambda, \omega > 0$. Let

$B_i = \min_{s=1,\ldots,i} B_{s-1} + L\left(\mathcal{T}_{(z_{\pi(s)}, z_{\pi(i)})}\right) + \omega$

2. Initialize $\tilde{\mathcal{A}} = \emptyset$, $\tilde{\mathcal{B}} = \emptyset$, $\tilde{\mathcal{A}}^* = \{n\}$, temporary variable $e' = n$, and $B_0 = -\omega$.

for $e \leftarrow 1$ *to* n **do**
  $B_e = \infty$;
  if $e/ngrid = c$ *for* $c \in \mathbb{Z}^+$ **then**
    for $s \leftarrow 1$ *to* e **do**
      $\tilde{\mathcal{B}} \leftarrow \tilde{\mathcal{B}} \cup \left\{B_{s-1} + \omega + L\left(\mathcal{T}_{(z_{\pi(s)}, z_{\pi(e)})}\right)\right\}$
    if $\min \tilde{\mathcal{B}} \leq B_e$ **then**
      $B_e \leftarrow \min \tilde{\mathcal{B}}$; $\tilde{\mathcal{A}} \leftarrow \tilde{\mathcal{A}} \cup \left\{\arg\min \tilde{\mathcal{B}} - 1\right\}$

while $e' \neq 0$ **do**
  $e' \leftarrow e'$th element in $\tilde{\mathcal{A}}$;
  $\tilde{\mathcal{A}}^* \leftarrow \{e'\} \cup \tilde{\mathcal{A}}^*$

3. Output: The set of estimated thresholds $\tilde{\mathcal{A}}^*$.

The key idea for this algorithm is to compute the minimum of $\sum_{\mathcal{T} \in \mathcal{P}'} L(\mathcal{T}) + \omega|\mathcal{P}'|$ recursively. For every $e \in 1, \ldots, n$, we enumerate the position of all the possible order of the thresholds that less than e, $s \in 1, \ldots, e-1$, then we change the minimization problem at e into the minimization problem at s, that is $L\left(\mathcal{T}_{(z_{\pi(s)}, z_{\pi(e)})}\right) + \omega$ and the cost on the previous regime, that is B_{s-1}. The optimal segmentation can be recovered by recursively taking the position of the minimum loss in $\tilde{\mathcal{B}}$.

The computational cost of this algorithm is of order $O\left(n^2 T_L(n)\right)$, where $T_L(n)$ is the computational cost of solving $L(\mathcal{T})$ with $|\mathcal{T}| = n$ (Friedrich et al. [2008]). Note that the dynamic programming approach is a type of exact search method that can find the global optimum (Equation (4.4)), but it is computationally expensive compared to other methods. (See Figure 4.4 for more details.)

4.4 Theory

In this section, we establish the consistency of the dynamic programming approach proposed in Section 4.3. We make the following assumptions.

Assumption C1. $\{\boldsymbol{\epsilon}_t\}$ *is a sequence of i.i.d. sub-Weibull random variables with bounded continuous and positive density and sub-Weibull constant* K_ϵ *and sub-Weibull parameter* $\varkappa_c > 0$*; specifically, there exist constants* K_ϵ *and* $\varkappa_c > 0$ *such that* $\|\boldsymbol{\epsilon}_t\|_\psi \leq K_\epsilon$ *where* $\|\boldsymbol{\epsilon}_t\|_\psi := \sup_{c \geq 1 \wedge \varkappa_c} c^{-\frac{1}{\varkappa_c}} \left(\mathbb{E}\left|\boldsymbol{\epsilon}_t\right|^c\right)^{1/c}$.

Assumption C2. *For each* $j = 1, 2, \ldots, m_0 + 1$*, the process*

$$\boldsymbol{x}_t = \sum_{k=1}^{K} \boldsymbol{A}^{(k,j)} \boldsymbol{x}_{t-k} + \boldsymbol{\epsilon}_t$$

is sub-Weibull with sub-Weibull parameter $\varkappa_1 > 0$ *and* β*-mixing stationary with a geometrically decaying mixing coefficient* b_n*; specifically, there exist constants* $c_b > 0$ *and* $\varkappa_2 > 0$ *such that for all* $n \in \mathbb{N}$*,* $b(n) \leq \exp(-c_b n^{\varkappa_2})$ *and for all* $t, \tau > 0$*,* $(\boldsymbol{x}_t, \ldots, \boldsymbol{x}_{t+n}) \overset{d}{=} (\boldsymbol{x}_{t+\tau}, \ldots \boldsymbol{x}_{t+\tau+n})$*, where* $\overset{d}{=}$ *denotes equality in distribution. Moreover,* $\mathbb{E}[\boldsymbol{x}_t] = \mathbf{0}_{p \times 1}$*. In addition, assume* $\varkappa_0 < 1$*, where* $\varkappa_0 := \left(\frac{2}{\varkappa_1} + \frac{1}{\varkappa_2}\right)^{-1}$.

Assumption C3. *The matrices* $\boldsymbol{A}^{(.,j)}$ *are sparse for* $j = 1, \ldots, m_0 + 1$*. More specifically, for all* $h = 1, 2, \ldots p$ *and* $j = 1, 2, \ldots m_0 + 1, d_{hj} \ll p$*, i.e.,* $d_{kj}/p = o(1)$*. Moreover, there exists a positive constant* $M_A > 0$ *such that*

$$\max_{1 \leq j \leq m_0 + 1} \left\| \boldsymbol{A}^{(.,j)} \right\|_\infty \leq M_A.$$

Assumption C4. *The minimal jump size* v *is defined as*

$$v := \min_{1 \leq j \leq m_0} \left\| \boldsymbol{A}^{(.,j+1)} - \boldsymbol{A}^{(.,j)} \right\|_2,$$

where $+\infty > v > 0$*. Moreover, there exist constants* l *and* u *such that* $r_j \in [l, u]$ *for* $1 \leq j \leq m_0$.

Assumption C5. $\{z_t\}$ *is a* β*-mixing stationary process with a geometric decaying mixing coefficient and positive density. In addition,* $\mathbb{E}|z_t|^{2+\iota} < \infty$ *for* $\iota > 0$.

Assumption C6. *Let* $\Delta_n = \min_{1 \leq j \leq m_0 + 1} |r_j - r_{j-1}|$*. Then,*

$$\Delta_n \geq C_\delta \left(\log\left(\max\left\{p^2 K, n\right\}\right)\right)^{2/\varkappa_0 + \xi} m_0 d_n^{*3} / \left(n v^2\right),$$

and ξ is a small positive constant.

The above assumptions are natural in high-dimensional settings and commonly used in the literature. Assumptions C1 and C2 are utilized to derive appropriate concentration inequalities that are necessary to verify the asymptotic properties of the proposed methodology and have been used in the literature [Li et al., 2012, Wong et al., 2020]. See more details in Zhang et al. [2022]. Assumption C3 ensures the sparsity of the model and is needed to quantify the effect of model misspecification, since exact recovery of threshold values is not possible. A similar assumption has been used in Safikhani and Shojaie [2020] in the context of change point detection. Further, Assumption C4 puts a minimum jump size on the transition matrices ensuring a detectable change occurred at threshold r_j.

Under the assumptions above, we provide the theoretical results of dpTAR. Before we prove the consistencies of the estimators, we first prove the following two propositions, and Theorem 11 follows from Propositions 9 and 10 immediately. Note that the restricted eigenvalue condition and deviations bounds, two inequalities that are important to prove Propositions 9 and 10, are provided in the appendix (Proposition 22).

Proposition 9. *Under Assumptions C1 to C6, there exist constants $C_0 > 0$, $c_2 > 5$, $c_4 > 0$, and $c_5 > 0$ such that with probability at least $1 - \delta_5$,*

(a) For each estimated regime $\hat{\mathcal{T}} = (s, e] \in \hat{\mathcal{P}}$ containing one and only one true threshold r, it holds that

$$\min\left\{\left|\mathcal{T}_{(s,r)}\right|, \left|\mathcal{T}_{(r,e)}\right|\right\} \leq C_0\left(\frac{d_n^* \lambda^2 + \omega}{v^2}\right);$$

(b) for each estimated regime $\hat{\mathcal{T}} = (s, e] \in \hat{\mathcal{P}}$ containing exactly two true thresholds, $r_1 < r_2$, it holds that

$$\max\left\{\left|\mathcal{T}_{(s,r_1)}\right|, \left|\mathcal{T}_{(r_2,e)}\right|\right\} \leq C_0\left(\frac{d_n^* \lambda^2 + \omega}{v^2}\right);$$

(c) for any two consecutive estimated regimes $\hat{\mathcal{T}}_1, \hat{\mathcal{T}}_2 \in \hat{\mathcal{P}}$, the regime $\hat{\mathcal{T}}_1 \cup \hat{\mathcal{T}}_2$ contains at least one true threshold; and

(d) no estimated regime $\hat{\mathcal{T}} \in \hat{\mathcal{P}}$ contains strictly more than two true thresholds,

where

$$\begin{aligned}\delta_5 = {} & 2\exp\left(-c_2 \log\left(\max\left\{p^2K, n\right\}\right) + 3\log n\right) \\ & + \exp\left(-c_4 \log\left(\max\left\{p^2K, n\right\}\right)\left(\log\left(\max\left\{p^2K, n\right\}\right)\right)^{1-\varkappa_0/2} + 3\log n\right) \\ & + \exp\left(-c_5\left(\log\left(\max\left\{p^2K, n\right\}\right)\right)^{2/\varkappa_0} + 3\log n\right).\end{aligned} \tag{4.5}$$

Proposition 9 demonstrates that we can only have at most one true threshold that closes to a given estimated threshold for a proper choice of λ and ω. In addition, no estimated regime contains more than two true thresholds.

Proposition 10. *Suppose Assumptions C1 to C4 hold and that $m_0 \leq |\hat{\mathcal{P}}| - 1 \leq 2m_0$. Then, $|\hat{\mathcal{P}}| = m_0 + 1$ with probability $1 - \delta_6$, where*

$$\begin{aligned}\delta_6 = {} & 2\exp\left(-c_2 \log\left(\max\left\{p^2K, n\right\}\right) + 5\log n\right) \\ & + \exp\left(-c_4 \log\left(\max\left\{p^2K, n\right\}\right)\left(\log\left(\max\left\{p^2K, n\right\}\right)\right)^{1-\varkappa_0/2} + 5\log n\right) \\ & + \exp\left(-c_5\left(\log\left(\max\left\{p^2K, n\right\}\right)\right)^{2/\varkappa_0} + 5\log n\right)\end{aligned} \tag{4.6}$$

for $c_2 > 5$, and $c_4, c_5 > 0$.

Proposition 10 states that the estimated number of thresholds is consistent to the number of true thresholds under certain conditions, which are verified in Proposition 9. With the results from Propositions 9 and 10, Theorem 11 shows the dpTAR consistently estimates the number and values of thresholds.

Theorem 11. *Under Assumptions C1 to C6, there exist estimated thresholds $\{\hat{r}_j\}_{j=1}^{\hat{m}}$ with tuning parameters*

$$\lambda = c_\lambda \left(\log\left(\max\left\{p^2K, n\right\}\right)\right)^{1/\varkappa_0} d_n^* \quad \text{and } \omega = C_\omega\left(m_0 + 1\right) d_n^{*3}\left(\log\left(\max\left\{p^2K, n\right\}\right)\right)^{2/\varkappa_0} \tag{4.7}$$

such that

$$\mathbb{P}\left(\hat{m} = m_0\right) \to 1, \tag{4.8}$$

and

$$\mathbb{P}\left(\max|\hat{r}_j - r_j| \leq \frac{m_0 C_0 d_n^{*3}\left(\log\left(\max\left\{p^2K, n\right\}\right)\right)^{2/\varkappa_0}}{nv^2}\right) \to 1, \tag{4.9}$$

where c_λ, C_ω, C_0 *are positive constants.*

Note that the minimizer of Equation (4.4) is not necessary to be unique, and the consistency rate in Equation (4.9) holds for any minimizer of Equation (4.4). When $p = cn^\kappa$, where $c > 0$ and $\kappa \in (0, 1)$, the dynamic programming approach can be applied to low-dimensional time series. The consistency results would be similar to those in Equation (4.9).

We also compare the results in Theorem 11 with theoretical results developed in Chapter 3. We first compare the main difference in the assumptions. Instead of assuming the minimal jump size v is a constant that is independent of sample size n in Chapter 3, we allow v changes with n in Assumption C6, more specifically $v \to 0$ when $n \to \infty$. This means we can handle more complex situations that the minimal jump size is small with sufficient large sample size, which can also be verified by the simulation results in Section 4.6. Moreover, the assumption on the minimal spacing in Assumption C6 is different from Assumption B6. Let $\Delta_n^{'}$ represent the minimal spacing, then $n\Delta_n^{'} > m_0^2 d_n^{*2}\left(d_n^* \log\left(p^2K\right)\right)^3$ for hdTAR and $n\Delta_n^{'} > m_0^2 d_n^{*2}\left(p^2K d_n^* \log\left(p^2K\right)\right)^3$ for mvTAR in Chapter 3. The dpTAR may need larger minimal spacing than hdTAR or mvTAR when $n >> p^2K$. This also matches with our simulation results in Tables C.1 to C.3. Next, we consider the consistency rate. Recall that the error bound established in Chapter 3 is of order $m_0\left(n\Delta_n^{'}\right)^{3/2} d_n^{*2}/n$. With the constraint on $n\Delta_n^{'}$, the consistency rate becomes $m_0^4 d_n^{*\frac{19}{2}}\left(\log\left(p^2K\right)\right)^{9/2}/n$ for hdTAR and $m_0^4 d_n^{*\frac{19}{2}}\left(p^2K\log\left(p^2K\right)\right)^{9/2}/n$ for mvTAR. It is obvious that the consistency rate of order $m_0 d_n^{*3}\left(\log\left(\max\left\{p^2K, n\right\}\right)\right)^{2/\varkappa_0}/\left(nv^2\right)$ achieved by the dpTAR is sharper when $\varkappa_0 \geq 2/3$ (the constraint in Chapter 3) even we assume v is a constant for the dpTAR.

4.5 Tuning Parameter Selection

We next provide guidance on selecting the tuning parameters for our dynamic programming approach.

λ_n λ_n is chosen by cross validation.

$\boldsymbol{\omega_n}$ Selecting ω_n is in general difficult. For all simulation studies, we use the similar methods proposed by Haynes et al. [2014] to choose ω_n. Let $m(\omega)$ be the number of thresholds with the given tuning parameter ω that is optimal for Equation (4.4). For all $m(\omega)$, we cluster the differences in the loss function Equation (4.2) into two subgroups, small and large. If removing a threshold only leads to a small decrease in loss Equation (4.2), then the removed threshold is likely redundant. In contrast, true thresholds lead to larger decrease. We choose the smallest decrease in the second group as the optimal value of ω. To this end, we first calculate the values of loss function Equation (4.2) for a range of ωs with $m(\omega)$s decreasing one by one, denoted as $L_0^*, L_1^*, \ldots, L_{m(\omega)}^*$. Then, ω is selected as the maximum values among $L_{j+1}^* - L_j'$ for $j = 0, 1, \ldots, m(\omega) - 1$.

z_t^* The true switching variable is selected by accumulated eBIC. Let p' be the number of switching variables. Set $\hat{m}_l$ be the number of estimated thresholds for each $z_{t,l}$. For each $z_{t,l}$, we use eBIC [Wang and Zhu, 2011] across all regimes. For each switching variable $z_{t,l}$, $l \in 1, 2, \ldots, p'$, and $j = 1, 2, \ldots, \hat{m}_l + 1$, the eBIC for interval $[\hat{r}_{j-1}, \hat{r}_j]$ is defined as

$$\text{eBIC}\left(j, z_{t,l}\right) = \log\left(\text{SSE}_{l,j} / \left|\mathcal{T}_{(\hat{r}_j - \hat{r}_{j-1})}\right|\right) + \frac{\left\|\hat{\boldsymbol{\theta}}^l_{\hat{r}_{j-1}, \hat{r}_j}\right\|_0}{\left|\mathcal{T}_{(\hat{r}_j - \hat{r}_{j-1})}\right|} \left(\gamma_2 \log\left(pK\right) + \log\left(\left|\mathcal{T}_{(\hat{r}_j - \hat{r}_{j-1})}\right|\right)\right),$$

where $\gamma_2 = 1.5$ that is within the recommended range in Wang and Zhu [2011] as well. z_t^* is selected as:

$$z_t^* = \arg\min_{z_{t,l}} \sum_j \text{eBIC}\left(j, z_{t,l}\right). \tag{4.10}$$

4.6 Simulations

In this section we present numerical experiments for dynamic programming approach. Below are the simulation scenarios we considered to evaluate the performance of our methods compared to existing methods. There are four switching variables, and only one switching

variable is used to generate data. The true switching variable is generated with AR(1) process with coefficient 0.6. The rest of switching variables are generated as: one with AR(1) process with the same coefficient as the true switching variable, one with AR(1) process with the different coefficient -0.5, one with $N(0, 0.02I)$.

Simulation Scenario 1 (Changes of minimal jump size) In this scenario, $T = 150$, $p = 15$, and $K = 1$. The auto-regressive coefficients are chosen to have the same structure but different values. There is only one threshold with value 4. The difference between non-zero element of auto-regressive coefficients in two regimes is denoted as M_δ, which is ranging from $(0.9, 1.2)$.

Simulation Scenario 2 (Changes of minimal spacing) In this scenario, $p = 15$, $T = 100$, and $K = 1$. The auto-regressive coefficients are chosen to have the same structure and same values. There are two thresholds r_1 and r_2, where the values of (r_1, r_2) are $(3.9, 6)$, $(4, 6)$, $(4.1, 6)$, and $(4.2, 6)$.

Simulation Scenario 3 (Simple high-dimensional $\boldsymbol{A}$ with uncorrelated error) In this scenario, $T = 80$, $p = 100$, and $K = 1$. There is only one threshold value $r_1 = 5$. The auto-regressive coefficients are chosen to have the same structure as in Scenario 1 but with different values.

We perform 100 simulations with all scenarios. Note that the number of thresholds is set to be known for Tsay (1998), Li (2016) and TVAR, and all these three methods are not applicable in the last scenario (Scenario 3) due to high-dimensionality of model. To compare all methods, we assume we know the true switching variable.

Settings	Threshold(s)	Methods	Mean	Std	Selection Rate
Scenario 1 $M_\delta = 1.2$		dpTAR	4.01	0.11	1.00
		hdTAR	4.04	0.09	1.00
	4	Tsay (1998)	3.99	0.10	1.00
		TVAR (2001)	4.32	0.88	0.97
		Li (2016)	3.76	0.36	1.00
Scenario 1 $M_\delta = 1.1$		dpTAR	4.00	0.09	1.00
		hdTAR	4.04	0.09	1.00
	4	Tsay (1998)	4.25	1.01	0.93
		TVAR (2001)	4.17	0.95	0.96
		Li (2016)	3.74	0.41	1.00
Scenario 1 $M_\delta = 1$		dpTAR	4.01	0.11	1.00
		hdTAR	4.04	0.10	1.00
	4	Tsay (1998)	5.51	1.82	0.58
		TVAR (2001)	4.07	1.05	0.94
		Li (2016)	3.70	0.44	1.00
Scenario 1 $M_\delta = 0.9$		dpTAR	4.02	0.24	1.00
		hdTAR	4.04	0.20	0.94
	4	Tsay (1998)	7.27	1.18	0.10
		TVAR (2001)	4.05	1.11	0.93
		Li (2016)	3.64	0.50	1.00

Table 4.1: Mean and standard deviation of estimated thresholds, the percentage of simulation runs where thresholds are correctly detected (selection rate) in simulation Scenario 1. If the estimated thresholds within one standard deviation of true threshold, we consider the estimated thresholds are correctly detected.

	Method	REE	SD(REE)	FPR	TPR
Scenario 1 $M_\delta = 1.2$	dpTAR	0.29	0.04	0.18	1.00
	hdTAR	0.30	0.05	0.16	1.00
	Tsay (1998)	0.32	0.06	0.05	0.93
	TVAR (2001)	0.72	0.48	1.00	1.00
	Li (2016)	0.22	0.07	1.00	1.00
Scenario 1 $M_\delta = 1.1$	dpTAR	0.36	0.07	0.06	0.93
	hdTAR	0.35	0.06	0.16	1.00
	Tsay (1998)	0.40	0.17	0.06	0.95
	TVAR (2001)	0.74	0.48	1.00	1.00
	Li (2016)	0.27	0.09	1.00	1.00
Scenario 1 $M_\delta = 1$	dpTAR	0.41	0.08	0.05	0.93
	hdTAR	0.40	0.07	0.16	1.00
	Tsay (1998)	0.65	0.33	0.10	0.96
	TVAR (2001)	0.81	0.86	1.00	1.00
	Li (2016)	0.31	0.09	1.00	1.00
Scenario 1 $M_\delta = 0.9$	dpTAR	0.47	0.07	0.17	1.00
	hdTAR	0.47	0.07	0.15	1.00
	Tsay (1998)	1.09	0.25	0.82	0.96
	TVAR (2001)	1.04	3.18	1.00	1.00
	Li (2016)	0.35	0.10	1.00	1.00

Table 4.2: Results of parameter estimation for simulation Scenario 1. The table shows mean and standard deviation of relative estimation error (REE), true positive rate (TPR), and false positive rate (FPR) for estimated coefficients.

Note that we put all results of Scenario 2 in the Appendix C.0.2, since hdTAR works better than dpTAR in Scenario 2. For all results shown in the tables or pictures in this section, the Mean and Std are computed only based on the cases that can correctly select the number of thresholds. In addition, the selection rate of Tsay (1998), Li (2016), TVAR,

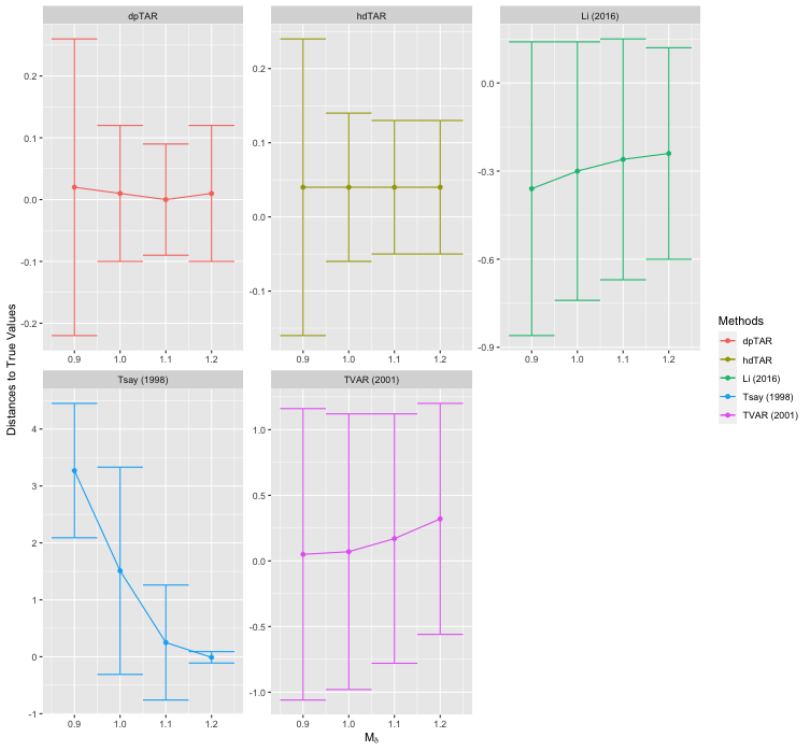


Fig 4.1: Distance between the estimated thresholds and true thresholds for simulation Scenario 1. The error bar represents one standard deviation.

and hdTAR is based on whether the estimated thresholds are within one standard deviation of true threshold. In Scenario 1, we gradually decrease the minimum jump size between two regimes. Table 4.1 and figures 4.1 and 4.2 summarize the results of threshold estimation in Scenario 1. Figure 4.1 shows the difference between the estimated thresholds and true thresholds $(\hat{r}_j - r_j)$ in simulation Scenario 1. The error bar represents one standard deviation. From Table 4.1, and are also displayed in Figure 4.1, dpTAR and hdTAR outperform other methods for all cases in Scenario 1. When the minimal jump size decreases, dpTAR and hdTAR tend to have larger standard deviations. Moreover, Table 4.1 and figure 4.2 demonstrate dpTAR is the only method that can correctly detect the number of thresholds without knowing the true number of thresholds for all cases in Scenario 1. When the minimum jump size is equal to 0.9, dpTAR achieves 100% threshold detection rate while the threshold detection rate of hdTAR is 0.94. This is expected since hdTAR uses l_1 penalty to

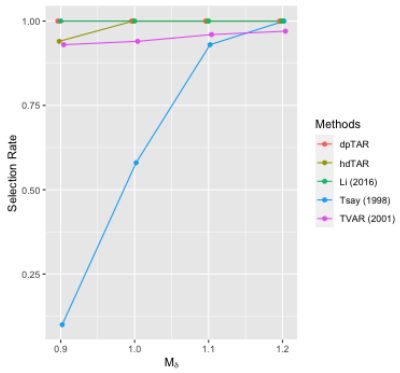


Fig 4.2: Results of selection rate for simulation Scenario 1. If the estimated thresholds within one standard deviation of true threshold, we consider the estimated thresholds are correctly detected.

select thresholds estimators, which may not be able to handle the cases that the minimal jump size is small (mentioned in Lin et al. [2017]). In Scenario 2, we gradually decrease the minimum spacing between two thresholds. When the minimum spacing between two thresholds decreases, the selection rate of all methods drops. In addition, the threshold detection rate of hdTAR is higher than the threshold detection rate of dpTAR (See Table C.1 for more details). Finally, both dpTAR and hdTAR successfully detect the threshold in the high-dimensional setting of Scenario 3 (results are in Table 4.3), and dpTAR has smaller standard deviation compared to hdTAR. In contrast, the other methods are not well suited for this scenario and are not included.

Since Tsay (1998) does not provide coefficient estimates, so we use the standard lasso approach to estimate the parameters given the thresholds obtained by Tsay (1998). Table 4.2 gives the results of auto-regressive parameter estimation in Scenario 1. The results indicate that dpTAR, hdTAR and Tsay (1998) perform well in the first three cases, as measured by their high true positive rates and low false positive rates, while Tsay (1998) does not work well when the minimum jump size is equal to 0.9. Recalling that TVAR does not perform variable selection, all estimated values of transition matrices using this method are non-zero. This leads to true positive and false positive rates that are both equal to 1, which are not meaningful. The auto-regressive parameter estimation results of Scenario 2

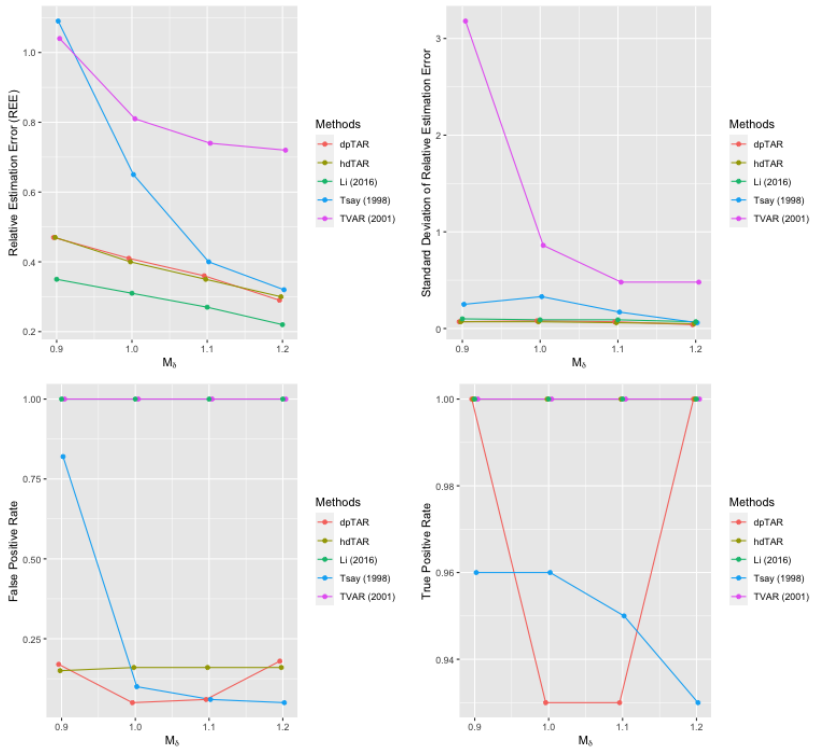


Fig 4.3: Results of transition matrices estimation for simulation Scenario 1.

in Appendix C.0.2 (Table C.2) are not discussed here, since the performance of dpTAR is no better than hdTAR. In Scenario 3, Table 4.4 summarizes the results of auto-regressive parameter estimation, hdTAR and dpTAR have very similar performance, and both of them perform satisfactorily in high dimensional settings.

Settings	Threshold(s)	Methods	Mean	Std	Selection Rate
Scenario 3	5	dpTAR	5.05	0.15	1.00
		hdTAR	5.04	0.20	1.00

Table 4.3: Mean and standard deviation of estimated thresholds, the percentage of simulation runs where thresholds are correctly detected (selection rate) in Scenario 3. If the estimated thresholds within one standard deviation of true threshold, we consider the estimated thresholds are correctly detected.

	Method	REE	SD(REE)	FPR	TPR
Scenario 3	dpTAR	0.58	0.03	0.17	0.87
	hdTAR	0.66	0.03	0.14	0.77

Table 4.4: Results of parameter estimation in Scenario 3. The table shows mean and standard deviation of relative estimation error (REE), true positive rate (TPR), and false positive rate (FPR) for estimated coefficients.

Recalling that we generate 4 switching variables, we use eBIC described in Section 4.5 to select the optimal switching variable for dpTAR. The selection rates for Scenario 1 are listed in Table 4.5. We successfully detect the true switching variable for all cases in both Scenarios. In Scenario 3, we also detect the true switching variable with a 100% detection rate. The results of the selection rate in Scenario 2 are in Appendix C (Table C.3), and they are comparably low since dpTAR cannot correctly detect the true thresholds even if the true switching variable is given.

$M_\delta = 1.1$	$M_\delta = 1.0$	$M_\delta = 0.9$	$M_\delta = 0.8$
1.00	1.00	1.00	1.00

Table 4.5: Results of selection rate in Scenario 1. The table shows the rates of selecting z_t correctly.

Finally, we compare the time spend according to the sample size T and the number of the concurrent time series p. Figure 4.4 shows that dpTAR takes much more time than other methods. In addition, the time cost of dpTAR grows when T or p increases. This demonstrates dpTAR is very time-consuming, which is a commonly known drawback of dynamic programming.

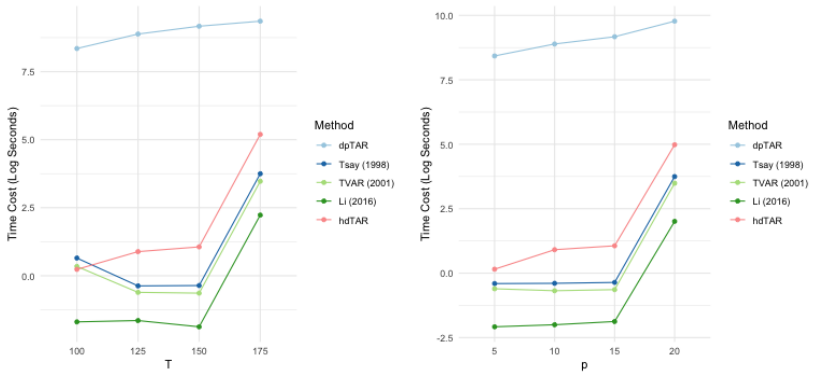


Fig 4.4: Time Cost for Each Method

4.7 Real Data Application

In this section, we utilize the stock data as a demonstration of our dynamic programming approach. The data consists of the top 15 stocks with maximum market capacity in 2005.

We fit a multivariate TAR model of the bi-weekly stock return over the period of Jun 2005 to Jun 2011 and consider the growth rate of the Dow Jones Industrial Average and S&P 500 Index as potential switching variables. For each switching variable y_t, $t = 1, 2, \ldots, T$ over T observations, the growth rate is defined as

$$z_t = \log y_t - \log y_{t-1}, \quad t = 2, 3, \ldots, T.$$

To reduce the non-stationarity, the stock data v_t, $t = 1, 2, \ldots, T$, is also transformed as

$$x_t = \log v_t - \log v_{t-1}, \quad t = 2, 3, \ldots, T.$$

The entire time series consist of $T = 156$ observations from 2005 to 2011. We apply both the dynamic programming approach and the three-step procedure for each potential switching variable and investigate the relationship between the stocks and the overall economic conditions. We first examine the Dow Jones Index as the switching variable. From Figure 4.5, both the dynamic programming approach and the three-step procedure divide the economic pattern into three types: recession, normal, and booming periods. In addition,

the thresholds detected by these two methods are very similar, and both of them successfully identify the great recession periods happen in 2008 and the economic recovery in the middle of 2009. Table 4.6 reports the thresholds detected by both methods.

Methods	Threshold 1	Threshold 2
Three-step Procedure	-0.02	0.01
Dynamic Programming	-0.02	0.04

Table 4.6: Results of detected thresholds based on the Dow Jones Index growth rate. The table shows the values of the selected thresholds by both the dynamic programming approach and the three-step procedure.

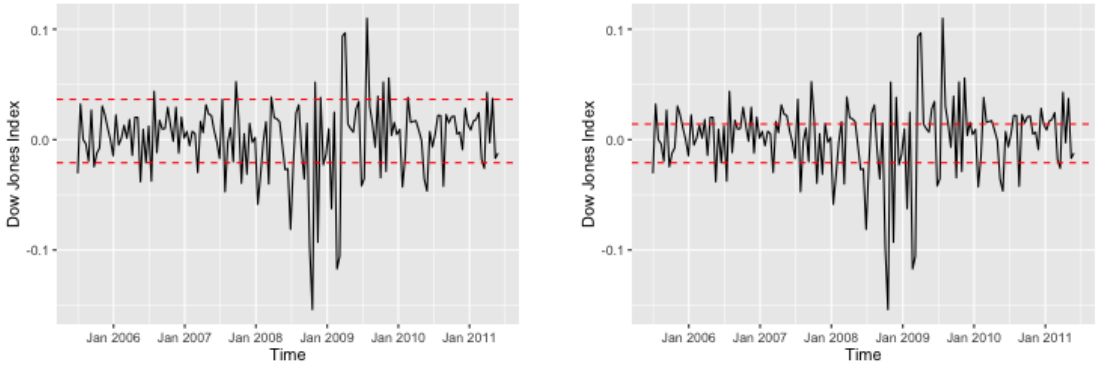


Fig 4.5: The Dow Jones growth rate and detected thresholds using data from 15 stocks. The red dash line shows the estimated threshold. The left panel shows the Dow Jones Index growth rate and detected thresholds based on the dynamic programming approach, while the right panel shows the Dow Jones Index growth rate and detected thresholds based on the three-step procedure. Both methods divide economic patterns into three conditions — recession, normal, and booming periods.

Next, we consider the S&P 500 Index as the switching variable. From Figure 4.6. Both the three-step procedure and the dynamic programming approach divide the economic pat-

tern again in three types, while the dynamic programming approach finds a finer partition. Specifically, the economic periods detected by the dynamic programming approach match more with the economic history. For example, the effects of the Great Recession continued for year 2010 and 2011, which are not correctly recognized by the three-step procedure. Note that the thresholds detected by these two methods are similar, and both of them successfully identify the great recession periods happen in 2008. Table 4.7 reports the thresholds detected by both methods.

Methods	Threshold 1	Threshold 2
Three-step Procedure	-0.01	0.01
Dynamic Programming	-0.01	0.03

Table 4.7: Results of detected thresholds based on the S&P 500 Index growth rate. The table shows the values of the selected thresholds by both the dynamic programming approach and the three-step procedure.

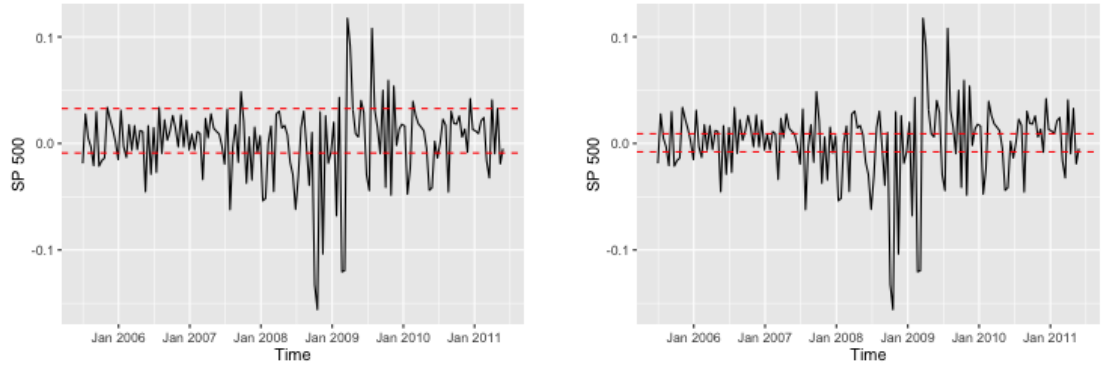


Fig 4.6: The S&P 500 Index growth rate and detected thresholds using data from 15 stocks. The red dash line shows the estimated threshold. The left panel shows the S&P 500 Index growth rate and detected thresholds based on the dynamic programming approach, while the right panel shows the S&P 500 Index growth rate and detected thresholds based on the three-step procedure.

Moreover, we select the switching variable that has more impact on the stocks. The final selected switching variable is S&P 500, which makes sense since the S&P 500 tracks 500 large publicly traded American stocks that are from all sectors of the economy, while the Dow Jones Index only tracks the stock prices of 30 of the biggest American companies.

The Granger causal networks of interactions detected by dynamic programming approach with S&P 500 Index as a switching variable are shown in Figure 4.7. The red links in each panel still represent the interactions that occur in that economic period only. The results show strong interactions among stocks during the recession periods, while there are less interactions during the economic expansion periods, which matches with our findings in Chapter 3.

(a) (b) (c)

Fig 4.7: The Granger causality graph for the top 15 stocks across time. Each vertex represents a stock, and the links display directed interactions between stocks. Figure 4.7a shows the interactions during the recession periods; Figure 4.7b shows the interactions during the normal periods; Figure 4.7c shows the interactions in booming periods. The red links in each panel represent the interactions that occur in that economic period only.

4.8 Discussion

Our key contribution is developing a dynamic programming approach to estimate the number and values of thresholds, as well as the auto-regressive parameters in a possibly high-dimensional TAR model. The proposed algorithm is more accurate than the three-step procedure proposed in Chapter 3 when the minimal jump size is small. We also provide a theoretical guarantee to verify that the dynamic programming approach works well in both

moderate dimension and high dimensions. The consistency rate is sharper than the rate of the three-step procedure. Moreover, we discuss a data-driven method to select the optimal switching variable, which opens the door to the future research direction of the switching variable selection.

A limitation of the proposed procedure is that it requires more time to estimate the thresholds. To speed up the algorithm may need future work. In addition, the tuning parameter selection is tricky. There exists much literature in the change point detection fields [Lavielle and Moulines, 2000, Picard et al., 2005, Haynes et al., 2014], but the problem remains when the data is limited. Moreover, how to choose the proper switching variable may require further investigation as well. Though we provide a data-driven method to identify the switching variable in the TAR model, this method becomes less powerful when two thresholds are very close and fewer observations lie in each regime. In addition, there is no theoretical guarantee that this method gives the optimal choice of switching variable, which is another future research direction.

Chapter 5

DISCUSSION

This chapter summarizes the main contributions together with possible future research directions.

5.1 Summary

In this dissertation, we mainly solve two statistical tasks, namely clustering and analysis of high-dimensional time series.

In the first task, the problem arises when performing clustering, that is, how to identify the underlying connectivity structures such as high-density regions and their connections. To solve this problem, inspired by the mode clustering, a new clustering method presented in Chapter 2 provides an additional attribute label for each cluster. In addition, the clustering results obtained by our method can be further extended to a two-sample tests and a visualization method. The two-sample tests proposed in Chapter 2 utilize local information within each cluster, which are more sensitive than other conventional approaches. The visualization method offers an informative way to display the structure of multi-dimensional data. Moreover, the performance of our improved method is assessed both empirically and theoretically. Both simulation and real data results indicate that our refined clustering method identifies meaningful clusters and provides fine structures within the data. We also derive both statistical and computational guarantees of the proposed method. Thus, the method discussed in Chapter 2 demonstrates good empirical performance and statistical and numerical properties.

The second task investigates the threshold autoregressive (TAR) models and their estimations. Chapter 3 and Chapter 4 investigate the TAR models by using different approaches. In Chapter 3, we develop a three-step procedure with two estimators (mvTAR and hdTAR) for identifying the (unknown) number and values of thresholds and estimating

regime-specific auto-regressive parameters in multivariate TAR models with many components. While both mvTAR and hdTAR are applicable for moderate dimensions, mvTAR is not appropriate for high-dimensional settings. Intuitively, mvTAR enforces all coefficients to change at the same thresholds, while hdTAR allows each coefficient to potentially change at different thresholds. Although this flexibility of hdTAR potentially comes at the cost of a larger number of thresholds in the TAR model, hdTAR yields better empirical performances in both moderate and high dimensions and better theoretical guarantees. Theoretically, mvTAR and hdTAR consistently estimate the model parameters under natural conditions on the distribution and on the level of temporal correlations in the model. As discussed in Section 3.8, when the total number of thresholds is finite, the consistency rates for both models depend explicitly on the effective number of time points, the number of time series components, the number of lags, and the total sparsity of the model. Theoretical results show that the convergence of mvTAR is only guaranteed in low to moderate dimensions and not in high dimensions, which means mvTAR is only suitable for moderate dimensions, while the convergence of hdTAR is guaranteed in both moderate and high dimensions. In Chapter 4, we introduce a dynamic programming approach, which can be used in more complex situations and yields a sharper consistency rate. Chapter 4 solves the questions left by Chapter 3. First, the computational cost of the selection criterion of the three-step procedure in Chapter 3 is exponentially growing with the number of the estimated thresholds, while the dpTAR can solve the problem in polynomial time. Second, hdTAR assumes the minimal jump size is a large enough constant while dpTAR allows the minimal jump size changes with the sample size, and simulation performances also demonstrate that the dpTAR is more accurate than the three-step procedure when the minimal jump size is small. Moreover, we propose a way to detect the optimal switching variable in Chapter 4 when we get more than one switching variable.

5.2 Future Work

In this section, we outline the limitations of the methods discussed in Chapters 2 to 4 and give suggestions for potential research directions based on the work presented.

While The refined clustering method introduced in Chapter 2 works well for the GvHD

data (d = 4), it may not be suitable for any higher dimensional data. When the dimensions get higher, our method fails due to the curse of dimensionality since our method is a nonparametric procedure involving derivative estimation. Future investigation is needed to extend the method to high dimensional settings.

One of the major limitations of the dpTAR is that it is computationally costly. Speeding up the dynamic algorithm, in general, may need more future work. One idea is to combine the algorithm in Chapters 3 and 4, that is, we use the first step to get the over-estimated thresholds and then apply the dynamic programming approach only with these over-estimated thresholds. It is only for computational purposes, and by doing that, we lose the tight theoretical bound. In addition, both the tuning parameter selection and switching variable selection are tricky. The number of estimated thresholds is greatly affected by choice of the tuning parameter. In change point detection fields, Lavielle and Moulines [2000] uses BIC to find the tuning parameter, but Picard et al. [2005] then argues that BIC tends to overestimate the number of change points sometimes. Instead of using BIC type of methods, the method we apply for selecting the tuning parameter is based on the fact that the optimization function is linear in the tuning parameter. However, there are some rooms to improve since not all simulation results work well, especially when the minimal spacing is small and the data is limited. For identifying the switching variable, simulation performances are usually good when there are enough data falls between two thresholds, which means the minimal spacing is large enough. However, there is no theoretical guarantee that this method gives the optimal choice of switching variable, where future investigation is needed.

BIBLIOGRAPHY

Elena Andreou and Eric Ghysels. Detecting multiple breaks in financial market volatility dynamics. *Journal of Applied Econometrics*, 17(5):579–600, 2002. ISSN 08837252, 10991255. URL `http://www.jstor.org/stable/4129273`.

Rosa Arboretti, Arne C Bathke, Eleonora Carrozzo, Fortunato Pesarin, and Luigi Salmaso. Multivariate permutation tests for two sample testing in presence of nondetects with application to microarray data. *Statistical Methods in Medical Research*, 29(1):258–271, 2020.

Ery Arias-Castro, David Mason, and Bruno Pelletier. On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. *Journal of Machine Learning Research*, 17(43):1–28, 2016.

Sipan Aslan, Ceylan Yozgatligil, and Cem Iyigun. Temporal clustering of time series via threshold autoregressive models: Application to commodity prices. *Annals of Operations Research*, in press, 01 2018.

Francis R. Bach and Michael I. Jordan. Learning spectral clustering, with application to speech separation. *Journal of Machine Learning Research*, 7:1963–2001, December 2006. ISSN 1532-4435.

Ali Bagherinia, Behrooz Minaei-Bidgoli, Mehdi Hossinzadeh, and Hamid Parvin. Elite fuzzy clustering ensemble based on clustering diversity and quality measures. *Applied Intelligence*, 49(5):1724–1747, May 2019.

A. Banyaga and D. Hurtubise. *Lectures on Morse Homology*. Texts in the Mathematical Sciences. Springer Netherlands, 2013. ISBN 9781402026966.

Sumanta Basu and George Michailidis. Regularized estimation in sparse high-dimensional

time series models. *Ann. Statist.*, 43(4):1535–1567, 08 2015a. doi: 10.1214/15-AOS1315. URL `https://doi.org/10.1214/15-AOS1315`.

Sumanta Basu and George Michailidis. Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4):1535–1567, 2015b.

Sumanta Basu, Ali Shojaie, and George Michailidis. Network granger causality with inherent grouping structure. *Journal of Machine Learning Research*, 16(13):417–453, 2015.

Sumanta Basu, Xianqi Li, and George Michailidis. Low rank and structured modeling of high-dimensional vector autoregressions. *IEEE Transactions on Signal Processing*, 67(5): 1207–1222, 2019.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009a.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, March 2009b.

Shai Ben-David, Ulrike von Luxburg, and Dávid Pál. *A Sober Look at Clustering Stability*, pages 5–19. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of statistics*, 37(4):1705–1732, 2009.

Kevin Bleakley and Jean-Philippe Vert. The group fused lasso for multiple change-point detection. *Arxiv preprint arXiv:1106.4199*, 06 2011.

J. Richard Bond, Lev Kofman, and Dmitry Pogosyan. How filaments of galaxies are woven into the cosmic web. *Nature*, 380:603, 1996.

Thomas Bonis and Steve Oudot. A fuzzy clustering algorithm for the mode-seeking framework. *Pattern Recognition Letters*, 102:37–43, 2018.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Richard C. Bradley. Basic properties of strong mixing conditions. a survey and some open questions. *Probab. Surveys*, 2:107–144, 2005.

J. V. Braun, R. K. Braun, and H. G. Muller. Multiple changepoint fitting via quasilikelihood, with application to dna sequence segmentation. *Biometrika*, 87(2):301–314, 2000. ISSN 00063444. URL `http://www.jstor.org/stable/2673465`.

G.E. Bredon. *Topology and Geometry*. Graduate texts in mathematics. Springer-Verlag, 1993.

Leo Breiman, William Meisel, and Edward Purcell. Variable kernel estimates of multivariate densities. *Technometrics*, 19(2):135–144, 1977.

Ryan Remy Brinkman, Maura Gasparetto, Shang-Jung Jessica Lee, Albert J. Ribickas, Janelle Perkins, William Janssen, Renee Smiley, and Clay Smith. High-content flow cytometry and temporal data analysis for defining a cellular signature of graft-versus-host disease. *Biology of Blood and Marrow Transplantation*, 13(6):691 – 700, 2007. ISSN 1083-8791.

Boris Brodsky and Boris Darkhovsky. Asymptotically optimal methods of early change-point detection. *Sequential Anal.*, 32(2):158–181, 2013. ISSN 0747-4946. doi: 10.1080/07474946.2013.774611.

Chris Brooks and Ian Garrett. Can we explain the dynamics of the uk ftse 100 stock and stock index futures markets? *Applied Financial Economics*, 12(1):25–31, 2002.

Hansen Bruce. Inference in TAR Models. *Studies in Nonlinear Dynamics & Econometrics*, 2(1):1–16, April 1997.

Sergio A Calderón V and Fabio H Nieto. Bayesian analysis of multivariate threshold autoregressive models with missing data. *Communications in Statistics-Theory and Methods*, 46(1):296–318, 2017.

Miguel Á Carreira-Perpiñán. A review of mean-shift algorithms for clustering. *arXiv preprint arXiv:1503.00687*, 2015.

E. José Chacón, Tarn Duong, and P. M. Wand. Asymptotics for general multivariate kernel density derivative estimators. *Statistica Sinica*, 21:807, 2011.

et al Chacón, José E. Asymptotics for general multivariate kernel density derivative estimators. *Statistica Sinica*, 21(2):807–840, 2011.

José E. Chacón. Clusters and water flows: a novel approach to modal clustering through Morse theory. *ArXiv e-prints*, December 2012.

José E. Chacón and Tarn Duong. Data-driven density derivative estimation, with applications to nonparametric clustering and bump hunting. *Electronic Journal of Statistics*, 7: 499–532, 2013. doi: 10.1214/13-EJS781.

José E Chacón et al. A population background for nonparametric density-based clustering. *Statistical Science*, 30(4):518–532, 2015a.

José E Chacón et al. A population background for nonparametric density-based clustering. *Statistical Science*, 30(4):518–532, 2015b.

K. S. Chan. Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. *The Annals of Statistics*, 21(1):520–533, 1993.

Ngai Hang Chan and Yury A. Kutoyants. On parameter estimation of threshold autoregressive models. *Statistical Inference for Stochastic Processes*, 15(1):81–104, Apr 2012.

Ngai Hang Chan, Chun Yip Yau, and Rong-Mao Zhang. Group lasso for structural break time series. *Journal of the American Statistical Association*, 109(506):590–599, 2014. doi: 10.1080/01621459.2013.866566.

Ngai Hang Chan, Chun Yip Yau, and Rong-Mao Zhang. LASSO estimation of threshold autoregressive models. *Journal of Econometrics*, 189(2):285–296, 2015.

Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for the cluster tree. In *Advances in Neural Information Processing Systems 23*, pages 343–351. Curran Associates, Inc., 2010.

Frédéric Chazal, Brittany Fasy, Fabrizio Lecci, Bertrand Michel, Alessandro Rinaldo, Alessandro Rinaldo, and Larry Wasserman. Robust topological inference: Distance to a measure and kernel distance. *The Journal of Machine Learning Research*, 18(1):5845–5884, 2017.

Cathy WS Chen, Richard H Gerlach, and Ann MH Lin. Falling and explosive, dormant, and rising markets via multiple-regime financial time series models. *Applied Stochastic Models in Business and Industry*, 26(1):28–49, 2010.

Cathy WS Chen, Feng-Chi Liu, and Mike KP So. A review of threshold time series models in finance. *Statistics and its Interface*, 4(2):167–181, 2011.

Hao Chen, Xu Chen, and Yi Su. A weighted edge-count two-sample test for multivariate and object data. *Journal of the American Statistical Association*, 113(523):1146–1155, 2018.

Rong Chen. Threshold variable selection in open-loop threshold autoregressive models. *Journal of Time Series Analysis*, 16(5):461–481, 1995.

Shizhe Chen, Ali Shojaie, and Daniela M Witten. Network reconstruction from high-dimensional ordinary differential equations. *Journal of the American Statistical Association*, 112(520):1697–1707, 2017a.

Y.-C. Chen, C. R. Genovese, and L. Wasserman. Statistical Inference using the Morse-Smale Complex. *arXiv e-prints*, June 2015.

Yen-Chi Chen. A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology*, 1(1):161–187, 2017a.

Yen-Chi Chen. A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology*, 1(1):161–187, 2017b.

Yen-Chi Chen, Christopher R. Genovese, and Larry Wasserman. A comprehensive approach to mode clustering. *Electronic Journal of Statistics*, 10(1):210–241, 2016. doi: 10.1214/15-EJS1102.

Yen-Chi Chen, Christopher R Genovese, Larry Wasserman, et al. Statistical inference using the morse-smale complex. *Electronic Journal of Statistics*, 11(1):1390–1433, 2017b.

Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 17(8):790?799, August 1995a.

Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE transactions on pattern analysis and machine intelligence*, 17(8):790–799, 1995b.

Kacper P Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton. Fast two-sample testing with analytic representations of probability measures. In *Advances in Neural Information Processing Systems*, pages 1981–1989, 2015.

D. Comaniciu and P. Meer. Mean shift analysis and applications. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1197–1203 vol.2, 1999.

Ori Davidov and Yuval Nov. Improving an estimator of hsieh and turnbull for the binormal roc curve. *Journal of Statistical Planning and Inference*, 142(4):872–877, 2012.

Nameirakpam Dhanachandra, Khumanthem Manglem, and Yambem Jina Chanu. Image segmentation using k-means clustering algorithm and subtractive clustering algorithm. *Procedia Computer Science*, 54:764 – 771, 2015. ISSN 1877-0509.

Michael J Dueker, Zacharias Psaradakis, Martin Sola, and Fabio Spagnolo. Multivariate contemporaneous-threshold autoregressive models. *Journal of Econometrics*, 160(2):311–325, 2011.

Tarn Duong. Local significant differences from nonparametric two-sample tests. *Journal of Nonparametric Statistics*, 25(3):635–645, 2013.

Walter Enders, Barry L Falk, and Pierre Siklos. A threshold model of real u.s. gdp and the problem of constructing confidence intervals in tar models. *Studies in Nonlinear Dynamics & Econometrics*, 11:1322–1322, 02 2007.

Moulines Eric, Francis R Bach, and Zaïd Harchaoui. Testing for homogeneity with kernel fisher discriminant analysis. In *Advances in Neural Information Processing Systems*, pages 609–616, 2008.

Ömer Esen. Threshold effects of energy consumption on economic growth in turkey. *Journal of Environmental Management and Tourism*, 3:3–370, 01 2016. doi: 10.14505/jemt.v7. 3(15).02.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pages 226–231. AAAI Press, 1996.

Jianqing Fan and Qiwei Yao. Nonlinear time series. nonparametric and parametric methods. page 384, 01 2005.

Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, Larry Wasserman, Sivaraman Balakrishnan, and Aarti Singh. Confidence sets for persistence diagrams. *The Annals of Statistics*, 42(6):2301–2339, 12 2014.

Christian Francq and Jean-Michel Zakoïan. Estimating the marginal law of a time series with applications to heavy-tailed distributions. *Journal of Business & Economic Statistics*, 31 (4):412–425, 2013. doi: 10.1080/07350015.2013.801776.

PE Freeman, I Kim, and AB Lee. Local two-sample testing: a new tool for analysing high-dimensional astronomical data. *Monthly Notices of the Royal Astronomical Society*, 471 (3):3273–3282, 2017.

Jerome Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshirani. Pathwise coordinate optimization. Technical report, Annals of Applied Statistics, 2007.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL `http://www.jstatsoft.org/v33/i01/`.

F Friedrich, A Kempe, V Liebscher, and G Winkler. Complexity penalized m-estimation. *Journal of Computational and Graphical Statistics*, 17(1):201–224, 2008. doi: 10.1198/106186008X285591. URL `https://doi.org/10.1198/106186008X285591`.

K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1): 32–40, 1975a.

Keinosuke Fukunaga and Larry Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on information theory*, 21(1):32–40, 1975b.

Christopher R Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. The geometry of nonparametric filament estimation. *Journal of the American Statistical Association*, 107(498):788–799, 2012.

Christopher R. Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. Nonparametric ridge estimation. *The Annals of Statistics*, 42(4):1511–1545, 08 2014.

Evarist Giné and Armelle Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 38(6):907 – 921, 2002.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(1):723–773, March 2012.

Thomas Grundy, Rebecca Killick, and Gueorgui Mihaylov. High-dimensional changepoint detection via a geometrically inspired mapping. *Statistics and Computing*, 30(4):1155–1166, 2020.

Fang Han and Han Liu. Transition matrix estimation in high dimensional time series. In *International Conference on Machine Learning*, pages 172–180, 2013.

Bruce E Hansen. Threshold autoregression in economics. *Statistics and its Interface*, 4(2): 123–127, 2011.

Bruce E Hansen and Byeongseon Seo. Testing for two-regime threshold cointegration in vector error-correction models. *Journal of econometrics*, 110(2):293–318, 2002.

Z. Harchaoui and C. Lévy-Leduc. Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association*, 105(492):1480–1493, 2010.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

Kaylea Haynes, Idris A Eckley, and Paul Fearnhead. Efficient penalty search for multiple changepoint problems. *arXiv preprint arXiv:1412.3617*, 2014.

Christian Hennig, Marina Meila, Fionn Murtagh, and Roberto Rocci. *Handbook of cluster analysis*. CRC Press, 2015.

Fushing Hsieh, Bruce W Turnbull, et al. Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *The annals of statistics*, 24(1):25–40, 1996.

Junzhou Huang and Tong Zhang. The benefit of group sparsity. *The Annals of Statistics*, 38(4):1978–2004, 2010.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2 (1):193–218, Dec 1985.

Heinrich Jiang and Samory Kpotufe. Modal-set estimation with an application to clustering. In *Artificial Intelligence and Statistics*, pages 1197–1206. PMLR, 2017.

Zhenyu Jiang, Chengan Du, Assen Jablensky, Hua Liang, Zudi Lu, Yang Ma, and Kok Lay Teo. Analysis of schizophrenia data using a nonlinear threshold index logistic model. *PloS one*, 9(10):e109454, 2014.

Wittawat Jitkrittum, Zoltán Szabó, Kacper P Chwialkowski, and Arthur Gretton. Interpretable distribution features with maximum testing power. *Advances in Neural Information Processing Systems*, 29:181–189, 2016.

Luciana Juvenal and Mark P. Taylor. Threshold adjustment of deviations from the law of one price. *Studies in Nonlinear Dynamics and Econometrics (Online)*, Vol.12(No.3): Article 8, 2008.

R. Killick, P. Fearnhead, and I. A. Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012. doi: 10.1080/01621459.2012.737745. URL `https://doi.org/10.1080/01621459.2012.737745`.

Ilmun Kim, Ann B. Lee, and Jing Lei. Global and local two-sample tests via regression. *Electron. J. Statist.*, 13(2):5253–5305, 2019. doi: 10.1214/19-EJS1648.

B. P. Koester, T. A. McKay, J. Annis, and R. H. et al. Wechsler. A maxbcg catalog of 13,823 galaxy clusters from the sloan digital sky survey. *Astrophysical Journal*, 660(1 I): 239–255, 5 2007a.

Benjamin P. Koester, Timothy A. McKay, James Annis, Risa H. Wechsler, August E. Evrard, Eduardo Rozo, Lindsey Bleem, Erin S. Sheldon, and David Johnston. MaxBCG: A red-sequence galaxy cluster finder. *The Astrophysical Journal*, 660(1):221–238, may 2007b.

Brian Ripley Kung-Sik Chan. *TSA: Time Series Analysis*, 2018. URL `http://homepage.divms.uiowa.edu/~kchan/TSA.htm`. R package version 1.2.

Clifford Lam and Qiwei Yao. Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics*, 40(2):694–726, 2012.

Marc Lavielle and Eric Moulines. Least-squares estimation of an unknown number of shifts in a time series. *Journal of Time Series Analysis*, 21(1):33–59, 2000. doi: https://doi.org/10.1111/1467-9892.00172. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9892.00172`.

Chien-Hui Lee, Bwo-Nung Huang, et al. The relationship between exports and economic growth in east asian countries: A multivariate threshold autoregressive approach. *Journal of Economic Development*, 27(2):45–68, 2002.

Florencia Leonardi and Peter Bühlmann. Computationally efficient change point detection for high-dimensional regression. *arXiv preprint arXiv:1601.03704*, 2016.

Dong Li and Shiqing Ling. On the least squares estimation of multiple-regime threshold autoregressive models. *Journal of Econometrics*, 167(1):240–253, 2012. doi: 10.1016/j.jeconom.2011.11.

Dong Li and Howell Tong. Nested sub-sample search algorithm for estimation of threshold models. LSE Research Online Documents on Economics 68880, London School of Economics and Political Science, LSE Library, October 2016.

Dong Li, Shiqing Ling, and Howell Tong. On moving-average models with feedback. *Bernoulli*, 18(2):735–745, 05 2012. URL `https://doi.org/10.3150/11-BEJ352`.

Jia Li, Surajit Ray, and Bruce G. Lindsay. A nonparametric statistical approach to clustering via mode identification. *Journal of Machine Learning Research*, 8:1687–1723, December 2007. ISSN 1532-4435.

Tao Li, Mitsunori Ogihara, and Sheng Ma. On combining multiple clusterings: An overview and a new perspective. *Applied Intelligence*, 33(2):207–219, October 2010.

Eckhard Liebscher. Towards a unified approach for proving geometric ergodicity and mixing properties of nonlinear autoregressive processes. *Journal of Time Series Analysis*, 26(5):669–689, 2005.

Jiahe Lin and George Michailidis. Regularized estimation and testing for high-dimensional multi-block vector-autoregressive models. *Journal of Machine Learning Research*, 18(117):1–49, 2017. URL `http://jmlr.org/papers/v18/17-055.html`.

Kevin Lin, James Sharpnack, Alessandro Rinaldo, and Ryan J. Tibshirani. A sharp error analysis for the fused lasso, with application to approximate changepoint screening. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6887–6896, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

X. Liu and R. Chen. Regime-switching factor models for high-dimensional time series. *Statistica Sinica*, 26(4):1427–1451, 2016.

Xialu Liu and Rong Chen. Threshold factor models for high-dimensional time series. *Journal of Econometrics*, 216(1):53–70, 2020.

Ming Chien Lo and Eric Zivot. Threshold cointegration and nonlinear adjustment to the law of one price. *Macroeconomic Dynamics*, 5(4):533–576, 2001.

Po-Ling Loh and Martin J Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In *Advances in Neural Information Processing Systems*, pages 2726–2734, 2011.

Frank J. Massey. The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.

Yukio Matsumoto. *An introduction to Morse theory*. American Mathematical Society, 2002.

Andrew McCallum, Kamal Nigam, and Lyle H. Ungar. *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Number 10. Association for Computing Machinery, New York, NY, USA, 2000.

Giovanna Menardi. A review on modal clustering. *International Statistical Review*, 84, 06 2015. doi: 10.1111/insr.12109.

Florence Merlevède, Magda Peligrad, and Emmanuel Rio. *Bernstein inequality and moderate deviations under strong mixing conditions*, volume Volume 5 of *Collections*, pages 273–292. Institute of Mathematical Statistics, 2009.

J. Milnor, M. SPIVAK, and R. WELLS. *Morse Theory. (AM-51), Volume 51*. Annals of mathematics studies. Princeton University Press, 1963.

Soumita Modak and Uttam Bandyopadhyay. A new nonparametric test for two sample multivariate location problem with application to astronomy. *arXiv preprint arXiv:1801.06809*, 2018.

Matthew C Modisett and Edgard M Maboudou-Tchao. Significantly lower estimates of volatility arise from the use of open-high-low-close price data. *North American Actuarial Journal*, 14(1):68–85, 2010.

Marston Morse. Relations between the critical points of a real function of n independent variables. *Transactions of the American Mathematical Society*, 27(3):345–396, 1925.

Y. Nardi and A. Rinaldo. Autoregressive process modeling via the lasso procedure. *Journal of Multivariate Analysis*, 102(3):528 – 549, 2011. ISSN 0047-259X. doi: https://doi.org/10.1016/j.jmva.2010.10.012.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Publishing Company, Incorporated, 1 edition, 2014. ISBN 1461346916.

Fabio H. Nieto. Modeling bivariate threshold autoregressive processes in the presence of missing data. *Communications in Statistics - Theory and Methods*, 34(4):905–930, 2005.

Lizet Viviana Romero Orjuela and Sergio Alejandro Calderón Villanueva. Bayesian estimation of a multivariate tar model when the noise process follows a student-t distribution. *Communications in Statistics-Theory and Methods*, 50(11):2508–2530, 2021.

Margaret Sullivan Pepe. An interpretation for the roc curve and inference using glm procedures. *Biometrics*, 56(2):352–359, 2000.

Margaret Sullivan Pepe. *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press, 2003.

Franck Picard, Stéphane Robin, Marc Lavielle, Christian Vaisse, and Jean-Jacques Daudin. A statistical approach for array cgh data analysis. *BMC bioinformatics*, 6:27, 02 2005. doi: 10.1186/1471-2105-6-27.

Gina-Maria Pomann, Ana-Maria Staicu, and Sujit Ghosh. A two-sample distribution-free test for functional data with application to a diffusion tensor imaging study of multiple sclerosis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 65(3): 395–414, 2016.

Jing Qin and Biao Zhang. Using logistic regression procedures for estimating receiver operating characteristic curves. *Biometrika*, 90(3):585–596, 2003.

Alessandro Rinaldo. Properties and refinements of the fused lasso. *The Annals of Statistics*, 37(5B):2922–2952, 2009.

Alessandro Rinaldo, Aarti Singh, Rebecca Nugent, and Larry Wasserman. Stability of density-based clustering. *The Journal of Machine Learning Research*, 13(1):905–948, 2012.

Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18:9 pp., 2013.

Sebastian Ruder. An overview of gradient descent optimization algorithms. *ArXiv*, abs/1609.04747, 2016.

Takumi Saegusa and Ali Shojaie. Joint estimation of precision matrices in heterogeneous populations. *Electronic journal of statistics*, 10(1):1341, 2016.

Abolfazl Safikhani and Ali Shojaie. Joint Structural Break Detection and Parameter Estimation in High-Dimensional Non-Stationary VAR Models. *Journal of American Statistical Association (Theory and Methods)*, page To Appear, 2020.

David W. Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, Inc, second edition edition, 2015a.

D.W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley Series in Probability and Statistics. Wiley, 2015b.

Luca Scrucca. Identifying connected components in gaussian finite mixture models for clustering. *Comput. Stat. Data Anal.*, 93(C):5–17, January 2016. ISSN 0167-9473.

B. W. Silverman. *Density estimation for statistics and data analysis*. Chapman and Hall, London, 1986. ISBN 9780412246203.

Tamar Sofer, Lee Dicker, and Xihong Lin. Variable selection for high dimensional multivariate outcomes, Oct 2014. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5478010/.

Jacopo Soriano, Li Ma, et al. Probabilistic multi-resolution scanning for two-sample differences. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2): 547–572, 2017.

Ingo Steinwart. Adaptive density level set clustering. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 703–738, 2011.

Gábor J. Székely and Maria L. Rizzo. Testing for equal distributions in high dimensions. *InterStat*, 2004.

Alex Tank, Emily B Fox, and Ali Shojaie. Granger causality networks for categorical time series. *arXiv preprint arXiv:1706.02781*, 2017.

Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

Howell Tong. On a threshold model. *In Chen, C, (ed.) Pattern Recognition and Signal Processing. NATO ASI Series E: Applied Sc.*, 29:575–586, 1978.

Howell Tong. Threshold models in time series analysis—30 years on. *Statistics and its Interface*, 4(2):107–118, 2011.

Howell Tong and Keng S Lim. Threshold autoregression, limit cycles and cyclical data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42:245–268, 07 1980.

Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, 2020. ISSN 0165-1684. doi: https://doi.org/10.1016/j.sigpro.2019.107299. URL https://www.sciencedirect.com/science/article/pii/S0165168419303494.

Ruey S. Tsay. Testing and modeling multivariate threshold models. *Journal of the American Statistical Association*, 93(443):1188–1202, 1998a.

Ruey S Tsay. Testing and modeling multivariate threshold models. *journal of the american statistical association*, 93(443):1188–1202, 1998b.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

Samuel Vaiter, Charles-Alban Deledalle, Gabriel Peyré, Mohamed-Jalal Fadili, and Charles Dossal. The degrees of freedom of the group lasso for a general design. *ArXiv*, abs/1212.6478, 2012.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices, 2010.

Philippe Vieu. A note on density mode estimation. *Statistics & probability letters*, 26(4): 297–307, 1996.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Is a correction for chance necessary? In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1073–1080, New York, NY, USA, 2009. ACM.

Mariia Vladimirova, Stéphane Girard, Hien Nguyen, and Julyan Arbel. Sub-weibull distributions: Generalizing sub-gaussian and sub-exponential properties to heavier tailed distributions. *Stat*, 9(1), Jan 2020.

Lyudmila Yur'evna Vostrikova. Detecting "disorder" in multidimensional random processes. 259(2):270–274, 1981.

Daren Wang, Yi Yu, and Alessandro Rinaldo. Optimal covariance change point localization in high dimension. *arXiv preprint arXiv:1712.09912*, 2017.

Daren Wang, Yi Yu, Alessandro Rinaldo, and Rebecca Willett. Localizing changes in high-dimensional vector autoregressive processes. *arXiv preprint arXiv:1909.06359*, 2019.

Tao Wang and Lixing Zhu. Consistent tuning parameter selection in high dimensional sparse linear regression. *Journal of Multivariate Analysis*, 102(7):1141 – 1151, 2011. ISSN 0047-259X.

Xiao-Feng Wang and De-Shuang Huang. A novel density-based clustering framework by using level set method. *IEEE Transactions on knowledge and data engineering*, 21(11): 1515–1531, 2009.

Larry Wasserman. *All of Nonparametric Statistics (Springer Texts in Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

Laurence Watier and Sylvia Richardson. Modelling of an epidemiological time series by a threshold autoregressive model. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 44(3):353–364, 1995.

Kam Chung Wong, Zifan Li, and Ambuj Tewari. Lasso guarantees for β-mixing heavy-tailed time series. *Annals of Statistics*, 48(2):1124–1142, 2020.

Fan Yang, Tao Li, Qifeng Zhou, and Han Xiao. Cluster ensemble selection with constraints. *Neurocomputing*, 235:59 – 70, 2017.

Yaxing Yang and Shiqing Ling. Self-weighted lad-based inference for heavy-tailed threshold autoregressive models. *Journal of Econometrics*, 197(2):368 – 381, 2017. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2016.11.009. URL `http://www.sciencedirect.com/science/article/pii/S0304407617300015`.

Chun Yip Yau, Chong Man Tang, and Thomas C. M. Lee. Estimation of multiple-regime threshold autoregressive models with structural breaks. *Journal of the American Statistical Association*, 110(511):1175–1186, 2015. doi: 10.1080/01621459.2014.954706.

Donald G. York, J. Adelman, and Jr. et al. John E. Anderson. The sloan digital sky survey: Technical summary. *The Astronomical Journal*, 120(3):1579–1587, sep 2000.

Kashif Yousuf. Variable screening for high dimensional time series. *Electron. J. Statist.*,

12(1):667–702, 2018. doi: 10.1214/18-EJS1402. URL https://doi.org/10.1214/18-EJS1402.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Kunhui Zhang, Abolfazl Safikhani, Alex Tank, and Ali Shojaie. Penalized estimation of threshold auto-regressive models with many components and thresholds. *Electronic Journal of Statistics*, 16(1):1891 – 1951, 2022. doi: 10.1214/22-EJS1982. URL https://doi.org/10.1214/22-EJS1982.

Jiayu Zhou, Jun Liu, Vaibhav A. Narayan, and Jieping Ye. Modeling disease progression via fused sparse group lasso. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012, 08 2012. doi: 10.1145/2339530.2339702.

Hui Zou, Trevor Hastie, and Robert Tibshirani. On the "degrees of freedom" of the lasso. *The Annals of Statistics*, 35(5):2173–2192, 10 2007.

Appendix A

SUPPLEMENTARY MATERIALS FOR CHAPTER 2

To explicitly describe the kernel assumption (K), we need to define a few notations first. A vector $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d)$ of non-negative integers is called a multi-index with $|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_d$ and the corresponding derivative operator is

$$D^\alpha = \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_d}}{\partial x_d^{\alpha_d}},$$

where $D^\alpha f$ is often written as $f^{(\alpha)}$. The assumption (K) requires the followings. Let

$$\mathcal{K} = \left\{ y \mapsto K^{(\alpha)}\left(\frac{x-y}{h}\right) : x \in \mathbb{R}^d, |\alpha| = l \right\},$$

where $K^{(\alpha)}$ is the partial derivative along $\alpha = (\alpha_1, \cdots, \alpha_d)$ direction and let $\mathcal{K}_r^* = \cup_{l=0}^r \mathcal{K}_l$. $\mathcal{K}_r^*$ is the partial derivatives of the kernel function up to fourth-order. We assume that $\mathcal{K}_4^*$ is a VC-type class. that is, there exists constants A, v, and constant envelope b_0 such that

$$\sup_Q N\left(\mathcal{K}_4^*, \mathcal{L}^2(Q), b_0\epsilon\right) \leq \left(\frac{A}{\epsilon}\right)^v,$$

where $N(T', d_T, \epsilon)$ is the ϵ-covering number for a semi-metric set T' with metric d_T and $\mathcal{L}^2(Q)$ is the $\mathcal{L}_2$ norm with respect to the probability measure Q. While this condition looks complicated, the Gaussian kernel and any smooth compactly supported kernel satisfy this condition; see Giné and Guillou [2002].

For simplicity, we describe some notations which will be used across all proofs. We denote $g_s(x) = \nabla s(x)$ be the gradient of $s(x)$ and $H_s(x) = \nabla^2 s(x)$ be the Hessian matrix. Denote $\hat{g}_s(x) = \nabla \hat{s}_n(x)$ and $\hat{H}_s(x) = \nabla^2 \hat{s}_n(x)$, where $\hat{s}_n$ is the estimator of function s. Let $g(x) = \nabla p(x)$ be the gradient of $p(x)$ and $H(x) = \nabla^2 p(x)$ be the Hessian matrix. Denote $\hat{g}_n(x) = \nabla \hat{p}_n(x)$ and $\hat{H}_n(x) = \nabla^2 \hat{p}_n(x)$, where $\hat{p}_n$ is the estimator of function p. For a

smooth function f, recall that we define $\|f\|_{l,\infty}$ be the $\mathcal{L}_\infty$-norm of l-th order derivative. For instance,

$$\|f\|_{0,\infty} = \sup_x \|f(x)\|, \ \|f\|_{1,\infty} = \sup_x \|\nabla f(x)\|_{\max}, \ \|f\|_{2,\infty} = \sup_x \|\nabla^2 f(x)\|_{\max}.$$

Proof of Lemma 1: Recall that $s(x) = \|g(x)\|^2$ and $\nabla s(x) = H(x)g(x)$. Thus, $\mathcal{C} \subset \mathcal{S}$, where $\mathcal{S}$ is the collection of critical points of $s(x)$. In addition, the Hessian matrix of $s(x)$ is

$$\nabla^2 s(x) = T(x),$$

where $T_{kk'}(x) = \left[H^2(x)\right]_{kk'} + \sum_{l=1}^d \frac{\partial H(x)}{\partial x_l} g_l(x)$ and $g_l(x)$ is the l-th component of $g(x)$. For any $m \in \mathcal{C},$, since $\mathcal{C}$ is the collection of critical points of the density p, we have $g(m) = 0$ and the Hessian of slope function $T(m) = H^2(m)$., since we assume s is a Morse function, the eigenvalues of $T(m)$ is non-zero, which implies the eigenvalues of $H(m)$ is non-zero, thus completes the proof. □

Proof of Theorem 2: We will prove the convergence rate and the one-one correspondence. The first assertion (estimated number of local minima equals the population number of local minima) follows from the one-one correspondence.

Our proof consists of two steps. First, we show that there is a one to one mapping between an estimated local minimum and the corresponding true local minimum. Then we can obtain the rate for the distance by using derivative estimation under assumption (K).

The one to one mapping assertion for local minima can be satisfied by modifying the result of Theorem 1 in Chen et al. [2016]. Recall that m is a local minimum of s, let $\hat{m}_n$ be a local minimum of $\hat{s}_n$. From the first two steps of the proof of Theorem 1 in Chen et al. [2016], we can get:

$$\min_{m \in \mathcal{S}} \|\hat{m}_n - m\| \leq \frac{\lambda_0'}{2dc_1}$$

when $\|\hat{p}_n - p\|_{4,\max}$ is sufficiently small. Such a local minimum $\hat{m}_n$ of $\hat{s}_n$ is unique, which means there cannot be another critical point for that given local minimum of s. In other words, each m only corresponds to one $\hat{m}_n$ and vice versa. This completes the proof of one

to one mapping assertion for local minima.

To derive the rate for the distance $\|\hat{m}_n - m\|$, note that $\hat{g}_s(\hat{m}_n) = g_s(m) = 0$. By Taylor's theorem,

$$\hat{g}_s(m) - g_s(m) = \hat{g}_s(m) - \hat{g}_s(\hat{m}_n) = \hat{H}_s(m)(m - \hat{m}_n) + O(\|\hat{m}_n - m\|^2).$$

After rearrangement, we obtain:

$$\hat{m}_n - m = -\hat{H}_s^{-1}(m)(\hat{g}_s(m) - g_s(m)) + O(\|\hat{m}_n - m\|^2) = -\hat{H}_s^{-1}(m)\hat{g}_s(m) + R_n,$$

where $R_n = O\left(\|H_s^{-1}(m) - \hat{H}_s^{-1}(m)\| \cdot \|\hat{g}_s(m)\| + \|\hat{m}_n - m\|^2\right)$, which is a second order term.

Since $H_s(m)$ is a positive definite matrix due to Lemma 1 and assumption (L), the rate of $\hat{m}_n - m$ is determined by the rate of $\hat{g}_s(m)$. By the definition of $\hat{s}$, $\hat{g}_s(x) = \nabla\hat{s}(x) = \hat{H}_n(x)\hat{g}_n(x)$. $\hat{g}_n(x) = \nabla\hat{p}_n(x)$ and $\hat{H}_n(x) = \nabla^2\hat{p}_n(x)$ are the gradient and Hessian matrix of kernel density estimator $\hat{p}(x)$, and $g(x) = \nabla p(x)$ and $H(x) = \nabla^2 p(x)$ are the gradient and Hessian matrix of true density function $p(x)$. Thus,

$$\begin{aligned}
\hat{g}_s(m) = \hat{g}_s(m) - g_s(m) &= \hat{H}_n(m)\hat{g}_n(m) - H_n(m)g_n(m) \\
&= \hat{H}_n(m)\hat{g}_n(m) - H_n(m)\hat{g}_n(m) + H_n(m)\hat{g}_n(m) - H_n(m)g_n(m) \\
&= \left(\hat{H}_n(m) - H_n(m)\right)\hat{g}_n(m) + H_n(m)\left(\hat{g}_n(m) - g_n(m)\right) \\
&= \left(\hat{H}_n(m) - H_n(m)\right)\left(\hat{g}_n(m) - g_n(m)\right) + H_n(m)\left(\hat{g}_n(m) - g_n(m)\right)
\end{aligned}$$

Let $[\beta] = (\beta_1, \beta_2, \ldots, \beta_d)$ be a multi-index (each $\beta_l \in [\beta]$ is a non-negative integer and $|[\beta]| = \sum_{l=1}^{d} \beta_l$). Define $D^{[\beta]} = \frac{\nabla^{\beta_1}}{\nabla x_1^{\beta_1}} \cdots \frac{\nabla^{\beta_d}}{\nabla x_d^{\beta_d}}$ to be the $[\beta]$-th order partial derivative operator Chen [2017b].

Under smoothness condition Chacón [2011],

$$D^{[\beta]}\hat{p}_n(x) - D^{[\beta]}p_n(x) = O(h^2) + O_P\left(\sqrt{\frac{1}{nh^{d+2|[\beta]|}}}\right).$$

Thus, under assumption (K), for a fixed point x,

$$\hat{H}_n(x) - H(x) = O(h^2) + O_P\left(\sqrt{\frac{1}{nh^{d+4}}}\right)$$

$$\hat{g}_n(x) - g(x) = O(h^2) + O_P\left(\sqrt{\frac{1}{nh^{d+2}}}\right)$$

So $\hat{g}_s(m) = O(h^2) + O_P\left(\sqrt{\frac{1}{nh^{d+4}}}\right)$, which leads to

$$\hat{m}_n - m = O(h^2) + O_P\left(\sqrt{\frac{1}{nh^{d+4}}}\right).$$

□

Before we discuss the proof of Theorem 4, we first recall a useful result:

Theorem 12 (Rate of convergence of KDE; page 17 of Genovese et al. [2014])**.** *Assume (P) and (K). Let $\hat{p}_n(x)$ be the kernel density estimator. For each $l = 0, 1, 2, 3, 4$, when $h \to 0$ and $\frac{nh^{d+2l}}{\log n} \to \infty$,*

$$\|\hat{p}_n - p\|_{l,\infty} = O(h^2) + O_P\left(\sqrt{\frac{\log n}{nh^{d+2l}}}\right)$$

Theorem 13 ((Modified) Theorem 2 in Arias-Castro et al. [2016])**.** *Suppose f and $\tilde{f}$ are two smooth functions that are three times differentiable. Given a point x_0, let $(x(t)\colon t > 0)$ be the gradient flow of f starting from x_0, and $(\tilde{x}(t)\colon t > 0)$ be the gradient flow of $\tilde{f}$ starting from the same point x_0. Assume that $x(t)$ ends at the local mode x^* and the eigenvalues of $\nabla^2 f(x^*)$ are in the interval $[v_1, v_2]$ where $\infty > v_2 \geq v_1 > 0$. Then there exists a constant C depends only on f, x_0, v_1, v_2 such that when $max\ \{\|f - \tilde{f}\|_{l,\infty} : l = 0, 1, 2, 3\} < max\ \{C, C^{-1}\}$,*

$$\sup_{t \geq 0} \|\tilde{x}(t) - x(t)\| \leq C max\left\{\sqrt{\|f - \tilde{f}\|_{0,\infty}}, \|f - \tilde{f}\|_{1,\infty}^{\alpha_0}\right\},$$

where $\alpha_0 = \frac{v_1}{v_1 + v_2}$.

Proof of Theorem 4: The main idea for this proof is to reverse the direction of the gradient flows described in Theorem 2 in Arias-Castro et al. [2016], which establish a stability result

for gradient flows of smooth functions f. To apply Theorem 13, the corresponded smooth function $f(x)$ is $s(x)$, and $s(x) = \|\nabla p(x)\|^2$ in our case. Thus, assumption (P) guarantees that $s(x)$ is three times differentiable., since in Theorem 13, it requires $\max\{\|f - \tilde{f}\|_{l,\infty} : l = 0, 1, 2, 3\} < \max\{C, C^{-1}\}$, which means $\max\{\|s(x) - \tilde{s}(x)\|_{l,\infty} : l = 0, 1, 2, 3\}$ is sufficient small. That is $\max\{\|p(x) - \tilde{p}(x)\|_{l,\infty} : l = 0, 1, 2, 3, 4\}$ should be small. By Theorem 12, we can get $\frac{\log n}{nh^{d+8}} \to 0$ if $h \to 0$, which guarantees our assumptions.

Recall that $\mu_{\min}(x)$ and $\mu_{\max}(x)$ are the smallest and largest eigenvalue of $H_s(\pi_x(\infty))$. Thus, all eigenvalues of $H_s(\pi_x(\infty))$ fall into $[\mu_{\min}(x), \mu_{\max}(x)]$, which means $\mu_{\min}(x)$ and $\mu_{\max}(x)$ are corresponding v_1 and v_2 in Theorem 13. Then, we can obtain

$$\sup_{t\geq 0} \|\hat{\pi}_x(t) - \pi_x(t)\| = \left\{ O(h^{2\alpha}) + O_P\left(\left(\frac{\log n}{nh^{d+4}} \right)^{\frac{\alpha}{2}} \right) \right\} \wedge \left\{ O(h) + O_P\left(\sqrt[4]{\frac{\log n}{nh^d}} \right) \right\},$$

where $\alpha = \frac{\mu_{\min}(x)}{\mu_{\max}(x) + \mu_{\min}(x)}$. □

Finally, the proof of Lemma 14 relies on some useful properties from convex optimization. We first recall a useful lemma.

Lemma 14. *According to Chapter 2 in Nesterov [2014], we have several properties below.*

- *Property 1: When a function $f(x)$ has an L-Lipschitz continuous gradient, then*

$$f(x) - f(y) \leq \langle x - y, \nabla f(y) \rangle + \frac{L}{2}\|x - y\|^2 \quad \text{for every } x,\ y \in \mathbb{R}^n. \tag{A.1}$$

  In addition, constant L is greater than or equal to the maximum eigenvalue of Hessian matrix of $f(x)$.

- *Property 2:*

  Let $f^* = f(x^*) = \min_x f(x)$, where x^* is the true minimum of the function $f(x)$. The function $f(x)$ is called C_m strongly convex if and only if there exists a constant $C_m > 0$ such that the $f(x) - \frac{C_m}{2}\|x\|^2$ is a convex function. In addition, for each step t, we have:

$$f^* - f(x_t) \geq (x^* - x_t)^T \nabla f(x_t) + \frac{C_m}{2}\|x^* - x_t\|^2, \tag{A.2}$$

which implies

$$(x_t - x^*)^T \nabla f(x_t) \geq f(x_t) - f^* + \frac{C_m}{2}\|x^* - x_t\|^2. \tag{A.3}$$

- *Property 3: Let $f^* = f(x^*) = 0$, where x^* is the true minimum of the function $f(x)$. Assume function $f(x)$ has an L-Lipschitz continuous gradient. Then, we have:*

$$f(x) \geq \frac{1}{2L}\|\nabla f(x)\|^2 + f^*. \tag{A.4}$$

- *Property 4: By the settings in Property 2 and Property 3, we have:*

$$\|\nabla f(x)\|^2 \geq C_m^2\|x - x^*\|^2 \geq \frac{2(f(x) - f^*)C_m^2}{L} \geq 2f(x)C_m^2/L. \tag{A.5}$$

Proof of Lemma 14: Property 1 can be directly obtained by the definition of L-Lipschitz continuity. For property 2, $f(x)$ is strongly convex, so $\|\nabla f(x)\| \geq C_m\|x - x^*\|$, where C_m is smaller than or equal to the minimum eigenvalue of Hessian matrix of $f(x)$. For property 3, $f(x)$ is L-Lipschitz, so $f(x) \leq \frac{L}{2}\|x - x^*\|^2 + f^*$. According to the fact that $f(x) \geq f^* = 0$, then,

$$\begin{aligned} -f(x_t) + f^* &\leq f(x_{t+1}) - f(x_t) \\ &= f(x_t - \gamma\nabla f(x_t)) - f(x_t) \\ &\leq f(x_t - \frac{1}{L}\nabla f(x_t)) - f(x_t) \\ &\leq -\frac{1}{L}\|\nabla f(x_t)\|^2 + \frac{1}{2L}\|\nabla f(x_t)\|^2 \\ &= -\frac{1}{2L}\|\nabla f(x_t)\|^2. \end{aligned} \tag{A.6}$$

Thus, the results are as desired. The C_m-strongly convexity implies $\|\nabla f(x)\| \geq C_m\|x - x^*\|$ and the L-Lipscthitz gradient implies $f(x) - f^* \leq \frac{L}{2}\|x - x^*\|$. Thus, the Property 4 holds. □

Proof of Theorem 5: From assumptions (A1) and (A2), there exists a ball with certain radius R_0 around each minimum of s such that all points within that ball have all positive

eigenvalues of the Hessian matrix. Let a starting point within a ball to be x_0. Note that within each ball, $s(x)$ is λ_0-strongly convex,, since the Hessian matrix has all of its eigenvalues bounded Nesterov [2014]. The constant λ_0 is from assumption (A2).

According to assumption (P) and (L), s is a continuously differentiable function with Lipschitz continuous gradient and Lipschitz constant L. Consider a minimum $m_j \in \mathcal{S}$ and let $s^* = s(m_j) = 0$. According to Property 3 and Property 4 in Lemma 14, we have:

$$s(x_t) \geq \frac{1}{2L}\|\nabla s(x_t)\|^2 \geq \frac{1}{2L} 2s(x_t)\lambda_0^2/L. \tag{A.7}$$

After rearrangement, we obtain:

$$1 \geq \frac{\lambda_0^2}{L^2}. \tag{A.8}$$

For step $t+1$,

$$\begin{aligned}
\|x_{t+1} - m_j\|^2 &= \|x_t - m_j - \gamma \nabla s(x_t)\| \\
&= \|x_t - m_j\|^2 - 2\gamma(x_t - m_j)^T \nabla s(x_t) + \gamma^2 \|\nabla s(x_t)\|^2 \\
&\leq \|x_t - m_j\|^2 - 2\gamma(s(x_t) - s^* + \frac{\lambda_0}{2}\|m_j - x_t\|^2) + \gamma^2 \|\nabla s(x_t)\|^2 \\
&\leq \|x_t - m_j\|^2(1 - \gamma\lambda_0) - 2\gamma s(x_t) + \gamma^2 \|\nabla s(x_t)\|^2 \\
&\leq \|x_t - m_j\|^2(1 - \gamma\lambda_0) - 2\gamma s(x_t) + \gamma^2 * 2Ls(x_t) \\
&\leq \|x_t - m_j\|^2(1 - \gamma\lambda_0) \\
&\leq \|x_0 - m_j\|^2(1 - \gamma\lambda_0)^{t+1}
\end{aligned} \tag{A.9}$$

The first and third inequalities are due to Equations (A.3) and (A.4). By Equation (A.8), $0 < \gamma\lambda_0 \leq \frac{\lambda_0}{L} \leq 1$. This proves the first statement.

Applying L-Lipschitz again and according to the Property 4 from Lemma 14, we have:

$$\begin{aligned} s(x_{t+1}) - s(x_t) &= s(x_t - \gamma \nabla s(x_t)) - s(x_t) \\ &\leq -\gamma \|\nabla s(x_t)\|^2 + \frac{L\gamma^2}{2}\|\nabla s(x_t)\|^2 \\ &= -\gamma(1 - \frac{L\gamma}{2})\|\nabla s(x_t)\|^2 \\ &\leq -\gamma(1 - \frac{L\gamma}{2})\frac{2(s(x_t) - s^*)\lambda_0^2}{L}. \end{aligned} \tag{A.10}$$

By rearrangements,

$$\begin{aligned} s(x_{t+1}) &\leq s(x_t)\left(1 - 2\gamma\left(1 - \frac{L\gamma}{2}\right)\frac{\lambda_0^2}{L}\right) \\ &= s(x_t)\left(1 - \frac{\lambda_0^2}{L^2} + \lambda_0^2\left(\gamma - \frac{1}{L}\right)^2\right). \end{aligned} \tag{A.11}$$

Recall that x_0 is the initial point. By telescoping, we can get:

$$\begin{aligned} s(x_{t+1}) - s(m) &= s(x_{t+1}) - s^* \\ &\leq s(x_0)\left(1 - \frac{\lambda_0^2}{L^2} + \lambda_0^2\left(\gamma - \frac{1}{L}\right)^2\right)^{t+1} \\ &= (s(x_0) - s(m))\left(1 - \frac{\lambda_0^2}{L^2} + \lambda_0^2\left(\gamma - \frac{1}{L}\right)^2\right)^{t+1}. \end{aligned}$$

, since $0 < \gamma \leq 1/L$, $-\frac{\lambda_0^2}{L^2} + \lambda_0^2\left(\gamma - \frac{1}{L}\right)^2$ lies in range $(0, \frac{\lambda_0^2}{L^2}]$. By Equation (A.8), $\frac{\lambda_0^2}{L^2} \leq 1$, $1 - \frac{\lambda_0^2}{L^2} + \lambda_0^2\left(\gamma - \frac{1}{L}\right)^2 < 1$. This completes the proof. □

Appendix B

SUPPLEMENTARY MATERIALS FOR CHAPTER 3

Notations. We first describe some notations which will be used across all proofs. For a symmetric matrix X, let $\lambda_{\max}(X)$ and $\lambda_{\max}(X)$ denote its maximum and minimum eigenvalues and $|||X|||$ denotes its operator norm $\sqrt{\lambda_{\max}(X'X)}$. For any matrix M, if $\{G_1, G_2, \ldots, G_{g_0}\}$ denote a partition of $\{1, 2, \ldots, |vec(M)|\}$ into g_0 non-overlapping groups, then we use $\|M\|_{2,\infty}$ to denote $\max_{g=1,2,\ldots,g_0} \|vec(M)_{G_g}\|_2$ and $\|M\|_{2,1}$ to denote $\sum_{g=1}^{g_0} \|vec(M)_{G_g}\|_2$, where $vec(M)_{G_g}$ represents all the elements of vectorized form M in G_g group. Let $\mathcal{S} = \{w_1, w_2, \ldots, w_{m_0}\}$, where w_j denotes the j-th order of true threshold. Set $m_0 = |\mathcal{S}|$. Let b_j denotes the order of the j-th estimated threshold in Step 2.

Appendix 3.1: Some Definitions

Sub-Weibull random variable: A random variable U is sub-Weibull [Rudelson and Vershynin, 2013] if there exist constants $K_U > 0$ and $\varkappa' > 0$ such that

$$\|U\|_\psi := \sup_{c \geq 1 \wedge \varkappa'} c^{-\frac{1}{\varkappa'}} \left(\mathbb{E}\,|U|^c\right)^{1/c} \leq K_U; \tag{B.1}$$

Moreover, K_U is called the sub-Weibull constant while $\varkappa' > 0$ is called the sub-Weibull parameter.

Mixing conditions: We follow the definitions in Bradley [2005]. Given the probability space $(\Omega, \mathcal{F}, P)$, for any σ-field $\mathcal{A} \subset \mathcal{F}$, define $L_2(\mathcal{A})$ to be the family of all square integrable $\mathcal{A}$-measurable random variables. For any two σ-fields $\mathcal{A}$ and $\mathcal{B} \subset \mathcal{F}$, we define:

$$\alpha'_n = \sup |P(A \cap B) - P(A)\,P(B)|, \quad A \in \mathcal{A}, B \in \mathcal{B}; \tag{B.2}$$

$$\beta'_n = \sup \frac{1}{2} \sum_{i_1=1}^{I} \sum_{i_2=1}^{J} |P(A_{i_1} \cap B_{i_2}) - P(A_{i_1})\,P(B_{i_2})|, \tag{B.3}$$

where the supremum is taken over all pairs of (finite) partitions $\{A_1, A_2, \ldots, A_I\}$ and $\{B_1, B_2, \ldots, B_J\}$ of Ω such that $A_{i_1} \in \mathcal{A}$ for each i_1 and $B_{i_2} \in \mathcal{B}$ for each i_2. The stochastic process is said to be α-mixing (strongly mixing) if $\alpha'_n \to 0$, and β-mixing if $\beta'_n \to 0$. Note that β-mixing implies α-mixing.

Appendix 3.2: Technical Lemmas

Lemma 15. *Under Assumptions B1, B2 and B5, for* $x \in \mathbb{R}, 1 \leq l, l' \leq p, 1 \leq k \leq K$,

$$x_{((t-k),l)} I\left(z_t \leq x\right) \epsilon_{(t,l')}$$

is sub-Weibull with parameter $\frac{1}{1/\varkappa_1 + 1/\varkappa_c}$;

$$x_{((t-k),l)} I\left(z_t \leq x\right) x_{(t,l')}$$

is sub-Weibull with parameter $\varkappa_1/2$.

Proof of Lemma 15: According to Assumptions B1 and B2, we know $x_{((t-k),l)}$ and $\epsilon_{(t,l')}$ are sub-Weibull with sub-Weibull parameter $\varkappa_1$ and $\varkappa_c$. From Proposition 3 in Vladimirova et al. [2020], we have $x_{((t-k),l)}\epsilon_{(t,l')}$ is sub-Weibull with parameter $\frac{1}{1/\varkappa_1 + 1/\varkappa_c}$. Similarly, $x_{((t-k),l)}x_{(t,l')}$ is sub-Weibull with parameter $\varkappa_1/2$.

Combined with above statement and based on Theorem 1 in Vladimirova et al. [2020], there exists $K_C > 0$ such that for all $y_x \geq 0$, we have:

$$\begin{aligned} \mathbb{P}\left(\left|x_{((t-k),l)} I\left(z_t \leq x\right) \epsilon_{(t,l')}\right| \geq y_x\right) &\leq \mathbb{P}\left(\left|x_{((t-k),l)}\epsilon_{(t,l')}\right| \geq y_x\right) \\ &\leq 2\exp\left(-\left(y_x/K_C\right)^{\left(\frac{\varkappa_1\varkappa_c}{\varkappa_1+\varkappa_c}\right)}\right). \end{aligned} \tag{B.4}$$

By Theorem 1 in Vladimirova et al. [2020] again, $x_{((t-k),l)} I\left(z_t \leq x\right) \epsilon_{(t,l')}$ is sub-Weibull with parameter $\frac{1}{1/\varkappa_1 + 1/\varkappa_c}$. By similar procedure, we can prove $x_{((t-k),l)} I\left(z_t \leq x\right) x_{(t,l')}$ is sub-Weibull with sub-Weibull parameter $\varkappa_1/2$.

Lemma 16. *Under Assumptions B1 to B4, there exist positive constants* C, c_0, c_1, c_2, c_3,

such that for

$$n \geq c_0 \left(\log(p^2K)\right)^{2/\varkappa_0 - 1},$$

with probability at least $1 - c_3\eta_1 - \eta_2$*, we have:*

$$\frac{1}{n}\|\boldsymbol{Z}'\boldsymbol{E}\|_\infty \leq C\frac{\log\left(p^2K\right)}{\sqrt{n}}, \tag{B.5}$$

where $\eta_1 = \exp\left(-c_1 \log\left(p^2K\right)\right)$ *and*

$$\eta_2 = \exp\left(-c_2 \frac{n^{\varkappa_1\varkappa_c/2(\varkappa_1+\varkappa_c)}}{(\log n)^{2\varkappa_1\varkappa_c/(\varkappa_1+\varkappa_c)}} + \log\left(np^2K\right)\right).$$

Proof of Lemma 16: First, we rewrite Equation (B.5) with respect to the switching variable z_t and t as:

$$\max_{1\leq i\leq n, 1\leq l,l'\leq p, 1\leq k\leq K} \frac{1}{n}\left|\sum_{t=1}^{n} x_{((t-k),l)} I\left(z_t \leq z_{\pi(i)}\right)\epsilon_{(t,l')}\right|. \tag{B.6}$$

The main goal is to find a proper rate for Equation (B.6). The indicator term $I(z_t < z_{\pi(i)})$ makes the proof more complicated, since we need to maximize Equation (B.6) w.r.t. t and we have no control on $z_{\pi(i)}$. Hence, we rewrite Equation (B.6) in the following form:

$$\max_{x\in\mathbb{R}, 1\leq l,l'\leq p, 1\leq k\leq K} \frac{1}{n}\left|\sum_{t=1}^{n} x_{((t-k),l)} I\left(z_t \leq x\right)\epsilon_{(t,l')}\right|. \tag{B.7}$$

Similar to Chan et al. [2015], we use the bracketing technique to bound Equation (B.7). To simplify the notation, we denote $x_{t,l}^k = x_{((t-k),l)}$, and let $W_n^{(l',l,k)}(x) = \frac{1}{\sqrt{n}}\sum_{t=1}^{n} x_{t,l}^k I\left(z_t \leq x\right)\epsilon_{(t,l')}$. Define $\Gamma_{(x)}(a) = a_1 I_{(-\infty,x)}(a_2)$ for $a \in \mathbb{R}^2$ and $\mathcal{F} = \left\{\Gamma_{(x)} : x \in \mathbb{R}\right\}$. Write $\Gamma_{(x)}$ as Γ. Let $M_t^{(l',l,k)} = x_{t,l}^k \epsilon_{(t,l')}$ and $Y_{nt}^{(l',l,k)} = \left(M_t^{(l',l,k)}/\sqrt{n}, z_t\right)$ for $l, l' \in 1, 2, \ldots, p$ and $k \in 1, 2, \ldots, K$, then $W_n^{(l',l,k)}(x) = \frac{1}{\sqrt{n}}\sum_{t=1}^{n} M_t^{(l',l,k)} I\left(z_t \leq x\right) = \sum_{t=1}^{n} \Gamma_{(x)}\left(Y_{nt}^{(l',l,k)}\right)$.

For any $x_1 < x_2$, we have:

$$\begin{aligned}
&\mathbb{E}[W_n^{(l',l,k)}(x_1) - W_n^{(l',l,k)}(x_2)]^2 \\
&= \frac{1}{n}\mathbb{E}\left[\sum_{t=1}^{n} M_t^{(l',l,k)} \left[I\left(z_t \le x_1\right) - I\left(z_t \le x_2\right)\right]\right]^2 \\
&= \mathbb{E}\left[\left(M_t^{(l',l,k)}\right)^2 I\left(x_1 < z_t \le x_2\right)\right] \\
&= \left(\mathbb{E}\left(M_t^{(l',l,k)}\right)^2\right)\left(G^{(l',l,k)}\left(x_2\right) - G^{(l',l,k)}\left(x_1\right)\right),
\end{aligned} \tag{B.8}$$

where $G^{(l',l,k)}(x) = \mathbb{E}\left[\left(M_t^{(l',l,k)}\right)^2 I\left(z_t \le x\right)\right] / \mathbb{E}\left(M_t^{(l',l,k)}\right)^2$. Then for fixed l', l, k, we can construct a psudo-metric

$$d\left(x_1, x_2\right) = \sqrt{\left(\mathbb{E}\left(M_t^{(l',l,k)}\right)^2\right)\left|G^{(l',l,k)}\left(x_2\right) - G^{(l',l,k)}\left(x_1\right)\right|}.$$

For any $0 < \delta < 1$, the integral of the bracketing entropy satisfies the following

$$\int_0^{\delta} \sqrt{\log N\left(\epsilon, \mathcal{F}, L_2\right)} d\epsilon \le C \int_0^{\delta} \sqrt{-\log \epsilon} d\epsilon < \infty, \tag{B.9}$$

where $N(\epsilon, \mathcal{F}, L_2)$ denotes the brackets number, that is, the minimum number of ϵ-brackets needed to cover $\mathcal{F}$. Choose a fixed integer q_0 such that $4\delta \le 2^{-q_0} \le 8\delta$. Then, choose a nested sequence of partitions $\mathcal{F}_{q_{u'}}$ of $\mathcal{F}$ indexed by the integer $q \ge q_0$. By the Chaining Lemma [Vaart, 1998], set $P_q = \{\Gamma_{(x)} : x \in B_{q_{u'}}, 1 \le u' \le N_q\}$ such that:

$$\begin{aligned}
&\sum_{q=q_0}^{\infty} 2^{-q} \sqrt{\log N_q} < \int_0^{\delta} \sqrt{\log N\left(\epsilon, \mathcal{F}, L_2\right)} d\epsilon, \\
&\mathbb{E}\Lambda^2\left(B_{q_{u'}}\right) := \frac{1}{n}\mathbb{E}\sum_{t=1}^{n} \sup_{(x_1,x_2)\in B_{q_{u'}}} \left(M_t^{(l',l,k)}\right)^2 \left|I\left(x_1 < z_t \le x_2\right)\right| \le 2^{-2q}.
\end{aligned} \tag{B.10}$$

Fix q for each level $q > q_0$ and each partition $\mathcal{F}_{q_{u'}}$. For $x \in B_{q_{u'}}$, select a fixed $x_{q_{u'}} \in B_{q_{u'}}$ and define:

$$\pi_q x = x_{q_{u'}};$$

$$B_q x = B_{q_{u'}}.$$

Note that

$$\begin{aligned}
&\sum_{t=1}^{n} \Gamma\left(Y_{nt}^{(l',l,k)}\right) \\
=&\sum_{t=1}^{n} \Gamma_{\left(\pi_{q_0} x\right)}\left(Y_{nt}^{(l',l,k)}\right) + \sum_{t=1}^{n}\left(\Gamma\left(Y_{nt}^{(l',l,k)}\right) - \Gamma_{\left(\pi_{q_0} x\right)}\left(Y_{nt}^{(l',l,k)}\right)\right) \\
=&H_1^{(l',l,k)} + H_2^{(l',l,k)},
\end{aligned} \tag{B.11}$$

where

$$H_1^{(l',l,k)} = \sum_{t=1}^{n} \Gamma_{\left(\pi_{q_0} x\right)}\left(Y_{nt}^{(l',l,k)}\right)$$

and

$$H_2^{(l',l,k)} = \sum_{t=1}^{n}\left(\Gamma\left(Y_{nt}^{(l',l,k)}\right) - \Gamma_{\left(\pi_{q_0} x\right)}\left(Y_{nt}^{(l',l,k)}\right)\right).$$

To bound $H_1^{(l',l,k)}$, we apply Proposition 7 in Wong et al. [2020], and take the union over N_{q_0} balls (which is a finite number). Let $\mathcal{K}$, c_0, c_1, and c_2 be positive constants. Then, for

$$n \geq c_0\left(\log(p^2 K)\right)^{2/\varkappa_0 - 1},$$

we can get:

$$\begin{aligned}
&\mathbb{P}\left(\max_{1\leq l,l'\leq p, 1\leq k\leq K} \sup_{\Gamma(x)\in\mathcal{F}} \sum_{t=1}^{n} \Gamma_{\left(\pi_{q_0} x\right)}\left(Y_{nt}^{(l',l,k)}\right) > c_1 \mathcal{K}\sqrt{n}\sqrt{\frac{\log(p^2 K)}{n}}\right) \\
\leq& N_{q_0}\left(2\exp\left(-c_2 \log\left(p^2 K\right)\right)\right).
\end{aligned} \tag{B.12}$$

To bound $H_2^{(l',l,k)}$, define

$$\begin{aligned}
a_q &= 2^{-q} / \left[(\log n)^2 \sqrt{\log N_{q+1}} \right], \\
\Omega_t^{(l',l,k)}(B) &= \sup_{(x_1,x_2)\in B} \left| \Gamma_{(x_1)}\left(Y_{nt}^{(l',l,k)}\right) - \Gamma_{(x_2)}\left(Y_{nt}^{(l',l,k)}\right) \right|, \\
A_{t,q_0}^{(l',l,k)} &= I\left(\Omega_t^{(l',l,k)}(B_{q_0}x) > a_{q_0}\right), \\
C_{t,q-1}^{(l',l,k)} &= I\left(\Omega_t^{(l',l,k)}(B_{q_0}x) \le a_{q_0}, \ldots, \Omega_t^{(l',l,k)}(B_{q-1}x) \le a_{q-1}^{(l',l,k)}\right), \\
D_{t,q}^{(l',l,k)} &= I\Big(\Omega_t^{(l',l,k)}(B_{q_0}x) \le a_{q_0}, \ldots, \\
&\qquad \Omega_t^{(l',l,k)}(B_{q-1}x) \le a_{q-1}, \Omega_t^{(l',l,k)}(B_q x) > a_q\Big).
\end{aligned} \tag{B.13}$$

Since $\mathbb{E}H_2^{(l',l,k)} = 0$, $H_2^{(l',l,k)}$ can be decomposed into three parts

$$\begin{aligned}
H_2^{(l',l,k)} &= \sum_{t=1}^{n} \Big\{ \left[\Gamma_{(x)}(Y_{nt}^{(l',l,k)}) - \Gamma_{(\pi_{q_0}x)}(Y_{nt}^{(l',l,k)})\right] A_{t,q_0}^{(l',l,k)} \\
&\qquad - \mathbb{E}\left[\left[\Gamma_{(x)}(Y_{nt}^{(l',l,k)}) - \Gamma_{(\pi_{q_0}x)}(Y_{nt}^{(l',l,k)})\right] A_{t,q_0}^{(l',l,k)}\right] \Big\} \\
&+ \sum_{t=1}^{n} \sum_{q=q_0+1}^{\infty} \Big\{ \left[\Gamma_{(\pi_q x)}(Y_{nt}^{(l',l,k)}) - \Gamma_{(\pi_{q-1}x)}(Y_{nt}^{(l',l,k)})\right] C_{t,q-1}^{(l',l,k)} \\
&\qquad - \mathbb{E}\left[\left[\Gamma_{(\pi_q x)}(Y_{nt}^{(l',l,k)}) - \Gamma_{(\pi_{q-1}x)}(Y_{nt}^{(l',l,k)})\right] C_{t,q-1}^{(l',l,k)}\right] \Big\} \\
&+ \sum_{t=1}^{n} \sum_{q=q_0+1}^{\infty} \Big\{ \left[\Gamma_{(x)}(Y_{nt}^{(l',l,k)}) - \Gamma_{(\pi_q x)}(Y_{nt}^{(l',l,k)})\right] D_{t,q}^{(l',l,k)} \\
&\qquad - \mathbb{E}\left[\left[\Gamma_{(x)}(Y_{nt}^{(l',l,k)}) - \Gamma_{(\pi_q x)}(Y_{nt}^{(l',l,k)})\right] D_{t,q}^{(l',l,k)}\right] \Big\} \\
&=: S_{n1}^{(l',l,k)} + S_{n2}^{(l',l,k)} + S_{n3}^{(l',l,k)}.
\end{aligned} \tag{B.14}$$

By Lemma 15 and Proposition 3 from Vladimirova et al. [2020], $\Gamma_{(x)}(Y_{nt}^{(l',l,k)}) - \Gamma_{(\pi_{q_0}x)}(Y_{nt}^{(l',l,k)})$ is sub-Weibull. So there exists a constant C_0 and $\varkappa_1 > 0$ such that

$$\left(\mathbb{E}\left|\Gamma_{(x)}(Y_{nt}^{(l',l,k)}) - \Gamma_{(\pi_{q_0}x)}(Y_{nt}^{(l',l,k)})\right|^c\right)^{1/c} \le C_0 c^{(1/\varkappa_1 + 1/\varkappa_c)}$$

for all $c \ge 1$. Then, using the sub-Weibull property (first part of Theorem 2.1 in Vladimirova

et al. [2020]), for any $\varepsilon > 0$, we can get

$$\begin{aligned}\mathbb{P}\left(\sup_{\Gamma(x)\in\mathcal{F}}\left|S_{n1}^{(l',l,k)}\right| > \varepsilon\right) &\leq \sum_{t=1}^{n}\mathbb{P}\left(\left|M_t^{(l',l,k)}\right| > \sqrt{n}a_{q_0}\right) \\ &\leq n\exp\left(-c_3(\sqrt{n}a_{q_0})^{\varkappa_1\varkappa_c/(\varkappa_1+\varkappa_c)}\right), \end{aligned} \tag{B.15}$$

where c_3 is a positive constant and $\varkappa_1, \varkappa_c > 0$. Now, take the union over p^2K. Then, for a positive constant c_4 such that:

$$\begin{aligned}&\mathbb{P}\left(\max_{1\leq l,l'\leq p,1\leq k\leq K}\sup_{\Gamma(x)\in\mathcal{F}}\left|S_{n1}^{(l',l,k)}\right| > \varepsilon\right) \\ &\leq np^2K\exp\left(-c_3\left(\sqrt{n}2^{-q_0}/\left[(\log n)^2\sqrt{\log N_{q_0+1}}\right]\right)^{\varkappa_1\varkappa_c/(\varkappa_1+\varkappa_c)}\right) \\ &\leq \exp\left(-c_4\frac{n^{\varkappa_1\varkappa_c/2(\varkappa_1+\varkappa_c)}}{(\log n)^{2\varkappa_1\varkappa_c/(\varkappa_1+\varkappa_c)}} + \log\left(np^2K\right)\right).\end{aligned} \tag{B.16}$$

The second inequality satisfies since q_0 is fixed and its value will be determined later. To bound $S_{n2}^{(l',l,k)}$ and $S_{n3}^{(l',l,k)}$, we apply a similar procedure as in Lemma A.1 in Chan et al. [2015]. Specifically, let

$$\begin{aligned}\zeta_{nt}^{(l',l,k)}\left(q,\Gamma_{(x)}\right) = &\Big\{\left(\Gamma_{\pi_q x}\left(Y_{nt}^{(l',l,k)}\right) - \Gamma_{(\pi_{q-1}x)}\left(Y_{nt}^{(l',l,k)}\right)\right)C_{t,q-1}^{(l',l,k)} \\ &- \mathbb{E}\left[\left(\Gamma_{\pi_q x}\left(Y_{nt}^{(l',l,k)}\right) - \Gamma_{(\pi_{q-1}x)}\left(Y_{nt}^{(l',l,k)}\right)\right)C_{t,q-1}^{(l',l,k)}\right]\Big\}.\end{aligned}$$

For any $q \geq q_0$, since $\Gamma_{(x_q)}(Y_{nt}^{(l',l,k)})$ and $\Gamma_{(x_{q-1})}(Y_{nt}^{(l',l,k)})$ are points lying in one of the balls $B_{q-1,u'}$, $u' \leq N_{q-1}$, $\{\zeta_{nt}^{(l',l,k)}(q,\Gamma_{(x)})\}$ is a centered β-mixing process. Since β-mixing implies α-mixing, we can use the results in Lemma A.1 in Chan et al. [2015]. For any $y \geq 0$, there exists a positive constant c_5 such that

$$\mathbb{P}\left(\sup_{\Gamma_{(x)}\in\mathcal{F}}\left|S_{n2}^{(l',l,k)}\right| \geq h_{q_0}y\right) \leq \sum_{q=q_0+1}^{\infty}N_q\exp\left\{-\frac{c_5y^2\log N_q}{2+y}\right\}, \tag{B.17}$$

where $h_{q_0} = \sum_{q=q_0}^{\infty}2^{-q/2}\sqrt{\log N_q}$, and $q_0, n \geq 3$. Since i and j are fixed, we take the union

over p^2K again and get the bound of $S_{n2}^{(l',l,k)}$ as

$$\begin{aligned}
&\mathbb{P}\left(\max_{1\leq l,l'\leq p,1\leq k\leq K}\sup_{\Gamma_{(x)}\in\mathcal{F}}\left|S_{n2}^{(l',l,k)}\right|\geq h_{q_0}y\right)\\
&\leq\sum_{q=q_0+1}^{\infty}N_q\exp\left\{-\frac{c_5y^2\log N_q}{2+y}+\log\left(p^2K\right)\right\}.
\end{aligned}\tag{B.18}$$

Use the same argument of $S_{n2}^{(l',l,k)}$, we have:

$$\begin{aligned}
&\mathbb{P}\left(\max_{1\leq l,l'\leq p,1\leq k\leq K}\sup_{\Gamma_{(x)}\in\mathcal{F}}\left|S_{n3}^{(l',l,k)}\right|\geq h_{q_0}y\right)\\
&\leq\sum_{q=q_0+1}^{\infty}N_q\exp\left\{-\frac{c_5y^2\log N_q}{2+y}+\log\left(p^2K\right)\right\}.
\end{aligned}\tag{B.19}$$

When n is large enough, we can combine $S_{n1}^{(l',l,k)}$, $S_{n2}^{(l',l,k)}$, and $S_{n3}^{(l',l,k)}$. Thus, we have:

$$\begin{aligned}
&\mathbb{P}\left\{\max_{1\leq l,l'\leq p,1\leq k\leq K}\sup_{\Gamma_{(x)}\in\mathcal{F}}H_2^{(l',l,k)}\geq 2h_{q_0}\left(y+1\right)\right\}\\
&\leq 2\sum_{q=q_0+1}^{\infty}N_q\exp\left\{-\frac{c_5y^2\log N_q}{2+y}+\log\left(p^2K\right)\right\}+\\
&\quad\exp\left(-c_4\frac{n^{\varkappa_1\varkappa_c/2(\varkappa_1+\varkappa_c)}}{(\log n)^{2\varkappa_1\varkappa_c/(\varkappa_1+\varkappa_c)}+\log\left(np^2K\right)}\right)\\
&\leq 2p^2K\sum_{q=q_0+1}^{\infty}N_q\exp\left(-\frac{c_5y^2\log N_q}{2+y}\right)+\\
&\quad np^2K\exp\left(-c_4\frac{n^{\varkappa_1\varkappa_c/2(\varkappa_1+\varkappa_c)}}{(\log n)^{2\varkappa_1\varkappa_c/(\varkappa_1+\varkappa_c)}}\right)\\
&\leq 2p^2K\sum_{q=q_0+1}^{\infty}N_q^{1-\frac{c_5y^2}{2+y}}+np^2K\exp\left(-c_4\frac{n^{\varkappa_1\varkappa_c/2(\varkappa_1+\varkappa_c)}}{(\log n)^{2\varkappa_1\varkappa_c/(\varkappa_1+\varkappa_c)}}\right)
\end{aligned}\tag{B.20}$$

Now, take $y=C_0'\frac{\log\left(p^2K\right)}{\sqrt{n}}\sqrt{n}$ and q_0 to be a smallest integer such that $q_0\geq 3$ and $h_{q_0}\leq 1$. Note that N_{q_0} is a constant since q_0 is selected to be fixed. Equation (B.12) and Equation (B.20) give Lemma 16 as desired. Specifically, if we choose C_0' large enough, there exist positive constants C, c_6, c_7, c_8, $c_9>1$, c_{10} and $\frac{c_5y^2}{2+y}>3$ such that

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{x\in\mathbb{R},1\le l,l'\le p,1\le k\le K}\frac{1}{n}\left|\sum_{t=1}^{n}x_{((t-k),l)}I\left(z_t\le x\right)\epsilon_{(t,l')}\right|\ge C\frac{\log\left(p^2K\right)}{\sqrt{n}}\right)\\
&\le N_{q_0}\left(2\exp\left(-c_2\log\left(p^2K\right)\right)\right)+2p^2K\sum_{q=q_0+1}^{\infty}N_q^{1-\frac{c_5y^2}{2+y}}\\
&\quad+np^2K\exp\left(-c_4\frac{n^{\varkappa_1\varkappa_c/2(\varkappa_1+\varkappa_c)}}{(\log n)^{2\varkappa_1\varkappa_c/(\varkappa_1+\varkappa_c)}}\right)\\
&\le N_{q_0}\left(2\exp\left(-c_2\log\left(p^2K\right)\right)\right)+c_7p^2K\exp\left(-c_8y\right)\\
&\quad+np^2K\exp\left(-c_4\frac{n^{\varkappa_1\varkappa_c/2(\varkappa_1+\varkappa_c)}}{(\log n)^{2\varkappa_1\varkappa_c/(\varkappa_1+\varkappa_c)}}\right)\\
&\le N_{q_0}\left(2\exp\left(-c_2\log\left(p^2K\right)\right)\right)+c_7\exp\left(-c_8C_0'\log\left(p^2K\right)+\log\left(p^2K\right)\right)\\
&\quad+\exp\left(-c_4\frac{n^{\varkappa_1\varkappa_c/2(\varkappa_1+\varkappa_c)}}{(\log n)^{2\varkappa_1\varkappa_c/(\varkappa_1+\varkappa_c)}}+\log\left(np^2K\right)\right)\\
&\le c_6\exp\left(-c_9\log\left(p^2K\right)+\log\left(p^2K\right)\right)\\
&\quad+\exp\left(-c_4\frac{n^{\varkappa_1\varkappa_c/2(\varkappa_1+\varkappa_c)}}{(\log n)^{2\varkappa_1\varkappa_c/(\varkappa_1+\varkappa_c)}}+\log\left(np^2K\right)\right)\\
&\le c_6\exp\left(-c_{10}\log\left(p^2K\right)\right)+\exp\left(-c_4\frac{n^{\varkappa_1\varkappa_c/2(\varkappa_1+\varkappa_c)}}{(\log n)^{2\varkappa_1\varkappa_c/(\varkappa_1+\varkappa_c)}}+\log\left(np^2K\right)\right).
\end{aligned}
\tag{B.21}
$$

Lemma 17. *Under Assumptions B1 to B4, there exist positive constants C, c_0, c_1, c_2, and c_3, such that for*

$$
n\ge c_0\left(\log(p^2K)\right)^{2/\varkappa_0-1},
$$

with probability at least $1-c_3\eta_1-\eta_2$, we have:

$$
\frac{1}{n}\|\boldsymbol{Z}'\boldsymbol{E}\|_{2,\infty}\le C\frac{\sqrt{p^2K}\log\left(p^2K\right)}{\sqrt{n}},
\tag{B.22}
$$

where $\eta_1=\exp\left(-c_1\log\left(p^2K\right)\right)$ and

$$
\eta_2=\exp\left(-c_2\frac{n^{\varkappa_1\varkappa_c/2(\varkappa_1+\varkappa_c)}}{(\log n)^{2\varkappa_1\varkappa_c/(\varkappa_1+\varkappa_c)}}+\log\left(np^2K\right)\right).
$$

Proof of Lemma 17: By combining Equation (B.12) and Equation (B.20) in Lemma 16, we have:

$$\begin{aligned}&\mathbb{P}\left(\frac{1}{n}\|\boldsymbol{Z}'\boldsymbol{E}\|_\infty \geq C\frac{\log\left(p^2K\right)}{\sqrt{n}}\right)\\&\leq c_3\exp\left(-c_1\log\left(p^2K\right)\right)+\exp\left(-c_2\frac{n^{\varkappa_1\varkappa_c/2(\varkappa_1+\varkappa_c)}}{(\log n)^{2\varkappa_1\varkappa_c/(\varkappa_1+\varkappa_c)}}+\log\left(np^2K\right)\right),\end{aligned} \tag{B.23}$$

where C, c_1, c_2, c_3 are positive constants.

Let G_g represents the group in group lasso for $g = 1, 2, \ldots, n$, where $G_1 = \{1, 2, \ldots, p^2K\}, G_2 = \{p^2K+1, p^2K+2, \ldots, 2p^2K\}, \ldots, G_n = \{(n-1)p^2K+1, (n-1)p^2K+2, \ldots, np^2K\}$. Note that for none overlapping group G_g with size p^2K, we have:

$$\frac{1}{\sqrt{p^2K}}\|vec(\boldsymbol{Z}_l'\boldsymbol{E}), l \in G_g\|_2 = \sqrt{\frac{1}{p^2K}\sum_{l\in G_g}|vec(\boldsymbol{Z}_l'\boldsymbol{E})|^2} \leq \max_{l\in G_g}|vec(\boldsymbol{Z}_l'\boldsymbol{E})|, \tag{B.24}$$

where $\boldsymbol{Z}_l'$ represents l-th row of $\boldsymbol{Z}'$. Thus,

$$\begin{aligned}\frac{1}{\sqrt{p^2K}}\|vec(\boldsymbol{Z}'\boldsymbol{E})\|_{2,\infty} &= \max_{g=1,\ldots,n}\frac{1}{\sqrt{p^2K}}\|vec(\boldsymbol{Z}_l'\boldsymbol{E}), l \in G_g\|_2\\&\leq \max_{g=1,\ldots,n}\max_{l\in G_g}|vec(\boldsymbol{Z}_l'\boldsymbol{E})|\\&= \|\boldsymbol{Z}'\boldsymbol{E}\|_\infty.\end{aligned} \tag{B.25}$$

Combining the Lemma 16 and Equation (B.25), we have:

$$\begin{aligned}&\mathbb{P}\left(\frac{1}{n}\|vec\left(\boldsymbol{Z}'\boldsymbol{E}\right)\|_{2,\infty} \geq C\sqrt{p^2K}\frac{\log\left(p^2K\right)}{\sqrt{n}}\right)\\&\leq \mathbb{P}\left(\frac{1}{n}\|\boldsymbol{Z}'\boldsymbol{E}\|_\infty \geq C\frac{\log\left(p^2K\right)}{\sqrt{n}}\right)\\&\leq c_3\exp\left(-c_1\log\left(p^2K\right)\right)+\exp\left(-c_2\frac{n^{\varkappa_1\varkappa_c/2(\varkappa_1+\varkappa_c)}}{(\log n)^{2\varkappa_1\varkappa_c/(\varkappa_1+\varkappa_c)}}+\log\left(np^2K\right)\right).\end{aligned} \tag{B.26}$$

Lemma 18. *Set* $\sigma^2(s) = \mathbb{E}\left(x_{(t-k,g')}I\left(r_j - s < z_t \leq r_j\right)\right)^2$ *for any given* $1 \leq g' \leq p$. *Under Assumptions B1 to B4, there exist positive constants* c_i, C, C', C_i', C_i'' *for* $i = 1, 2, \ldots$, *such*

that for any given j-th threshold, with probability at least $1-\delta_1$,

$$\begin{aligned}
&\sup_{\substack{1\le k\le K,\\ |s|\ge\gamma_n}} \left(n\sigma^2(s)\right)^{-1} \Big\| \sum_{t=1}^{n} \boldsymbol{x}_{(t-k)}\boldsymbol{x}_{(t-k)}^{'} I\left(r_j - s < z_t \le r_j\right) \\
&\qquad - \mathbb{E}\left(\boldsymbol{x}_{(t-k)}\boldsymbol{x}_{(t-k)}^{'} I\left(r_j - s < z_t \le r_j\right)\right) \Big\|_{\infty} \\
&\le C\left(\frac{\left(\log(p^2K)\right)^{1/\varkappa_0 - 1/2}}{\sqrt{n\gamma_n}}\right),
\end{aligned} \tag{B.27}$$

where

$$\begin{aligned}
\delta_1 = \max\Bigg\{ &\exp\left(-C_1'\left(\frac{n}{\gamma_n}\right)^{\varkappa_0/2}\left(\log\left(p^2K\right)\right)^{1-\varkappa_0/2} + \log\left(p^2K\right) + \log(n)\right) \\
&+ \exp\left(-C_2'\frac{1}{\gamma_n}\log\left(p^2K\right)^{2/\varkappa_0-1} + \log\left(p^2K\right)\right), \\
&\exp\left(-C_3'\left(n\gamma_n\right)^{\varkappa_0/2}\left(\log\left(p^2K\right)\right)^{1-\varkappa_0/2} + \log\left(p^2K\right) + \log\left(n\gamma_n\right)\right) \\
&+ \exp\left(-C_4'\log\left(p^2K\right)^{2/\varkappa_0-1} + \log\left(p^2K\right)\right)\Bigg\}.
\end{aligned}$$

In addition, with probability at least $1-\delta_2$,

$$\sup_{\substack{1\le l,l'\le p,\\ 1\le k\le K,\\ |s|\ge\gamma_n}} \left(n\sigma^2(s)\right)^{-1} \left|\sum_{t=1}^{n} x_{(t-k,l)} I\left(r_j - s < z_t \le r_j\right)\epsilon_{(t,l')}\right| \le C'\frac{\log(p^2K)}{\sqrt{n\gamma_n}}, \tag{B.28}$$

where

$$\delta_2 = c_3\exp\left(-c_4\log\left(p^2K\right)\right) + \exp\left(-c_5\frac{n^{\varkappa_1\varkappa_c/2(\varkappa_1+\varkappa_c)}}{(\log n)^{2\varkappa_1\varkappa_c/(\varkappa_1+\varkappa_c)}} + \log\left(np^2K\right)\right). \tag{B.29}$$

Proof of Lemma 18: The proof for this lemma is along the lines of the proof of Lemma A.2 in Chan et al. [2015]. Assume a fixed small number $D > 0$. Since $\sigma^2(s)$ is non-decreasing in given distance s, and $\boldsymbol{\epsilon}_t$ and z_t both have bounded positive density based on Assumptions B1 and B5,

$$\sigma^2(s) \ge \sigma^2(D) \ge CD \quad \text{if} \quad s \ge D, \tag{B.30}$$

where C is a constant greater than 0. Similar to Lemma 16, for $s \geq D \geq \gamma_n$, according to Equation (B.12) and Equation (B.20), for a given j-th threshold, there exist large enough positive constant C_0, and positive constants C', $c_{h'}$, $C'_{h'}$ for $h' = 1, 2, \ldots, 12$ such that

$$\begin{aligned}
&\mathbb{P}\Bigg(\sup_{\substack{1\leq l,l'\leq p,\\ 1\leq k\leq K,\\ |s|\geq D}} \left(n\sigma^2(s)\right)^{-1} \left| \sum_{t=1}^{n} x_{(t-k,l)} I\left(r_j - s < z_t \leq r_j\right) \epsilon_{(t,l')} \right| \\
&\qquad\qquad \geq \left(\frac{C_0}{CD}\right) \frac{\log\left(p^2K\right)}{\sqrt{n}} \Bigg) \\
&\leq \mathbb{P}\Bigg(\sup_{\substack{1\leq l,l'\leq p,\\ 1\leq k\leq K,\\ |s|\geq D}} \left(n\sigma^2(s)\right)^{-1} \left| \sum_{t=1}^{n} x_{(t-k,l)} I\left(r_j - s < z_t \leq r_j\right) \epsilon_{(t,l')} \right| \\
&\qquad\qquad \geq C_0 \frac{\log\left(p^2K\right)}{\sqrt{n}} / \sigma^2(s) \Bigg) \\
&\leq c_3 \exp\left(-c_1 \log\left(p^2K\right)\right) + \exp\left(-c_2 \frac{n^{\varkappa_1\varkappa_c/2(\varkappa_1+\varkappa_c)}}{(\log n)^{2\varkappa_1\varkappa_c/(\varkappa_1+\varkappa_c)}} + \log\left(np^2K\right) \right).
\end{aligned} \tag{B.31}$$

Thus, with high probability, we obtain Equation (B.28) when $s \geq D$.

When $s \in [\gamma_n, D]$, we want to partition the interval into small pieces, and prove the

consistency in each piece. Let $M = [\log(D/\gamma_n)/\log b]$, where $b > 1$ is a constant. Now,

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{\substack{1\le l,l'\le p,\\ 1\le k\le K,\\ s\in[\gamma_n, D]}} \left(n\sigma^2(s)\right)^{-1} \left| \sum_{t=1}^{n} x_{(t-k,l)} I\left(r_{j-1} - s < z_t \le r_{j-1}\right) \epsilon_{(t,l')} \right| \ge y_1 \right) \\
&\le \sum_{g=0}^{M} \mathbb{P}\Bigg(\sup_{\substack{1\le l,l'\le p,\\ 1\le k\le K,\\ s\in[b^g\gamma_n, b^{g+1}\gamma_n]}} \left(n\sigma^2(b^g\gamma_n)\right)^{-1} \\
&\qquad\qquad \left| \sum_{t=1}^{n} x_{(t-k,l)} I\left(r_j - s < z_t \le r_j - b^g\gamma_n\right) \epsilon_{(t,l')} \right| \ge y_1/2 \Bigg) \\
&\quad + \sum_{g=0}^{M} \mathbb{P}\Bigg(\sup_{\substack{1\le l,l'\le p,\\ 1\le k\le K}} \left(n\sigma^2(b^g\gamma_n)\right)^{-1} \\
&\qquad\qquad \left| \sum_{t=1}^{n} x_{(t-k,l)} I\left(r_j - b^g\gamma_n < z_t \le r_j\right) \epsilon_{(t,j)} \right| \ge y_1/2 \Bigg) \\
&\le \sum_{g=0}^{M} \mathbb{P}\Bigg(\sup_{\substack{1\le l,l'\le p,\\ 1\le k\le K}} \left(n\sigma^2(b^g\gamma_n)\right)^{-1} \\
&\qquad\qquad \sum_{t=1}^{n} \left| x_{(t-k,l)} I\left(r_j - b^{g+1}\gamma_n < z_t \le r_j - b^g\gamma_n\right) \epsilon_{(t,l')} \right| \ge y_1/2 \Bigg) \\
&\quad + \sum_{g=0}^{M} \mathbb{P}\Bigg(\sup_{\substack{1\le l,l'\le p,\\ 1\le k\le K}} \left(n\sigma^2(b^g\gamma_n)\right)^{-1} \\
&\qquad\qquad \left| \sum_{t=1}^{n} x_{(t-k,l)} I\left(r_j - b^g\gamma_n < z_t \le r_j\right) \epsilon_{(t,l')} \right| \ge y_1/2 \Bigg) \\
&\le \sum_{g=0}^{M} \mathbb{P}\Bigg(\sup_{\substack{1\le l,l'\le p,\\ 1\le k\le K}} \left(C_1' n\gamma_n b^g\right)^{-1} \\
&\qquad\qquad \sum_{t=1}^{n} \left| x_{(t-k,l)} I\left(r_j - b^{g+1}\gamma_n < z_t \le r_j - b^g\gamma_n\right) \epsilon_{(t,l')} \right| \ge y_1/2 \Bigg) \\
&\quad + \sum_{g=0}^{M} \mathbb{P}\Bigg(\sup_{\substack{1\le l,l'\le p,\\ 1\le k\le K}} \left(C_1' n\gamma_n b^g\right)^{-1} \\
&\qquad\qquad \left| \sum_{t=1}^{n} x_{(t-k,l)} I\left(r_j - b^g\gamma_n < z_t \le r_j\right) \epsilon_{(t,l')} \right| \ge y_1/2 \Bigg) \\
&=: \sum_{g=0}^{M} H_{ng} + \sum_{g=0}^{M} I_{ng}.
\end{aligned}
\tag{B.32}
$$

Recall that $1 > \varkappa_0 > 0$, and the fact that the function of a β-mixing process is also a β-mixing. Since $\boldsymbol{x}_t$ and z_t are β-mixing,

$$x_{(t-k,l)} I\left(r_j - b^g \gamma_n < z_t \leq r_j\right) \epsilon_{(t,l')}$$

and

$$x_{(t-k,l)} I\left(r_j - b^{g+1} \gamma_n < z_t \leq r_j - b^g \gamma_n\right) \epsilon_{(t,l')}$$

are β-mixing. To bound H_{ng} and I_{ng}, we can apply Proposition 7 from Wong et al. [2020]. Set

$$y_1/2 = C_2' \frac{\log(p^2 K)}{\sqrt{n\gamma_n}}.$$

For

$$C_1' b^g n \gamma_n \geq C_3' \left(\log\left(p^2 K\right)\right)^{2/\varkappa_0 - 1},$$

we can get

$$\begin{aligned} I_{ng} < \mathbb{P}\Bigg(& \left(C_1' n \gamma_n b^g\right)^{-1} \left| \sum_{t=1}^{n} x_{(t-k,l)} I\left(r_j - b^g \gamma_n < z_t \leq r_j\right) \epsilon_{(t,l')} \right| \\ & \geq C_2' \sqrt{\frac{\log(p^2 K)}{n\gamma_n}} \Bigg) \\ \leq 2 \exp & \left(-C_4' \log\left(p^2 K\right)\right). \end{aligned} \tag{B.33}$$

Then, we can get

$$\sum_{g=0}^{M} I_{ng} \leq C_5' \exp\left(-C_4' \log\left(p^2 K\right)\right). \tag{B.34}$$

Similarly, we can get

$$\sum_{g=0}^{M} H_{ng} \leq C_6' \exp\left(-C_4' \log\left(p^2 K\right)\right). \tag{B.35}$$

By Equation (B.35), Equation (B.34), and Equation (B.31), we then can get:

$$\sup_{\substack{1 \leq l,l' \leq p, \\ 1 \leq k \leq K, \\ |s| \geq \gamma_n}} \left(n\sigma^2(s)\right)^{-1} \left| \sum_{t=1}^{n} x_{(t-k,l)} I\left(r_j - s < z_t \leq r_j\right) \epsilon_{(t,l')} \right| \leq C' \frac{\log(p^2 K)}{\sqrt{n\gamma_n}}, \tag{B.36}$$

with probability $1-\delta_3$, where

$$\delta_3 = c_4 \exp\left(-c_5 \log\left(p^2 K\right)\right) + \exp\left(-c_2 \frac{n^{\varkappa_1\varkappa_c/2(\varkappa_1+\varkappa_c)}}{(\log n)^{2\varkappa_1\varkappa_c/(\varkappa_1+\varkappa_c)}} + \log\left(np^2K\right)\right). \tag{B.37}$$

Thus, this proves Equation (B.28). Similarly, we can prove Equation (B.27). Note that

$$\mathbb{E}\left(\boldsymbol{x}_{(t-k)}\boldsymbol{x}_{(t-k)}^{'} I\left(r_j - s < z_t \leq r_j\right)\right)$$

is non-decreasing in s and $\mathbb{E}|\boldsymbol{x}_t|^2$ is positive and bounded from Assumption B2. For $s \geq D \geq \gamma_n$, we first fix l, l' in $1, 2, \ldots, p$ and k in $1, 2, \ldots, K$. Recall that

$$\sigma^2(s) \geq \sigma^2(D) \geq CD \quad \text{if} \quad s \geq D, \tag{B.38}$$

where C is a constant greater than 0. Note that $\varkappa_0 = (1/(\varkappa_1/2) + 1/\varkappa_2)^{-1} < 1$. By Assumption B2, Lemma 15, and Fact 1 and Lemma 13 in Wong et al. [2020], for $n > 4$,

$$\begin{aligned}
&\mathbb{P}\Bigg(\sup_{|s|\geq D}\left(n\sigma^2(s)\right)^{-1}\Bigg|\sum_{t=1}^{n} x_{(t-k,l)}x_{(t-k,l')} I\left(r_j - s < z_t \leq r_j\right) - \\
&\quad \mathbb{E}\left(x_{(t-k,l)}x_{(t-k,l')} I\left(r_j - s < z_t \leq r_j\right)\right)\Bigg| \geq y_2\Bigg) \\
&\leq \mathbb{P}\Bigg(\sup_{|s|\geq D}(CDn)^{-1}\Bigg|\Bigg(\sum_{t=1}^{n} x_{(t-k,l)}x_{(t-k,l')} I\left(r_j - s < z_t \leq r_j\right) - \\
&\quad \mathbb{E}\left(x_{(t-k,l)}x_{(t-k,l')} I\left(r_j - s < z_t \leq r_j\right)\right)\Bigg)\Bigg| \geq y_2\Bigg) \\
&\leq n\exp\left(-c_6(ny_2)^{\varkappa_0}\right) + \exp\left(-c_7 n y_2^2\right).
\end{aligned} \tag{B.39}$$

Then, we take the union over p^2K:

$$
\begin{aligned}
&\mathbb{P}\Bigg(\sup_{\substack{1\le k\le K,\\ |s|\ge D}} \left(n\sigma^2(s)\right)^{-1} \Bigg\| \sum_{t=1}^{n} \boldsymbol{x}_{(t-k)}\boldsymbol{x}'_{(t-k)} I\left(r_j - s < z_t \le r_j\right) - \\
&\qquad \mathbb{E}\left(\boldsymbol{x}_{(t-k)}\boldsymbol{x}'_{(t-k)} I\left(r_j - s < z_t \le r_j\right)\right) \Bigg\|_\infty \ge y_2 \Bigg) \\
&= \mathbb{P}\Bigg(\sup_{\substack{1\le l,l'\le p,\\ 1\le k\le K,\\ |s|\ge D}} \left(n\sigma^2(s)\right)^{-1} \Bigg| \sum_{t=1}^{n} x_{(t-k,l)}x_{(t-k,l')} I\left(r_j - s < z_t \le r_j\right) - \\
&\qquad \mathbb{E}\left(x_{(t-k,l)}x_{(t-k,l')} I\left(r_j - s < z_t \le r_j\right)\right) \Bigg| \ge y_2 \Bigg) \\
&\le n\exp\left(-c_6(ny_2)^{\varkappa_0} + \log\left(p^2K\right)\right) + \exp\left(-c_7 n y_2^2 + \log\left(p^2K\right)\right).
\end{aligned}
\tag{B.40}
$$

For $s \in [\gamma_n, D]$, we have:

$$
\begin{aligned}
&\mathbb{P}\Bigg(\sup_{\substack{1\le k\le K,\\ D\ge|s|\ge\gamma_n}} \left(n\sigma^2(s)\right)^{-1} \Bigg\| \sum_{t=1}^{n} \boldsymbol{x}_{(t-k)}\boldsymbol{x}_{(t-k)}^{'} I\left(r_j - s < z_t \le r_j\right) - \\
&\qquad \mathbb{E}\left(\boldsymbol{x}_{(t-k)}\boldsymbol{x}_{(t-k)}^{'} I\left(r_j - s < z_t \le r_j\right)\right) \Bigg\|_\infty \ge y_2 \Bigg) \\
\le &\sum_{g=0}^{M} \mathbb{P}\Bigg(\sup_{\substack{1\le k\le K,\\ s\in[b^g\gamma_n, b^{g+1}\gamma_n]}} (n\sigma^2(b^g\gamma_n))^{-1} \\
&\Bigg\| \sum_{t=1}^{n} \boldsymbol{x}_{(t-k)}\boldsymbol{x}_{(t-k)}^{'} I\left(r_j - s < z_t \le r_j - b^g\gamma_n\right) - \\
&\qquad \mathbb{E}\left(\boldsymbol{x}_{(t-k)}\boldsymbol{x}_{(t-k)}^{'} I\left(r_j - s < z_t \le r_j - b^g\gamma_n\right)\right) \Bigg\|_\infty \ge y_2/2 \Bigg) \\
&+ \sum_{g=0}^{M} \mathbb{P}\Bigg(\sup_{\substack{1\le k\le K,\\ s\in[b^g\gamma_n, b^{g+1}\gamma_n]}} \left(n\sigma^2\left(b^g\gamma_n\right)\right)^{-1} \\
&\Bigg\| \sum_{t=1}^{n} \boldsymbol{x}_{(t-k)}\boldsymbol{x}_{(t-k)}^{'} I\left(r_j - b^g\gamma_n < z_t \le r_j\right) - \\
&\qquad \mathbb{E}\left(\boldsymbol{x}_{(t-k)}\boldsymbol{x}_{(t-k)}^{'} I\left(r_j - b^g\gamma_n < z_t \le r_j\right)\right) \Bigg\|_\infty \ge y_2/2 \Bigg) \\
\le &\sum_{g=0}^{M} \mathbb{P}\Bigg(\left(n\sigma^2\left(b^g\gamma_n\right)\right)^{-1} \Bigg\| \sum_{t=1}^{n} \boldsymbol{x}_{(t-k)}\boldsymbol{x}_{(t-k)}^{'} I\left(r_j - b^{g+1}\gamma_n < z_t \le r_j - b^g\gamma_n\right) - \\
&\qquad \mathbb{E}\left(\boldsymbol{x}_{(t-k)}\boldsymbol{x}_{(t-k)}^{'} I\left(r_j - b^{g+1}\gamma_n < z_t \le r_j - b^g\gamma_n\right)\right) \Bigg\|_\infty \ge y_2/2 \Bigg) \\
&+ \sum_{g=0}^{M} \mathbb{P}\Bigg(\sup_{\substack{1\le k\le K,\\ s\in[b^g\gamma_n, b^{g+1}\gamma_n]}} \left(n\sigma^2\left(b^g\gamma_n\right)\right)^{-1} \\
&\Bigg\| \sum_{t=1}^{n} \boldsymbol{x}_{(t-k)}\boldsymbol{x}_{(t-k)}^{'} I\left(r_j - b^g\gamma_n < z_t \le r_j\right) - \\
&\qquad \mathbb{E}\left(\boldsymbol{x}_{(t-k)}\boldsymbol{x}_{(t-k)}^{'} I\left(r_j - b^g\gamma_n < z_t \le r_j\right)\right) \Bigg\|_\infty \ge y_2/2 \Bigg) \\
=: &\sum_{g=0}^{M} I_{1ng} + \sum_{g=0}^{M} I_{2ng}.
\end{aligned}
\tag{B.41}
$$

Note that $\mathbb{E}\left(\boldsymbol{x}_{(t-k)}\boldsymbol{x}_{(t-k)}^{'}I\left(r_j - s < z_t \leq r_j\right)\right)$ is non-decreasing in s and $\mathbb{E}|\boldsymbol{x}_t|^2$ is positive and bounded from Assumption B2. By Lemma 15, Lemma 13 in Wong et al. [2020] and taking union over p^2K, for $n\gamma_n > 4$, we can obtain

$$\begin{aligned}\sum_{g=0}^{M} I_{1ng} \leq n\gamma_n \exp\left(-c_8(n\gamma_n y_2)^{\varkappa_0} + \log\left(p^2K\right)\right) \\ + \exp\left(-c_9 n\gamma_n y_2^2 + \log\left(p^2K\right)\right).\end{aligned} \tag{B.42}$$

Similarly, we have

$$\begin{aligned}\sum_{g=0}^{M} I_{2ng} \leq n\gamma_n \exp\left(-c_{10}(n\gamma_n y_2)^{\varkappa_0} + \log\left(p^2K\right)\right) \\ + \exp\left(-c_{11} n\gamma_n y_2^2 + \log\left(p^2K\right)\right).\end{aligned} \tag{B.43}$$

Combining Equation (B.42), Equation (B.43) and Equation (B.40) and setting

$$y_2/2 = c_{12}\left(\frac{\left(\log(p^2K)\right)^{1/\varkappa_0 - 1/2}}{\sqrt{n\gamma_n}}\right)$$

with large enough c_{12}, we have

$$\begin{aligned}&\sup_{\substack{1\leq k\leq K,\\ |s|\geq\gamma_n}} (n\sigma^2(s))^{-1}\Bigg\| \sum_{t=1}^{n} \boldsymbol{x}_{(t-k)}\boldsymbol{x}_{(t-k)}^{'} I\left(r_j - s < z_t \leq r_j\right) \\ &\quad - \mathbb{E}\left(\boldsymbol{x}_{(t-k)}\boldsymbol{x}_{(t-k)}^{'} I\left(r_j - s < z_t \leq r_j\right)\right)\Bigg\|_{\infty} \\ &\leq c_{12}\left(\frac{\left(\log(p^2K)\right)^{1/\varkappa_0 - 1/2}}{\sqrt{n\gamma_n}}\right),\end{aligned} \tag{B.44}$$

with probability $1-\delta_4$, where

$$
\begin{aligned}
\delta_4 = \max\Bigg\{ & n\exp\left(-c_6(ny_2)^{\varkappa_0} + \log\left(p^2K\right)\right) + \exp\left(-c_7 n y_2^2 + \log\left(p^2K\right)\right), \\
& n\gamma_n \exp\left(-c_8(n\gamma_n y_2)^{\varkappa_0} + \log\left(p^2K\right)\right) + \exp\left(-c_9 n\gamma_n y_2^2 + \log\left(p^2K\right)\right)\Bigg\} \\
= \max\Bigg\{ & n\exp\left(-c_6\left(nc_{12}\left(\frac{\left(\log(p^2K)\right)^{1/\varkappa_0-1/2}}{\sqrt{n\gamma_n}}\right)\right)^{\varkappa_0} + \log\left(p^2K\right)\right) \\
& + \exp\left(-c_7 n\left(c_{12}\left(\frac{\left(\log(p^2K)\right)^{1/\varkappa_0-1/2}}{\sqrt{n\gamma_n}}\right)\right)^{2} + \log\left(p^2K\right)\right), \\
& n\gamma_n\exp\left(-c_8\left(n\gamma_n c_{12}\left(\frac{\left(\log(p^2K)\right)^{1/\varkappa_0-1/2}}{\sqrt{n\gamma_n}}\right)\right)^{\varkappa_0} + \log\left(p^2K\right)\right) \\
& + \exp\left(-c_9 n\gamma_n\left(c_{12}\left(\frac{\left(\log(p^2K)\right)^{1/\varkappa_0-1/2}}{\sqrt{n\gamma_n}}\right)\right)^{2} + \log\left(p^2K\right)\right)\Bigg\} \\
= \max\Bigg\{ & \exp\left(-C_7'\left(\frac{n}{\gamma_n}\right)^{\varkappa_0/2}\left(\log\left(p^2K\right)\right)^{1-\varkappa_0/2} + \log\left(p^2K\right) + \log(n)\right) \\
& + \exp\left(-C_8'\frac{1}{\gamma_n}\log\left(p^2K\right)^{2/\varkappa_0-1} + \log\left(p^2K\right)\right), \\
& \exp\left(-C_9'\left(n\gamma_n\right)^{\varkappa_0/2}\left(\log\left(p^2K\right)\right)^{1-\varkappa_0/2} + \log\left(p^2K\right) + \log\left(n\gamma_n\right)\right) \\
& + \exp\left(-C_{10}'\log\left(p^2K\right)^{2/\varkappa_0-1} + \log\left(p^2K\right)\right)\Bigg\}.
\end{aligned}
\tag{B.45}
$$

Note that $C'\left(n\gamma_n\right)^{\varkappa_0/2} > \left(\log\left(p^2K\right)\right)^{\varkappa_0}$ by Assumption B4. So

$$
C' n^{\varkappa_0/2}\left(\log\left(p^2K\right)\right)^{1-\varkappa_0/2} \geq \log(p^2K).
$$

Thus, both

$$
\exp\left(-C_7'\left(\frac{n}{\gamma_n}\right)^{\varkappa_0/2}\left(\log\left(p^2K\right)\right)^{1-\varkappa_0/2} + \log\left(p^2K\right) + \log(n)\right)
$$

and

$$
\exp\left(-C_9'\left(n\gamma_n\right)^{\varkappa_0/2}\left(\log\left(p^2K\right)\right)^{1-\varkappa_0/2} + \log\left(p^2K\right) + \log\left(n\gamma_n\right)\right)
$$

will converge to zero as sample size n tends to infinity. According to Assumption B2, $\varkappa_0 < 1$, so $2/\varkappa_0 - 1 > 1$. As a result, δ_4 will converge to 0.

Lemma 19. *Set* $\sigma^2(s) = \mathbb{E}\left(x_{(t-k,g')} I(r_j - s < z_t < r_j)\right)^2$ *for any given* $1 \le g' \le p$. *Let* $\mathbb{I}_s \in \mathbb{R}^{np \times np}$ *be the diagonal matrix with diagonal* $I_1(s), \ldots, I_n(s)$, *where* $I_t(s)$ *is a* $p \times p$ *diagonal matrix with all diagonal elements equal to* $I(r_j - s < z_t \le r_j)$ *for* $t = 1, 2, \ldots, n$. *Under Assumptions B1 to B4, there exist positive constants* c_3, c_4, c_5, C', C_1'', *such that for any given* j*-th threshold, with probability at least* $1 - \delta_2'$,

$$\sup_{|s| \ge \gamma_n} \left(n\sigma^2(s)\right)^{-1} \left\| \boldsymbol{Z}' \mathbb{I}_{r_j}(s) \boldsymbol{E} \right\|_{2,\infty} \le C' \frac{\sqrt{p^2 K} \log(p^2 K)}{\sqrt{n\gamma_n}}, \tag{B.46}$$

where

$$\delta_2' = c_3 \exp\left(-c_4 \log\left(p^2 K\right)\right) + \exp\left(-c_5 \frac{n^{\varkappa_1 \varkappa_c / 2(\varkappa_1 + \varkappa_c)}}{(\log n)^{2\varkappa_1 \varkappa_c / (\varkappa_1 + \varkappa_c)}} + \log\left(np^2 K\right)\right).$$

Proof of Lemma 19: By Lemma 18, we have:

$$\begin{aligned} &\mathbb{P}\left(\sup_{\substack{1 \le l, l' \le p, \\ 1 \le k \le K, \\ |s| \ge \gamma_n}} \left(n\sigma^2(s)\right)^{-1} \left| \sum_{t=1}^{n} x_{(t-k,l)} I(r_j - s < z_t \le r_j) \epsilon_{(t,l')} \right| \le C' \frac{\log(p^2 K)}{\sqrt{n\gamma_n}} \right) \\ &\le c_3 \exp\left(-c_4 \log\left(p^2 K\right)\right) + \exp\left(-c_5 \frac{n^{\varkappa_1 \varkappa_c / 2(\varkappa_1 + \varkappa_c)}}{(\log n)^{2\varkappa_1 \varkappa_c / (\varkappa_1 + \varkappa_c)}} + \log\left(np^2 K\right)\right) \end{aligned} \tag{B.47}$$

where C', c_3, c_4, c_5 are positive constants.

Note $\sup_{\substack{1 \le l, l' \le p, \\ 1 \le k \le K, \\ |s| \ge \gamma_n}} \left(n\sigma^2(s)\right)^{-1} \left| \sum_{t=1}^{n} x_{(t-k,l)} I(r_j - s < z_t \le r_j) \epsilon_{(t,l')} \right|$ can be written as $\sup_{|s| \ge \gamma_n} \left(n\sigma^2(s)\right)^{-1} \left\| \boldsymbol{Z}' \mathbb{I}_{r_j}(s) \boldsymbol{E} \right\|_{\infty}$ for a given r_j.

Recall that G_g represents the group in group lasso for $g = 1, 2, \ldots, n$, where $G_1 = \{1, 2, \ldots, p^2 K\}, G_2 = \{p^2 K + 1, p^2 K + 2, \ldots, 2p^2 K\}, \ldots, G_n = \{(n-1)p^2 K + 1, (n-1)p^2 K +$

$2, \ldots, np^2K\}$. For none overlapping group G_g with size p^2K, we have:

$$\begin{aligned}\frac{1}{\sqrt{p^2K}}\|vec(\boldsymbol{Z}_l'\mathbb{I}_{r_j}(s)\boldsymbol{E}), l \in G_g\|_2 &= \sqrt{\frac{1}{p^2K}\sum_{l\in G_g}|vec(\boldsymbol{Z}_l'\mathbb{I}_{r_j}(s)\boldsymbol{E})|^2} \\ &\leq \max_{l\in G_g}|vec(\boldsymbol{Z}_l'\mathbb{I}_{r_j}(s)\boldsymbol{E})|,\end{aligned} \tag{B.48}$$

where $\boldsymbol{Z}_l'$ represents the l-th row in $\boldsymbol{Z}'$. Thus,

$$\begin{aligned}\frac{1}{\sqrt{p^2K}}\|vec(\boldsymbol{Z}'\mathbb{I}_{r_j}(s)\boldsymbol{E})\|_{2,\infty} &= \max_{g=1,\ldots,n}\frac{1}{\sqrt{p^2K}}\|vec(\boldsymbol{Z}_l'\mathbb{I}_{r_j}(s)\boldsymbol{E}), l \in G_g\|_2 \\ &\leq \max_{g=1,\ldots,n}\max_{l\in G_g}|vec(\boldsymbol{Z}_l'\mathbb{I}_{r_j}(s)\boldsymbol{E})| \\ &= \|\boldsymbol{Z}'\mathbb{I}_{r_j}(s)\boldsymbol{E}\|_\infty.\end{aligned} \tag{B.49}$$

Combining the Lemma 18 and Equation (B.49), we have:

$$\begin{aligned}&\mathbb{P}\left(\sup_{|s|\geq\gamma_n}\left(n\sigma^2(s)\right)^{-1}\|vec\left(\boldsymbol{Z}'\mathbb{I}_{r_j}(s)\boldsymbol{E}\right)\|_{2,\infty} \geq C\sqrt{p^2K}\frac{\log\left(p^2K\right)}{\sqrt{n}}\right) \\ &\leq \mathbb{P}\left(\sup_{|s|\geq\gamma_n}\left(n\sigma^2(s)\right)^{-1}\|\boldsymbol{Z}'\mathbb{I}_{r_j}(s)\boldsymbol{E}\|_\infty \geq C\frac{\log\left(p^2K\right)}{\sqrt{n}}\right) \\ &\leq c_3\exp\left(-c_1\log\left(p^2K\right)\right) + \exp\left(-c_2\frac{n^{\varkappa_1\varkappa_c/2(\varkappa_1+\varkappa_c)}}{(\log n)^{2\varkappa_1\varkappa_c/(\varkappa_1+\varkappa_c)}} + \log\left(np^2K\right)\right).\end{aligned} \tag{B.50}$$

Lemma 20. *Under the Assumptions B1 to B5, for $m < m_0$, there exist constants $c_1, c_2 > 0$ such that*

$$\begin{aligned}\mathbb{P}\Bigg(&\min_{\{s_1,s_2,\ldots,s_m\}\subset\{1,2,\ldots,n\}} L_n(s_1, s_2, \ldots, s_m\ \eta_n) \\ &> \sum_{i=1}^{n}\left\|\epsilon_{\pi(i)}\right\|_2^2 + c_1n\Delta_n - c_2md_n^{*2}\left(n\gamma_n\right)^{3/2}\Bigg) \to 1.\end{aligned} \tag{B.51}$$

where $\Delta_n = \min_{1\leq j\leq m_0+1}|r_j - r_{j-1}|$.

Proof of Lemma 20: The road-map for the proof of Lemma 20 is similar to that of Lemma 4 in Safikhani and Shojaie [2020], once adapted to the TAR modeling framework.

Denote $b_{j'}$ as the order of the given j'-th estimated threshold $s_{j'}$. Since $m < m_0$, there

exists a threshold r_j such that $|s_{j'} - r_j| > \Delta_n/4$. In order to find a lower bound on the sum of the least squares, without loss of generality, we try to find a lower bound for the sum of squared errors plus penalty term in the following three cases: (a) $|s_{j'} - s_{j'-1}| \leq \gamma_n$; (b) there exist two true thresholds r_j, r_{j+1} such that $|s_{j'-1} - r_j| \leq \gamma_n$ and $|s_{j'} - r_{j+1}| \leq \gamma_n$; and (c) otherwise.

Based on the Assumption B5, $\{z_t\}$ is a β-mixing process, then $I(u_0 < z_t \leq u_1)$ is an β-mixing process for fixed u_0 and u_1. By the second inequality of Theorem 1 in [Merlevède et al., 2009], there exists a certain positive constant c_B such that:

$$\left|\sum_{t=1}^{n} I\left(u_0 < z_t \leq u_1\right)\right| < c_B n \mathbb{E}\left|I\left(u_0 < z_t \leq u_1\right)\right| \tag{B.52}$$

with high probability. Since $n\mathbb{E}\left|I\left(u_0 < z_t \leq u_1\right)\right| \leq n\left(u_1 - u_0\right)$,

$$\left|\sum_{t=1}^{n} I\left(u_0 < z_t \leq u_1\right)\right| < n c_B \mathbb{E}\left|I\left(u_0 < z_t \leq u_1\right)\right| \leq n c_B \left|u_1 - u_0\right| \tag{B.53}$$

with high probability. Recall that according to Assumptions B1 and B5, the density of $\{\boldsymbol{\epsilon}_t\}$ and z_t are positive, so

$$\sigma^2\left(s_{j'} - s_{j'-1}\right) \geq c_0\left|s_{j'} - s_{j'-1}\right|, \tag{B.54}$$

where c_0 is certain positive constant.

Use $\hat{\boldsymbol{\theta}}_{s_{j'-1}, s_{j'}}$ to denote the estimated parameter in the estimated regime $(s_{j'} - 1, s_{j'}]$. Recall that $b_{j'}$ represents the order of the given j'-th estimated threshold $s_{j'}$. To simplify the notation, set $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{s_{j'-1}, s_{j'}}$. For case (a), consider the case where $r_j < s_{j'-1} < s_{j'} < r_{j+1}$.

Then,

$$
\begin{aligned}
&\sum_{i=b_{j'-1}+1}^{b_{j'}} \left\| \boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{\theta}} \boldsymbol{Y}_{\pi(i)} \right\|_2^2 \\
=& \sum_{i=b_{j'-1}+1}^{b_{j'}} \left\| \boldsymbol{\epsilon}_{\pi(i)} \right\|_2^2 + \sum_{i=b_{j'-1}}^{b_{j'}-1} \left\| \left(\boldsymbol{A}^{(\cdot,j+1)} - \hat{\boldsymbol{\theta}} \right) \boldsymbol{Y}_{\pi(i)} \right\|_2^2 \\
&+2 \sum_{i=b_{j'-1}+1}^{b_{j'}} \boldsymbol{Y}'_{\pi(i)} \left(\boldsymbol{A}^{(\cdot,j+1)} - \hat{\boldsymbol{\theta}} \right)' \boldsymbol{\epsilon}_{\pi(i)} \\
=& \sum_{i=b_{j'-1}+1}^{b_{j'}} \left\| \boldsymbol{\epsilon}_{\pi(i)} \right\|_2^2 + \sum_{t=1}^{n} \left\| \left(\boldsymbol{A}^{(\cdot,j+1)} - \hat{\boldsymbol{\theta}} \right) \boldsymbol{Y}_t \right\|_2^2 I\left(s_{j'-1} < z_t \leq s_{j'} - 1 \right) \\
&+2 \sum_{i=b_{j'-1}+1}^{b_{j'}} \boldsymbol{Y}'_{\pi(i)} \left(\boldsymbol{A}^{(\cdot,j+1)} - \hat{\boldsymbol{\theta}} \right)' \boldsymbol{\epsilon}_{\pi(i)} \\
\geq& \sum_{i=b_{j'-1}+1}^{b_{j'}} \left\| \boldsymbol{\epsilon}_{\pi(i)} \right\|_2^2 - 2 \left| \sum_{i=b_{j'-1}+1}^{b_{j'}} \boldsymbol{Y}'_{\pi(i)} \left(\boldsymbol{A}^{(\cdot,j+1)} - \hat{\boldsymbol{\theta}} \right)' \boldsymbol{\epsilon}_{\pi(i)} \right| \\
\geq& \sum_{i=b_{j'-1}+1}^{b_{j'}} \left\| \boldsymbol{\epsilon}_{\pi(i)} \right\|_2^2 - c_2 \left| \sum_{i=b_{j'-1}+1}^{b_{j'}} \boldsymbol{Y}'_{\pi(i)} \boldsymbol{\epsilon}_{\pi(i)} \right|_\infty \left\| \boldsymbol{A}^{(\cdot,j+1)} - \hat{\boldsymbol{\theta}} \right\|_1 .
\end{aligned}
\tag{B.55}
$$

In case (a), $|s_{j'} - s_{j'-1}| \leq \gamma_n$. Based on Lemma 16 and Equation (B.53), we can get

$$
\begin{aligned}
&\sum_{i=b_{j'-1}+1}^{b_{j'}} \left\| \boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{\theta}} \boldsymbol{Y}_{\pi(i)} \right\|_2^2 \\
\geq& \sum_{i=b_{j'-1}+1}^{b_{j'}} \left\| \boldsymbol{\epsilon}_{\pi(i)} \right\|_2^2 - c_2 \sqrt{n\gamma_n} \log\left(p^2 K\right) \left\| \boldsymbol{A}^{(\cdot,j+1)} - \hat{\boldsymbol{\theta}} \right\|_1 .
\end{aligned}
\tag{B.56}
$$

According to Assumption B7, we obtain

$$
\begin{aligned}
&\sum_{i=b_{j'-1}+1}^{b_{j'}} \left\| \boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{\theta}} \boldsymbol{Y}_{\pi(i)} \right\|_2^2 + \eta_{(s_{j'-1}, s_{j'})} \left\| \hat{\boldsymbol{\theta}} \right\|_1 \\
\geq& \sum_{i=b_{j'-1}+1}^{b_{j'}} \left\| \boldsymbol{\epsilon}_{\pi(i)} \right\|_2^2 - c_2 \sqrt{n\gamma_n} \log\left(p^2 K\right) \left\| \boldsymbol{A}^{(\cdot,j+1)} \right\|_1 .
\end{aligned}
\tag{B.57}
$$

For case (b), consider the case where $s_{j'-1} < r_j$ and $s_{j'} < r_{j+1}$.

$$\begin{aligned}
&\frac{1}{b_{j'} - b_{j'-1}} \sum_{i=b_{j'-1}+1}^{b_{j'}} \left\| \boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{\theta}} \boldsymbol{Y}_{\pi(i)} \right\|_2^2 + \eta_{(s_{j'-1}, s_{j'})} \left\| \hat{\boldsymbol{\theta}} \right\|_1 \\
\leq &\frac{1}{b_{j'} - b_{j'-1}} \sum_{i=b_{j'-1}+1}^{b_{j'}} \left\| \boldsymbol{x}_{\pi(i)} - \boldsymbol{A}^{(.,j+1)} \boldsymbol{Y}_{\pi(i)} \right\|_2^2 + \eta_{(s_{j'-1}, s_{j'})} \left\| \boldsymbol{A}^{(.,j+1)} \right\|_1 .
\end{aligned} \tag{B.58}$$

By rearrangement, there exist constants $c' > 0$, $c_1 > 0$, $c_2 > 0$, $c_3 > 0$, and $c_4 > 0$ that satisfy

$$\begin{aligned}
0 \leq & c' \left\| \boldsymbol{A}^{(.,j+1)} - \hat{\boldsymbol{\theta}} \right\|_2^2 \\
\leq & \frac{1}{n\sigma^2 (s_{j'} - s_{j'-1})} \sum_{i=b_{j'-1}+1}^{b_{j'}} \boldsymbol{Y}'_{\pi(i)} \left(\boldsymbol{A}^{(.,j+1)} - \hat{\boldsymbol{\theta}} \right)' \left(\boldsymbol{A}^{(.,j+1)} - \hat{\boldsymbol{\theta}} \right) \boldsymbol{Y}_{\pi(i)} \\
\leq & \frac{2}{n\sigma^2 (s_{j'} - s_{j'-1})} \sum_{i=b_{j'-1}+1}^{b_{j'}} \boldsymbol{Y}'_{\pi(i)} \left(\hat{\boldsymbol{\theta}} - \boldsymbol{A}^{(.,j+1)} \right)' \left(\boldsymbol{x}_{\pi(i)} - \boldsymbol{A}^{(.,j+1)} \boldsymbol{Y}'_{\pi(i)} \right) \\
& + \frac{\left| b_{j'} - b_{j'-1} \right|}{n\sigma^2 (s_{j'} - s_{j'-1})} \eta_{(s_{j'-1}, s_{j'})} \left(\left\| \boldsymbol{A}^{(.,j+1)} \right\|_1 - \left\| \hat{\boldsymbol{\theta}} \right\|_1 \right) \\
\leq & \left(c_1 \frac{\log(p^2 K)}{\sqrt{n \left(s_{j'} - s_{j'-1} \right)}} + c_2 M_A d_n^* \frac{\left| b_{j'} - b_{j'-1} \right|}{n\sigma^2 \left(s_{j'} - s_{j'-1} \right)} \right) \left\| \boldsymbol{A}^{(.,j+1)} - \hat{\boldsymbol{\theta}} \right\|_1 \\
& + c_3 \eta_{(s_{j'-1}, s_{j'})} \left(\left\| \boldsymbol{A}^{(.,j+1)} \right\|_1 - \left\| \hat{\boldsymbol{\theta}} \right\|_1 \right) \\
\leq & \left(c_1 \frac{\log(p^2 K)}{\sqrt{n \left(s_{j'} - s_{j'-1} \right)}} + c_2 M_A d_n^* \frac{n\gamma_n}{n \left(s_{j'} - s_{j'-1} \right)} \right) \left\| \boldsymbol{A}^{(.,j+1)} - \hat{\boldsymbol{\theta}} \right\|_1 \\
& + c_3 \eta_{(s_{j'-1}, s_{j'})} \left(\left\| \boldsymbol{A}^{(.,j+1)} \right\|_1 - \left\| \hat{\boldsymbol{\theta}} \right\|_1 \right) \\
\leq & \frac{c_3 \eta_{(s_{j'-1}, s_{j'})}}{2} \left\| \boldsymbol{A}^{(.,j+1)} - \hat{\boldsymbol{\theta}} \right\|_1 + c_3 \eta_{(s_{j'-1}, s_{j'})} \left(\left\| \boldsymbol{A}^{(.,j+1)} \right\|_1 - \left\| \hat{\boldsymbol{\theta}} \right\|_1 \right) \\
\leq & \frac{3 c_3 \eta_{(s_{j'-1}, s_{j'})}}{2} \left\| \boldsymbol{A}^{(.,j+1)} - \hat{\boldsymbol{\theta}} \right\|_{1,\mathcal{I}} - \frac{c_3 \eta_{(s_{j'-1}, s_{j'})}}{2} \left\| \boldsymbol{A}^{(.,j+1)} - \hat{\boldsymbol{\theta}} \right\|_{1,\mathcal{I}^c} \\
\leq & 2 c_3 \eta_{(s_{j'-1}, s_{j'})} \left\| \boldsymbol{A}^{(.,j+1)} - \hat{\boldsymbol{\theta}} \right\|_1 .
\end{aligned} \tag{B.59}$$

This implies $3 \left\| \boldsymbol{A}^{(.,j+1)} - \hat{\boldsymbol{\theta}} \right\|_{1,\mathcal{I}} \geq \left\| \boldsymbol{A}^{(.,j+1)} - \hat{\boldsymbol{\theta}} \right\|_{1,\mathcal{I}^c}$, thus, $4 \left\| \boldsymbol{A}^{(.,j+1)} - \hat{\boldsymbol{\theta}} \right\|_{1,\mathcal{I}} \geq$

$\left\|\boldsymbol{A}^{(.,j+1)}-\hat{\boldsymbol{\theta}}\right\|_1$. By Cauchy–Schwarz inequality, we can get $4\left\|\boldsymbol{A}^{(.,j+1)}-\hat{\boldsymbol{\theta}}\right\|_{1,\mathcal{I}} \leq 4\sqrt{d_n^*}\left\|\boldsymbol{A}^{(.,j+1)}-\hat{\boldsymbol{\theta}}\right\|_2$. In addition, we can get $\left\|\boldsymbol{A}^{(.,j+1)}-\hat{\boldsymbol{\theta}}\right\|_2 \leq c_4\sqrt{d_n^*}\eta_{(b_{j'-1},b_{j'})}$ from Equation (B.59).

Recall that w_j denotes the j-th order of the true threshold. By Equation (B.59), we can use the same procedure as in the case (a). For some constants $c_{h'}>0$ for $h'=5,6,\ldots,11$, we have:

$$
\begin{aligned}
\sum_{i=w_j+1}^{b_{j'}}\left\|\boldsymbol{x}_{\pi(i)}-\hat{\boldsymbol{\theta}}\boldsymbol{Y}_{\pi(i)}\right\|_2^2 &\geq \sum_{i=w_j+1}^{b_{j'}}\left\|\boldsymbol{\epsilon}_{\pi(i)}\right\|_2^2+c_5\left|b_{j'}-w_j\right|\left\|\boldsymbol{A}^{(.,j+1)}-\hat{\boldsymbol{\theta}}\right\|_2^2 \\
&\quad -c_6\sqrt{\left|n(s_{j'}-r_j)\right|}\log(p^2K)\|\boldsymbol{A}^{(.,j+1)}-\hat{\boldsymbol{\theta}}\|_1 \\
&\geq \sum_{i=w_j+1}^{b_{j'}}\left\|\boldsymbol{\epsilon}_{\pi(i)}\right\|_2^2+c_5\left|n(s_{j'}-r_j)\right|\left\|\boldsymbol{A}^{(.,j+1)}-\hat{\boldsymbol{\theta}}\right\|_2 \\
&\quad \left(\left\|\boldsymbol{A}^{(.,j+1)}-\hat{\boldsymbol{\theta}}\right\|_2-\frac{c_6}{c_5}\frac{\log\left(p^2K\right)\sqrt{d_n^*}}{\sqrt{\left|n(s_{j'}-r_j)\right|}}\right) \\
&\geq \sum_{i=w_j+1}^{b_{j'}}\left\|\boldsymbol{\epsilon}_{\pi(i)}\right\|_2^2-c_7\left|n(s_{j'}-r_j)\right|\left\|\boldsymbol{A}^{(.,j+1)}-\hat{\boldsymbol{\theta}}\right\|_2 \\
&\quad \left(c_8\sqrt{d_n^*}\eta_{(s_{j'-1},s_{j'})}+\frac{\log\left(p^2K\right)\sqrt{d_n^*}}{\sqrt{\left|n(s_{j'}-r_j)\right|}}\right) \\
&\geq \sum_{i=w_j+1}^{b_{j'}}\left\|\boldsymbol{\epsilon}_{\pi(i)}\right\|_2^2-c_7\left|n(s_{j'}-r_j)\right|c_9\sqrt{d_n^*}\eta_{(s_{j'-1},s_{j'})} \\
&\quad \left(c_8\sqrt{d_n^*}\eta_{(s_{j'-1},s_{j'})}+\frac{\log\left(p^2K\right)\sqrt{d_n^*}}{\sqrt{\left|n(s_{j'}-r_j)\right|}}\right) \\
&\geq \sum_{i=w_j+1}^{b_{j'}}\left\|\boldsymbol{\epsilon}_{\pi(i)}\right\|_2^2-c_{10}d_n^*\left(\log(p^2K)\right)^2.
\end{aligned}
\tag{B.60}
$$

For the threshold interval $(s_{j'-1},r_j)$, there exist positive constants $C_{h'}$ for $h'=1,2,\ldots,9$

such that

$$
\begin{aligned}
&\sum_{i=b_{j'-1}+1}^{w_j} \left\| \boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{\theta}} \boldsymbol{Y}_{\pi(i)} \right\|_2^2 \\
\geq & \sum_{i=b_{j'-1}+1}^{w_j} \left\| \boldsymbol{\epsilon}_{\pi(i)} \right\|_2^2 - C_2 \sqrt{n\gamma_n} \log\left(p^2 K\right) \left\| \boldsymbol{A}^{(.,j)} - \hat{\boldsymbol{\theta}} \right\|_1 \\
\geq & \sum_{i=b_{j'-1}+1}^{w_j} \left\| \boldsymbol{\epsilon}_{\pi(i)} \right\|_2^2 - C_2 \sqrt{n\gamma_n} \log(p^2 K) \left(\left\| \boldsymbol{A}^{(.,j+1)} - \hat{\boldsymbol{\theta}} \right\|_1 \right. \\
& \left. + \left\| \boldsymbol{A}^{(.,j+1)} - \boldsymbol{A}^{(.,j)} \right\|_1 \right) \\
\geq & \sum_{i=b_{i-1}+1}^{w_i} \left\| \boldsymbol{\epsilon}_{\pi(i)} \right\|_2^2 - C_2 \sqrt{n\gamma_n} \log(p^2 K) \left(d_n^* \eta_{(s_{j'-1}, s_{j'})} + \left\| \boldsymbol{A}^{(.,j+1)} - \boldsymbol{A}^{(.,j)} \right\|_1 \right) \\
\geq & \sum_{i=b_{j'-1}+1}^{w_j} \left\| \boldsymbol{\epsilon}_{\pi(i)} \right\|_2^2 - C_4 d_n^* \left(C_1 \frac{\log(p^2 K)}{\sqrt{n \left| s_{j'} - s_{j'-1} \right|}} + C_3 M_A \frac{\gamma_n}{\left| s_{j'} - s_{j'-1} \right|} \right) \\
& \qquad \sqrt{n\gamma_n} \log(p^2 K) \\
\geq & \sum_{i=b_{j'-1}+1}^{w_j} \left\| \boldsymbol{\epsilon}_{\pi(i)} \right\|_2^2 - C_5 d_n^* \sqrt{n\gamma_n} \left(\log(p^2 K) \right)^2 .
\end{aligned}
\tag{B.61}
$$

By Equation (B.60) and Equation (B.61), for certain constant $C_1' > 0$, we have:

$$
\sum_{i=b_{j'-1}+1}^{b_{j'}} \left\| \boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{\theta}} \boldsymbol{Y}_{\pi(i)} \right\|_2^2 \geq \sum_{i=b_{j'-1}+1}^{b_{j'}} \left\| \boldsymbol{\epsilon}_{\pi(i)} \right\|_2^2 - C_1' d_n^* \sqrt{n\gamma_n} \left(\log\left(p^2 K\right) \right)^2 . \tag{B.62}
$$

In case (c), $s_{j'-1} < r_j < s_{j'}$ with $|s_{j'-1} - r_j| > \Delta_n/4$ and $|s_{j'} - r_j| > \Delta_n/4$. In this case, the restricted eigenvalue condition does not hold, since the distance between two consecutive true thresholds is very large, which leads to a large distance to the intersection of the estimated thresholds. However, if the tuning parameters are chosen properly, then the deterministic part of the deviation bound argument holds. Consider threshold intervals

$(s_{j'-1}, r_j)$ and $(r_j, s_{j'})$

$$\begin{aligned}
&\sum_{i=b_{j'-1}+1}^{w_j} \left\| \boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{\theta}} \boldsymbol{Y}_{\pi(i)} \right\|_2^2 \\
\geq & \sum_{i=b_{j'-1}+1}^{w_j} \left\| \boldsymbol{\epsilon}_{\pi(i)} \right\|_2^2 + \sum_{i=b_{j'-1}+1}^{w_j} \left\| \boldsymbol{A}^{(.,j)} - \hat{\boldsymbol{\theta}} \right\|_2^2 \left\| \boldsymbol{Y}_{\pi(i)} \right\|_2^2 \\
& - 2 \sum_{i=b_{j'-1}+1}^{w_j} \left\| \boldsymbol{Y}_{\pi(i)} \boldsymbol{\epsilon}_{\pi(i)} \left(\boldsymbol{A}^{(.,j)} - \hat{\boldsymbol{\theta}} \right) \right\|_1 .
\end{aligned} \tag{B.63}$$

According to the results from Lemma 18 and Equation (B.63), we have

$$\begin{aligned}
&\sum_{i=b_{j'-1}+1}^{w_j} \left\| \boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{\theta}} \boldsymbol{Y}_{\pi(i)} \right\|_2^2 \\
\geq & \sum_{i=b_{j'-1}+1}^{w_j} \left\| \boldsymbol{\epsilon}_{\pi(i)} \right\|_2^2 + \left\| \boldsymbol{A}^{(.,j)} - \hat{\boldsymbol{\theta}} \right\|_2 \left(\sum_{i=b_{j'-1}+1}^{w_j} \left\| \boldsymbol{Y}_{\pi(i)} \right\|_2^2 \left\| \boldsymbol{A}^{(.,j)} - \hat{\boldsymbol{\theta}} \right\|_2 \right. \\
& \left. - C_6 n \sigma^2 (r_j - s_{j'-1}) \frac{\log(p^2 K) \sqrt{d_n^*}}{\sqrt{n \gamma_n}} \right) \\
\geq & \sum_{i=b_{j'-1}+1}^{w_j} \left\| \boldsymbol{\epsilon}_{\pi(i)} \right\|_2^2 + C_7' \left\| \boldsymbol{A}^{(.,j)} - \hat{\boldsymbol{\theta}} \right\|_2 n \mathbb{E} \left(x_{(t-k,l)} x_{(t-k,l)}^{'} I \left(s_{j'-1} < z_t \leq r_j \right) \right) \\
& \left(\left\| \boldsymbol{A}^{(.,j)} - \hat{\boldsymbol{\theta}} \right\|_2 - C_7 \frac{\log(p^2 K) \sqrt{d_n^*}}{\sqrt{n \gamma_n}} \right) \\
\geq & \sum_{i=b_{j'-1}+1}^{w_j} \left\| \boldsymbol{\epsilon}_{\pi(i)} \right\|_2^2 + C_7' \left\| \boldsymbol{A}^{(.,j)} - \hat{\boldsymbol{\theta}} \right\|_2^2 n \mathbb{E} \left(x_{(t-k,l)} x_{(t-k,l)}^{'} I \left(s_{j'-1} < z_t \leq r_j \right) \right) \\
\geq & \sum_{i=b_{j'-1}+1}^{w_j} \left\| \boldsymbol{\epsilon}_{\pi(i)} \right\|_2^2 + C_8 n \left(r_j - s_{j'-1} \right) \left\| \boldsymbol{A}^{(.,j)} - \hat{\boldsymbol{\theta}} \right\|_2^2 \\
\geq & \sum_{i=b_{j'-1}+1}^{w_j} \left\| \boldsymbol{\epsilon}_{\pi(i)} \right\|_2^2 + C_9 n \Delta_n,
\end{aligned} \tag{B.64}$$

Similarly, we have

$$\begin{aligned}\sum_{i=w_j+1}^{b_{j'}}\left\|\boldsymbol{x}_{\pi(i)}-\hat{\boldsymbol{\theta}}\boldsymbol{Y}_{\pi(i)}\right\|_2^2 &\geq \sum_{i=w_j+1}^{b_{j'}}\left\|\boldsymbol{\epsilon}_{\pi(i)}\right\|_2^2+c_1'\left|b_{j'}-w_j\right|\left\|\boldsymbol{A}^{(.,j+1)}-\hat{\boldsymbol{\theta}}\right\|_2^2 \\ &\quad -c_2\sqrt{\left|b_{j'}-w_j\right|}\log(p^2K)\left\|\boldsymbol{A}^{(.,j+1)}-\hat{\boldsymbol{\theta}}\right\|_1 \\ &\geq \sum_{i=w_j+1}^{b_{j'}}\left\|\boldsymbol{\epsilon}_{\pi(i)}\right\|_2^2+c_1'\left|b_{j'}-w_j\right|\left\|\boldsymbol{A}^{(.,j+1)}-\hat{\boldsymbol{\theta}}\right\|_2 \\ &\quad \left(\left\|\boldsymbol{A}^{(.,j+1)}-\hat{\boldsymbol{\theta}}\right\|_2-\frac{c_2'}{c_1'}\frac{\log(p^2K)\sqrt{d_n^*}}{\sqrt{\left|b_{j'}-w_j\right|}}\right),\end{aligned} \tag{B.65}$$

where c_1', c_2' are some positive constants.

Based on the Assumption B4, $\left\|\boldsymbol{A}^{(.,j+1)}-\boldsymbol{A}^{(.,j)}\right\|_2\geq v>0$, then either $\left\|\boldsymbol{A}^{(.,j+1)}-\hat{\boldsymbol{\theta}}\right\|_2\geq v/4$ or $\left\|\boldsymbol{A}^{(.,j)}-\hat{\boldsymbol{\theta}}\right\|_2\geq v/4$. Assume that $\left\|\boldsymbol{A}^{(.,j)}-\hat{\boldsymbol{\theta}}\right\|_2\geq v/4$. Based on Equation (B.63) and Equation (B.65), when $|s_{j'}-r_j|\leq\gamma_n$, there exist positive constants $c_{h'}'$for $h'=3,4,\ldots,8$ such that:

$$\sum_{i=b_{j'-1}}^{w_j-1}\left\|\boldsymbol{x}_{\pi(i)}-\hat{\boldsymbol{\theta}}\boldsymbol{Y}_{\pi(i)}\right\|_2^2\geq\sum_{i=b_{j'-1}+1}^{w_j}\left\|\boldsymbol{\epsilon}_{\pi(i)}\right\|_2^2+c_3'n\Delta_n, \tag{B.66}$$

and

$$\sum_{i=w_j+1}^{b_{j'}}\left\|\boldsymbol{x}_{\pi(i)}-\hat{\boldsymbol{\theta}}\boldsymbol{Y}_{\pi(i)}\right\|_2^2\geq\sum_{i=w_j+1}^{b_{j'}}\left\|\boldsymbol{\epsilon}_{\pi(i)}\right\|_2^2-c_4'd_n^*\left(\log(p^2K)\right)^2. \tag{B.67}$$

According to Equation (B.66) and Equation (B.67), we can get:

$$\sum_{i=b_{j'-1}+1}^{b_{j'}}\left\|\boldsymbol{x}_{\pi(i)}-\hat{\boldsymbol{\theta}}\boldsymbol{Y}_{\pi(i)}\right\|_2^2\geq\sum_{i=b_{j'-1}+1}^{b_{j'}}\left\|\boldsymbol{\epsilon}_{\pi(i)}\right\|_2^2+c_5'n\Delta_n-c_6'd_n^*\left(\log(p^2K)\right)^2. \tag{B.68}$$

When both $|s_{j'-1}-r_j|>\gamma_n$ and $|s_{j'}-r_j|>\gamma_n$, we can follow the similar steps that obtain Equation (B.68) and obtain:

$$\sum_{i=b_{j'-1}+1}^{b_{j'}}\left\|\boldsymbol{x}_{\pi(i)}-\hat{\boldsymbol{\theta}}\boldsymbol{Y}_{\pi(i)}\right\|_2^2\geq\sum_{i=b_{j'-1}+1}^{b_{j'}}\left\|\boldsymbol{\epsilon}_{\pi(i)}\right\|_2^2+c_7'n\Delta_n-c_8'\left(n\gamma_n\right)^{3/2}d_n^{*2}. \tag{B.69}$$

Combining above three cases (a),(b), and (c), we can prove the results.

Appendix 3.3: Proof of Main Results

B.0.1 Proof of Theorem 6

The idea is similar to Theorem 2 in Safikhani and Shojaie [2020]. However, in our case, we do not have the assumptions related to spectral density matrices. Instead, we assume the random variables are β-mixing, sub-Weibull and stationary. For a matrix $\boldsymbol{A} \in \mathbb{R}^{pK\times p}$, let $\|\boldsymbol{A}\|_{1,\mathcal{I}} = \sum_{(i',l')\in\mathcal{I}} |a_{i'l'}|$, where $\mathcal{I}$ is the set of non-zero indices of $\boldsymbol{A}$ and $a_{i'l'}$ is the element at i'-th row and l'-th column. First, we prove the second part. For some $j = 1, 2, \ldots, m_0$, given the estimated threshold $\hat{r}_j$, $|r_j - \hat{r}_j| > \gamma_n$, there exists a true threshold point r_{j_0} which is isolated from all the estimated thresholds, i.e., $\min_{1\le j\le m_0} |\hat{r}_j - r_{j_0}| > \gamma_n$. In other words, there exists an estimated threshold $\hat{r}_j$ such that, $r_{j_0} - r_{j_0-1} \vee \hat{r}_j \ge \gamma_n$ and $r_{j_0+1} \vee \hat{r}_{j+1} - r_{j_0} \ge \gamma_n$. The idea of the proof is to show the estimated parameters in the interval $[r_{j_0-1} \vee \hat{r}_j, r_{j_0+1} \wedge \hat{r}_{j+1}]$ converges in ℓ_2 to both $\boldsymbol{A}^{(.,j_0)}$ and $\boldsymbol{A}^{(.,j_0+1)}$ which contradicts with the Assumption B4. The length of the interval is large enough to verify restricted eigenvalue and deviation bound inequalities needed to show parameter estimation consistency.

Define a new parameter sequence $\boldsymbol{\varphi}_q$, $q = 1, 2, \ldots, n$ with $\boldsymbol{\varphi}_q = \hat{\boldsymbol{\theta}}_q$ except for two thresholds $q = \hat{r}_j$ and $q = r_{j_0}$. For these two points, set $\boldsymbol{\varphi}_{\hat{r}_j} = \boldsymbol{A}^{(.,j_0)} - \hat{\boldsymbol{A}}_j$ and $\boldsymbol{\varphi}_{r_{j_0}} = \hat{\boldsymbol{A}}_{j+1} - \boldsymbol{A}^{(.,j_0)}$, where $\hat{\boldsymbol{A}}_j = \sum_{q=1}^{w_{j_0-1}\vee\hat{w}_j-1} \hat{\boldsymbol{\theta}}_q$ and $\hat{\boldsymbol{A}}_{j+1} = \sum_{q=1}^{w_{j_0}\vee\hat{w}_j} \hat{\boldsymbol{\theta}}_q$, i.e. $\hat{\boldsymbol{\theta}}_{w_{j_0}\vee\hat{w}_j} = \hat{\boldsymbol{A}}_{j+1} - \hat{\boldsymbol{A}}_j$. Denote $\Psi = \text{vector}(\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, \ldots, \boldsymbol{\varphi}_n) \in \mathbb{R}^{np^2K\times 1}$. By Equation (3.5) and focusing on the case of lasso (i.e. hdTAR), we have

$$\begin{aligned} &\frac{1}{n}\left\|\boldsymbol{Y} - \boldsymbol{Z}\hat{\boldsymbol{\Theta}}\right\|_2^2 + \lambda_{1,n}\left\|\hat{\boldsymbol{\Theta}}\right\|_1 + \lambda_{2,n}\sum_{i=1}^{n}\left\|\sum_{i'=1}^{i}\hat{\boldsymbol{\theta}}_{i'}\right\|_1 \\ &\le \frac{1}{n}\|\boldsymbol{Y} - \boldsymbol{Z}\Psi\|_2^2 + \lambda_{1,n}\|\Psi\|_1 + \lambda_{2,n}\sum_{i=1}^{n}\left\|\sum_{i'=1}^{i}\boldsymbol{\varphi}_{i'}\right\|_1. \end{aligned} \tag{B.70}$$

By rearrangement, for a constant c, we can get

$$
\begin{aligned}
0 \leq & c \left\| \boldsymbol{A}^{(.,j_0)} - \hat{\boldsymbol{A}}_{j+1} \right\|_2^2 \\
\leq & \frac{1}{n\sigma^2(r_{j_0} - r_{j_0-1} \vee \hat{r}_j)} \left\| \sum_{i=(w_{j_0-1} \vee \hat{w}_j)+1}^{w_{j_0}} \boldsymbol{Y}'_{\pi(i)} \left(\boldsymbol{A}^{(.,j_0)} - \hat{\boldsymbol{A}}_{j+1} \right) \right\|_2^2 \\
\leq & \frac{2}{n\sigma^2(r_{j_0} - r_{j_0-1} \vee \hat{r}_j)} \sum_{i=(w_{j_0-1} \vee \hat{w}_j)+1}^{w_{j_0}} \boldsymbol{Y}'_{\pi(i)} \left(\boldsymbol{A}^{(.,j_0)} - \hat{\boldsymbol{A}}_{j+1} \right) \boldsymbol{\epsilon}_{\pi(i)} \\
& + \frac{n\lambda_{1,n}}{n\sigma^2(r_{j_0} - r_{j_0-1} \vee \hat{r}_j)} \Bigg(\left\| \boldsymbol{A}^{(.,j_0)} - \hat{\boldsymbol{A}}_{j+1} \right\|_1 + \left\| \boldsymbol{A}^{(.,j_0)} - \hat{\boldsymbol{A}}_j \right\|_1 \\
& \qquad - \left\| \hat{\boldsymbol{A}}_{j+1} - \hat{\boldsymbol{A}}_j \right\|_1 \Bigg) \\
& + \frac{n\lambda_{2,n}}{n\sigma^2 \left(r_{j_0} - r_{j_0-1} \vee \hat{r}_j\right)} n\sigma^2 \left(r_{j_0} - r_{j_0-1} \vee \hat{r}_j\right) \left(\left\| \boldsymbol{A}^{(.,j_0)} \right\|_1 - \left\| \hat{\boldsymbol{A}}_{j+1} \right\|_1 \right) \\
\leq & \left(\frac{2n\lambda_{1,n}}{n\sigma^2 \left(r_{j_0} - r_{j_0-1} \vee \hat{r}_j\right)} + C \frac{\log(p^2 K)}{\sqrt{n\gamma_n}} \right) \left\| \boldsymbol{A}^{(.,j_0)} - \hat{\boldsymbol{A}}_{j+1} \right\|_1 \\
& + n\lambda_{2,n} \left(\left\| \boldsymbol{A}^{(.,j_0)} \right\|_1 - \left\| \hat{\boldsymbol{A}}_{j+1} \right\|_1 \right) \\
\leq & \frac{n\lambda_{2,n}}{2} \left\| \boldsymbol{A}^{(.,j_0)} - \hat{\boldsymbol{A}}_{j+1} \right\|_1 + n\lambda_{2,n} \left(\left\| \boldsymbol{A}^{(.,j_0)} \right\|_1 - \left\| \hat{\boldsymbol{A}}_{j+1} \right\|_1 \right) \\
\leq & \frac{3n\lambda_{2,n}}{2} \left\| \boldsymbol{A}^{(.,j_0)} - \hat{\boldsymbol{A}}_{j+1} \right\|_{1,\mathcal{I}} - \frac{n\lambda_{2,n}}{2} \left\| \boldsymbol{A}^{(.,j_0)} - \hat{\boldsymbol{A}}_{j+1} \right\|_{1,\mathcal{I}^c}.
\end{aligned}
\tag{B.71}
$$

According to Lemma 18 and the fact that $r_{j_0} - r_{j_0-1} \vee \hat{r}_j \geq \gamma_n$, the second inequality holds with high probability converging to 1 in Equation (B.71). The third inequality holds because $w_{j_0} - w_{j_0-1} \vee \hat{w}_j \leq c_1 n\sigma^2(s)$ for a certain positive constant c_1. The fourth inequality holds with high probability converging to 1 according to second part of Lemma 18 and triangular inequality. The fifth inequality is based on Assumption B4 and the selection for $\lambda_{2,n}$ in the statement of the theorem. The last inequality holds by Assumption B3. Thus,

$$
\left\| \boldsymbol{A}^{(.,j_0)} - \hat{\boldsymbol{A}}_{j+1} \right\|_2 = o_p \left(d_n^* \frac{\log\left(p^2 K\right)}{\sqrt{n\gamma_n}} \right), \tag{B.72}
$$

which means that it converges to zero in probability based on Assumption B4. The convergence of $\left\| \boldsymbol{A}^{(.,j_0+1)} - \hat{\boldsymbol{A}}_{j+1} \right\|_2$ can be proved in the same procedure in the interval $[r_{j_0}, r_{j_0+1} \wedge \hat{r}_{j+1}]$, which contradicts Assumption B4. Thus, the results are as desired.

Similarly, we can prove the first part. Assume $\left|\hat{\mathcal{A}}_n\right| < m_0$. There exist an isolated true threshold r_{j_0} such that $r_{j_0} - r_{j_0-1} \vee \hat{r}_j \geq \gamma_n/3$ and $r_{j_0+1} \wedge \hat{r}_{j+1} - r_{j_0} \geq \gamma_n/3$. Similar procedure in the second part is applied to the interval $[r_{j_0-1} \vee \hat{r}_j, r_{j_0}]$ and $[r_{j_0}, r_{j_0-1} \wedge \hat{r}_{j+1}]$, then the proof is completed for the hdTAR case.

Similar procedure can be applied to mvTAR which is briefly described next. We obtain:

$$
\begin{aligned}
0 \leq\ & c\left\|\boldsymbol{A}^{(.,j_0)}-\hat{\boldsymbol{A}}_{j+1}\right\|_2^2 \\
\leq\ & \frac{1}{n\sigma^2(r_{j_0}-r_{j_0-1}\vee\hat{r}_j)}\left\|\sum_{i=(w_{j_0-1}\vee\hat{w}_j)+1}^{w_{j_0}}\boldsymbol{Y}'_{\pi(i)}\left(\boldsymbol{A}^{(.,j_0)}-\hat{\boldsymbol{A}}_{j+1}\right)\right\|_2^2 \\
\leq\ & \frac{2}{n\sigma^2(r_{j_0}-r_{j_0-1}\vee\hat{r}_j)}\sum_{i=(w_{j_0-1}\vee\hat{w}_j)+1}^{w_{j_0}}\boldsymbol{Y}'_{\pi(i)}\left(\boldsymbol{A}^{(.,j_0)}-\hat{\boldsymbol{A}}_{j+1}\right)\boldsymbol{\epsilon}_{\pi(i)} \\
& +\frac{n\lambda_{1,n}}{n\sigma^2(r_{j_0}-r_{j_0-1}\vee\hat{r}_j)}\Big(\left\|\boldsymbol{A}^{(.,j_0)}-\hat{\boldsymbol{A}}_{j+1}\right\|_2+\left\|\boldsymbol{A}^{(.,j_0)}-\hat{\boldsymbol{A}}_{j}\right\|_2 \\
& \qquad\qquad -\left\|\hat{\boldsymbol{A}}_{j+1}-\hat{\boldsymbol{A}}_{j}\right\|_2\Big) \\
& +\frac{n\lambda_{2,n}}{n\sigma^2\left(r_{j_0}-r_{j_0-1}\vee\hat{r}_j\right)}n\sigma^2\left(r_{j_0}-r_{j_0-1}\vee\hat{r}_j\right)\left(\left\|\boldsymbol{A}^{(.,j_0)}\right\|_1-\left\|\hat{\boldsymbol{A}}_{j+1}\right\|_1\right) \\
\leq\ & \frac{2}{n\sigma^2(r_{j_0}-r_{j_0-1}\vee\hat{r}_j)}\left\|\boldsymbol{Z}'\mathbb{I}_{r_j}(r_{j_0}-r_{j_0-1}\vee\hat{r}_j))\boldsymbol{E}\right\|_{2,\infty}\left\|\left(\boldsymbol{A}^{(.,j_0)}-\hat{\boldsymbol{A}}_{j+1}\right)\right\|_{2,1} \\
& +\frac{n\lambda_{1,n}}{n\sigma^2(r_{j_0}-r_{j_0-1}\vee\hat{r}_j)}\Big(\left\|\boldsymbol{A}^{(.,j_0)}-\hat{\boldsymbol{A}}_{j+1}\right\|_2+\left\|\boldsymbol{A}^{(.,j_0)}-\hat{\boldsymbol{A}}_{j}\right\|_2 \\
& \qquad\qquad -\left\|\hat{\boldsymbol{A}}_{j+1}-\hat{\boldsymbol{A}}_{j}\right\|_2\Big) \\
& +\frac{n\lambda_{2,n}}{n\sigma^2\left(r_{j_0}-r_{j_0-1}\vee\hat{r}_j\right)}n\sigma^2\left(r_{j_0}-r_{j_0-1}\vee\hat{r}_j\right)\left(\left\|\boldsymbol{A}^{(.,j_0)}\right\|_1-\left\|\hat{\boldsymbol{A}}_{j+1}\right\|_1\right) \\
\leq\ & \left(\frac{2n\lambda_{1,n}}{n\sigma^2\left(r_{j_0}-r_{j_0-1}\vee\hat{r}_j\right)}+C\frac{\sqrt{p^2K}\log(p^2K)}{\sqrt{n\gamma_n}}\right)\left\|\boldsymbol{A}^{(.,j_0)}-\hat{\boldsymbol{A}}_{j+1}\right\|_2 \\
& +n\lambda_{2,n}\left(\left\|\boldsymbol{A}^{(.,j_0)}\right\|_1-\left\|\hat{\boldsymbol{A}}_{j+1}\right\|_1\right) \\
\leq\ & \left(\frac{2n\lambda_{1,n}}{n\sigma^2\left(r_{j_0}-r_{j_0-1}\vee\hat{r}_j\right)}+C\frac{\sqrt{p^2K}\log(p^2K)}{\sqrt{n\gamma_n}}\right)\left\|\boldsymbol{A}^{(.,j_0)}-\hat{\boldsymbol{A}}_{j+1}\right\|_1 \\
& +n\lambda_{2,n}\left(\left\|\boldsymbol{A}^{(.,j_0)}\right\|_1-\left\|\hat{\boldsymbol{A}}_{j+1}\right\|_1\right) \\
\leq\ & \frac{n\lambda_{2,n}}{2}\left\|\boldsymbol{A}^{(.,j_0)}-\hat{\boldsymbol{A}}_{j+1}\right\|_1+n\lambda_{2,n}\left(\left\|\boldsymbol{A}^{(.,j_0)}\right\|_1-\left\|\hat{\boldsymbol{A}}_{j+1}\right\|_1\right) \\
\leq\ & \frac{3n\lambda_{2,n}}{2}\left\|\boldsymbol{A}^{(.,j_0)}-\hat{\boldsymbol{A}}_{j+1}\right\|_{1,\mathcal{I}}-\frac{n\lambda_{2,n}}{2}\left\|\boldsymbol{A}^{(.,j_0)}-\hat{\boldsymbol{A}}_{j+1}\right\|_{1,\mathcal{I}^c}.
\end{aligned}
\tag{B.73}
$$

Note that there is only one group in $\left\|\boldsymbol{A}^{(.,j_0)}-\hat{\boldsymbol{A}}_{j+1}\right\|_{2,1}$, so it is $\left\|\boldsymbol{A}^{(.,j_0)}-\hat{\boldsymbol{A}}_{j+1}\right\|_2$. According to the first part in Lemma 18 and the fact that $r_{j_0}-r_{j_0-1}\vee\hat{r}_j\geq\gamma_n$, the second

inequality holds with high probability converging to 1 in Equation (B.73). The third inequality holds because $w_{j_0} - w_{j_0-1} \vee \hat{w}_j \leq c_1 n\sigma^2(s)$ for a certain positive constant c_1. The fifth inequality holds with high probability converging to 1 according to Lemma 19 and triangular inequality. The sixth inequality holds due to Minkowski inequality. The seventh inequality is based on Assumption B4 and the selection for $\lambda_{2,n}$ in the statement of the theorem. The last inequality holds by Assumption B3. Thus,

$$\left\| \boldsymbol{A}^{(\cdot,j_0)} - \hat{\boldsymbol{A}}_{j+1} \right\|_2 = o_p\left(d_n^* \frac{\sqrt{p^2 K}\log\left(p^2 K\right)}{\sqrt{n\gamma_n}} \right), \tag{B.74}$$

The remaining parts are similar to hdTAR case, hence details are omitted to avoid duplication. This completes the proof for both cases.

B.0.2 Proof of Theorem 7

For the first part, we want to prove that $\mathbb{P}(\tilde{m} < m_0) \to 0$, and $\mathbb{P}(\tilde{m} > m_0) \to 0$. From Theorem 6, we know that there exist estimated thresholds $\hat{r}_j \in \hat{\mathcal{A}}_n$ such that $\max_{1\leq j\leq m_0} |\hat{r}_j - r_j| \leq \gamma_n$, where $r_j \in \mathcal{A}_n$. Recall that w_j denotes the j-th order of the thresholds, and $b_{j'}$ denotes the j'-th order of the estimated threshold.

Without loss of generality, we only show one of the estimated regimes. For $s_{j'-1} < r_j < s_i$ with $\left|r_j - s_{j'-1}\right| \leq \gamma_n$, the estimated coefficient is denoted as $\hat{\boldsymbol{\theta}}$ in $\left(s_{j'-1}, s_{j'}\right)$. Similar to case (b) in the proof of Lemma 20, recall that $\left|b_{j'} - w_j\right| \leq nc_B \left|s_{j'} - r_j\right|$ according to

Equation (B.53). For the threshold interval $\left(r_j, s_{j'}\right)$, we have

$$
\begin{aligned}
&\sum_{i=w_j+1}^{b_{j'}} \left\| \boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{\theta}} \boldsymbol{Y}_{\pi(i)} \right\|_2^2 \\
\leq & \sum_{i=w_j+1}^{b_{j'}} \left\| \boldsymbol{\epsilon}_{\pi(i)} \right\|_2^2 + c_3 \left| b_{j'} - w_j \right| \left\| \boldsymbol{A}^{(.,j+1)} - \hat{\boldsymbol{\theta}} \right\|_2^2 \\
& + c_4 \sqrt{\left| b_{j'} - w_j \right|} \log(p^2 K) \left\| \boldsymbol{A}^{(.,j+1)} - \hat{\boldsymbol{\theta}} \right\|_1 \\
\leq & \sum_{i=w_j+1}^{b_{j'}} \left\| \boldsymbol{\epsilon}_{\pi(i)} \right\|_2^2 \\
& + c_5 n \left| s_{j'} - r_j \right| d_n^* \left(c_1 \frac{\log(p^2 K)}{\sqrt{n \left(s_{j'} - s_{j'-1} \right)}} + c_2 M_A d_n^* \frac{\gamma_n}{s_{j'} - s_{j'-1}} \right)^2 \\
& + c_6 \sqrt{\left| b_i - w_j \right|} \log \left(p^2 K \right) d_n^* \left(c_1 \frac{\log(p^2 K)}{\sqrt{n \left(s_{j'} - s_{j'-1} \right)}} + c_2 M_A d_n^* \frac{\gamma_n}{s_{j'} - s_{j'-1}} \right) \\
\leq & \sum_{i=w_j+1}^{b_{j'}} \left\| \boldsymbol{\epsilon}_{\pi(i)} \right\|_2^2 + c_7 \sqrt{n \gamma_n} \left(\log \left(p^2 K \right) d_n^* \right)^2 \\
\leq & \sum_{i=w_j+1}^{b_{j'}} \left\| \boldsymbol{\epsilon}_{\pi(i)} \right\|_2^2 + c \left(n \gamma_n \right)^{3/2} d_n^{*2},
\end{aligned}
\tag{B.75}
$$

where c, $c_{h'}$ are positive constants for $h' = 1, 2, \ldots, 7$.

Let $c'_{h'}$ be positive constants for $h' = 1, 2, \ldots, 6$. For regime (s_{i-1}, r_j), we can get

$$
\begin{aligned}
&\sum_{i=b_{j'-1}+1}^{w_j} \left\| \boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{\theta}} \boldsymbol{Y}_{\pi(i)} \right\|_2^2 \\
\leq &\sum_{i=b_{j'-1}+1}^{w_j} \left\| \boldsymbol{\epsilon}_{\pi(i)} \right\|_2^2 + c'_1 \left| w_j - b_{j'-1} \right| \left\| \boldsymbol{A}^{(.,j)} - \hat{\boldsymbol{\theta}} \right\|_2^2 \\
&+ c'_2 \sqrt{\left| w_j - b_{j'-1} \right|} \log\left(p^2 K\right) \left\| \boldsymbol{A}^{(.,j)} - \hat{\boldsymbol{\theta}} \right\|_1 \\
\leq &\sum_{i=b_{j'-1}+1}^{w_j} \left\| \boldsymbol{\epsilon}_{\pi(i)} \right\|_2^2 + c'_1 \left| w_j - b_{j'-1} \right| \left(\left\| \boldsymbol{A}^{(.,j+1)} - \hat{\boldsymbol{\theta}} \right\|_2^2 + \left\| \boldsymbol{A}^{(.,j+1)} - \boldsymbol{A}^{(.,j)} \right\|_2^2 \right) \\
&+ c'_2 \sqrt{\left| w_j - b_{j'-1} \right|} \log\left(p^2 K\right) \left(\left\| \boldsymbol{A}^{(.,j+1)} - \hat{\boldsymbol{\theta}} \right\|_1 + \left\| \boldsymbol{A}^{(.,j+1)} - \boldsymbol{A}^{(.,j)} \right\|_1 \right) \\
\leq &\sum_{i=b_{j'-1}+1}^{w_j} \left\| \boldsymbol{\epsilon}_{\pi(i)} \right\|_2^2 + c'_3 d_n^{*2} \sqrt{n\gamma_n} \left(\log\left(p^2 K\right) \right)^2 \\
\leq &\sum_{i=w_j+1}^{b_{j'}} \left\| \boldsymbol{\epsilon}_{\pi(i)} \right\|_2^2 + c'_4 \left(n\gamma_n\right)^{3/2} d_n^{*2}.
\end{aligned}
\tag{B.76}
$$

Since

$$
\eta_{(s_{j'-1}, s_{j'})} \left\| \hat{\boldsymbol{\theta}} \right\|_1 \leq \eta_{(s_{j'-1}, s_{j'})} \left(\left\| \boldsymbol{A}^{(.,j+1)} - \hat{\boldsymbol{\theta}} \right\|_1 + \left\| \boldsymbol{A}^{(.,j+1)} \right\|_1 \right) \leq c'_5 d_n^*, \tag{B.77}
$$

we combine Equation (B.75) to Equation (B.77) and get

$$
\sum_{i=b_{j'-1}}^{b_{j'}-1} \left\| \boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{\theta}} \boldsymbol{Y}_{\pi(i)} \right\|_2^2 + \eta_{(b_{j'-1}, b_{j'})} \left\| \hat{\boldsymbol{\theta}} \right\|_1 \leq \sum_{i=b_{j'-1}}^{b_{j'}-1} \left\| \boldsymbol{\epsilon}_{\pi(i)} \right\|_2^2 + c'_6 \left(n\gamma_n\right)^{3/2} d_n^{*2}. \tag{B.78}
$$

Since there are $m_0 + 1$ regimes, we can get:

$$
L_n(\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_{m_0}\, ; \eta_n) \leq \sum_{i=1}^{n} \left\| \boldsymbol{\epsilon}_{\pi(i)} \right\|_2^2 + c'_7 m_0 \left(n\gamma_n\right)^{3/2} d_n^{*2}. \tag{B.79}
$$

Given subset from the candidate thresholds found in Step 1. Let $C_{h'}$ be positive constants

for $h' = 1, 2, \ldots, 6$. Assume $\tilde{m} < m_0$. By Lemma 20, we can get

$$\begin{aligned}
&\mathrm{IC}\left(\tilde{r}_1, \ldots, \tilde{r}_{\tilde{m}}\right) \\
&= L_n\left(\tilde{r}_1, \ldots, \tilde{r}_{\tilde{m}}\,; \eta_n\right) + \tilde{m}\omega_n \\
&> \sum_{i=1}^{n} \left\|\boldsymbol{\epsilon}_{\pi(i)}\right\|_2^2 + C_1 n\Delta_n - C_2\tilde{m}d_n^{*2}\left(n\gamma_n\right)^{3/2} + \tilde{m}\omega_n \\
&\geq L_n(\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_{m_0}\,; \eta_n) + m_0\omega_n + C_1 n\Delta_n - C_2\tilde{m}d_n^{*2}\left(n\gamma_n\right)^{3/2} \\
&\quad - C_3 m_0\left(n\gamma_n\right)^{3/2} d_n^{*2} - \left(m_0 - \tilde{m}\right)\omega_n \\
&\geq L_n(\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_{m_0}\,; \eta_n) + m_0\omega_n + C_1 n\Delta_n - C_4 m_0\left(n\gamma_n\right)^{3/2} d_n^{*2} - (m_0 - \tilde{m})\omega_n.
\end{aligned} \tag{B.80}$$

According to Assumption B6, we have

$$m_0\left(n\gamma_n\right)^{3/2} d_n^{*2}/\omega_n \to 0 \text{ and } m_0\omega_n/n\Delta_n \to 0.$$

Then,

$$C_1 n\Delta_n - C_4 m_0\left(n\gamma_n\right)^{3/2} d_n^{*2} - \left(m_0 - \tilde{m}\right)\omega_n \geq 0. \tag{B.81}$$

Thus, $\mathrm{IC}(\tilde{r}_1, \tilde{r}_2, \ldots, \tilde{r}_{\tilde{m}}) \geq L_n\left(\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_{m_0}\,; \eta_n\right) + m_0\omega_n$, which proves

$$\mathbb{P}\left(\tilde{m} < m_0\right) \to 0.$$

Next, we want to prove $\mathbb{P}(\tilde{m} > m_0) \to 0$. Similar procedure in Lemma 20 can be used to get:

$$L_n\left(\tilde{r}_1, \tilde{r}_2, \ldots, \tilde{r}_{\tilde{m}}\,; \eta_n\right) \geq \sum_{i=1}^{n} \left\|\boldsymbol{\epsilon}_{\pi(i)}\right\|_2^2 - C_5\tilde{m}d_n^{*2}\left(n\gamma_n\right)^{3/2}. \tag{B.82}$$

Then,

$$\begin{aligned}\sum_{i=1}^{n}\left\|\boldsymbol{\epsilon}_{\pi(i)}\right\|_2^2-C_5\tilde{m}d_n^{*2}\left(n\gamma_n\right)^{3/2}+\tilde{m}\omega_n &\leq \mathrm{IC}(\tilde{r}_1,\tilde{r}_2,\cdots,\tilde{r}_{\tilde{m}})\\ &\leq \mathrm{IC}(\hat{r}_1,\hat{r}_2,\cdots,\hat{r}_{m_0})\\ &\leq \sum_{i=1}^{n}\left\|\boldsymbol{\epsilon}_{\pi(i)}\right\|_2^2+C_6m_0\left(n\gamma_n\right)^{3/2}d_n^{*2}\\ &\quad +m_0\omega_n.\end{aligned} \tag{B.83}$$

Thus,

$$\left(\tilde{m}-m_0\right)\omega_n\leq C_5\left(n\gamma_n\right)^{3/2}\tilde{m}d_n^{*2}+C_6m_0\left(n\gamma_n\right)^{3/2}d_n^{*2}. \tag{B.84}$$

If $\tilde{m}>m_0$, it contradicts assumption that $m_0\left(n\gamma_n\right)^{3/2}d_n^{*2}/\omega_n\to 0$. Then, we can get $\mathbb{P}\left(\tilde{m}-m_0\right)\to 1$.

Next, we prove $\mathbb{P}\left(\max_{1\leq j\leq m_0}\left|\tilde{r}_j-r_j\right|\leq Bm_0\left(\gamma_n\right)^{3/2}d_n^{*2}\sqrt{n}\right)$. Given certain two constants $C_7>0$ and $c'>0$, let $B=2C_7/c'$. Suppose there exists a threshold r_j such that $\min_{1\leq j\leq m_0}\left|\tilde{r}_j-r_j\right|\geq Bm_0\left(\gamma_n\right)^{3/2}d_n^{*2}\sqrt{n}$. Applying similar procedure to Lemma 20, we can get:

$$\begin{aligned}\sum_{i=1}^{n}\left\|\boldsymbol{\epsilon}_{\pi(i)}\right\|_2^2+c'Bm_0\left(n\gamma_n\right)^{3/2}d_n^{*2} &\leq L_n(\tilde{r}_1,\tilde{r}_2,\ldots,\tilde{r}_{m_0};\eta_n)\\ &\leq L_n(\hat{r}_1,\hat{r}_2,\ldots,\hat{r}_{m_0};\eta_n)\\ &\leq \sum_{i=1}^{n}\left\|\boldsymbol{\epsilon}_{\pi(i)}\right\|_2^2+C_7m_0\left(n\gamma_n\right)^{3/2}d_n^{*2},\end{aligned} \tag{B.85}$$

which contradicts the value of B.

B.0.3 Proof of Theorem 8

Theorem 8 can be proved according to the modification of Corollary 9 in Wong et al. [2020]. In Corollary 9 [Wong et al., 2020], we know that for regime j, we have

$$\|\text{vec}(\hat{\boldsymbol{\beta}}^{(\cdot,j)}) - \text{vec}(\boldsymbol{A}^{(\cdot,j)})\|_2 \leq c_1 \alpha_j \sqrt{d_n^*} \tag{B.86}$$

with high probability, where α_j represent a tuning parameter determined by p, K, and the number of observations in each regime j. Let $\tilde{w}_j$ be the order of estimated threshold $\tilde{r}_j$. What is left is to find a lower bound on the number of observations. Due to Assumption B5, z_t has positive density. Combining Assumption B5 and Corollary 9 of Wong et al. [2020], we have:

$$\begin{aligned} \tilde{w}_{j-1} - \tilde{w}_j &= n\mathbb{P}\left(\tilde{r}_{j-1} < z_t \leq \tilde{r}_j\right) \\ &\geq c_2 n \left|\tilde{r}_j - \tilde{r}_{j-1}\right|, \end{aligned} \tag{B.87}$$

where $c_2 > 0$ is a constant. Now, by plugging in the optimal value of α_j, we get

$$\begin{aligned} \left\|\hat{\boldsymbol{\beta}}^{(\cdot,j)} - \boldsymbol{A}^{(\cdot,j)}\right\|_2 &\leq c_3 \sqrt{\frac{d_n^* \log\left(p^2 K\right)}{\left(\tilde{w}_{j-1} - \tilde{w}_j\right)}} \\ &\leq c_4 \sqrt{\frac{d_n^* \log\left(p^2 K\right)}{n\gamma_n}}, \end{aligned} \tag{B.88}$$

where $c_3, c_4 > 0$ are constants.

Appendix 3.4: A Sufficient Condition for β-mixing

In this section, we provide a sufficient condition for the TAR process $\boldsymbol{x}_t$ to be β-mixing by imposing a restriction on the operator norm of transition matrices. To that end, note that the TAR process,

$$\boldsymbol{x}_t = \sum_{k=1}^{K} \boldsymbol{A}^{(k,j)} \boldsymbol{x}_{t-k} + \boldsymbol{\epsilon}_t, \tag{B.89}$$

can be rewritten as a Kp-dimensional TAR process with lag 1; that is,

$$\tilde{\boldsymbol{X}}_t = \tilde{\boldsymbol{B}}^{(j)} \tilde{\boldsymbol{X}}_{t-1} + \tilde{\boldsymbol{U}}_t, \tag{B.90}$$

where $\tilde{\boldsymbol{X}}_t = \begin{pmatrix} \boldsymbol{x}_t' & \boldsymbol{x}_{t-1}' & \dots & \boldsymbol{x}_{t-K+1}' \end{pmatrix}' \in \mathbb{R}^{Kp \times 1}$,

$$\tilde{\boldsymbol{B}}^{(j)} = \begin{pmatrix} \boldsymbol{A}^{(1,j)} & \boldsymbol{A}^{(k,j)} & \dots & \boldsymbol{A}^{(K-1,j)} & \boldsymbol{A}^{(K,j)} \\ \boldsymbol{I}_p & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_p & & \boldsymbol{0} & \boldsymbol{0} \\ \vdots & & \ddots & \vdots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{I}_p & \boldsymbol{0} \end{pmatrix} \in \mathbb{R}^{Kp \times Kp}$$

for $\boldsymbol{I}_p$ is a $p \times p$ identity matrix, and $\tilde{\boldsymbol{U}}_t = \begin{pmatrix} \boldsymbol{\epsilon}_t' & \boldsymbol{0} & \dots & \boldsymbol{0} \end{pmatrix}' \in \mathbb{R}^{KP \times 1}$. Let $\tilde{\boldsymbol{B}}_{\max} = \arg\max_{j=1,\dots,m_0+1} \left|\!\left|\!\left| \tilde{\boldsymbol{B}}^{(j)} \right|\!\right|\!\right|$ where $\left|\!\left|\!\left| \tilde{\boldsymbol{B}} \right|\!\right|\!\right|$ denotes the operator norm of matrix $\tilde{\boldsymbol{B}}$; that is $\left|\!\left|\!\left| \tilde{\boldsymbol{B}} \right|\!\right|\!\right| = \sqrt{\lambda_{\max}\left(\tilde{\boldsymbol{B}}' \tilde{\boldsymbol{B}}\right)}$.

Lemma 21. *For the TAR model in Equation (B.89), if $\left|\!\left|\!\left| \tilde{\boldsymbol{B}}_{\max} \right|\!\right|\!\right| < 1$, then $\boldsymbol{x}_t$ is β-mixing with a geometrically decaying mixing coefficient. If, in addition, $\boldsymbol{\epsilon}_t$ follows a sub-Weibull distribution, then $\boldsymbol{x}_t$ is also sub-Weibull. In other words, Assumption B2 holds.*

Remark 3. *The condition based on the operator norm of transition matrices may not be the optimal for $\boldsymbol{x}_t$ to be β-mixing and sub-Weibull, and a condition based on the spectral norm could be less restrictive. However, a condition based on the spectral norm does not seem achievable as the argument used for VAR models does not hold in this case. Specifically, in VAR models, we have a sufficient condition based on the spectral norm according to*

Lemma 8.2 in Fan and Yao [2005] stating that the geometric Ergodicity of any subsequence with deterministic index entails the geometric Ergodicity of the original series. But this result does not hold for the TAR models, as the index of the sub-sequence in the TAR model is not deterministic.

Proof of Lemma 21: The proof of Lemma 21 is similar to that in Appendix E.1 of Wong et al. [2020]. We mainly need to apply Proposition 1 and Proposition 2 in Liebscher [2005] and the fact that any measurable function of a stationary process is β-mixing if the original stationary process is β-mixing. Proposition 1 in Liebscher [2005] gives the result that the sequence is geometrically Ergodic based on certain conditions, and we can show that the sequence will be β-mixing with geometrically decaying mixing coefficients, by using Proposition 2 in Liebscher [2005]. Finally, we verify the sub-Weibull assumption by using the definition of sub-Weibull distributions.

To apply Proposition 1 in Liebscher [2005], we check the three conditions, where we set the corresponding parameters $E = \mathbb{R}^p$, and μ as the Lebesgue measure. Condition (i) is satisfied if we set the parameter m in the Proposition 1 of Liebscher [2005] to 1. (Note that here m is not the number of thresholds.) For condition (ii), we set $\bar{m} = \lceil \inf_{u\in C, v\in A} \|u-v\|_2 \rceil$ the minimum "distance" between the sets C and set A in Liebscher [2005], where A is any set that $A \in \mathcal{B}$ where $\mathcal{B}$ is the σ-algebra of Borel sets of E, and C is any compact set that $C \subset E$. Since C is bounded and A is Borel, $\bar{m}$ is finite. For condition (iii), the function $Q(\cdot) = \|\cdot\|$ and set $K_c = \{x \in \mathbb{R}^p : \|x\| \leq \frac{4C_{ac}}{c}\}$ where $c = 1 - \left|\left|\left| \tilde{\boldsymbol{B}}_{\max} \right|\right|\right|$ and $C_{ac} := \mathbb{E}\|\boldsymbol{\epsilon}_t\|$. Since $\max_{j=1,2,\ldots,m_0+1} \left|\left|\left| \tilde{\boldsymbol{B}}^{(j)} \right|\right|\right| < 1$,

- For all $\tilde{y} \in E \setminus K_c$; i.e. $\tilde{y}$ such that $\|\tilde{y}\| > \frac{4C_{ac}}{c}$,

$$\begin{aligned}
\mathbb{E}\left[\|\tilde{\boldsymbol{X}}_{t+1}\| | \tilde{\boldsymbol{X}}_t = \tilde{y}\right] &= \mathbb{E}_{z_t}\left[\mathbb{E}\left[\|\tilde{\boldsymbol{X}}_{t+1}\| | \tilde{\boldsymbol{X}}_t = \tilde{y}, z_t\right]\right] \\
&\leq \mathbb{E}_{z_t}\left[\left|\left|\left| \tilde{\boldsymbol{B}}_{\max} \right|\right|\right| \|\tilde{y}\| + \mathbb{E}\|\boldsymbol{\epsilon}_t\|\right] \\
&= \left|\left|\left| \tilde{\boldsymbol{B}}_{\max} \right|\right|\right| \|\tilde{y}\| + \mathbb{E}\|\boldsymbol{\epsilon}_t\| \\
&\equiv (1-c)\|\tilde{y}\| + C_{ac} \\
&< (1 - \frac{c}{2})\|\tilde{y}\| - C_{ac}.
\end{aligned}$$

- For all $\tilde{y} \in K_c$,

$$\begin{aligned}\mathbb{E}\left[\|\tilde{\boldsymbol{X}}_{t+1}\||\tilde{\boldsymbol{X}}_t = \tilde{y}\right] &= \mathbb{E}_{z_t}\left[\mathbb{E}\left[\|\tilde{\boldsymbol{X}}_{t+1}\||\tilde{\boldsymbol{X}}_t = \tilde{y}, z_t\right]\right] \\ &< \mathbb{E}_{z_t}\left[\left|\left|\left|\tilde{\boldsymbol{B}}_{\max}\right|\right|\right|\|\tilde{y}\| + C_{ac}\right] \\ &\leq \frac{4C_{ac}(1-c)}{c} + C_{ac}.\end{aligned}$$

- For all $\tilde{y} \in K_C$,

$$0 \leq \|\tilde{y}\| \leq \frac{4C_{ac}}{c}.$$

By Proposition 1 in Liebscher [2005], $\tilde{\boldsymbol{X}}_t$ is geometrically Ergodic and stationary. By Proposition 2 in Liebscher [2005], the sequence will be β-mixing with geometrically decaying mixing coefficients.

Next, we verify the sub-Weibull distribution. Let $\varkappa$ be the sub-Weibull parameter associated with $\tilde{\boldsymbol{U}}_t$ in (B.90). Since

$$\|\tilde{\boldsymbol{X}}_t\|_\psi \leq \left|\left|\left|\tilde{\boldsymbol{B}}_{\max}\right|\right|\right|\|\tilde{\boldsymbol{X}}_{t-1}\|_\psi + \|\tilde{\boldsymbol{U}}_{t-1}\|_\psi,$$

and $\left|\left|\left|\tilde{\boldsymbol{B}}_{\max}\right|\right|\right| < 1$,

$$\|\tilde{\boldsymbol{X}}_t\|_\psi \leq \frac{\|\boldsymbol{\epsilon}_t\|_\psi}{1 - \left|\left|\left|\tilde{\boldsymbol{B}}_{\max}\right|\right|\right|} < \infty.$$

Now, given that $\tilde{\boldsymbol{X}}_t$ is an (equivalent) representation for $\boldsymbol{x}_t$, it follows that $\boldsymbol{x}_t$ is also sub-Weibull. Therefore, Assumption B2 holds.

Appendix 3.5: Algorithms

In this section, we present two algorithms for solving the optimization Equation (3.5). In high dimension, we use Algorithm 5, while in moderate dimension, we use Algorithm 6. Let $S(\cdot;\lambda)$ be the element-wise soft thresholding operator. Recall that throughout the paper, for a $m \times n$ matrix A, $\|A\|_\infty = \max_{1\le i\le m, 1\le j\le n} |a_{ij}|$. The algorithms are as follows:

Algorithm 5: The fused lasso algorithms

Initialize $\boldsymbol{\theta}_i = 0$, for $i = 1, \cdots, n$.;

while $h <$ *maximum iteration* **do**

  for $i = 1, \cdots, n$ **do**

    Calculate the $(h+1)$th iteration of $\boldsymbol{\theta}_i^{h+1}$ by KKT condition:

$$\boldsymbol{\theta}_i^{'(h+1)} = \left(\sum_{l=i}^{n} \boldsymbol{Y}_{\pi(l)} \boldsymbol{Y}'_{\pi(l)}\right)^{-1} S\left(\sum_{l=i}^{n} \boldsymbol{Y}_{\pi(l)} \boldsymbol{x}'_{\pi(l)} - \sum_{j\neq i} \left(\sum_{l=\max(i,j)} \boldsymbol{Y}_{\pi(l)} \boldsymbol{Y}'_{\pi(l)}\right) \boldsymbol{\theta}_j^{'(h)}\,;\lambda_1\right),$$

    where $\boldsymbol{Y}'_{\pi(l)} = \left(\boldsymbol{x}_{\pi(l)}, \cdots, \boldsymbol{x}_{\pi(l)-K+1}\right)_{1\times pK}$ and

$$S(y\,;\lambda) = \begin{cases} y-\lambda & \text{if } y > \lambda \\ y+\lambda & \text{if } y < -\lambda \\ 0 & \text{otherwise} \end{cases}.$$

  if $\max_{1\le i\le n} \|\boldsymbol{\theta}_i^{(h+1)} - \boldsymbol{\theta}_i^{(h)}\|_\infty < \delta$, *where* δ *is the tolerance set to* $2e^{-4}$ *in the paper* **then**

    Stop and denote the final estimate by $\boldsymbol{\Theta}^{(\text{intermediate})}$.

Apply soft-thresholding to the partial sums of $\boldsymbol{\Theta}^{(\text{intermediate})}$, i.e. $\sum_{i=1}^{k} \boldsymbol{\theta}_i^{(\text{intermediate})}$ to find the optimizer in Equation (3.5). In other words, $\widehat{\boldsymbol{\theta}}_1 = S\left(\boldsymbol{\theta}_1^{(\text{intermediate})}; \lambda_2\right)$ and $\widehat{\boldsymbol{\theta}}_k = S\left(\sum_{i=1}^{k} \boldsymbol{\theta}_i^{(\text{intermediate})}; \lambda_2\right) - S\left(\sum_{i=1}^{k-1} \boldsymbol{\theta}_i^{(\text{intermediate})}; \lambda_2\right)$ for $k = 2, 3, \ldots, n$. Finally $\widehat{\boldsymbol{\Theta}} = \left(\widehat{\boldsymbol{\theta}}_1, \ldots, \widehat{\boldsymbol{\theta}}_n\right)$.

Algorithm 6: The group lasso algorithms

Initialize $\boldsymbol{\theta}_i = 0$, for $i = 1, \cdots, n.$;

while $h <$ *maximum iteration* **do**

  for $i = 1, \cdots, n$ **do**

    Calculate the $(h+1)$th iteration of $\boldsymbol{\theta}_i^{h+1}$: Let

$$\begin{aligned}\Omega =& \boldsymbol{\theta}_i^{'(h)} + \gamma \left(\sum_{l=i}^{n} \boldsymbol{Y}_{\pi(l)} \boldsymbol{x}'_{\pi(l)} - \sum_{j \neq i} \left(\sum_{l=\max(i,j)} \boldsymbol{Y}_{\pi(l)} \boldsymbol{Y}'_{\pi(l)} \right) \boldsymbol{\theta}_j^{'(h)} \right. \\ & \left. - \sum_{l=i}^{n} \left(\boldsymbol{Y}_{\pi(l)} \boldsymbol{Y}'_{\pi(l)} \right) \boldsymbol{\theta}_i^{'(h)} \right)\end{aligned}$$

$$\begin{aligned}\boldsymbol{\theta}_i^{'(h+1)} &= \frac{1}{2\gamma} \arg\min_{U} \|U - \Omega\|_2^2 + \|\Omega\|_2 \\ &= \left(1 - \frac{\gamma\lambda_1}{\|U\|_2}\right)_+ \Omega\end{aligned} \tag{B.91}$$

    .

  if $\max_{1 \leq i \leq n} \|\boldsymbol{\theta}_i^{(h+1)} - \boldsymbol{\theta}_i^{(h)}\|_\infty < \delta$, *where* δ *is the tolerance set to* $2e^{-4}$ *in the paper* **then**

    Stop and denote the final estimate by $\boldsymbol{\Theta}^{(\text{final})}$.

Appendix 3.6: Extended Literature Review

In this section, we summarise the existing methods for estimating multivariate TARs, along with their treatment of the number of thresholds m_0 and dimension of the TAR model.

Paper	m_0	m_0 assumed known?	Dimension
Tsay [1998b]	finite (at most three)	No	low
Lo and Zivot [2001] (TVAR)	at most 2	Yes	low
Hansen and Seo [2002]	1	Yes	low (bi-variate)
Nieto [2005]	finite	No	low (bi-variate)
Dueker et al. [2011]	finite	Yes	low
Li and Tong [2016]	1	Yes	low
Calderón V and Nieto [2017]	at most 3	No	low
Orjuela and Villanueva [2021]	finite	No	low
Our method	diverging with T	No	moderate & high

Table B.1: Comparison of existing methods for estimating multivariate TAR models. Here m_0 represents the number of thresholds and T the length of the time series.

Table B.1 highlights the limitations of existing approaches and the fact that our methods are the only available approach that can handle both low- and high-dimensional settings, while allowing for an *unknown* number of thresholds that could diverge with the number of observations T. Allowing for an unknown number of thresholds amounts significantly complicates the problem as previous approaches for multivariate TAR models need to first estimate the number of thresholds and then proceed with estimating the location of thresholds. An incorrect estimation of number of thresholds in the first step may result in biased estimation of thresholds due to having misspecified components in the estimation procedure.

As seen in Table B.1, many methods assume that the number of thresholds is known, even though this information is often not available in practice. Thus, in the remainder of this section we discuss how existing approaches treat the number of thresholds.

Utilizing this assumption, Tsay [1998b] performs a grid search, estimating the coefficient using simple linear model for each interval, and selecting the threshold based on the Akaike information criterion (AIC). Lo and Zivot [2001] instead assume the model as at most 2 thresholds. While a relaxation compared to a known number of thresholds, this assumption still considerably simplifies the problem. Using this assumption, Lo and Zivot [2001] use nested hypothesis tests (testing whether the data can be modeled by the linear model versus a TAR model) to detect the thresholds, and apply the grid search method to estimate the values of the thresholds based on the results of the hypothesis testing. As an alternative, Hansen and Seo [2002] couples the grid search with a maximum likelihood estimation (MLE) of the model parameters. However, the algorithm is difficult to implement in higher dimensions, and the consistency and/or distribution of the MLE estimator is not investigated. Dueker et al. [2011] restricts the switching variable to be constructed based on the lags of the original time series that is being modeled and performs a grid search with respected to certain log likelihood function. The key advantage of this method is that it allows for multiple switching variables, but with only one threshold for each switching variable. Li and Tong [2016] provides a nested sub-sample search algorithm to reduce the time complexity of the grid search.

A few methods have recently tried to estimate multivariate TAR models under less restrictive assumptions on the number of thresholds. However, these methods can only handle finite number of thresholds or only work in low-dimensional settings. To our knowledge, Nieto [2005], Calderón V and Nieto [2017] and Calderón V and Nieto [2017] are the only methods that do not require a known number of thresholds or a bound on the number of thresholds. This is achieved by utilizing a Bayesian estimation framework. However, the consistency of the number of estimated thresholds is not investigated for these Bayesian methods, which could be a challenging problem. Our proposed methods and the corresponding theory thus bridge a gap in the existing literature, as the only methods that allow for an unknown and diverging number of thresholds, m_0, while also facilitating estimation of moderate and high-dimensional time series.

Appendix 3.7: Simulation Settings

In all simulation scenarios, the switching variable is generated from an AR(1) process with coefficient 0.6. The error term follows normal distribution with mean 0 and standard deviation 2.

Simulation Scenario 1 (Simple $\boldsymbol{A}$ with uncorrelated error) In this scenario, $T = 300$, $p = 20$, and $K = 2$. There is only one threshold value $r_1 = 4$, which is not close to the boundary. The auto-regressive coefficients are chosen to have the same structure but different values (see Figure B.1a).

Simulation Scenario 2 (Simple $\boldsymbol{A}$ with correlated error) This is the same settings as in Scenario 1, but the covariance matrix of the error term is changed. Specifically, we set $\boldsymbol{\Sigma}_\epsilon = 0.02(\sigma_{ij})_{n\times n}$ with $\sigma_{ij} = \rho^{|i-j|}$, where $\rho = 0.5$.

Simulation Scenario 3 (Random $\boldsymbol{A}$ with uncorrelated error) This setting is also similar to Scenario 1. However, the auto-regressive coefficients are chosen at random (see Figure B.1b).

Simulation Scenario 4 (Simple $\boldsymbol{A}$ with correlated error allowing changes in different regimes) In this scenario, $T = 600$, $p = 20$, and $K = 1$. There are two threshold values $r_1 = 4$ and $r_2 = 6$. The auto-regressive coefficients are chosen to have the same structure as in Scenario 1 but the values change at different thresholds (see Figure B.1c). We also include an additional simulation setting with $T = 300$ and threshold points $r_1 = 4$ and $r_2 = 6$ for this scenario.

Simulation Scenario 5 (Simple high-dimensional $\boldsymbol{A}$ with uncorrelated error) . In this scenario, $T = 80$, $p = 100$, and $K = 2$. There is only one threshold value $r_1 = 5$. The auto-regressive coefficients are chosen to have the same structure as in Scenario 1 but with different values (see Figure B.1d).

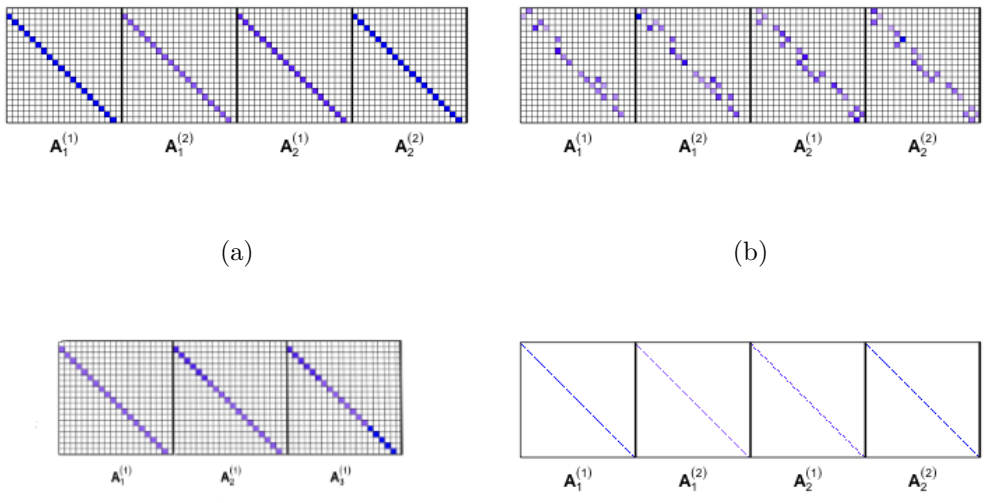


Fig B.1: Images of true auto-regressive coefficients in different simulation scenarios considered. (a): The two regimes in Simulation Scenario 1 and 2. (b): The two regimes in Simulation Scenario 3. (c): The three regimes in Simulation Scenario 4. (d): The two regimes in Simulation Scenario 5.

The auto-regressive coefficients for the above simulation scenarios are visualized in Figure B.1, where different coefficient values are represented by different colors. For Scenarios 1, 2 and 5, the 1-off diagonal values for the two lags in the two regimes are 0.49, -0.3, -0.4, and 0.49, respectively. For Scenario 4, the auto-regressive coefficients are allowed to change in different regimes. The 1-off diagonal values for one lag in the first regime are 0.25. In the second regime, the first $p/3$ values are decreased to -0.2. In the third regime, the last $p/4$ values are increased to 0.49. For Scenario 3, the auto-regressive coefficients are chosen at random.

Appendix C

SUPPLEMENTARY MATERIALS FOR CHAPTER 4

Appendix 0: Setup and Notations

C.0.1 Transforming TAR(K) process to TAR(1) process

The TAR process can be rewritten as a corresponding Kp-dimensional TAR process with 1 lag for regime j, that is

$$\boldsymbol{X}_t = \boldsymbol{B}^{(j)}\boldsymbol{X}_{t-1} + \boldsymbol{U}_t, \tag{C.1}$$

where $\boldsymbol{X}_t = \begin{pmatrix} \boldsymbol{x}_t' & \boldsymbol{x}_{t-1}' & \dots & \boldsymbol{x}_{t-K+1}' \end{pmatrix}' \in \mathbb{R}^{Kp\times 1}$, $\boldsymbol{B}^{(j)} = \begin{pmatrix} \boldsymbol{A}^{(1,j)} & \boldsymbol{A}^{(k,j)} & \dots & \boldsymbol{A}^{(K-1,j)} & \boldsymbol{A}^{(K,j)} \\ \boldsymbol{I}_p & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_p & & \boldsymbol{0} & \boldsymbol{0} \\ \vdots & & \ddots & \vdots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{I}_p & \boldsymbol{0} \end{pmatrix} \in \mathbb{R}^{Kp\times Kp}$ for $\boldsymbol{I}_p$ is a $p \times p$ identity matrix, and $\boldsymbol{U}_t = \begin{pmatrix} \boldsymbol{\epsilon}_t' & \boldsymbol{0} & \dots & \boldsymbol{0} \end{pmatrix}' \in \mathbb{R}^{KP\times 1}$.

C.0.2 Setup

Let $\boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}$ be defined as the solution of

$$\left(\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}} \mathbb{E}\left[\boldsymbol{Y}_{\pi(i)}\boldsymbol{Y}_{\pi(i)}'\right] \right) \boldsymbol{A}^{*'}_{\mathcal{T}_{(s,e)}} = \sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}} \mathbb{E}\left[\boldsymbol{Y}_{\pi(i)}\boldsymbol{Y}_{\pi(i)}'\right] \boldsymbol{A}_{\pi(i)}^{'}, \tag{C.2}$$

where $\boldsymbol{A}_{\pi(i)}^{'} = \left(\boldsymbol{A}_{(1,\pi(i))}, \boldsymbol{A}_{(2,\pi(i))}, \dots, \boldsymbol{A}_{(K,\pi(i))}\right) \in \mathbb{R}^{p\times pK}$ and $\boldsymbol{A}_{(k,\pi(i))}$ is the transition matrix of the k-th lag at $\pi(i)$-th time point. Note that when there are no thresholds in

$\mathcal{T}_{(s,e)}$, $\boldsymbol{A}^*_{\mathcal{T}_{(s,e)}} = \boldsymbol{A}_{\pi(i)}$; otherwise,

$$\boldsymbol{A}^{*'}_{\mathcal{T}_{(s,e)}} = \left(\sum_{z_{\pi(i)} \in \mathcal{T}_{(s,e)}} \mathbb{E}\left[\boldsymbol{Y}_{\pi(i)} \boldsymbol{Y}'_{\pi(i)} \right] \right)^{-1} \sum_{z_{\pi(i)} \in \mathcal{T}_{(s,e)}} \mathbb{E}\left[\boldsymbol{Y}_{\pi(i)} \boldsymbol{Y}'_{\pi(i)} \right] \boldsymbol{A}'_{\pi(i)}.$$

With a permutation of the rows and columns in $\mathbb{E}\left[\boldsymbol{Y}_{\pi(i)} \boldsymbol{Y}'_{\pi(i)} \right] \boldsymbol{A}'_{\pi(i)}$ for $z_{\pi(i)} \in \mathcal{T}_{(s,e)}$ and by Assumption C3, without loss of generality, we have $\|\boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}\|_0 \leq C_0 d_n^*$, where C_0 is a positive constant.

Appendix 1: Technical Lemmas

Throughout the proof, we use z_t to represent all the $z_{t,l}$ for simplicity. For a matrix M, denote $\|M\|_2$ as its Frobenius norm. $\text{vec}(M)_l$ represents the l-th element of $\text{vec}(M)$, where $\text{vec}(M)$ is the vector obtained from the matrix M by concatenating the rows of M. Let $\boldsymbol{I}_{pK}$ be the $pK \times pK$ diagonal matrix with all diagonal elements equal to 1. Let $L^*\left(\mathcal{T}_{(s,e)}\right)$ be the population counterpart of $L\left(\mathcal{T}_{(s,e)}\right)$ obtained by replacing the coefficient matrix estimator $\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}$ with its population counterpart $\boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}$.

Proposition 22. *Let $|\mathcal{T}_{(s,e)}|$ be the number of $z_{\pi(i)}$s that fall into the given interval $(s,e]$. Under Assumptions C1 to C4, there exist positive constants C, c_1, $c_2 > 5$, c_3, c_4, and c_5 such that, for $\left|\mathcal{T}_{(s,e)}\right| \geq C\left(\log\left(\max\left\{p^2K, n\right\}\right)\right)^{2/\varkappa_0 - 1}$,*

$$\mathbb{P}\left(\frac{1}{\left|\mathcal{T}_{(s,e)}\right|} \|\boldsymbol{\epsilon}_{\pi(i)} \boldsymbol{Y}'_{\pi(i)}\|_\infty > c_1 \sqrt{\frac{\log\left(\max\left\{p^2K, n\right\}\right)}{\left|\mathcal{T}_{(s,e)}\right|}} \right) \leq \delta_1, \tag{C.3}$$

where $\delta_1 = 2\exp\left(-c_2 \log\left(\max\left\{p^2K, n\right\}\right)\right)$. In addition,

$$\begin{aligned} &\mathbb{P}\Bigg(\left\| \sum_{z_{\pi(i)} \in \mathcal{T}_{(s,e)}} \boldsymbol{Y}_{\pi(i)} \boldsymbol{Y}'_{\pi(i)} - \mathbb{E}\left(\sum_{i \in \mathcal{T}_{(s,e)}} \boldsymbol{Y}_{\pi(i)} \boldsymbol{Y}'_{\pi(i)} \right) \right\|_\infty \\ &> c_3 \left(\log\left(\max\left\{p^2K, n\right\}\right)\right)^{1/\varkappa_0} \left|\mathcal{T}_{(s,e)}\right|^{1/2} \Bigg) \leq \delta_2, \end{aligned} \tag{C.4}$$

where

$$\delta_2 = \exp\left\{-c_4 \log\left(\max\left\{p^2K, n\right\}\right)\left(\log\left(\max\left\{p^2K, n\right\}\right)\right)^{1-\varkappa_0/2}\right\} + \exp\left\{-c_5\left(\log\left(\max\left\{p^2K, n\right\}\right)\right)^{2/\varkappa_0}\right\}.$$

Proof of Proposition 22: The proof of Equations (C.3) and (C.4) are directly obtained from the proof of Propositions 7 and 8 in Wong et al. [2020].

Given that a sub-sequence of a $\beta-$mixing process is $\beta-$mixing and Assumption C2, then

$$(\boldsymbol{Y}_t I(s \leq z_t < e), \boldsymbol{x}_t I(s \leq z_t < e))$$

is $\beta-$mixing. After rearrangement, $(\boldsymbol{Y}_t I(s \leq z_t < e), \boldsymbol{x}_t I(s \leq z_t < e))$ becomes $\left(\boldsymbol{Y}_{\pi(i)}, \boldsymbol{x}_{\pi(i)}\right)$ and the consistency bounds in Propositions 7 and 8 of Wong et al. [2020] will continue to hold after this rearrangement.

From the proof on page 47 in Wong et al. [2020], set $t = c_1\sqrt{\frac{\log(\max\{p^2K,n\})}{|\mathcal{T}_{(s,e)}|}}$. Note that c_2 is a positive constant that depends on $c_1 > 0$. Thus, we can choose a large enough constant c_1 such that $c_2 > 5$. Then, Equation (C.3) is as desired. Similarly, select $t = c_3\left(\log\left(\max\left\{p^2K, n\right\}\right)\right)^{1/\varkappa_0}\left|\mathcal{T}_{(s,e)}\right|^{1/2}$ in the proof on page 48 in Wong et al. [2020]. Then, Equation (C.4) is as desired.

Lemma 23. *Under Assumptions C1 to C4, assume that the interval $(s, e]$ has one and only one true threshold r. If*

$$L(\mathcal{T}_{(s,e)}) \leq L\left(\mathcal{T}_{(s,r)}\right) + L\left(\mathcal{T}_{(r,e)}\right) + \omega \tag{C.5}$$

and $\lambda = c_\lambda\left(\log\left(\max\left\{p^2K, n\right\}\right)\right)^{1/\varkappa_0} d_n^*$, then there exists a positive constant C_0 such that with probability

$$(1-\delta_1)(1-\delta_2),$$

we have

$$\min\left\{\left|\mathcal{T}_{(s,r)}\right|, \left|\mathcal{T}_{(r,e)}\right|\right\} \leq C_0\left(\frac{\lambda^2 d_n^* + \omega}{v^2}\right). \tag{C.6}$$

Proof of Lemma 23:

The proof for this lemma is along the lines of the proof of Lemma 5 in Wang et al. [2019]. If $\left|\mathcal{T}_{(s,e)}\right| \leq \omega$, then Equation (C.6) holds. If $\max\left\{\left|\mathcal{T}_{(s,r)}\right|, \left|\mathcal{T}_{(r,e)}\right|\right\} \leq \omega$, then Equation (C.6) also holds. Now for the case $\max\left\{\left|\mathcal{T}_{(s,r)}\right|, \left|\mathcal{T}_{(r,e)}\right|\right\} > \omega$, we prove by contradiction. Assume that

$$\min\left\{\left|\mathcal{T}_{(s,r)}\right|, \left|\mathcal{T}_{(r,e)}\right|\right\} > C_0\left(\frac{\lambda^2 d_n^* + \omega}{v^2}\right). \tag{C.7}$$

Based on Equation (C.5), we get

$$\begin{aligned}
&\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left(\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right)^2 \\
\leq &\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,r)}}\left(\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,r)}}\boldsymbol{Y}_{\pi(i)}\right)^2 + \sum_{z_{\pi(i)}\in\mathcal{T}_{(r,e)}}\left(\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(r,e)}}\boldsymbol{Y}_{\pi(i)}\right)^2 + \omega.
\end{aligned} \tag{C.8}$$

Then, together with Lemma 29, we can obtain

$$\begin{aligned}
&\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,r)}}\left(\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right)^2 + \sum_{z_{\pi(i)}\in\mathcal{T}_{(r,e)}}\left(\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right)^2 \\
\leq &\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,r)}}\left(\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,r)}}\boldsymbol{Y}_{\pi(i)}\right)^2 + \sum_{z_{\pi(i)}\in\mathcal{T}_{(r,e)}}\left(\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(r,e)}}\boldsymbol{Y}_{\pi(i)}\right)^2 + \omega \\
\leq &\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,r)}}\left(\boldsymbol{x}_{\pi(i)} - \boldsymbol{A}_{\mathcal{T}_{(s,r)}}\boldsymbol{Y}_{\pi(i)}\right)^2 + \sum_{z_{\pi(i)}\in\mathcal{T}_{(r,e)}}\left(\boldsymbol{x}_{\pi(i)} - \boldsymbol{A}_{\mathcal{T}_{(r,e)}}\boldsymbol{Y}_{\pi(i)}\right)^2 + \omega + 2C_4\lambda^2 d_n^*.
\end{aligned} \tag{C.9}$$

Note that

$$\begin{aligned}
&\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,r)}}\left(\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right)^2 + \sum_{z_{\pi(i)}\in\mathcal{T}_{(r,e)}}\left(\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right)^2 \\
= &\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,r)}}\left(\left(\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} - \boldsymbol{A}_{\mathcal{T}_{(s,r)}}\right)\boldsymbol{Y}_{\pi(i)}\right)^2 + \sum_{i\in\mathcal{T}_{(r,e)}}\left(\left(\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} - \boldsymbol{A}_{\mathcal{T}_{(r,e)}}\right)\boldsymbol{Y}_{\pi(i)}\right)^2 \\
&+ \sum_{z_{\pi(i)}\in\mathcal{T}_{(s,r)}}\left(\boldsymbol{x}_{\pi(i)} - \boldsymbol{A}_{\mathcal{T}_{(s,r)}}\boldsymbol{Y}_{\pi(i)}\right)^2 + \sum_{z_{\pi(i)}\in\mathcal{T}_{(r,e)}}\left(\boldsymbol{x}_{\pi(i)} - \boldsymbol{A}_{\mathcal{T}_{(r,e)}}\boldsymbol{Y}_{\pi(i)}\right)^2 \\
&- 2\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,r)}}\boldsymbol{\epsilon}'_{\pi(i)}\left(\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} - \boldsymbol{A}_{\mathcal{T}_{(s,r)}}\right)\boldsymbol{Y}_{\pi(i)} - 2\sum_{z_{\pi(i)}\in\mathcal{T}_{(r,e)}}\boldsymbol{\epsilon}'_{\pi(i)}\left(\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} - \boldsymbol{A}_{\mathcal{T}_{(r,e)}}\right)\boldsymbol{Y}_{\pi(i)}.
\end{aligned} \tag{C.10}$$

Combining Equation (C.9) and Equation (C.10), we get

$$
\begin{aligned}
&\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,r)}}\left(\left(\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}-\boldsymbol{A}_{\mathcal{T}_{(s,r)}}\right)\boldsymbol{Y}_{\pi(i)}\right)^2+\sum_{i\in\mathcal{T}_{(r,e)}}\left(\left(\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}-\boldsymbol{A}_{\mathcal{T}_{(r,e)}}\right)\boldsymbol{Y}_{\pi(i)}\right)^2\\
\leq&2\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,r)}}\boldsymbol{\epsilon}'_{\pi(i)}\left(\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}-\boldsymbol{A}_{\mathcal{T}_{(s,r)}}\right)\boldsymbol{Y}_{\pi(i)}+2\sum_{z_{\pi(i)}\in\mathcal{T}_{(r,e)}}\boldsymbol{\epsilon}'_{\pi(i)}\left(\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}-\boldsymbol{A}_{\mathcal{T}_{(r,e)}}\right)\boldsymbol{Y}_{\pi(i)}\\
&+\omega+2C_4\lambda^2d_n^*\\
\leq&2\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}-\boldsymbol{A}_{\mathcal{T}_{(s,r)}}\right\|_1\left\|\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,r)}}\boldsymbol{\epsilon}_{\pi(i)}\boldsymbol{Y}'_{\pi(i)}\right\|_\infty\\
&+2\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}-\boldsymbol{A}_{\mathcal{T}_{(r,e)}}\right\|_1\left\|\sum_{z_{\pi(i)}\in\mathcal{T}_{(r,e)}}\boldsymbol{\epsilon}_{\pi(i)}\boldsymbol{Y}'_{\pi(i)}\right\|_\infty+\omega+2C_4\lambda^2d_n^*\\
\leq&2\left\|\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,r)}}\boldsymbol{\epsilon}_{\pi(i)}\boldsymbol{Y}'_{\pi(i)}\right\|_\infty\left(\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}-\boldsymbol{A}_{\mathcal{T}_{(s,r)}}\right\|_{1,\mathcal{I}}+\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}-\boldsymbol{A}_{\mathcal{T}_{(s,r)}}\right\|_{1,\mathcal{I}^c}\right)\\
&+2\left\|\sum_{z_{\pi(i)}\in\mathcal{T}_{(r,e)}}\boldsymbol{\epsilon}_{\pi(i)}\boldsymbol{Y}'_{\pi(i)}\right\|_\infty\left(\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}-\boldsymbol{A}_{\mathcal{T}_{(r,e)}}\right\|_{1,\mathcal{I}}+\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}-\boldsymbol{A}_{\mathcal{T}_{(r,e)}}\right\|_{1,\mathcal{I}^c}\right)\\
&+\omega+2C_4\lambda^2d_n^*.
\end{aligned}
\tag{C.11}
$$

By the Cauchy-Schwarz inequality (see, e.g., Equation (C.60)), we obtain

$$
\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}-\boldsymbol{A}_{\mathcal{T}_{(r,e)}}\right\|_{1,\mathcal{I}}\leq\sqrt{d_n^*}\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}-\boldsymbol{A}_{\mathcal{T}_{(r,e)}}\right\|_{2,\mathcal{I}}
$$

and

$$
\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}-\boldsymbol{A}_{\mathcal{T}_{(s,r)}}\right\|_{1,\mathcal{I}}\leq\sqrt{d_n^*}\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}-\boldsymbol{A}_{\mathcal{T}_{(s,r)}}\right\|_{2,\mathcal{I}}.
$$

Denote $\boldsymbol{B}_1=\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}-\boldsymbol{A}_{\mathcal{T}_{(s,r)}}$ and $\boldsymbol{B}_2=\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}-\boldsymbol{A}_{\mathcal{T}_{(r,e)}}$. Then, we can rewrite Equa-

tion (C.11) to obtain

$$
\begin{aligned}
&\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,r)}}\left(\boldsymbol{B}_1\boldsymbol{Y}_{\pi(i)}\right)^2+\sum_{i\in\mathcal{T}_{(r,e)}}\left(\boldsymbol{B}_2\boldsymbol{Y}_{\pi(i)}\right)^2\\
\leq& 2\left\|\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,r)}}\boldsymbol{\epsilon}_{\pi(i)}\boldsymbol{Y}'_{\pi(i)}\right\|_\infty\left(\|\boldsymbol{B}_1\|_{1,\mathcal{I}}+\|\boldsymbol{B}_1\|_{1,\mathcal{I}^c}\right)\\
&+2\left\|\sum_{z_{\pi(i)}\in\mathcal{T}_{(r,e)}}\boldsymbol{\epsilon}_{\pi(i)}\boldsymbol{Y}'_{\pi(i)}\right\|_\infty\left(\|\boldsymbol{B}_2\|_{1,\mathcal{I}}+\|\boldsymbol{B}_2\|_{1,\mathcal{I}^c}\right)+\omega+2C_4\lambda^2d_n^* \qquad \text{(C.12)}\\
\leq& 2\left\|\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,r)}}\boldsymbol{\epsilon}_{\pi(i)}\boldsymbol{Y}'_{\pi(i)}\right\|_\infty\left(\sqrt{d_n^*}\,\|\boldsymbol{B}_1\|_{2,\mathcal{I}}+\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_{1,\mathcal{I}^c}\right)\\
&+2\left\|\sum_{z_{\pi(i)}\in\mathcal{T}_{(r,e)}}\boldsymbol{\epsilon}_{\pi(i)}\boldsymbol{Y}'_{\pi(i)}\right\|_\infty\left(\sqrt{d_n^*}\,\|\boldsymbol{B}_2\|_{2,\mathcal{I}}+\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_{1,\mathcal{I}^c}\right)+\omega+2C_4\lambda^2d_n^*.
\end{aligned}
$$

By Proposition 22 (Equation (C.3)) and the choice of $\lambda = 2C_{12}\left(\log\left(\max\left\{p^2K,n\right\}\right)\right)^{1/\varkappa_0}d_n^* \geq \sqrt{\log\left(\max\left\{p^2K,n\right\}\right)}$, we get

$$
\begin{aligned}
&\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,r)}}\left(\boldsymbol{B}_1\boldsymbol{Y}_{\pi(i)}\right)^2+\sum_{i\in\mathcal{T}_{(r,e)}}\left(\boldsymbol{B}_2\boldsymbol{Y}_{\pi(i)}\right)^2\\
\leq&2c_1\sqrt{\left|\mathcal{T}_{(s,r)}\right|\log\left(\max\left\{p^2K,n\right\}\right)}\left(\sqrt{d_n^*}\left\|\boldsymbol{B}_1\right\|_{2,\mathcal{I}}+\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_{1,\mathcal{I}^c}\right)\\
&+2c_1\sqrt{\left|\mathcal{T}_{(r,e)}\right|\log\left(\max\left\{p^2K,n\right\}\right)}\left(\sqrt{d_n^*}\left\|\boldsymbol{B}_2\right\|_{2,\mathcal{I}}+\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_{1,\mathcal{I}^c}\right)+\omega+2C_4\lambda^2d_n^*\\
\leq&c_7\lambda\left(\sqrt{\left|\mathcal{T}_{(s,r)}\right|d_n^*}\left\|\boldsymbol{B}_1\right\|_{2,\mathcal{I}}+\sqrt{\left|\mathcal{T}_{(s,r)}\right|}\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_{1,\mathcal{I}^c}\right.\\
&\left.+\sqrt{\left|\mathcal{T}_{(r,e)}\right|d_n^*}\left\|\boldsymbol{B}_2\right\|_{2,\mathcal{I}}+\sqrt{\left|\mathcal{T}_{(r,e)}\right|}\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_{1,\mathcal{I}^c}\right)+\omega+2C_4\lambda^2d_n^*\\
\leq&2c_7^2\lambda^2d_n^*/c_x+\frac{c_x\left\|\boldsymbol{B}_1\right\|_2^2\left|\mathcal{T}_{(s,r)}\right|}{4}+2c_7^2\lambda^2d_n^*/c_x+\frac{c_x\left\|\boldsymbol{B}_2\right\|_2^2\left|\mathcal{T}_{(r,e)}\right|}{4}\\
&+\lambda\left(\sqrt{\left|\mathcal{T}_{(s,r)}\right|}+\sqrt{\left|\mathcal{T}_{(r,e)}\right|}\right)\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_{1,\mathcal{I}^c}+\omega+2C_4\lambda^2d_n^*\\
\leq&c_8\lambda^2d_n^*+\frac{c_x\left\|\boldsymbol{B}_1\right\|_2^2\left|\mathcal{T}_{(s,r)}\right|}{4}+\frac{c_x\left\|\boldsymbol{B}_2\right\|_2^2\left|\mathcal{T}_{(r,e)}\right|}{4}+2C_8\lambda^2\sqrt{d_n^*}+\omega\\
\leq&\frac{c_x\left\|\boldsymbol{B}_1\right\|_2^2\left|\mathcal{T}_{(s,r)}\right|}{4}+\frac{c_x\left\|\boldsymbol{B}_2\right\|_2^2\left|\mathcal{T}_{(r,e)}\right|}{4}+\omega+c_9\lambda^2d_n^*,
\end{aligned}
\tag{C.13}
$$

where the third and forth inequalities follow from Hölder's inequality and Lemma 30, respectively.

Now, by Equation (C.7) and the choice of λ and ω, we have

$$
\begin{aligned}
\min\left\{\left|\mathcal{T}_{(s,r)}\right|,\left|\mathcal{T}_{(r,e)}\right|\right\}>&C_0\left(\frac{\lambda^2d_n^*+\omega}{v^2}\right)\\
=&C_0\frac{c_\lambda^2\left(\log\left(\max\left\{p^2K,n\right\}\right)\right)^{2/\varkappa_0}d_n^{*3}+\omega}{v^2}\\
>&c_{10}\left(\log\left(\max\left\{p^2K,n\right\}\right)\right)^{2/\varkappa_0}d_n^{*3}.
\end{aligned}
\tag{C.14}
$$

Recalling that $\boldsymbol{I}_{pK}$ is a $pK \times pK$ diagonal matrix with all diagonal elements equal to 1,

$$
\begin{aligned}
&\sum_{z_{\pi(i)} \in \mathcal{T}_{(s,r)}} \left(\left(\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} - \boldsymbol{A}_{\mathcal{T}_{(s,r)}} \right) \boldsymbol{Y}_{\pi(i)} \right)^2 \\
=& \left(\operatorname{vec}\left(\boldsymbol{B}_1 \right) \right)' \left(\left(\sum_{z_{\pi(i)} \in \mathcal{T}_{(s,r)}} \boldsymbol{Y}_{\pi(i)} \boldsymbol{Y}'_{\pi(i)} \right) \otimes \boldsymbol{I}_{pK} \right) \operatorname{vec}\left(\mathrm{B}_1 \right) \\
\geq& \left(\operatorname{vec}\left(\boldsymbol{B}_1 \right) \right)' \left(\mathbb{E} \left(\sum_{z_{\pi(i)} \in \mathcal{T}_{(s,r)}} \boldsymbol{Y}_{\pi(i)} \boldsymbol{Y}'_{\pi(i)} \right) \otimes \boldsymbol{I}_{pK} \right) \operatorname{vec}\left(\boldsymbol{B}_1 \right) \\
&- \left(\operatorname{vec}\left(\boldsymbol{B}_1 \right) \right)' \left(\left(\sum_{z_{\pi(i)} \in \mathcal{T}_{(s,r)}} \boldsymbol{Y}_{\pi(i)} \boldsymbol{Y}'_{\pi(i)} - \mathbb{E} \left(\sum_{z_{\pi(i)} \in \mathcal{T}_{(s,r)}} \boldsymbol{Y}_{\pi(i)} \boldsymbol{Y}'_{\pi(i)} \right) \right) \otimes \boldsymbol{I}_{pK} \right) \operatorname{vec}\left(\mathrm{B}_1 \right) \\
\geq& c_{11} \left| \mathcal{T}_{(s,r)} \right| \left\| \boldsymbol{B}_1 \right\|_2^2 - \left\| \left(\sum_{z_{\pi(i)} \in \mathcal{T}_{(s,r)}} \boldsymbol{Y}_{\pi(i)} \boldsymbol{Y}'_{\pi(i)} - \mathbb{E} \left(\sum_{z_{\pi(i)} \in \mathcal{T}_{(s,r)}} \boldsymbol{Y}_{\pi(i)} \boldsymbol{Y}'_{\pi(i)} \right) \right) \otimes \boldsymbol{I}_{pK} \right\|_\infty \left\| \boldsymbol{B}_1 \right\|_1^2 \\
\geq& c_{11} \left| \mathcal{T}_{(s,r)} \right| \left\| \boldsymbol{B}_1 \right\|_2^2 - c_3 \left(\log\left(\max\left\{ p^2 K, n \right\} \right) \right)^{1/\varkappa_0} \sqrt{\left| \mathcal{T}_{(s,r)} \right|} \left\| \boldsymbol{B}_1 \right\|_1^2 \\
\geq& c_{11} \left| \mathcal{T}_{(s,r)} \right| \left\| \boldsymbol{B}_1 \right\|_2^2 - c_3 \left(\log\left(\max\left\{ p^2 K, n \right\} \right) \right)^{1/\varkappa_0} \sqrt{\left| \mathcal{T}_{(s,r)} \right|} d_n^* \left\| \boldsymbol{B}_1 \right\|_{2,\mathcal{I}}^2 \\
&- c_3 \left(\log\left(\max\left\{ p^2 K, n \right\} \right) \right)^{1/\varkappa_0} \sqrt{\left| \mathcal{T}_{(s,r)} \right|} \left\| \boldsymbol{B}_1 \right\|_{1,\mathcal{I}^c}^2 \\
\geq& c_{11} \left| \mathcal{T}_{(s,r)} \right| \left\| \boldsymbol{B}_1 \right\|_2^2 - c_3 \left(\log\left(\max\left\{ p^2 K, n \right\} \right) \right)^{1/\varkappa_0} \sqrt{\left| \mathcal{T}_{(s,r)} \right|} d_n^* \left(C_9 \lambda d_n^* / \sqrt{\left| \mathcal{T}_{(s,r)} \right|} \right)^2 \\
&- c_3 \left(\log\left(\max\left\{ p^2 K, n \right\} \right) \right)^{1/\varkappa_0} \sqrt{\left| \mathcal{T}_{(s,r)} \right|} \left(C_8 \lambda d_n^* / \sqrt{\left| \mathcal{T}_{(s,r)} \right|} \right)^2 \\
\geq& c_{11} \left| \mathcal{T}_{(s,r)} \right| \left\| \boldsymbol{B}_1 \right\|_2^2 - c_{12} \lambda^2 \left(\log\left(\max\left\{ p^2 K, n \right\} \right) \right)^{1/\varkappa_0} d_n^{*2} / \sqrt{\left| \mathcal{T}_{(s,r)} \right|} \\
\geq& c_x \left| \mathcal{T}_{(s,r)} \right| \left\| \boldsymbol{B}_1 \right\|_2^2 - c_{13} \lambda^2 d_n^*.
\end{aligned}
\tag{C.15}
$$

Here, the second and the third inequalities are according to the Cauchy-Schwarz inequality and Proposition 22 (Equation (C.4)), respectively; the forth inequality follows from the triangle inequality and Assumption C3; the fifth inequality is according to Equation (C.14) and Lemma 30; and the last inequality is due to the fact that $\sqrt{\left| \mathcal{T}_{(s,r)} \right|} > \sqrt{c_{10}} \left(\log\left(\max\left\{ p^2 K, n \right\} \right) \right)^{1/\varkappa_0} d_n^*$ by Equation (C.14).

By a similar procedure, we get

$$\sum_{z_{\pi(i)}\in\mathcal{T}_{(r,e)}}\left(\left(\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}-\boldsymbol{A}_{\mathcal{T}_{(r,e)}}\right)\boldsymbol{Y}_{\pi(i)}\right)^2\geq c_x\left|\mathcal{T}_{(r,e)}\right|\|\boldsymbol{B}_2\|_2^2-c_{13}\lambda^2 d_n^*. \tag{C.16}$$

Combining Equation (C.13), Equation (C.15) and Equation (C.16), we get

$$\begin{aligned}&\frac{c_x\|\boldsymbol{B}_1\|_2^2\left|\mathcal{T}_{(s,r)}\right|}{4}+\frac{c_x\|\boldsymbol{B}_2\|_2^2\left|\mathcal{T}_{(r,e)}\right|}{4}+\omega+c_9\lambda^2 d_n^*\\ \geq& c_x\left|\mathcal{T}_{(s,r)}\right|\|\boldsymbol{B}_1\|_2^2-2c_{13}\lambda^2 d_n^*+c_x\left|\mathcal{T}_{(r,e)}\right|\|\boldsymbol{B}_2\|_2^2,\end{aligned} \tag{C.17}$$

which leads to

$$\omega+c_{14}\lambda^2 d_n^*\geq\frac{3c_x}{4}\left|\mathcal{T}_{(s,r)}\right|\|\boldsymbol{B}_1\|_2^2+\frac{3c_x}{4}\left|\mathcal{T}_{(r,e)}\right|\|\boldsymbol{B}_2\|_2^2. \tag{C.18}$$

Since by Assumption C4,

$$\begin{aligned}\left|\mathcal{T}_{(s,r)}\right|\|\boldsymbol{B}_1\|_2^2+\left|\mathcal{T}_{(r,e)}\right|\|\boldsymbol{B}_2\|_2^2\geq&\inf_{\boldsymbol{M}}\left\{\left|\mathcal{T}_{(s,r)}\right|\|\boldsymbol{A}^*_{\mathcal{T}_{(s,r)}}-\boldsymbol{M}\|_2^2+\left|\mathcal{T}_{(r,e)}\right|\|\boldsymbol{A}^*_{\mathcal{T}_{(r,e)}}-\boldsymbol{M}\|_2^2\right\}\\ \geq& v^2\frac{\left|\mathcal{T}_{(s,r)}\right|\left|\mathcal{T}_{(r,e)}\right|}{\left|\mathcal{T}_{(s,r)}\right|+\left|\mathcal{T}_{(r,e)}\right|}\\ \geq&\min\left\{\left|\mathcal{T}_{(s,r)}\right|,\left|\mathcal{T}_{(r,e)}\right|\right\}v^2/2,\end{aligned} \tag{C.19}$$

we have $\min\left\{\left|\mathcal{T}_{(s,r)}\right|,\left|\mathcal{T}_{(r,e)}\right|\right\}\leq\frac{c_{16}\omega+c_{15}\lambda^2 d_n^*}{v^2}\leq\frac{c_{17}\left(\omega+\lambda^2 d_n^*\right)}{v^2}$, which contradicts Equation (C.7), proving that Equation (C.6) holds.

Lemma 24. *Suppose Assumptions C1 to C6 hold, and that the interval $(s,e]$ has exactly two true thresholds r_1 and r_2. Let δ_1 and δ_2 be defined in Proposition 22. Then, if*

$$L\left(\mathcal{T}_{(s,e)}\right)\leq L\left(\mathcal{T}_{(s,r_1)}\right)+L\left(\mathcal{T}_{(r_1,r_2)}\right)+L\left(\mathcal{T}_{(r_2,e)}\right)+2\omega, \tag{C.20}$$

there exist a positive constant C_1 such that with probability $(1-\delta_1)(1-\delta_2)$, we get

$$\max\left\{\left|\mathcal{T}_{(s,r_1)}\right|,\left|\mathcal{T}_{(r_2,e)}\right|\right\}\leq C_0\left(\frac{\lambda^2 d_n^*+\omega}{v^2}\right). \tag{C.21}$$

Proof of Lemma 24: The proof for this lemma is along the lines of the proof of Lemma 6 in Wang et al. [2019]. Due to Assumption C5, z_t has positive density. Thus,

$$\left|\mathcal{T}_{(s,e)}\right| = n\mathbb{P}\left(s < z_t \leq e\right) \geq c_e n \left|e - s\right| \geq c_e n \Delta_n, \tag{C.22}$$

where $c_e > 0$. By Equation (C.22), $\left|\mathcal{T}_{(s,e)}\right| \geq \left|\mathcal{T}_{(r_1,r_2)}\right| \geq c_e n \Delta_n$. By Assumption C6, it holds that $\left|\mathcal{T}_{(s,e)}\right| \geq \left|\mathcal{T}_{(r_1,r_2)}\right| \geq \omega$. Without loss of generality, we assume that $\left|\mathcal{T}_{(s,r_1)}\right| \geq \left|\mathcal{T}_{(r_2,e)}\right|$ ($\left|\mathcal{T}_{(s,r_1)}\right| \leq \left|\mathcal{T}_{(r_2,e)}\right|$ is similar). Similar to the proof of Lemma 23, we prove by contradiction. Assume

$$\max\left\{\left|\mathcal{T}_{(s,r_1)}\right|, \left|\mathcal{T}_{(r_2,e)}\right|\right\} > C_0 \left(\frac{\lambda^2 d_n^* + \omega}{v^2}\right). \tag{C.23}$$

Equation (C.23) implies

$$\left|\mathcal{T}_{(s,r_1)}\right| > C_0 \left(\frac{\lambda^2 d_n^* + \omega}{v^2}\right). \tag{C.24}$$

Here, we consider two cases: $\left|\mathcal{T}_{(r_2,e)}\right| \geq \omega$ and $\left|\mathcal{T}_{(r_2,e)}\right| < \omega$. First, considering the case $\left|\mathcal{T}_{(r_2,e)}\right| \geq \omega$, according to Equation (C.20), we get

$$\begin{aligned}
&\sum_{z_{\pi(i)} \in \mathcal{T}_{(s,e)}} \left(\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} \boldsymbol{Y}_{\pi(i)}\right)^2 \\
\leq &\sum_{z_{\pi(i)} \in \mathcal{T}_{(s,r_1)}} \left(\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,r_1)}} \boldsymbol{Y}_{\pi(i)}\right)^2 + \sum_{z_{\pi(i)} \in \mathcal{T}_{(r_1,r_2)}} \left(\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(r_1,r_2)}} \boldsymbol{Y}_{\pi(i)}\right)^2 \\
&+ \sum_{z_{\pi(i)} \in \mathcal{T}_{(r_2,e)}} \left(\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(r_2,e)}} \boldsymbol{Y}_{\pi(i)}\right)^2 + 2\omega.
\end{aligned} \tag{C.25}$$

Then, combining Lemmas 29 and 31, we obtain

$$\begin{aligned}
&\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}} \left(\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right)^2 \\
&\le \sum_{z_{\pi(i)}\in\mathcal{T}_{(s,r_1)}} \left(\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,r_1)}}\boldsymbol{Y}_{\pi(i)}\right)^2 + \sum_{z_{\pi(i)}\in\mathcal{T}_{(r_1,r_2)}} \left(\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(r_1,r_2)}}\boldsymbol{Y}_{\pi(i)}\right)^2 \\
&\quad + \sum_{z_{\pi(i)}\in\mathcal{T}_{(r_2,e)}} \left(\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(r_2,e)}}\boldsymbol{Y}_{\pi(i)}\right)^2 + 2\omega \\
&\le \sum_{z_{\pi(i)}\in\mathcal{T}_{(s,r_1)}} \left(\boldsymbol{x}_{\pi(i)} - \boldsymbol{A}^*_{\mathcal{T}_{(s,r_1)}}\boldsymbol{Y}_{\pi(i)}\right)^2 + \sum_{z_{\pi(i)}\in\mathcal{T}_{(r_1,r_2)}} \left(\boldsymbol{x}_{\pi(i)} - \boldsymbol{A}^*_{\mathcal{T}_{(r_1,r_2)}}\boldsymbol{Y}_{\pi(i)}\right)^2 \\
&\quad + \sum_{z_{\pi(i)}\in\mathcal{T}_{(r_2,e)}} \left(\boldsymbol{x}_{\pi(i)} - \boldsymbol{A}^*_{\mathcal{T}_{(r_2,e)}}\boldsymbol{Y}_{\pi(i)}\right)^2 + 2\omega + 3C_4\lambda^2 d_n^*.
\end{aligned} \tag{C.26}$$

Let $r_0 = s$ and $r_3 = e$. Note that

$$\begin{aligned}
&\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}} \left(\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right)^2 \\
&= \sum_{j=1}^{3} \sum_{z_{\pi(i)}\in\mathcal{T}_{(r_{j-1},r_j)}} \left(\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right)^2 \\
&= \sum_{j=1}^{3} \sum_{z_{\pi(i)}\in\mathcal{T}_{(r_{j-1},r_j)}} \left(\left(\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} - \boldsymbol{A}^*_{\mathcal{T}_{(r_{j-1},r_j)}}\right)\boldsymbol{Y}_{\pi(i)}\right)^2 \\
&\quad + \sum_{j=1}^{3} \sum_{z_{\pi(i)}\in\mathcal{T}_{(r_{j-1},r_j)}} \left(\boldsymbol{x}_{\pi(i)} - \boldsymbol{A}^*_{\mathcal{T}_{(r_{j-1},r_j)}}\boldsymbol{Y}_{\pi(i)}\right)^2 \\
&\quad - 2\sum_{j=1}^{3} \sum_{z_{\pi(i)}\in\mathcal{T}_{(r_{j-1},r_j)}} \boldsymbol{\epsilon}'_{\pi(i)}\left(\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} - \boldsymbol{A}^*_{\mathcal{T}_{(r_{j-1},r_j)}}\right)\boldsymbol{Y}_{\pi(i)}.
\end{aligned} \tag{C.27}$$

Recalling that $\boldsymbol{A}^*_{\mathcal{T}_{(r_{j-1},r_j)}} = \boldsymbol{A}_{\mathcal{T}_{(r_{j-1},r_j)}}$ when there are no thresholds in $(r_{j-1}, r_j]$ and

combining Equations (C.26) and (C.27), we get

$$
\begin{aligned}
&\sum_{j=1}^{3} \sum_{z_{\pi(i)} \in \mathcal{T}_{(r_{j-1}, r_j)}} \left\| \left(\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} - \boldsymbol{A}^*_{\mathcal{T}_{(r_{j-1}, r_j)}} \right) \boldsymbol{Y}_{\pi(i)} \right\|_2^2 \\
\leq & 2 \sum_{j=1}^{3} \sum_{z_{\pi(i)} \in \mathcal{T}_{(r_{j-1}, r_j)}} \boldsymbol{\epsilon}'_{\pi(i)} \left(\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} - \boldsymbol{A}^*_{\mathcal{T}_{(r_{j-1}, r_j)}} \right) \boldsymbol{Y}_{\pi(i)} + 2\omega + 3C_4 \lambda^2 d_n^* \\
\leq & 2 \sum_{j=1}^{3} \left\| \sum_{z_{\pi(i)} \in \mathcal{T}_{(r_{j-1}, r_j)}} \boldsymbol{\epsilon}_{\pi(i)} \boldsymbol{Y}'_{\pi(i)} \right\|_\infty \left\| \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} - \boldsymbol{A}^*_{\mathcal{T}_{(r_{j-1}, r_j)}} \right\|_1 + 2\omega + 3C_4 \lambda^2 d_n^* \\
\leq & 2c_1 \sum_{j=1}^{3} \left(\sqrt{d_n^* \log\left(\max\left\{p^2 K, n\right\}\right) \left| \mathcal{T}_{(r_{j-1}, r_j)} \right|} \left\| \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} - \boldsymbol{A}^*_{\mathcal{T}_{(r_{j-1}, r_j)}} \right\|_{2,\mathcal{I}} \right. \\
& \left. + \sqrt{\log\left(\max\left\{p^2 K, n\right\}\right) \left| \mathcal{T}_{(r_{j-1}, r_j)} \right|} \left\| \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} - \boldsymbol{A}^*_{\mathcal{T}_{(r_{j-1}, r_j)}} \right\|_{1,\mathcal{I}^c} \right) + 2\omega + 3C_4 \lambda^2 d_n^* \\
\leq & \lambda \sum_{j=1}^{3} \left(\sqrt{d_n^* \left| \mathcal{T}_{(r_{j-1}, r_j)} \right|} \left\| \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} - \boldsymbol{A}^*_{\mathcal{T}_{(r_{j-1}, r_j)}} \right\|_{2,\mathcal{I}} + \sqrt{\left| \mathcal{T}_{(r_{j-1}, r_j)} \right|} \left\| \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} - \boldsymbol{A}^*_{\mathcal{T}_{(r_{j-1}, r_j)}} \right\|_{1,\mathcal{I}^c} \right) \\
& + 2\omega + 3C_4 \lambda^2 d_n^*,
\end{aligned}
\tag{C.28}
$$

where the third inequality holds by Proposition 22 (Equation (C.3)) and Cauchy–Schwarz inequality. Then, using similar steps as in Lemma 23, we get

$$
\min\left\{ \left| \mathcal{T}_{(s,r_1)} \right|, \left| \mathcal{T}_{(r_1,r_2)} \right| \right\} \leq C_0 \left(\frac{\lambda^2 d_n^* + \omega}{v^2} \right).
$$

Since $\left| \mathcal{T}_{(r_1,r_2)} \right| \geq c_e n \Delta_n > C_0 \left(\frac{\lambda^2 d_n^* + \omega}{v^2} \right)$ by Assumption C6 and Equation (C.22), we get

$$
\left| \mathcal{T}_{(s,r_1)} \right| \leq C_0 \left(\frac{\lambda^2 d_n^* + \omega}{v^2} \right),
$$

which contradicts Equation (C.23).

For the case $\left|\mathcal{T}_{(r_2,e)}\right| < \omega$, according to Equation (C.20), we get

$$\begin{aligned}\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left(\boldsymbol{x}_{\pi(i)}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right)^2 \leq & \sum_{z_{\pi(i)}\in\mathcal{T}_{(s,r_1)}}\left(\boldsymbol{x}_{\pi(i)}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,r_1)}}\boldsymbol{Y}_{\pi(i)}\right)^2 \\ & + \sum_{z_{\pi(i)}\in\mathcal{T}_{(r_1,r_2)}}\left(\boldsymbol{x}_{\pi(i)}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(r_1,r_2)}}\boldsymbol{Y}_{\pi(i)}\right)^2 + 2\omega.\end{aligned} \tag{C.29}$$

Then, using Lemmas 29 and 31, we obtain

$$\begin{aligned}&\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left(\boldsymbol{x}_{\pi(i)}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right)^2 \\ \leq & \sum_{z_{\pi(i)}\in\mathcal{T}_{(s,r_1)}}\left(\boldsymbol{x}_{\pi(i)}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,r_1)}}\boldsymbol{Y}_{\pi(i)}\right)^2 + \sum_{z_{\pi(i)}\in\mathcal{T}_{(r_1,r_2)}}\left(\boldsymbol{x}_{\pi(i)}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(r_1,r_2)}}\boldsymbol{Y}_{\pi(i)}\right)^2 + 2\omega \\ \leq & \sum_{z_{\pi(i)}\in\mathcal{T}_{(s,r_1)}}\left(\boldsymbol{x}_{\pi(i)}-\boldsymbol{A}^*_{\mathcal{T}_{(s,r_1)}}\boldsymbol{Y}_{\pi(i)}\right)^2 + \sum_{z_{\pi(i)}\in\mathcal{T}_{(r_1,r_2)}}\left(\boldsymbol{x}_{\pi(i)}-\boldsymbol{A}^*_{\mathcal{T}_{(r_1,r_2)}}\boldsymbol{Y}_{\pi(i)}\right)^2 + 2\omega + 2C_4\lambda^2 d_n^*.\end{aligned} \tag{C.30}$$

Similar to Equation (C.26), we set $r_0 = s$ and $r_3 = e$ and rearrange Equation (C.30). By the similar steps as in Equation (C.28), we get

$$\begin{aligned}&\sum_{j=1}^{2}\sum_{z_{\pi(i)}\in\mathcal{T}_{(r_{j-1},r_j)}}\left\|\left(\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}-\boldsymbol{A}^*_{\mathcal{T}_{(r_{j-1},r_j)}}\right)\boldsymbol{Y}_{\pi(i)}\right\|_2^2 \\ \leq & \lambda\sum_{j=1}^{2}\left(\sqrt{d_n^*\left|\mathcal{T}_{(r_{j-1},r_j)}\right|}\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}-\boldsymbol{A}^*_{\mathcal{T}_{(r_{j-1},r_j)}}\right\|_{2,\mathcal{I}} + \sqrt{\left|\mathcal{T}_{(r_{j-1},r_j)}\right|}\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}-\boldsymbol{A}^*_{\mathcal{T}_{(r_{j-1},r_j)}}\right\|_{1,\mathcal{I}}\right) \\ & + 2\omega + 2C_4\lambda^2 d_n^*.\end{aligned} \tag{C.31}$$

Finally, using similar arguments as in Lemma 23, we have

$$\min\left\{\left|\mathcal{T}_{(s,r_1)}\right|, \left|\mathcal{T}_{(r_1,r_2)}\right|\right\} \leq C_0\left(\frac{\lambda^2 d_n^* + \omega}{v^2}\right).$$

Since $\left|\mathcal{T}_{(r_1,r_2)}\right| \geq c_e n \Delta_n$ by Assumption C6 and Equation (C.22), this gives

$$\left|\mathcal{T}_{(s,r_1)}\right| \leq C_0 \left(\frac{\lambda^2 d_n^* + \omega}{v^2}\right),$$

which contradicts Equation (C.23), hence proving the result.

Lemma 25. *Suppose Assumptions C1 to C4 hold, and that there are no thresholds in $\mathcal{T}_{(s,e)}$. Then, with probability $1-\delta_4$,*

$$L\left(\mathcal{T}_{(s,e)}\right) < \min_{r' \in \mathcal{T}_{(s,e)}} \left\{L\left(\mathcal{T}_{(s,r')}\right) + L\left(\mathcal{T}_{(r',e)}\right)\right\} + \omega, \tag{C.32}$$

where

$$\begin{aligned}\delta_4 = {} & 2n \exp\left(-c_2 \log\left(\max\left\{p^2 K, n\right\}\right)\right) \\ & + n \exp\left\{-c_4 \log\left(\max\left\{p^2 K, n\right\}\right) \left(\log\left(\max\left\{p^2 K, n\right\}\right)\right)^{1-\varkappa_0/2}\right\} \\ & + n \exp\left\{-c_5 \left(\log\left(\max\left\{p^2 K, n\right\}\right)\right)^{2/\varkappa_0}\right\}.\end{aligned} \tag{C.33}$$

.

Proof of Lemma 25: The proof for this lemma is along the lines of the proof of Lemma 7 in Wang et al. [2019]. For any fixed $r' \in (s,e]$, let $\mathcal{T}_1 = \mathcal{T}_{(s,r')}$ and $\mathcal{T}_2 = \mathcal{T}_{(r',e)}$. Recall that $L^*\left(\mathcal{T}\right)$ is the population counterpart of $L\left(\mathcal{T}\right)$. By Lemma 32 and the choice of ω, it holds that with probability $\left(1-\delta_1\right)\left(1-\delta_2\right)$ that

$$\max_{s' \in \left\{\mathcal{T}_{(s,e)}, \mathcal{T}_{(s,r')}, \mathcal{T}_{(r',e)}\right\}} \left|L\left(s'\right) - L^*\left(s'\right)\right| \leq C d_n^* \lambda^2 < \omega/3.$$

When there are no thresholds in $(s,e]$, we have $\boldsymbol{A}^*_{\mathcal{T}_{(s,e)}} = \boldsymbol{A}^*_{\mathcal{T}_{(s,r')}} = \boldsymbol{A}^*_{\mathcal{T}_{(r',e)}}$. Thus, $L\left(\mathcal{T}_{(s,e)}\right) < L\left(\mathcal{T}_{(s,r')}\right) + L\left(\mathcal{T}_{(r',e)}\right) + \omega$ with probability $\left(1-\delta_1\right)\left(1-\delta_2\right)$. Since $r' \in (s,e]$ is

fixed, we have

$$\begin{aligned}
&\mathbb{P}\left(L\left(\mathcal{T}_{(s,e)}\right) \geq \min_{r'\in(s,e]}\left\{L\left(\mathcal{T}_{(s,r')}\right)+L\left(\mathcal{T}_{(r',e)}\right)\right\}+\omega\right)\\
&\leq \sum_{r'\in(s,e]}\mathbb{P}\left(L\left(\mathcal{T}_{(s,e)}\right) \geq L\left(\mathcal{T}_{(s,r')}\right)+L\left(\mathcal{T}_{(r',e)}\right)+\omega\right)\\
&\leq n\left(1-\left(1-\delta_1\right)\left(1-\delta_2\right)\right)\\
&\leq n\left(\delta_1+\delta_2\right)\\
&\leq 2n\exp\left(-c_2\log\left(\max\left\{p^2K,n\right\}\right)\right)\\
&\quad + n\exp\left\{-c_4\log\left(\max\left\{p^2K,n\right\}\right)\left(\log\left(\max\left\{p^2K,n\right\}\right)\right)^{1-\varkappa_0/2}\right\}\\
&\quad + n\exp\left\{-c_5\left(\log\left(\max\left\{p^2K,n\right\}\right)\right)^{2/\varkappa_0}\right\}.
\end{aligned} \tag{C.34}$$

Then,

$$L\left(\mathcal{T}_{(s,e)}\right) < \min_{r'\in(s,e]}\left\{L\left(\mathcal{T}_{(s,r')}\right)+L\left(\mathcal{T}_{(r',e)}\right)\right\}+\omega,$$

with probability $1-\delta_4$ for

$$\begin{aligned}
\delta_4 = {} & 2n\exp\left(-c_2\log\left(\max\left\{p^2K,n\right\}\right)\right)\\
& + n\exp\left\{-c_4\log\left(\max\left\{p^2K,n\right\}\right)\left(\log\left(\max\left\{p^2K,n\right\}\right)\right)^{1-\varkappa_0/2}\right\}\\
& + n\exp\left\{-c_5\left(\log\left(\max\left\{p^2K,n\right\}\right)\right)^{2/\varkappa_0}\right\}.
\end{aligned} \tag{C.35}$$

Recalling that $c_2 > 5$ by the definition of δ_1 and δ_2 in Proposition 22, δ_4 converges to zero as $p, n \to \infty$.

Lemma 26. *Suppose Assumptions C1 to C4 hold, and that there are J thresholds in the interval $(s,e]$ and $J \geq 3$. Use the same notations as in Proposition 22 and Lemma 25. Let $r'_0 = s$, $r'_j = r_j$, and $r'_{J+1} = e$ for $j = 1, 2, ..., J$. Then, with probability $1-\delta_4$,*

$$L\left(\mathcal{T}_{(s,e)}\right) > \sum_{j=1}^{J+1} L\left(\mathcal{T}_{(r'_{j-1},r'_j)}\right) + J\omega. \tag{C.36}$$

Proof of Lemma 26: The proof for this lemma is along the lines of the proof of Lemma 8 in Wang et al. [2019]. For $J \geq 3$, $\left|\mathcal{T}_{(s,e)}\right| \geq \omega$ and $L\left(\mathcal{T}_{(s,e)}\right) =$

$\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}} \left\|\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right\|_2^2$. To prove by contradiction, we assume

$$L\left(\mathcal{T}_{(s,e)}\right) \leq \sum_{j=1}^{J+1} L\left(\mathcal{T}_{(r'_{j-1},r'_j)}\right) + J\omega. \tag{C.37}$$

Equation (C.37) gives

$$\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}} \left(\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right)^2 \leq \sum_{j=1}^{J+1} L\left(\mathcal{T}_{(r'_{j-1},r'_j)}\right) + J\omega. \tag{C.38}$$

By Lemma 32, we get

$$\begin{aligned}&\mathbb{P}\left(\sum_{j=1}^{J+1}\left(L\left(\mathcal{T}_{(r'_{j-1},r'_j)}\right) - L^*\left(\mathcal{T}_{(r'_{j-1},r'_j)}\right)\right) > (J+1)C_{20}\lambda^2 d_n^*\right)\\ \leq& n\left(1-(1-\delta_1)(1-\delta_2)\right) \leq n\left(\delta_1+\delta_2\right).\end{aligned} \tag{C.39}$$

By Assumptions C5 and C6 and Equation (C.22), $\left|\mathcal{T}_{(r_{j-1},r_j)}\right| \geq c_e n\Delta_n$ for $j = 2,\ldots,J$. By the choice of ω, $\left|\mathcal{T}_{(r_{j-1},r_j)}\right| \geq \omega$ for $j = 2,\ldots,J$. Using Equations (C.38) and (C.39), with high probability,

$$\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}} \left(\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right)^2 \leq \sum_{j=1}^{J+1} L^*\left(\mathcal{T}_{(r'_{j-1},r'_j)}\right) + (J+1)C_{20}\lambda^2 d_n^* + J\omega. \tag{C.40}$$

Without loss of generality, we assume $\left|\mathcal{T}_{(r'_0,r'_1)}\right| \leq \left|\mathcal{T}_{(r'_J,r'_{J+1})}\right|$ ($\left|\mathcal{T}_{(r'_0,r'_1)}\right| \geq \left|\mathcal{T}_{(r'_J,r'_{J+1})}\right|$ is similar). There are three cases to be considered: $\left|\mathcal{T}_{(r'_0,r'_1)}\right| \geq \omega$, $\left|\mathcal{T}_{(r'_0,r'_1)}\right| < \omega \leq \left|\mathcal{T}_{(r'_J,r'_{J+1})}\right|$ and $\left|\mathcal{T}_{(r'_J,r'_{J+1})}\right| < \omega$.

First, we prove $\left|\mathcal{T}_{(r'_0,r'_1)}\right| \geq \omega$. By Equation (C.40),

$$\begin{aligned}&\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}} \left(\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right)^2\\ \leq&\sum_{j=1}^{J+1}\sum_{z_{\pi(i)}\in\mathcal{T}_{(r'_{j-1},r'_j)}} \left(\boldsymbol{x}_{\pi(i)} - \boldsymbol{A}^*_{\mathcal{T}_{(r'_{j-1},r'_j)}}\boldsymbol{Y}_{\pi(i)}\right)^2 + (J+1)C_{20}\lambda^2 d_n^* + J\omega.\end{aligned} \tag{C.41}$$

Note that

$$
\begin{aligned}
&\sum_{z_{\pi(i)} \in \mathcal{T}_{(s,e)}} \left(\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} \boldsymbol{Y}_{\pi(i)} \right)^2 \\
&= \sum_{j=1}^{J+1} \sum_{z_{\pi(i)} \in \mathcal{T}_{(r_{j-1}, r_j)}} \left(\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} \boldsymbol{Y}_{\pi(i)} \right)^2 \\
&= \sum_{j=1}^{J+1} \sum_{z_{\pi(i)} \in \mathcal{T}_{(r_{j-1}, r_j)}} \left(\left(\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} - \boldsymbol{A}^*_{\mathcal{T}_{(r_{j-1}, r_j)}} \right) \boldsymbol{Y}_{\pi(i)} \right)^2 \\
&\quad + \sum_{j=1}^{J+1} \sum_{z_{\pi(i)} \in \mathcal{T}_{(r_{j-1}, r_j)}} \left(\boldsymbol{x}_{\pi(i)} - \boldsymbol{A}^*_{\mathcal{T}_{(r_{j-1}, r_j)}} \boldsymbol{Y}_{\pi(i)} \right)^2 \\
&\quad - 2 \sum_{j=1}^{J+1} \sum_{z_{\pi(i)} \in \mathcal{T}_{(r_{j-1}, r_j)}} \boldsymbol{\epsilon}'_{\pi(i)} \left(\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} - \boldsymbol{A}^*_{\mathcal{T}_{(r_{j-1}, r_j)}} \right) \boldsymbol{Y}_{\pi(i)}.
\end{aligned}
\tag{C.42}
$$

Recalling that $\boldsymbol{A}^*_{\mathcal{T}_{(r_{j-1}, r_j)}} = \boldsymbol{A}_{\mathcal{T}_{(r_{j-1}, r_j)}}$ when there are no thresholds in $(r_{j-1}, r_j]$ and combining Equations (C.41) and (C.42), we get

$$
\begin{aligned}
&\sum_{j=1}^{J+1} \sum_{z_{\pi(i)} \in \mathcal{T}_{(r'_{j-1}, r'_j)}} \left\| \left(\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} - \boldsymbol{A}^*_{\mathcal{T}_{(r'_{j-1}, r'_j)}} \right) \boldsymbol{Y}_{\pi(i)} \right\|_2^2 \\
\leq & 2 \sum_{j=1}^{J+1} \sum_{z_{\pi(i)} \in \mathcal{T}_{(r'_{j-1}, r'_j)}} \boldsymbol{\epsilon}'_{\pi(i)} \left(\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} - \boldsymbol{A}^*_{\mathcal{T}_{(r'_{j-1}, r'_j)}} \right) \boldsymbol{Y}_{\pi(i)} + (J+1) C_{20} \lambda^2 d_n^* + J\omega \\
\leq & 2 \sum_{j=1}^{J+1} \left\| \sum_{z_{\pi(i)} \in \mathcal{T}_{(r'_{j-1}, r'_j)}} \boldsymbol{\epsilon}_{\pi(i)} \boldsymbol{Y}'_{\pi(i)} \right\|_\infty \left\| \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} - \boldsymbol{A}^*_{\mathcal{T}_{(r'_{j-1}, r'_j)}} \right\|_1 + (J+1) C_{20} \lambda^2 d_n^* + J\omega \\
\leq & 2c_1 \sum_{j=1}^{J+1} \left(\sqrt{d_n^* \log\left(\max\left\{p^2 K, n\right\}\right) \left| \mathcal{T}_{(r'_{j-1}, r'_j)} \right|} \left\| \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} - \boldsymbol{A}^*_{\mathcal{T}_{(r'_{j-1}, r'_j)}} \right\|_{2,\mathcal{I}} \right. \\
& \left. + \sqrt{\log\left(\max\left\{p^2 K, n\right\}\right) \left| \mathcal{T}_{(r'_{j-1}, r'_j)} \right|} \left\| \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} - \boldsymbol{A}^*_{\mathcal{T}_{(r'_{j-1}, r'_j)}} \right\|_{1,\mathcal{I}^c} \right) + (J+1) C_{20} \lambda^2 d_n^* + J\omega \\
\leq & \lambda \sum_{j=1}^{J+1} \left(\sqrt{d_n^* \left| \mathcal{T}_{(r'_{j-1}, r'_j)} \right|} \left\| \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} - \boldsymbol{A}^*_{\mathcal{T}_{(r'_{j-1}, r'_j)}} \right\|_{2,\mathcal{I}} + \sqrt{\left| \mathcal{T}_{(r'_{j-1}, r'_j)} \right|} \left\| \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} - \boldsymbol{A}^*_{\mathcal{T}_{(r'_{j-1}, r'_j)}} \right\|_{1,\mathcal{I}^c} \right) \\
& + (J+1) C_{20} \lambda^2 d_n^* + J\omega,
\end{aligned}
\tag{C.43}
$$

where the third inequality holds by Proposition 22 (Equation (C.3)) and Cauchy-Schwarz inequality, and the forth inequality holds by the choice of λ.

Set $\boldsymbol{B}_j = \left(\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} - \boldsymbol{A}^*_{\mathcal{T}_{(r'_{j-1}, r'_j)}} \right)$. By Proposition 22 (Equation (C.4)) and the choice of λ, that is,

$$
\lambda = 2C_{12} \left(\log\left(\max\left\{ p^2 K, n \right\} \right) \right)^{1/\varkappa_0} d_n^* \geq \sqrt{\log\left(\max\left\{p^2 K, n\right\}\right)},
$$

we get

$$
\begin{aligned}
&\sum_{j=1}^{J+1} \sum_{z_{\pi(i)} \in \mathcal{T}_{(r'_{j-1}, r'_j)}} \left\| \left(\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} - \boldsymbol{A}^*_{\mathcal{T}_{(r'_{j-1}, r'_j)}} \right) \boldsymbol{Y}_{\pi(i)} \right\|_2^2 \\
\leq & \sum_{j=1}^{J+1} c_1 \sqrt{\left| \mathcal{T}_{(r'_{j-1}, r'_j)} \right| \log\left(\max\left\{p^2 K, n\right\}\right)} \left(\sqrt{d_n^*} \left\| \boldsymbol{B}_j \right\|_{2,\mathcal{I}} + \left\| \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} \right\|_{1,\mathcal{I}^c} \right) + J\omega \\
& + (J+1) C_{20} \lambda^2 d_n^* \\
\leq & \sum_{j=1}^{J+1} c_{19} \lambda \left(\sqrt{\left| \mathcal{T}_{(r'_{j-1}, r'_j)} \right| d_n^*} \left\| \boldsymbol{B}_j \right\|_{2,\mathcal{I}} + \sqrt{\left| \mathcal{T}_{(r'_{j-1}, r'_j)} \right|} \left\| \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} \right\|_{1,\mathcal{I}^c} \right) + J\omega + (J+1) C_{20} \lambda^2 d_n^* \\
\leq & \sum_{j=1}^{J+1} 2 c_{19}^2 \lambda^2 d_n^* / c_x + \sum_{j=1}^{J+1} \frac{c_x \left\| \boldsymbol{B}_j \right\|_2^2 \left| \mathcal{T}_{(r'_{j-1}, r'_j)} \right|}{4} + \sum_{j=1}^{J+1} \lambda / 2 \sqrt{\left| \mathcal{T}_{(r'_{j-1}, r'_j)} \right|} \left\| \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} \right\|_{1,\mathcal{I}^c} \\
& + J\omega + (J+1) C_{20} \lambda^2 d_n^* \\
\leq & (J+1) c_{20} \lambda^2 d_n^* + \sum_{j=1}^{J+1} \frac{c_x \left\| \boldsymbol{B}_j \right\|_2^2 \left| \mathcal{T}_{(r'_{j-1}, r'_j)} \right|}{4} + \sum_{j=1}^{J+1} \lambda / 2 \sqrt{\left| \mathcal{T}_{(r'_{j-1}, r'_j)} \right|} C_8 \lambda \sqrt{d_n^*} / \sqrt{\left| \mathcal{T}_{(s,e)} \right|} + J\omega \\
\leq & \sum_{j=1}^{J+1} \frac{c_x \left\| \boldsymbol{B}_j \right\|_2^2 \left| \mathcal{T}_{(r'_{j-1}, r'_j)} \right|}{4} + J\omega + (J+1) c_{21} \lambda^2 d_n^*,
\end{aligned}
\tag{C.44}
$$

where the third and forth inequalities follow from Hölder's inequality and Lemma 30, respectively.

Recalling that the case we consider is $\left| \mathcal{T}_{(r'_0, r'_1)} \right| \geq \omega$, $\left| \mathcal{T}_{(r'_{j-1}, r'_j)} \right| \geq \omega$ for $j = 1, \ldots, J+1$. With the choice of ω and λ, we get

$$
\left| \mathcal{T}_{(r'_{j-1}, r'_j)} \right| \geq c_{22}(m_0 + 1) \left(\log\left(\max\left\{p^2 K, n\right\}\right) \right)^{2/\varkappa_0} d_n^{*3} \geq c_{22} \left(\log\left(\max\left\{p^2 K, n\right\}\right) \right)^{2/\varkappa_0} d_n^{*3}
\tag{C.45}
$$

for $j = 1, \ldots, J+1$.

Recalling that $\boldsymbol{I}_{pK}$ is a $pK \times pK$ diagonal matrix with all diagonal elements equal to 1,

$$
\begin{aligned}
&\sum_{z_{\pi(i)} \in \mathcal{T}_{(r'_{j-1}, r'_j)}} \left\| \left(\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} - \boldsymbol{A}^*_{\mathcal{T}_{(r'_{j-1}, r'_j)}} \right) \boldsymbol{Y}_{\pi(i)} \right\|_2^2 \\
=& \left(\operatorname{vec}(\boldsymbol{B}_j) \right)' \left(\left(\sum_{z_{\pi(i)} \in \mathcal{T}_{(r'_{j-1}, r'_j)}} \boldsymbol{Y}_{\pi(i)} \boldsymbol{Y}'_{\pi(i)} \right) \otimes \boldsymbol{I}_{pK} \right) \operatorname{vec}(\mathrm{B_j}) \\
\geq& \left(\operatorname{vec}(\boldsymbol{B}_j) \right)' \left(\mathbb{E} \left(\sum_{z_{\pi(i)} \in \mathcal{T}_{(r'_{j-1}, r'_j)}} \boldsymbol{Y}_{\pi(i)} \boldsymbol{Y}'_{\pi(i)} \right) \otimes \boldsymbol{I}_{pK} \right) \operatorname{vec}(\boldsymbol{B}_j) \\
&- \left(\operatorname{vec}(\boldsymbol{B}_j) \right)' \left(\left(\sum_{z_{\pi(i)} \in \mathcal{T}_{(r'_{j-1}, r'_j)}} \boldsymbol{Y}_{\pi(i)} \boldsymbol{Y}'_{\pi(i)} - \mathbb{E} \left(\sum_{z_{\pi(i)} \in \mathcal{T}_{(r'_{j-1}, r'_j)}} \boldsymbol{Y}_{\pi(i)} \boldsymbol{Y}'_{\pi(i)} \right) \right) \otimes \boldsymbol{I}_{pK} \right) \operatorname{vec}(\mathrm{B_j}) \\
\geq& c_{11} \left| \mathcal{T}_{(r'_{j-1}, r'_j)} \right| \|\boldsymbol{B}_j\|_2^2 \\
&- \left\| \left(\sum_{z_{\pi(i)} \in \mathcal{T}_{(r'_{j-1}, r'_j)}} \boldsymbol{Y}_{\pi(i)} \boldsymbol{Y}'_{\pi(i)} - \mathbb{E} \left(\sum_{z_{\pi(i)} \in \mathcal{T}_{(r'_{j-1}, r'_j)}} \boldsymbol{Y}_{\pi(i)} \boldsymbol{Y}'_{\pi(i)} \right) \right) \otimes \boldsymbol{I}_{pK} \right\|_\infty \|\boldsymbol{B}_j\|_1^2 \\
\geq& c_{11} \left| \mathcal{T}_{(r'_{j-1}, r'_j)} \right| \|\boldsymbol{B}_j\|_2^2 - c_3 \left(\log\left(\max\left\{ p^2 K, n \right\} \right) \right)^{1/\varkappa_0} \sqrt{\left| \mathcal{T}_{(r'_{j-1}, r'_j)} \right|} \|\boldsymbol{B}_j\|_1^2 \\
\geq& c_{11} \left| \mathcal{T}_{(r'_{j-1}, r'_j)} \right| \|\boldsymbol{B}_j\|_2^2 - c_3 \left(\log\left(\max\left\{ p^2 K, n \right\} \right) \right)^{1/\varkappa_0} \sqrt{\left| \mathcal{T}_{(r'_{j-1}, r'_j)} \right|} d_n^* \|\boldsymbol{B}_j\|_{2,\mathcal{I}}^2 \\
&- c_3 \left(\log\left(\max\left\{ p^2 K, n \right\} \right) \right)^{1/\varkappa_0} \sqrt{\left| \mathcal{T}_{(r'_{j-1}, r'_j)} \right|} \|\boldsymbol{B}_j\|_{1,\mathcal{I}^c}^2 \\
\geq& c_{11} \left| \mathcal{T}_{(r'_{j-1}, r'_j)} \right| \|\boldsymbol{B}_j\|_2^2 - c_3 \left(\log\left(\max\left\{ p^2 K, n \right\} \right) \right)^{1/\varkappa_0} \sqrt{\left| \mathcal{T}_{(r'_{j-1}, r'_j)} \right|} \left(C_8 \lambda d_n^* / \sqrt{\left| \mathcal{T}_{(s,e)} \right|} \right)^2 \\
\geq& c_{11} \left| \mathcal{T}_{(r'_{j-1}, r'_j)} \right| \|\boldsymbol{B}_j\|_2^2 - c_{12} \lambda^2 \left(\log\left(\max\left\{ p^2 K, n \right\} \right) \right)^{1/\varkappa_0} d_n^{*2} / \sqrt{\left| \mathcal{T}_{(r'_{j-1}, r'_j)} \right|} \\
\geq& c_x \left| \mathcal{T}_{(r'_{j-1}, r'_j)} \right| \|\boldsymbol{B}_j\|_2^2 - c_{13} \lambda^2 d_n^*,
\end{aligned}
\tag{C.46}
$$

where the second and the third inequalities follow from the Cauchy-Schwarz inequality and Proposition 22 (Equation (C.4)), respectively. The forth inequality follows from the Cauchy-Schwarz inequality; the fifth inequality holds by Lemma 30 and the fact that $\left| \mathcal{T}_{(s,e)} \right| \geq \left| \mathcal{T}_{(r'_{j-1}, r'_j)} \right| \geq c_{22} \left(\log\left(\max\left\{ p^2 K, n \right\} \right) \right)^{2/\varkappa_0} d_n^{*3}$ by Equation (C.45); the last inequality is due

to the fact that

$$\sqrt{\left|\mathcal{T}_{(r'_{j-1},r'_j)}\right|} > \sqrt{c_{22}}\left(\log\left(\max\left\{p^2K,n\right\}\right)\right)^{1/\varkappa_0} d_n^*$$

by Equation (C.45).

Combining Equations (C.44) and (C.46), we get

$$\begin{aligned}&\sum_{j=1}^{J+1}\frac{c_x\left\|\boldsymbol{B}_j\right\|_2^2\left|\mathcal{T}_{(r'_{j-1},r'_j)}\right|}{4}+J\omega+(J+1)c_{21}\lambda^2 d_n^*\\ \geq&\sum_{j=1}^{J+1}c_x\left|\mathcal{T}_{(r'_{j-1},r'_j)}\right|\left\|\boldsymbol{B}_j\right\|_2^2-(J+1)c_{13}\lambda^2 d_n^*,\end{aligned}\tag{C.47}$$

which leads to

$$J\omega+c_{23}(J+1)\lambda^2 d_n^*\geq\sum_{j=1}^{J+1}\frac{3c_x}{4}\left|\mathcal{T}_{(r'_{j-1},r'_j)}\right|\left\|\boldsymbol{B}_j\right\|_2^2\geq\sum_{j=2}^{J}\frac{3c_x}{4}\left|\mathcal{T}_{(r'_{j-1},r'_j)}\right|\left\|\boldsymbol{B}_j\right\|_2^2.\tag{C.48}$$

By Assumption C4, for $j=2,...,J-1$,

$$\begin{aligned}&\left|\mathcal{T}_{(r'_{j-1},r'_j)}\right|\left\|\boldsymbol{B}_j\right\|_2^2+\left|\mathcal{T}_{(r'_j,r'_{j+1})}\right|\left\|\boldsymbol{B}_{j-1}\right\|_2^2\\ \geq&\inf_{\boldsymbol{M}}\left\{\left|\mathcal{T}_{(r'_{j-1},r'_j)}\right|\left\|\boldsymbol{A}^*_{\mathcal{T}_{(r'_{j-1},r'_j)}}-\boldsymbol{M}\right\|_2^2+\left|\mathcal{T}_{(r'_j,r'_{j+1})}\right|\left\|\boldsymbol{A}^*_{\mathcal{T}_{(r'_j,r'_{j+1})}}-\boldsymbol{M}\right\|_2^2\right\}\\ \geq&v^2\frac{\left|\mathcal{T}_{(r'_{j-1},r'_j)}\right|\left|\mathcal{T}_{(r'_j,r'_{j+1})}\right|}{\left|\mathcal{T}_{(r'_{j-1},r'_j)}\right|+\left|\mathcal{T}_{(r'_j,r'_{j+1})}\right|}\\ \geq&\min\left\{\left|\mathcal{T}_{(r'_{j-1},r'_j)}\right|,\left|\mathcal{T}_{(r'_j,r'_{j+1})}\right|\right\}v^2/2,\end{aligned}\tag{C.49}$$

Applying similar arguments as in Lemma 23 and combining Equations (C.48) and (C.49), with probability at least $1-n\delta_1-n\delta_2$,

$$\min_{j=2,...,J}\left|\mathcal{T}_{(r'_{j-1},r'_j)}\right|\leq C_0\left(\frac{\lambda^2 d_n^*+\omega}{v^2}\right),$$

which contradicts the fact that $\left|\mathcal{T}_{(r'_{j-1},r'_j)}\right|\geq c_e n\Delta_n\geq c_e C_\delta\left(\log\left(p^2K\right)\right)^{2/\varkappa_0+\xi}m_0 d_n^{*3}/v^2$ for $j=2,...,J$ by Assumption C6 and Equation (C.22). Thus, if $\left|\mathcal{T}_{(r'_0,r'_1)}\right|\geq\omega$, then $L\left(\mathcal{T}_{(s,e)}\right)>\sum_{j=1}^{J+1}L\left(\mathcal{T}_{(r'_{j-1},r'_j)}\right)+J\omega$.

Similarly, we can prove the rest of two cases. For the case $\left|\mathcal{T}_{(r'_0,r'_1)}\right| < \omega \leq \left|\mathcal{T}_{(r'_J,r'_{J+1})}\right|$, Equation (C.40) becomes

$$\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}} \left(\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right)^2 \leq \sum_{j=2}^{J+1} L^*\left(\mathcal{T}_{(r'_{j-1},r'_j)}\right) + JC_{20}\lambda^2 d_n^* + J\omega. \tag{C.50}$$

The rest of the proof is identical to the case $\left|\mathcal{T}_{(r'_0,r'_1)}\right| \geq \omega$.

For the case $\left|\mathcal{T}_{(r'_J,r'_{J+1})}\right| < \omega$, Equation (C.40) becomes

$$\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}} \left(\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right)^2 \leq \sum_{j=2}^{J} L^*\left(\mathcal{T}_{(r'_{j-1},r'_j)}\right) + (J-1)C_{20}\lambda^2 d_n^* + J\omega. \tag{C.51}$$

The rest of the proof is identical to the case $\left|\mathcal{T}_{(r'_0,r'_1)}\right| \geq \omega$.

Lemma 27. *Denote* $\boldsymbol{B}_{\mathcal{T}_{(s,e)}} = \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} - \boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}$ *and use the same notations as in Proposition 22. Suppose Assumptions C1 to C4 hold, for any given interval* $(s, e]$,

$$\begin{aligned}\mathbb{P}\Bigg(\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}} \left(\left(\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} - \boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}\right)\boldsymbol{Y}_{\pi(i)}\right)^2 \geq C_6\left|\mathcal{T}_{(s,e)}\right|\left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\right\|_2^2 \\ -c_3\left(\log\left(\max\left\{p^2K, n\right\}\right)\right)^{1/\varkappa_0}\sqrt{\left|\mathcal{T}_{(s,e)}\right|}\left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\right\|_1^2\Bigg) = 1-\delta_2,\end{aligned} \tag{C.52}$$

where constants $C_6 > 0, c_3 > 0$.

Proof of Lemma 27: Recalling that $\boldsymbol{I}_{pK}$ is a $pK \times pK$ diagonal matrix with all diagonal

elements equal to 1,

$$
\begin{aligned}
&\sum_{z_{\pi(i)} \in \mathcal{T}_{(s,e)}} \left(\left(\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} - \boldsymbol{A}^*_{\mathcal{T}_{(s,e)}} \right) \boldsymbol{Y}_{\pi(i)} \right)^2 \\
=& \left(\operatorname{vec} \left(\boldsymbol{B}_{\mathcal{T}_{(s,e)}} \right) \right)' \left(\left(\sum_{z_{\pi(i)} \in \mathcal{T}_{(s,e)}} \boldsymbol{Y}_{\pi(i)} \boldsymbol{Y}'_{\pi(i)} \right) \otimes \boldsymbol{I}_{pK} \right) \operatorname{vec} \left(\boldsymbol{B}_{\mathcal{T}_{(s,e)}} \right) \\
\geq& \left(\operatorname{vec} \left(\boldsymbol{B}_{\mathcal{T}_{(s,e)}} \right) \right)' \left(\mathbb{E} \left(\sum_{z_{\pi(i)} \in \mathcal{T}_{(s,e)}} \boldsymbol{Y}_{\pi(i)} \boldsymbol{Y}'_{\pi(i)} \right) \otimes \boldsymbol{I}_{pK} \right) \operatorname{vec} \left(\boldsymbol{B}_{\mathcal{T}_{(s,e)}} \right) \\
&- \left| \left(\operatorname{vec} \left(\boldsymbol{B}_{\mathcal{T}_{(s,e)}} \right) \right)' \left(\left(\sum_{z_{\pi(i)} \in \mathcal{T}_{(s,e)}} \boldsymbol{Y}_{\pi(i)} \boldsymbol{Y}'_{\pi(i)} - \mathbb{E} \left(\sum_{z_{\pi(i)} \in \mathcal{T}_{(s,e)}} \boldsymbol{Y}_{\pi(i)} \boldsymbol{Y}'_{\pi(i)} \right) \right) \otimes \boldsymbol{I}_{pK} \right) \right. \\
&\left. \operatorname{vec} \left(\boldsymbol{B}_{\mathcal{T}_{(s,e)}} \right) \right| \\
\geq& C_6 \left| \mathcal{T}_{(s,e)} \right| \left\| \boldsymbol{B}_{\mathcal{T}_{(s,e)}} \right\|_2^2 \\
&- \left\| \left(\sum_{z_{\pi(i)} \in \mathcal{T}_{(s,e)}} \boldsymbol{Y}_{\pi(i)} \boldsymbol{Y}'_{\pi(i)} - \mathbb{E} \left(\sum_{z_{\pi(i)} \in \mathcal{T}_{(s,e)}} \boldsymbol{Y}_{\pi(i)} \boldsymbol{Y}'_{\pi(i)} \right) \right) \otimes \boldsymbol{I}_{pK} \right\|_\infty \left\| \boldsymbol{B}_{\mathcal{T}_{(s,e)}} \right\|_1^2 \\
\geq& C_6 \left| \mathcal{T}_{(s,e)} \right| \left\| \boldsymbol{B}_{\mathcal{T}_{(s,e)}} \right\|_2^2 - c_3 \left(\log \left(\max \left\{ p^2 K, n \right\} \right) \right)^{1/\varkappa_0} \sqrt{\left| \mathcal{T}_{(s,e)} \right|} \left\| \boldsymbol{B}_{\mathcal{T}_{(s,e)}} \right\|_1^2,
\end{aligned}
\tag{C.53}
$$

where the second and the third inequalities follow from the Cauchy-Schwarz inequality and Proposition 22 (Equation (C.4)), respectively.

Lemma 28. *Suppose Assumptions C1 to C4 hold and that the interval* $(s, e] \subseteq (r_{j-1}, r_j]$. *Then, there exist certain positive constants* C, C_1, *and* C_2 *such that with probability* $(1 - \delta_1)(1 - \delta_2)$, *for* $\left| \mathcal{T}_{(s,e)} \right| \geq C \left(\log \left(\max \left\{ p^2 K, n \right\} \right) \right)^{2/\varkappa_0} d_n^{*2}$ *and* $\lambda \geq 2C_1 \sqrt{\log \left(\max \left\{ p^2 K, n \right\} \right)}$,

$$
\left\| \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} \right\|_{1, \mathcal{I}^c} \leq C_2 \lambda d_n^* / \sqrt{\left| \mathcal{T}_{(s,e)} \right|}, \tag{C.54}
$$

and

$$
\left\| \boldsymbol{A}_{\mathcal{T}_{(s,e)}} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} \right\|_1 \leq C_2 \lambda d_n^* / \sqrt{\left| \mathcal{T}_{(s,e)} \right|}. \tag{C.55}
$$

In addition,

$$\begin{aligned}\left\|\boldsymbol{A}_{\mathcal{T}_{(s,e)}}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_{1,\mathcal{I}} &\leq \left\|\boldsymbol{A}_{\mathcal{T}_{(s,e)}}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_{1} \\ &\leq 4\left\|\boldsymbol{A}_{\mathcal{T}_{(s,e)}}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_{1,\mathcal{I}} \\ &\leq 4\sqrt{d_n^*}\left\|\boldsymbol{A}_{\mathcal{T}_{(s,e)}}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_{2,\mathcal{I}}.\end{aligned} \tag{C.56}$$

Proof of Lemma 28: By Equation (4.3),

$$\begin{aligned}&\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left\|\boldsymbol{x}_{\pi(i)}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right\|_2^2+\lambda\sqrt{\left|\mathcal{T}_{(s,e)}\right|}\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_1 \\ \leq &\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left\|\boldsymbol{x}_{\pi(i)}-\boldsymbol{A}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right\|_2^2+\lambda\sqrt{\left|\mathcal{T}_{(s,e)}\right|}\left\|\boldsymbol{A}_{\mathcal{T}_{(s,e)}}\right\|_1.\end{aligned} \tag{C.57}$$

Then, with probability $1-\delta_1$, we can rewrite Equation (C.57) to obtain

$$\begin{aligned}&\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}-\boldsymbol{A}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right\|_2^2 \\ \leq &2\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left(\boldsymbol{x}_{\pi(i)}-\boldsymbol{A}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right)'\left(\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}-\boldsymbol{A}_{\mathcal{T}_{(s,e)}}\right)\boldsymbol{Y}_{\pi(i)} \\ &+\lambda\sqrt{\left|\mathcal{T}_{(s,e)}\right|}\left(\left\|\boldsymbol{A}_{\mathcal{T}_{(s,e)}}\right\|_1-\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_1\right) \\ \leq &2\left\|\boldsymbol{A}_{\mathcal{T}_{(s,e)}}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_1\left\|\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\boldsymbol{\epsilon}_{\pi(i)}\boldsymbol{Y}'_{\pi(i)}\right\|_\infty+\lambda\sqrt{\left|\mathcal{T}_{(s,e)}\right|}\left(\left\|\boldsymbol{A}_{\mathcal{T}_{(s,e)}}\right\|_1-\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_1\right) \\ \leq &C_1\left\|\boldsymbol{A}_{\mathcal{T}_{(s,e)}}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_1\left|\mathcal{T}_{(s,e)}\right|\sqrt{\frac{\log\left(\max\left\{p^2K,n\right\}\right)}{\left|\mathcal{T}_{(s,e)}\right|}}+\lambda\sqrt{\left|\mathcal{T}_{(s,e)}\right|}\left(\left\|\boldsymbol{A}_{\mathcal{T}_{(s,e)}}\right\|_1-\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_1\right) \\ \leq &C_1\sqrt{\left|\mathcal{T}_{(s,e)}\right|\log\left(\max\left\{p^2K,n\right\}\right)}\left\|\boldsymbol{A}_{\mathcal{T}_{(s,e)}}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_1+\lambda\sqrt{\left|\mathcal{T}_{(s,e)}\right|}\left(\left\|\boldsymbol{A}_{\mathcal{T}_{(s,e)}}\right\|_1-\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_1\right) \\ \leq &\frac{\lambda}{2}\sqrt{\left|\mathcal{T}_{(s,e)}\right|}\left\|\boldsymbol{A}_{\mathcal{T}_{(s,e)}}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_1+\lambda\sqrt{\left|\mathcal{T}_{(s,e)}\right|}\left(\left\|\boldsymbol{A}_{\mathcal{T}_{(s,e)}}\right\|_1-\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_1\right),\end{aligned} \tag{C.58}$$

where the third inequality is due to Proposition 22 (Equation (C.3)).

Since

$$\left\|\boldsymbol{A}_{\mathcal{T}_{(s,e)}}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_1=\left\|\boldsymbol{A}_{\mathcal{T}_{(s,e)}}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_{1,\mathcal{I}}+\left\|\boldsymbol{A}_{\mathcal{T}_{(s,e)}}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_{1,\mathcal{I}^c}$$

and Equation (C.58), we get

$$\left\|\boldsymbol{A}_{\mathcal{T}_{(s,e)}}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_{1,\mathcal{I}^c}=\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_{1,\mathcal{I}^c}\leq 3\left\|\boldsymbol{A}_{\mathcal{T}_{(s,e)}}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_{1,\mathcal{I}}. \tag{C.59}$$

By Cauchy-Schwarz inequality and Assumption C3, we can obtain

$$\begin{aligned}
&\left\|\boldsymbol{A}_{\mathcal{T}_{(s,e)}}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_{1,\mathcal{I}}\\
\leq&\sum_{l\in\mathcal{I}}\left(\left|\operatorname{vec}\left(\boldsymbol{A}_{\mathcal{T}_{(s,e)}}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right)_l\right|\right)\\
\leq&\left(\sum_{l\in\mathcal{I}}\left|\operatorname{vec}\left(\boldsymbol{A}_{\mathcal{T}_{(s,e)}}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right)_l\right|^2\right)^{1/2}\left(\sum_{l\in\mathcal{I}}1\right)^{1/2}\\
=&\sqrt{d_n^*}\left\|\boldsymbol{A}_{\mathcal{T}_{(s,e)}}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_{2,\mathcal{I}},
\end{aligned} \tag{C.60}$$

where $\operatorname{vec}\left(\boldsymbol{A}_{\mathcal{T}_{(s,e)}}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right)_l$ represents the l-th element of $\operatorname{vec}\left(\boldsymbol{A}_{\mathcal{T}_{(s,e)}}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right)$ and $\operatorname{vec}\left(\boldsymbol{A}_{\mathcal{T}_{(s,e)}}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right)$ is the vector obtained from $\left(\boldsymbol{A}_{\mathcal{T}_{(s,e)}}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right)$ by concatenating the rows of $\left(\boldsymbol{A}_{\mathcal{T}_{(s,e)}}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right)$. Combining Equations (C.59) and (C.60), we obtain

$$\begin{aligned}
&\left\|\boldsymbol{A}_{\mathcal{T}_{(s,e)}}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_{1,\mathcal{I}}\leq\left\|\boldsymbol{A}_{\mathcal{T}_{(s,e)}}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_1\\
\leq&4\left\|\boldsymbol{A}_{\mathcal{T}_{(s,e)}}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_{1,\mathcal{I}}\\
\leq&4\sqrt{d_n^*}\left\|\boldsymbol{A}_{\mathcal{T}_{(s,e)}}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_{2,\mathcal{I}}.
\end{aligned} \tag{C.61}$$

Recalling that if there are no thresholds in the interval $(s,e]$, then $\boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}=\boldsymbol{A}_{\mathcal{T}_{(s,e)}}$. Since $(s,e]\subseteq(r_{j-1},r_j]$, $\boldsymbol{B}_{\mathcal{T}_{(s,e)}}=\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}-\boldsymbol{A}_{\mathcal{T}_{(s,e)}}$. Directly followed by Lemma 27, with

probability $1-\delta_2$, we have

$$\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left(\left(\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}-\boldsymbol{A}_{\mathcal{T}_{(s,e)}}\right)\boldsymbol{Y}_{\pi(i)}\right)^2 \geq C_6\left|\mathcal{T}_{(s,e)}\right|\left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\right\|_2^2 - c_3\left(\log\left(\max\left\{p^2K,n\right\}\right)\right)^{1/\varkappa_0}\sqrt{\left|\mathcal{T}_{(s,e)}\right|}\left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\right\|_1^2. \tag{C.62}$$

Since Equation (C.61), we can get

$$\begin{aligned}\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left(\left(\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}-\boldsymbol{A}_{\mathcal{T}_{(s,e)}}\right)\boldsymbol{Y}_{\pi(i)}\right)^2 &\geq C_6\left|\mathcal{T}_{(s,e)}\right|\left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\right\|_2^2 \\ &\quad - c_3\left(\log\left(\max\left\{p^2K,n\right\}\right)\right)^{1/\varkappa_0}\sqrt{\left|\mathcal{T}_{(s,e)}\right|}\left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\right\|_1^2 \\ &\geq \left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\right\|_2^2\Bigg(C_6\left|\mathcal{T}_{(s,e)}\right| \\ &\quad -C_{13}\left(\log\left(\max\left\{p^2K,n\right\}\right)\right)^{1/\varkappa_0}\sqrt{\left|\mathcal{T}_{(s,e)}\right|}d_n^*\Bigg).\end{aligned} \tag{C.63}$$

Since $\left|\mathcal{T}_{(s,e)}\right| \geq C\left(\log\left(\max\left\{p^2K,n\right\}\right)\right)^{2/\varkappa_0} d_n^{*2}$, then $\sqrt{C}\left(\log\left(\max\left\{p^2K,n\right\}\right)\right)^{1/\varkappa_0} d_n^*\sqrt{\left|\mathcal{T}_{(s,e)}\right|} \leq \left|\mathcal{T}_{(s,e)}\right|$. Recalling that C_{13} depends on the constant $c_3 > 5$ chosen in Proposition 22, then we can choose c_3 to be small enough so that $C_{13}/\sqrt{C} < C_6$. Next, Equation (C.63) leads to $\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left(\left(\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}-\boldsymbol{A}_{\mathcal{T}_{(s,e)}}\right)\boldsymbol{Y}_{\pi(i)}\right)^2 \geq C_{14}\left|\mathcal{T}_{(s,e)}\right|\left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\right\|_2^2$ for $C_{14}>0$.

Then, we can rewrite Equation (C.58) to obtain

$$\begin{aligned}C_{14}\left|\mathcal{T}_{(s,e)}\right|\left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\right\|_2^2 &\leq \frac{\lambda}{2}\sqrt{\left|\mathcal{T}_{(s,e)}\right|}\left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\right\|_1 + \lambda\sqrt{\left|\mathcal{T}_{(s,e)}\right|}\left(\left\|\boldsymbol{A}_{\mathcal{T}_{(s,e)}}\right\|_1 - \left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_1\right) \\ &\leq C_{15}\lambda\sqrt{\left|\mathcal{T}_{(s,e)}\right|}\left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\right\|_{1,\mathcal{I}} \\ &\leq C_{16}\lambda\sqrt{\left|\mathcal{T}_{(s,e)}\right|}\left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\right\|_2\sqrt{d_n^*},\end{aligned} \tag{C.64}$$

where the second and the third inequalities follow from Equation (C.61) and the triangle inequality.

Finally, by Equation (C.64), we get

$$\left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\right\|_2 \leq C_{17}\lambda\sqrt{d_n^*}/\sqrt{\left|\mathcal{T}_{(s,e)}\right|}. \tag{C.65}$$

In addition, by Equation (C.61), we obtain

$$\left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\right\|_1 \leq C_{17}\lambda d_n^*/\sqrt{\left|\mathcal{T}_{(s,e)}\right|}. \tag{C.66}$$

Lemma 29. *Suppose Assumptions C1 to C4 hold and that the interval* $(s,e] \subseteq (r_{j-1}, r_j]$. *For*

$$\left|\mathcal{T}_{(s,e)}\right| \geq C\left(\log\left(\max\left\{p^2K, n\right\}\right)\right)^{2/\varkappa_0} d_n^{*2}$$

and $\lambda \geq C_3\sqrt{\log\left(\max\left\{p^2K, n\right\}\right)}$,

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left\|\boldsymbol{x}_{\pi(i)} - \boldsymbol{A}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right\|_2^2 - \sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left\|\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right\|_2^2\right| \leq C_4\lambda^2 d_n^*\right) \\
&= (1-\delta_1)(1-\delta_2),
\end{aligned} \tag{C.67}$$

where C, C_3 *and* C_4 *are positive constants.*

Proof of Lemma 29: By rewriting Equation (C.57), we can get

$$\begin{aligned}
&\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left\|\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right\|_2^2 - \sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left\|\boldsymbol{x}_{\pi(i)} - \boldsymbol{A}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right\|_2^2 \\
&\leq \lambda\sqrt{\left|\mathcal{T}_{(s,e)}\right|}\left(\left\|\boldsymbol{A}_{\mathcal{T}_{(s,e)}}\right\|_1 - \left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_1\right) \\
&\leq \lambda\sqrt{\left|\mathcal{T}_{(s,e)}\right|}\left\|\boldsymbol{A}_{\mathcal{T}_{(s,e)}} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_1.
\end{aligned} \tag{C.68}$$

By Equation (C.68) and Lemma 28, we obtain

$$\begin{aligned}
&\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left\|\boldsymbol{x}_{\pi(i)}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right\|_2^2-\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left\|\boldsymbol{x}_{\pi(i)}-\boldsymbol{A}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right\|_2^2\\
&\leq\lambda\sqrt{|\mathcal{T}_{(s,e)}|}\left\|\boldsymbol{A}_{\mathcal{T}_{(s,e)}}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_1\\
&\leq\lambda\sqrt{|\mathcal{T}_{(s,e)}|}C_2\lambda d_n^*/\sqrt{|\mathcal{T}_{(s,e)}|}\\
&\leq C_2\lambda^2 d_n^*.
\end{aligned}\tag{C.69}$$

In addition,

$$\begin{aligned}
&\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left\|\boldsymbol{x}_{\pi(i)}-\boldsymbol{A}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right\|_2^2-\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left\|\boldsymbol{x}_{\pi(i)}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right\|_2^2\\
&=-\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}-\boldsymbol{A}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right\|_2^2\\
&\quad+2\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left(\boldsymbol{x}_{\pi(i)}-\boldsymbol{A}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right)'\left(\boldsymbol{A}_{\mathcal{T}_{(s,e)}}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right)\boldsymbol{Y}_{\pi(i)}\\
&\leq 2\left\|\boldsymbol{A}_{\mathcal{T}_{(s,e)}}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_1\left\|\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\boldsymbol{\epsilon}_{\pi(i)}\boldsymbol{Y}'_{\pi(i)}\right\|_\infty\\
&\leq C_2\lambda d_n^*/\sqrt{|\mathcal{T}_{(s,e)}|}c_1\frac{|\mathcal{T}_{(s,e)}|\sqrt{\log\left(\max\left\{p^2K,n\right\}\right)}}{\sqrt{|\mathcal{T}_{(s,e)}|}}\\
&\leq C_2c_1\lambda d_n^*\sqrt{\log\left(\max\left\{p^2K,n\right\}\right)}\\
&\leq C_5\lambda^2 d_n^*,
\end{aligned}\tag{C.70}$$

where C_2 and C_5 are positive constants. The second inequality holds with high probability by Proposition 22 (Equation (C.3)) and Lemma 28; the last inequality follows from the choice of λ. Combining Equations (C.69) and (C.70), Equation (C.67) holds.

Lemma 30. *Suppose Assumptions C1 to C4 holds. For any given the interval $(s,e]$, with probability $(1-\delta_1)(1-\delta_2)$, $|\mathcal{T}_{(s,e)}|\geq C\left(\log\left(\max\left\{p^2K,n\right\}\right)\right)^{2/\varkappa_0}d_n^{*2}$ and $\lambda=$*

$2C_{12}\left(\log\left(\max\left\{p^2K,n\right\}\right)\right)^{1/\varkappa_0} d_n^*$,

$$\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_{1,\mathcal{I}^c} \leq C_8\lambda\sqrt{d_n^*}/\sqrt{\left|\mathcal{T}_{(s,e)}\right|}, \tag{C.71}$$

where C, c_λ and C_8 are positive constants.

In addition, for $C_9 > 0$,

$$\begin{aligned}\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} - \boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}\right\|_2 &\leq C_9\lambda\sqrt{d_n^*}/\sqrt{\left|\mathcal{T}_{(s,e)}\right|},\\ \left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} - \boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}\right\|_1 &\leq C_9\lambda d_n^*/\sqrt{\left|\mathcal{T}_{(s,e)}\right|},\end{aligned} \tag{C.72}$$

and

$$\begin{aligned}\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} - \boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}\right\|_{1,\mathcal{I}} &\leq \left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} - \boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}\right\|_1\\ &\leq 4\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} - \boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}\right\|_{1,\mathcal{I}}\\ &\leq 4\sqrt{d_n^*}\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} - \boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}\right\|_{2,\mathcal{I}}.\end{aligned} \tag{C.73}$$

Proof of Lemma 30: Set $\boldsymbol{B}_{\mathcal{T}_{(s,e)}} = \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} - \boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}$. By Equation (4.3), we have

$$\begin{aligned}&\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right\|_2^2 + 2\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left(\boldsymbol{x}_{\pi(i)} - \boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right)'\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\\ &\leq \lambda\sqrt{\left|\mathcal{T}_{(s,e)}\right|}\left(\left\|\boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}\right\|_1 - \left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_1\right).\end{aligned} \tag{C.74}$$

Note that

$$\begin{aligned}&\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left(\boldsymbol{x}_{\pi(i)} - \boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right)'\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\\ =&\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\boldsymbol{\epsilon}'_{\pi(i)}\left(\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right) + \sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left(\left(\boldsymbol{A}_{\pi(i)} - \boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}\right)\boldsymbol{Y}_{\pi(i)}\right)'\left(\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right)\\ :=&H_1 + H_2,\end{aligned} \tag{C.75}$$

where $\boldsymbol{A}_{\pi(i)}$ is the true coefficient at the time point $\pi(i)$.

For H_1, we have $|H_1| \leq \left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\right\|_1 \left\|\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}} \boldsymbol{\epsilon}_{\pi(i)}\boldsymbol{Y}'_{\pi(i)}\right\|_\infty$. By Proposition 22, with probability $1-\delta_1$,

$$\begin{aligned} |H_1| &\leq c_3\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\|_1 \left(\left|\mathcal{T}_{(s,e)}\right| \sqrt{\log\left(\max\left\{p^2K, n\right\}\right) / \left|\mathcal{T}_{(s,e)}\right|}\right) \\ &= c_3\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\|_1 \sqrt{\log\left(\max\left\{p^2K, n\right\}\right)\left|\mathcal{T}_{(s,e)}\right|}. \end{aligned} \tag{C.76}$$

For H_2, we have

$$\begin{aligned} |H_2| &\leq \left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\right\|_1 \left\|\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}} \boldsymbol{Y}_{\pi(i)}\boldsymbol{Y}'_{\pi(i)}\left(\boldsymbol{A}_{\pi(i)} - \boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}\right)\right\|_\infty \\ &\leq \left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\right\|_1 \max_{l,l'\in\{1,2,\cdots,p\}} \left|\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}} \left(\boldsymbol{Y}_{\pi(i)}\boldsymbol{Y}'_{\pi(i)}\right)_{(l,.)} \left(\boldsymbol{A}_{\pi(i)} - \boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}\right)_{(.,l')}\right|, \end{aligned} \tag{C.77}$$

where $\left(\boldsymbol{Y}_{\pi(i)}\boldsymbol{Y}'_{\pi(i)}\right)_{(l,.)}$ represents the l-th row of $\boldsymbol{Y}_{\pi(i)}\boldsymbol{Y}'_{\pi(i)}$, and $\left(\boldsymbol{A}_{\pi(i)} - \boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}\right)_{(.,l')}$ represents the l'-th column of $\boldsymbol{A}_{\pi(i)} - \boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}$.

By Assumption C3, $\max\|\boldsymbol{A}_{\pi(i)}\|_\infty \leq M_A$. Then, by Equation (C.2), we obtain $\max\|\boldsymbol{A}_{\pi(i)} - \boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}\|_\infty \leq C_{10}M_A$. Note that

$$\mathbb{E}\left(\boldsymbol{x}_{(t-k)}\boldsymbol{x}'_{(t-k)} I\left(s < z_t \leq e\right)\right) \leq \mathbb{E}\left(\boldsymbol{x}_{(t-k)}\boldsymbol{x}'_{(t-k)}\right)$$

and that $\mathbb{E}|\boldsymbol{x}_t|^2$ is positive and bounded by Assumption C2. Combining Assumption C3 and Proposition 22 (Equation (C.4)), with probability $1-\delta_2$,

$$\begin{aligned} |H_2| &\leq \left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\right\|_1 \max_{l\in\{1,2,\cdots,p\}} \left|\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}} \left(\boldsymbol{Y}_{\pi(i)}\boldsymbol{Y}'_{\pi(i)}\right)_{(l,.)}\right| C_{10}d_n^*M_A \\ &\leq C_{10}d_n^*M_A \left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\right\|_1 \left\|\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}} \left(\boldsymbol{Y}_{\pi(i)}\boldsymbol{Y}'_{\pi(i)}\right)_{(l,.)}\right\|_\infty \\ &\leq C_{11}\left(\log\left(\max\left\{p^2K, n\right\}\right)\right)^{1/\varkappa_0} \sqrt{\left|\mathcal{T}_{(s,e)}\right|} d_n^* \left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\right\|_1. \end{aligned} \tag{C.78}$$

Since $\varkappa_0 < 1$, $\left(\log\left(\max\left\{p^2K, n\right\}\right)\right)^{1/\varkappa_0} \geq \left(\log\left(\max\left\{p^2K, n\right\}\right)\right)^{1/2}$. By Equa-

tions (C.75), (C.76) and (C.78), Equation (C.74) leads to

$$\begin{aligned}
&\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}} \left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right\|_2^2 \\
\leq& 2\Bigg(c_3\left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\right\|_1\sqrt{\log\left(\max\left\{p^2K,n\right\}\right)\left|\mathcal{T}_{(s,e)}\right|} \\
&+C_{11}\left(\log\left(\max\left\{p^2K,n\right\}\right)\right)^{1/\varkappa_0}\sqrt{\left|\mathcal{T}_{(s,e)}\right|}d_n^*\left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\right\|_1\Bigg) \\
&+\lambda\sqrt{\left|\mathcal{T}_{(s,e)}\right|}\left(\left\|\boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}\right\|_1-\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_1\right) \\
\leq& C_{12}\left(\log\left(\max\left\{p^2K,n\right\}\right)\right)^{1/\varkappa_0}d_n^*\sqrt{\left|\mathcal{T}_{(s,e)}\right|}\left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\right\|_1+\lambda\sqrt{\left|\mathcal{T}_{(s,e)}\right|}\left(\left\|\boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}\right\|_1-\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_1\right) \\
\leq& \frac{\lambda}{2}\sqrt{\left|\mathcal{T}_{(s,e)}\right|}\left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\right\|_1+\lambda\sqrt{\left|\mathcal{T}_{(s,e)}\right|}\left(\left\|\boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}\right\|_1-\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_1\right).
\end{aligned} \tag{C.79}$$

Next, using similar arguments as in the proof of Lemma 28, we can get

$$\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}-\boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}\right\|_{1,\mathcal{I}^c}=\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_{1,\mathcal{I}^c}\leq 3\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}-\boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}\right\|_{1,\mathcal{I}}. \tag{C.80}$$

In addition,

$$\begin{aligned}
\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}-\boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}\right\|_{1,\mathcal{I}} &\leq \left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}-\boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}\right\|_1 \\
&\leq 4\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}-\boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}\right\|_{1,\mathcal{I}} \\
&\leq 4\sqrt{d_n^*}\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}-\boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}\right\|_{2,\mathcal{I}}.
\end{aligned} \tag{C.81}$$

By Lemma 27, with $1-\delta_2$,

$$\begin{aligned}
\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left(\left(\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}-\boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}\right)\boldsymbol{Y}_{\pi(i)}\right)^2 \geq& C_6\left|\mathcal{T}_{(s,e)}\right|\left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\right\|_2^2 \\
&-c_3\left(\log\left(\max\left\{p^2K,n\right\}\right)\right)^{1/\varkappa_0}\sqrt{\left|\mathcal{T}_{(s,e)}\right|}\left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\right\|_1^2.
\end{aligned} \tag{C.82}$$

Then, by Equation (C.81),

$$\begin{aligned}\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left(\left(\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}-\boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}\right)\boldsymbol{Y}_{\pi(i)}\right)^2 &\geq C_6\left|\mathcal{T}_{(s,e)}\right|\left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\right\|_2^2\\ &\quad-c_3\left(\log\left(\max\left\{p^2K,n\right\}\right)\right)^{1/\varkappa_0}\sqrt{\left|\mathcal{T}_{(s,e)}\right|}\left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\right\|_1^2\\ &\geq\left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\right\|_2^2\Bigg(C_6\left|\mathcal{T}_{(s,e)}\right|\\ &\quad-C_{13}\left(\log\left(\max\left\{p^2K,n\right\}\right)\right)^{1/\varkappa_0}\sqrt{\left|\mathcal{T}_{(s,e)}\right|}d_n^*\Bigg).\end{aligned}\tag{C.83}$$

Since $\left|\mathcal{T}_{(s,e)}\right| \geq C\left(\log\left(\max\left\{p^2K,n\right\}\right)\right)^{2/\varkappa_0} d_n^{*2}$, then $\sqrt{C}\left(\log\left(\max\left\{p^2K,n\right\}\right)\right)^{1/\varkappa_0} d_n^*\sqrt{\left|\mathcal{T}_{(s,e)}\right|} \leq \left|\mathcal{T}_{(s,e)}\right|$. Recalling that C_{13} depends on the constant c_3 chosen in Proposition 22, we can choose c_3 to be small enough so that $C_{13}/\sqrt{C} < C_6$. Thus, $\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left(\left(\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}-\boldsymbol{A}_{\mathcal{T}_{(s,e)}}\right)\boldsymbol{Y}_{\pi(i)}\right)^2 \geq C_{14}\left|\mathcal{T}_{(s,e)}\right|\left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\right\|_2^2$ for $C_{14}>0$.

Then, we can rewrite Equation (C.79) to obtain

$$\begin{aligned}C_{14}\left|\mathcal{T}_{(s,e)}\right|\left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\right\|_2^2 &\leq \frac{\lambda}{2}\sqrt{\left|\mathcal{T}_{(s,e)}\right|}\left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\right\|_1+\lambda\sqrt{\left|\mathcal{T}_{(s,e)}\right|}\left(\left\|\boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}\right\|_1-\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_1\right)\\ &\leq C_{15}\lambda\sqrt{\left|\mathcal{T}_{(s,e)}\right|}\left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\right\|_{1,\mathcal{I}}\\ &\leq C_{16}\lambda\sqrt{\left|\mathcal{T}_{(s,e)}\right|}\left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\right\|_2\sqrt{d_n^*},\end{aligned}\tag{C.84}$$

where the second and the third inequalities follow from Equation (C.81) and the triangle inequality, respectively.

Finally, by Equations (C.81) and (C.84), we get

$$\begin{aligned}\left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\right\|_2 &\leq C_{17}\lambda\sqrt{d_n^*}/\sqrt{\left|\mathcal{T}_{(s,e)}\right|},\\ \left\|\boldsymbol{B}_{\mathcal{T}_{(s,e)}}\right\|_1 &\leq C_{17}\lambda d_n^*/\sqrt{\left|\mathcal{T}_{(s,e)}\right|}.\end{aligned}\tag{C.85}$$

Lemma 31. *Suppose Assumptions C1 to C4 hold and that the interval $(s,e]$ contains only one threshold r_j. For $\left|\mathcal{T}_{(s,e)}\right| \geq C\left(\log\left(\max\left\{p^2K,n\right\}\right)\right)^{2/\varkappa_0} d_n^{*2}$ and $\lambda =$*

$$2C_{12}\left(\log\left(\max\left\{p^2K,n\right\}\right)\right)^{1/\varkappa_0}d_n^*,$$

$$\mathbb{P}\left(\left|\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left\|\boldsymbol{x}_{\pi(i)}-\boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right\|_2^2-\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left\|\boldsymbol{x}_{\pi(i)}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right\|_2^2\right|\leq C_{18}\lambda^2d_n^*\right)$$
$$=(1-\delta_1)(1-\delta_2).$$
(C.86)

Proof of Lemma 31: Rearranging Equation (4.3), we can obtain

$$\begin{aligned}&\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left\|\boldsymbol{x}_{\pi(i)}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right\|_2^2-\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left\|\boldsymbol{x}_{\pi(i)}-\boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right\|_2^2\\&\leq\lambda\sqrt{|\mathcal{T}_{(s,e)}|}\left(\left\|\boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}\right\|_1-\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_1\right)\\&\leq\lambda\sqrt{|\mathcal{T}_{(s,e)}|}\left\|\boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_1.\end{aligned}$$
(C.87)

Then, by Equation (C.87), we obtain

$$\begin{aligned}&\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left\|\boldsymbol{x}_{\pi(i)}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right\|_2^2-\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left\|\boldsymbol{x}_{\pi(i)}-\boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right\|_2^2\\&\leq\lambda\sqrt{|\mathcal{T}_{(s,e)}|}\left\|\boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_1\\&\leq\lambda\sqrt{|\mathcal{T}_{(s,e)}|}C_9\lambda d_n^*/\sqrt{|\mathcal{T}_{(s,e)}|}\\&\leq C_9\lambda^2d_n^*.\end{aligned}$$
(C.88)

The second inequality follows from Lemma 30.

In addition, we rearrange Equation (4.3) again and get

$$
\begin{aligned}
&\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left\|\boldsymbol{x}_{\pi(i)}-\boldsymbol{A}^{*}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right\|_2^2-\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left\|\boldsymbol{x}_{\pi(i)}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right\|_2^2\\
&=-\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}-\boldsymbol{A}^{*}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right\|_2^2\\
&\quad+2\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left(\boldsymbol{x}_{\pi(i)}-\boldsymbol{A}^{*}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right)'\left(\boldsymbol{A}^{*}_{\mathcal{T}_{(s,e)}}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right)\boldsymbol{Y}_{\pi(i)}\\
&\leq 2\left\|\boldsymbol{A}^{*}_{\mathcal{T}_{(s,e)}}-\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\right\|_1\left\|\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\boldsymbol{\epsilon}_{\pi(i)}\boldsymbol{Y}'_{\pi(i)}\right\|_\infty\\
&\leq 2C_9\lambda d_n^{*}/\sqrt{\left|\mathcal{T}_{(s,e)}\right|}c_1\frac{\left|\mathcal{T}_{(s,e)}\right|\sqrt{\log\left(\max\left\{p^2K,n\right\}\right)}}{\sqrt{\left|\mathcal{T}_{(s,e)}\right|}}\\
&\leq C_{19}\lambda d_n^{*}\sqrt{\log\left(\max\left\{p^2K,n\right\}\right)}\\
&\leq C_{19}\lambda^2 d_n^{*},
\end{aligned}
\tag{C.89}
$$

where C_{19} is a positive constant. The second inequality holds with high probability by Proposition 22 (Equation (C.3)) and Lemma 30; the last inequality holds, since $\lambda \geq c_1\sqrt{\log\left(\max\left\{p^2K,n\right\}\right)}$. Finally, combining Equations (C.88) and (C.89), Equation (C.86) holds.

Lemma 32. *Suppose Assumptions C1 to C4 hold and that the interval $(s,e]\subseteq(r_{j-1},r_j]$. For $\left|\mathcal{T}_{(s,e)}\right|\geq C\left(\log\left(\max\left\{p^2K,n\right\}\right)\right)^{2/\varkappa_0}d_n^{*2}$ and $\lambda\geq C_3\sqrt{\log\left(\max\left\{p^2K,n\right\}\right)}$, there exists a positive constant C_{20} such that, with probability $(1-\delta_1)(1-\delta_2)$,*

$$
\left|L^{*}\left(\mathcal{T}_{(s,e)}\right)-L\left(\mathcal{T}_{(s,e)}\right)\right|\leq C_{20}d_n^{*}\lambda^2, \tag{C.90}
$$

where δ_1 and δ_2 are denoted the same as in Proposition 22.

Proof of Lemma 32: We proof this lemma by considering two cases: $\left|\mathcal{T}_{(s,e)}\right|<\omega$ and $\left|\mathcal{T}_{(s,e)}\right|\geq\omega$. By Equation (4.2), we have $L^{*}\left(\mathcal{T}_{(s,e)}\right)=L\left(\mathcal{T}_{(s,e)}\right)=0$ if $\left|\mathcal{T}_{(s,e)}\right|<\omega$, which gives Equation (C.90).

When $\left|\mathcal{T}_{(s,e)}\right| \geq \omega$,

$$\begin{aligned}
&\left|L^*\left(\mathcal{T}_{(s,e)}\right) - L\left(\mathcal{T}_{(s,e)}\right)\right| \\
\leq &\left| \sum_{z_{\pi(i)} \in \mathcal{T}_{(s,e)}} \left\| \boldsymbol{x}_{\pi(i)} - \boldsymbol{A}^*_{\mathcal{T}_{(s,e)}} \boldsymbol{Y}_{\pi(i)} \right\|_2^2 - \sum_{z_{\pi(i)} \in \mathcal{T}_{(s,e)}} \left\| \boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} \boldsymbol{Y}_{\pi(i)} \right\|_2^2 \right| \\
= &\left| \sum_{z_{\pi(i)} \in \mathcal{T}_{(s,e)}} \left\| \boldsymbol{x}_{\pi(i)} - \boldsymbol{A}_{\mathcal{T}_{(s,e)}} \boldsymbol{Y}_{\pi(i)} \right\|_2^2 - \sum_{z_{\pi(i)} \in \mathcal{T}_{(s,e)}} \left\| \boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}} \boldsymbol{Y}_{\pi(i)} \right\|_2^2 \right| \\
\leq &C_4 \lambda^2 d_n^*,
\end{aligned} \tag{C.91}$$

where the last inequality follows from Lemma 29.

Lemma 33. *Suppose Assumptions C1 to C4 hold and that the interval $(s,e] \subseteq (r_{j-1}, r_j]$. Set $\lambda = 2C_{12}\left(\log\left(\max\left\{p^2K, n\right\}\right)\right)^{1/\varkappa_0} d_n^*$ and denote δ_4 the same as in Lemma 25. For any interval $(s',e']$ that $(s,e] \subseteq (s',e']$, with probability $1-\delta_4$,*

$$L^*\left(\mathcal{T}_{(s,e)}\right) - \sum_{z_{\pi(i)} \in \mathcal{T}_{(s,e)}} \left(\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s',e')}} \boldsymbol{Y}_{\pi(i)} \right)^2 \leq C_{21} d_n^* \lambda^2, \tag{C.92}$$

where C_{21} is a positive constant.

Proof of Lemma 33: If $\left|\mathcal{T}_{(s,e)}\right| < \omega$, then

$$\begin{aligned}
&L^*\left(\mathcal{T}_{(s,e)}\right) - \sum_{z_{\pi(i)} \in \mathcal{T}_{(s,e)}} \left(\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s',e')}} \boldsymbol{Y}_{\pi(i)} \right)^2 \\
= &- \sum_{z_{\pi(i)} \in \mathcal{T}_{(s,e)}} \left(\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s',e')}} \boldsymbol{Y}_{\pi(i)} \right)^2 \leq C_{21} d_n^* \lambda^2.
\end{aligned}$$

For $\left|\mathcal{T}_{(s,e)}\right| \geq \omega$, then $\left|\mathcal{T}_{(s',e')}\right| \geq \omega$, and

$$
\begin{aligned}
& L^*\left(\mathcal{T}_{(s,e)}\right) - \sum_{z_{\pi(i)} \in \mathcal{T}_{(s,e)}} \left(\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s',e')}} \boldsymbol{Y}_{\pi(i)}\right)^2 \\
=& L^*\left(\mathcal{T}_{(s,e)}\right) - \sum_{z_{\pi(i)} \in \mathcal{T}_{(s,e)}} \left(\boldsymbol{x}_{\pi(i)} - \boldsymbol{A}_{\mathcal{T}_{(s,e)}} \boldsymbol{Y}_{\pi(i)} + \boldsymbol{A}_{\mathcal{T}_{(s,e)}} \boldsymbol{Y}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s',e')}} \boldsymbol{Y}_{\pi(i)}\right)^2 \\
=& 2 \sum_{z_{\pi(i)} \in \mathcal{T}_{(s,e)}} \left(\boldsymbol{x}_{\pi(i)} - \boldsymbol{A}_{\mathcal{T}_{(s,e)}} \boldsymbol{Y}_{\pi(i)}\right)' \left(\hat{\boldsymbol{A}}_{\mathcal{T}_{(s',e')}} - \boldsymbol{A}_{\mathcal{T}_{(s,e)}}\right) \boldsymbol{Y}_{\pi(i)} \\
& - \sum_{z_{\pi(i)} \in \mathcal{T}_{(s,e)}} \left\| \hat{\boldsymbol{A}}_{\mathcal{T}_{(s',e')}} \boldsymbol{Y}_{\pi(i)} - \boldsymbol{A}_{\mathcal{T}_{(s,e)}} \boldsymbol{Y}_{\pi(i)} \right\|_2^2 \\
\leq & 2 \left\| \boldsymbol{A}_{\mathcal{T}_{(s,e)}} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s',e')}} \right\|_1 \left\| \sum_{z_{\pi(i)} \in \mathcal{T}_{(s,e)}} \boldsymbol{\epsilon}_{\pi(i)} \boldsymbol{Y}'_{\pi(i)} \right\|_\infty - \sum_{z_{\pi(i)} \in \mathcal{T}_{(s,e)}} \left\| \hat{\boldsymbol{A}}_{\mathcal{T}_{(s',e')}} \boldsymbol{Y}_{\pi(i)} - \boldsymbol{A}_{\mathcal{T}_{(s,e)}} \boldsymbol{Y}_{\pi(i)} \right\|_2^2 \\
\leq & C_1 \left\| \boldsymbol{A}_{\mathcal{T}_{(s,e)}} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s',e')}} \right\|_1 \left|\mathcal{T}_{(s,e)}\right| \sqrt{\frac{\log\left(\max\left\{p^2 K, n\right\}\right)}{\left|\mathcal{T}_{(s,e)}\right|}} \\
& - \sum_{z_{\pi(i)} \in \mathcal{T}_{(s,e)}} \left\| \hat{\boldsymbol{A}}_{\mathcal{T}_{(s',e')}} \boldsymbol{Y}_{\pi(i)} - \boldsymbol{A}_{\mathcal{T}_{(s,e)}} \boldsymbol{Y}_{\pi(i)} \right\|_2^2 \\
\leq & C_1 \sqrt{\left|\mathcal{T}_{(s,e)}\right| \log\left(\max\left\{p^2 K, n\right\}\right)} \left\| \boldsymbol{A}_{\mathcal{T}_{(s,e)}} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s',e')}} \right\|_1 \\
& - \sum_{z_{\pi(i)} \in \mathcal{T}_{(s,e)}} \left\| \hat{\boldsymbol{A}}_{\mathcal{T}_{(s',e')}} \boldsymbol{Y}_{\pi(i)} - \boldsymbol{A}_{\mathcal{T}_{(s,e)}} \boldsymbol{Y}_{\pi(i)} \right\|_2^2,
\end{aligned}
\tag{C.93}
$$

where the first inequality follows from the Cauchy–Schwarz inequality, and the second inequality follows from Proposition 22 (Equation (C.3)).

Using similar arguments as in the proof of Lemma 27, with probability $1 - \delta_1 - \delta_2$, we

get

$$
\begin{aligned}
&\sum_{z_{\pi(i)} \in \mathcal{T}_{(s,e)}} \left(\left(\hat{\boldsymbol{A}}_{\mathcal{T}_{(s',e')}} - \boldsymbol{A}_{\mathcal{T}_{(s,e)}} \right) \boldsymbol{Y}_{\pi(i)} \right)^2 \\
&\geq C_6 \left| \mathcal{T}_{(s,e)} \right| \left\| \hat{\boldsymbol{A}}_{\mathcal{T}_{(s',e')}} - \boldsymbol{A}_{\mathcal{T}_{(s,e)}} \right\|_2^2 \\
&\quad - C_7 \left(\log \left(\max \left\{ p^2 K, n \right\} \right) \right)^{1/\varkappa_0} \sqrt{\left| \mathcal{T}_{(s,e)} \right|} \left\| \hat{\boldsymbol{A}}_{\mathcal{T}_{(s',e')}} - \boldsymbol{A}_{\mathcal{T}_{(s,e)}} \right\|_1^2 \\
&= C_6 \left| \mathcal{T}_{(s,e)} \right| \left\| \hat{\boldsymbol{A}}_{\mathcal{T}_{(s',e')}} - \boldsymbol{A}_{\mathcal{T}_{(s,e)}} \right\|_2^2 \\
&\quad - C_7 \left(\log \left(\max \left\{ p^2 K, n \right\} \right) \right)^{1/\varkappa_0} \sqrt{\left| \mathcal{T}_{(s,e)} \right|} \left\| \hat{\boldsymbol{A}}_{\mathcal{T}_{(s',e')}} - \boldsymbol{A}_{\mathcal{T}_{(s,e)}} \right\|_{1,\mathcal{I}}^2 \\
&\quad - C_7 \left(\log \left(\max \left\{ p^2 K, n \right\} \right) \right)^{1/\varkappa_0} \sqrt{\left| \mathcal{T}_{(s,e)} \right|} \left\| \hat{\boldsymbol{A}}_{\mathcal{T}_{(s',e')}} - \boldsymbol{A}_{\mathcal{T}_{(s,e)}} \right\|_{1,\mathcal{I}^c}^2 \\
&\geq C_6 \left| \mathcal{T}_{(s,e)} \right| \left\| \hat{\boldsymbol{A}}_{\mathcal{T}_{(s',e')}} - \boldsymbol{A}_{\mathcal{T}_{(s,e)}} \right\|_2^2 \\
&\quad - C_7 \left(\log \left(\max \left\{ p^2 K, n \right\} \right) \right)^{1/\varkappa_0} \sqrt{\left| \mathcal{T}_{(s,e)} \right|} d_n^* \left\| \hat{\boldsymbol{A}}_{\mathcal{T}_{(s',e')}} - \boldsymbol{A}_{\mathcal{T}_{(s,e)}} \right\|_2^2 \\
&\quad - C_7 \left(\log \left(\max \left\{ p^2 K, n \right\} \right) \right)^{1/\varkappa_0} \sqrt{\left| \mathcal{T}_{(s,e)} \right|} \left\| \hat{\boldsymbol{A}}_{\mathcal{T}_{(s',e')}} \right\|_{1,\mathcal{I}^c}^2 \\
&\geq C_{24} \left| \mathcal{T}_{(s,e)} \right| \left\| \hat{\boldsymbol{A}}_{\mathcal{T}_{(s',e')}} - \boldsymbol{A}_{\mathcal{T}_{(s,e)}} \right\|_2^2 \\
&\quad - C_8 C_7 \left(\log \left(\max \left\{ p^2 K, n \right\} \right) \right)^{1/\varkappa_0} \sqrt{\left| \mathcal{T}_{(s,e)} \right|} \lambda^2 d_n^* / \left| \mathcal{T}_{(s',e')} \right| \\
&\geq C_{24} \left| \mathcal{T}_{(s,e)} \right| \left\| \hat{\boldsymbol{A}}_{\mathcal{T}_{(s',e')}} - \boldsymbol{A}_{\mathcal{T}_{(s,e)}} \right\|_2^2 - C_{22} \lambda^2,
\end{aligned}
\tag{C.94}
$$

where the second inequality follows from the Cauchy–Schwarz inequality (using the similar arguments as in Equation (C.60)), and the third inequality is due to the fact that $\left| \mathcal{T}_{(s,e)} \right| \geq \omega \geq C_\omega d_n^{*3} \log \left(p^2 K \right)^{2/\varkappa_0}$ and Lemma 30. The last inequality is due to the fact that $\left| \mathcal{T}_{(s',e')} \right| \geq \left| \mathcal{T}_{(s,e)} \right| \geq \omega \geq C_\omega d_n^{*3} \log \left(p^2 K \right)^{2/\varkappa_0}$ and the choice of λ, that is $\lambda = 2 C_{12} \left(\log \left(\max \left\{ p^2 K, n \right\} \right) \right)^{1/\varkappa_0} d_n^*$.

Then, combining Equations (C.93) and (C.94), we get

$$
\begin{aligned}
&L^*\left(\mathcal{T}_{(s,e)}\right) - \sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}} \left(\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s',e')}}\boldsymbol{Y}_{\pi(i)}\right)^2 \\
\leq& C_1\sqrt{\left|\mathcal{T}_{(s,e)}\right|\log\left(\max\left\{p^2K,n\right\}\right)}\left\|\boldsymbol{A}_{\mathcal{T}_{(s,e)}} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s',e')}}\right\|_1 - C_{24}\left|\mathcal{T}_{(s,e)}\right|\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s',e')}} - \boldsymbol{A}_{\mathcal{T}_{(s,e)}}\right\|_2^2 + C_{22}\lambda^2 \\
\leq& C_1\lambda\sqrt{\left|\mathcal{T}_{(s,e)}\right|}\left\|\boldsymbol{A}_{\mathcal{T}_{(s,e)}} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s',e')}}\right\|_1 - C_{24}\left|\mathcal{T}_{(s,e)}\right|\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s',e')}} - \boldsymbol{A}_{\mathcal{T}_{(s,e)}}\right\|_2^2 + C_{22}\lambda^2 \\
\leq& C_{23}\lambda\sqrt{\left|\mathcal{T}_{(s,e)}\right|}\left(\sqrt{d_n^*}\left\|\boldsymbol{A}_{\mathcal{T}_{(s,e)}} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s',e')}}\right\|_{2,\mathcal{I}} + \left\|\boldsymbol{A}_{\mathcal{T}_{(s,e)}} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s',e')}}\right\|_{1,\mathcal{I}^c}\right) \\
&- C_{24}\left|\mathcal{T}_{(s,e)}\right|\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s',e')}} - \boldsymbol{A}_{\mathcal{T}_{(s,e)}}\right\|_2^2 + C_{22}\lambda^2 \\
\leq& C_{23}\lambda\sqrt{\left|\mathcal{T}_{(s,e)}\right|}\sqrt{d_n^*}\left\|\boldsymbol{A}_{\mathcal{T}_{(s,e)}} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s',e')}}\right\|_{2,\mathcal{I}} + C_{23}C_8\lambda\sqrt{\left|\mathcal{T}_{(s,e)}\right|}\lambda\sqrt{d_n^*}/\sqrt{\left|\mathcal{T}_{(s',e')}\right|} \\
&- C_{24}\left|\mathcal{T}_{(s,e)}\right|\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s',e')}} - \boldsymbol{A}_{\mathcal{T}_{(s,e)}}\right\|_2^2 + C_{22}\lambda^2 \\
\leq& C_{24}\left|\mathcal{T}_{(s,e)}\right|\left\|\boldsymbol{A}_{\mathcal{T}_{(s,e)}} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s',e')}}\right\|_2^2 + C_{23}^2/\left(2C_{24}\right)\lambda^2 d_n^* + C_{23}C_8\lambda^2\sqrt{d_n^*} \\
&- C_{24}\left|\mathcal{T}_{(s,e)}\right|\left\|\hat{\boldsymbol{A}}_{\mathcal{T}_{(s',e')}} - \boldsymbol{A}_{\mathcal{T}_{(s,e)}}\right\|_2^2 + C_{22}\lambda^2 \\
\leq& C_{25}\lambda^2 d_n^*,
\end{aligned}
\tag{C.95}
$$

where the second inequality holds by the choice of λ, the third inequality holds by Cauchy–Schwarz inequality (using the similar arguments as in Equation (C.60)), the forth inequality follows from Lemma 30, and the forth inequality follows from Hölder's inequality.

Lemma 34. *Suppose Assumptions C1 to C4 hold and that the interval $(s,e] \in \hat{\mathcal{P}}$ contains J true thresholds for $J \geq 1$. Let $\lambda = 2C_{12}\left(\log\left(\max\left\{p^2K,n\right\}\right)\right)^{1/\varkappa_0} d_n^*$, $r'_0 = s, r'_j = r_j$, and $r'_{J+1} = e$ for $j = 1,2,\ldots,J$. Then, with probability $1 - n\delta_4$,*

$$
L\left(\mathcal{T}_{(s,e)}\right) \geq \sum_{j=1}^{J+1} L\left(\mathcal{T}_{(r'_{j-1},r'_j)}\right) - C_{26}\left(J+1\right)d_n^*\lambda^2, \tag{C.96}
$$

where C_{26} is a positive constant and δ_4 is defined in Lemma 25.

Proof of Lemma 34: We prove this lemma by considering two cases: $\left|\mathcal{T}_{(s,e)}\right| \geq \omega$ and $\left|\mathcal{T}_{(s,e)}\right| < \omega$.

For the case $\left|\mathcal{T}_{(s,e)}\right| \geq \omega$, we take the union bound of Lemma 32,

$$\mathbb{P}\left(\sum_{j=1}^{J+1}\left\{\left|L^*\left(\mathcal{T}_{(r'_{j-1},r'_j)}\right) - L\left(\mathcal{T}_{(r'_{j-1},r'_j)}\right)\right|\right\} > C_{20}\left(J+1\right)d_n^*\lambda^2\right) \leq n\delta_4, \tag{C.97}$$

which implies

$$\sum_{j=1}^{J+1} L\left(\mathcal{T}_{(r'_{j-1},r'_j)}\right) - \sum_{j=1}^{J+1} L^*\left(\mathcal{T}_{(r'_{j-1},r'_j)}\right) \leq C_{20}\left(J+1\right)d_n^*\lambda^2 \tag{C.98}$$

with probability $1 - n\delta_4$.

Taking the union bound of Lemma 33, we can get

$$\mathbb{P}\left(\sum_{j=1}^{J+1}\left(L^*\left(\mathcal{T}_{(r'_{j-1},r'_j)}\right) - \sum_{z_{\pi(i)}\in\mathcal{T}_{(r'_{j-1},r'_j)}}\left(\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right)^2\right) > C_{21}\left(J+1\right)d_n^*\lambda^2\right) \leq n\delta_4, \tag{C.99}$$

which implies, with probability $1 - n\delta_4$,

$$\begin{aligned}\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left(\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right)^2 &\geq \sum_{j=1}^{J+1}\sum_{z_{\pi(i)}\in\mathcal{T}_{(r'_{j-1},r'_j)}}\left(\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right)^2 \\ &\geq \sum_{j=1}^{J+1} L^*\left(\mathcal{T}_{(r'_{j-1},r'_j)}\right) - C_{21}\left(J+1\right)d_n^*\lambda^2.\end{aligned} \tag{C.100}$$

Combining Equations (C.98) and (C.100), we can obtain

$$\sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left(\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}\boldsymbol{Y}_{\pi(i)}\right)^2 \geq \sum_{j=1}^{J+1} L\left(\mathcal{T}_{(r'_{j-1},r'_j)}\right) - \left(C_{20}+C_{21}\right)\left(J+1\right)d_n^*\lambda^2. \tag{C.101}$$

For the case $\left|\mathcal{T}_{(s,e)}\right| < \omega$, we have $L\left(\mathcal{T}_{(s,e)}\right) = \sum_{j=1}^{J+1} L\left(\mathcal{T}_{(r'_{j-1},r'_j)}\right) = 0$ by Equation (4.2), which proves Equation (C.96).

Appendix 2: Proof of Main Results

Proof of Proposition 9: This proposition can be proved by taking the union bound over all possible choice of s, e for Lemma 23, Lemma 24, Lemma 25, and Lemma 26.

For case (a), Equation (C.5) holds, since $\hat{\mathcal{P}}$ is a minimizer of Equation (4.4). Then, we can apply Lemma 23 and get

$$\begin{aligned}
&\mathbb{P}\left(\max_{(s,e]\in\hat{\mathcal{P}}} \min\left\{\left|\mathcal{T}_{(s,r)}\right|, \left|\mathcal{T}_{(r,e)}\right|\right\} > C_0\left(\frac{\lambda^2 d_n^* + \omega}{v^2}\right)\right) \\
\leq & n^2\mathbb{P}\left(\min\left\{\left|\mathcal{T}_{(s,r)}\right|, \left|\mathcal{T}_{(r,e)}\right|\right\} > C_0\left(\frac{\lambda^2 d_n^* + \omega}{v^2}\right)\right) \\
\leq & n^2\delta_1 + n^2\delta_2.
\end{aligned} \tag{C.102}$$

For case (b), Equation (C.20) holds, since $\hat{\mathcal{P}}$ is a minimizer of Equation (4.4). Similarly, we use Lemma 24 and obtain

$$\mathbb{P}\left(\max_{(s,e]\in\hat{\mathcal{P}}} \max\left\{\left|\mathcal{T}_{(s,r_1)}\right|, \left|\mathcal{T}_{(r_2,e)}\right|\right\} > C_0\left(\frac{d_n^*\lambda^2 + \omega}{v^2}\right)\right) \leq n^2\delta_1 + n^2\delta_2.$$

For case (c), we prove by contradiction. Assume that there are no thresholds in any two estimated consecutive regimes $\mathcal{T}_{(\hat{r}_{j-1},\hat{r}_j)}$ and $\mathcal{T}_{(\hat{r}_j,\hat{r}_{j+1})}$. By Lemma 25 we can obtain

$$L\left(\mathcal{T}_{(s,e)}\right) < \min_{r'\in(s,e]}\left\{L\left(\mathcal{T}_{(s,r')}\right) + L\left(\mathcal{T}_{(r',e)}\right)\right\} + \omega, \tag{C.103}$$

for a fixed s, e with probability $1-\delta_4$. Fix a $\hat{r}_j$ and the union of two estimated consecutive regimes $\mathcal{T}_{(\hat{r}_{j-1},\hat{r}_j)} \cup \mathcal{T}_{(\hat{r}_j,\hat{r}_{j+1})}$, then

$$\begin{aligned}
L\left(\mathcal{T}_{(\hat{r}_{j-1},\hat{r}_{j+1})}\right) < & \min_{r'\in(\hat{r}_{j-1},\hat{r}_j]}\left\{L\left(\mathcal{T}_{(\hat{r}_{j-1},r')}\right) + L\left(\mathcal{T}_{(r',\hat{r}_{j+1})}\right)\right\} + \omega \\
\leq & L\left(\mathcal{T}_{(\hat{r}_{j-1},\hat{r}_j)}\right) + L\left(\mathcal{T}_{(\hat{r}_j,\hat{r}_{j+1})}\right) + \omega,
\end{aligned} \tag{C.104}$$

which contradicts the fact that Equation (4.4) can be minimized by $\hat{\mathcal{P}}$. Thus, case (c) is proved.

For case (d), we assume that there are J thresholds in the interval $(s, e]$ for $J \geq 3$. Let

$r'_0 = s$, $r'_j = r_j$, and $r'_{J+1} = e$ for $j = 1, 2, ..., J$. If $J \geq 3$, we take the uion bound of Lemma 26 and then,

$$\mathbb{P}\left(\bigcup_{(s,e]\in\hat{\mathcal{P}}}\left\{L\left(\mathcal{T}_{(s,e)}\right) \leq \sum_{j=1}^{J+1} L\left(\mathcal{T}_{(r'_{j-1},r'_j)}\right) + J\omega\right\}\right) \leq n^2\delta_4, \tag{C.105}$$

which contradict the fact that Equation (4.4) can be minimized by $\hat{\mathcal{P}}$. Thus, case (d) is proved.

Proof of Proposition 10: Let $L^*\left(\mathcal{T}_{(s,e)}\right)$ be the population counterpart of $L\left(\mathcal{T}_{(s,e)}\right)$ by replacing the coefficient matrix estimator $\hat{\boldsymbol{A}}_{\mathcal{T}_{(s,e)}}$ with its population counterpart $\boldsymbol{A}^*_{\mathcal{T}_{(s,e)}}$. If $(s,e] \subseteq (r_{j-1}, r_j]$, and $(s,e] \subseteq (s',e']$, then we can take an union bound of Lemma 33, which gives

$$\mathbb{P}\left(\max_{(s,e],(s',e']}\left\{L^*\left(\mathcal{T}_{(s,e)}\right) - \sum_{z_{\pi(i)}\in\mathcal{T}_{(s,e)}}\left(\boldsymbol{x}_{\pi(i)} - \hat{\boldsymbol{A}}_{\mathcal{T}_{(s',e')}}\boldsymbol{Y}_{\pi(i)}\right)^2\right\} > C_{21}d_n^*\lambda^2\right) \leq n^4\delta_4. \tag{C.106}$$

By the union bound of Lemma 32, we get

$$\mathbb{P}\left(\max_{(s,e]}\left\{\left|L^*\left(\mathcal{T}_{(s,e)}\right) - L\left(\mathcal{T}_{(s,e)}\right)\right|\right\} > C_{20}d_n^*\lambda^2\right) \leq n^2\delta_4. \tag{C.107}$$

Let $r'_{j'}$ be the j'-th value from the sorted set $\{r_1, r_2, \ldots, r_{m_0}, \hat{r}_0, \hat{r}_1, \hat{r}_2, \ldots, r_{\hat{m}+1}\}$, where $r_0 = \hat{r}_0 = -\infty$ and $r_{m_0+1} = \hat{r}_{\hat{m}+1} = \infty$.

By Equation (C.107), we can obtain

$$
\begin{aligned}
&\sum_{j=1}^{m_0+1} L^*\left(\mathcal{T}_{(r_{j-1},r_j)}\right) + m_0\omega \\
\geq &\sum_{j=1}^{m_0+1} \left(L\left(\mathcal{T}_{(r_{j-1},r_j)}\right) - C_{20} d_n^* \lambda^2\right) + m_0\omega \\
\geq &\sum_{j=1}^{m_0+1} L\left(\mathcal{T}_{(r_{j-1},r_j)}\right) - C_{20}\left(m_0+1\right) d_n^* \lambda^2 + m_0\omega \\
\geq &\sum_{j=1}^{\hat{m}+1} L\left(\mathcal{T}_{(\hat{r}_{j-1},\hat{r}_j)}\right) - C_{20}\left(m_0+1\right) d_n^* \lambda^2 + \hat{m}\omega \\
\geq &\sum_{j=1}^{\hat{m}+m_0+1} L\left(\mathcal{T}_{\left(r'_{j'-1},r'_{j'}\right)}\right) - C_{26}\left(\hat{m}+1+m_0\right) d_n^* \lambda^2 - C_{20}\left(m_0+1\right) d_n^* \lambda^2 + \hat{m}\omega \\
\geq &\sum_{j=1}^{\hat{m}+m_0+1} L\left(\mathcal{T}_{\left(r'_{j'-1},r'_{j'}\right)}\right) - C_{27}\left(\hat{m}+m_0+1\right) d_n^* \lambda^2 + \hat{m}\omega,
\end{aligned}
\tag{C.108}
$$

where the first inequality is due to Equation (C.107); the third inequality holds, since Equation (4.4) is minimized by $\hat{\mathcal{P}}$; the forth inequality follows from Lemma 34.

In addition,

$$
\begin{aligned}
\sum_{j=1}^{m_0+1} L^*\left(\mathcal{T}_{(r_{j-1},r_j)}\right) &\leq \sum_{j=1}^{\hat{m}+m_0+1} L^*\left(\mathcal{T}_{(r'_{j-1},r'_j)}\right) \\
&\leq \sum_{j=1}^{\hat{m}+m_0+1} L\left(\mathcal{T}_{(r'_{j-1},r'_j)}\right) + C_{20}\left(m_0+\hat{m}+1\right) d_n^* \lambda^2,
\end{aligned}
\tag{C.109}
$$

where the first inequality is due to the fact that $L^*(\cdot)$ is the counterpart of $L(\cdot)$ and the last inequality follows from Equation (C.107).

Then, we prove by contradiction. Assume that $\hat{m} \geq m_0 + 1$. Note that $\hat{m} = \left|\hat{\mathcal{P}}\right| - 1$.

Combining Equations (C.108) and (C.109), we get

$$\begin{aligned}-C_{27}\left(\hat{m}+m_0+1\right)d_n^*\lambda^2+\hat{m}\omega-m_0\omega\leq&\sum_{j=1}^{m_0+1}L^*\left(\mathcal{T}_{(r_{j-1},r_j)}\right)-\sum_{j=1}^{\hat{m}+m_0+1}L\left(\mathcal{T}_{(r'_{j-1},r'_j)}\right)\\ \leq&C_{20}\left(m_0+\hat{m}+1\right)d_n^*\lambda^2,\end{aligned}\tag{C.110}$$

which implies $C_{28}\left(m_0+\hat{m}+1\right)d_n^*\lambda^2\geq\left(\hat{m}-m_0\right)\omega$. Since we assume $\hat{m}\geq m_0+1$ and $\hat{m}\leq 2m_0$, then Equation (C.110) leads to $C_{28}\left(3m_0+1\right)d_n^*\lambda^2\geq\omega$, which contradicts the choice of ω. Thus, $\hat{m}=m_0$ with high probability.

Proof of Theorem 11: The proof of Theorem 11 is according to Proposition 9 and Proposition 10.

By part (d) in Proposition 9, we can only have at most two true thresholds in each estimated regime. For any given j-th estimated threshold $\hat{r}_j$, suppose the estimated regime $(\hat{r}_{j-1},\hat{r}_j]$ contains a true threshold r_a such that $|\hat{r}_{j-1}-r_a|<|\hat{r}_j-r_a|$. If the nearest estimated regime $(\hat{r}_{j-2},\hat{r}_{j-1}]$ contains a true threshold r_c such that $|\hat{r}_{j-1}-r_c|<|\hat{r}_{j-2}-r_c|$. By Assumption C5, z_t has positive density. Thus,

$$\begin{aligned}\left|\mathcal{T}_{(r_c,r_a)}\right|=&n\mathbb{P}\left(r_c<z_t\leq r_a\right)\geq c_e n\left|r_a-r_c\right|\geq c_e n\Delta_n\\ \geq&C'_\delta\left(\log\left(\max\left\{p^2K,n\right\}\right)\right)^{2/\varkappa_0+\xi}m_0d_n^{*3}/v^2,\end{aligned}\tag{C.111}$$

where $c_e,C'_\delta>0$ are constant, and ξ is a small positive constant. Since we can only have at most two true thresholds in each estimated regime, by part (a) and part (b) in Proposition 9, we have:

$$\begin{aligned}\left|\mathcal{T}_{(r_c,r_a)}\right|=&\left|\mathcal{T}_{(r_c,\hat{r}_{j-1})}\right|+\left|\mathcal{T}_{(\hat{r}_{j-1},r_a)}\right|\\ \leq&2C_0\left(\frac{d_n^*\lambda^2+\omega}{v^2}\right)\\ \leq&\frac{C'_0\left(m_0+1\right)d_n^{*3}\left(\log\left(\max\left\{p^2K,n\right\}\right)\right)^{2/\varkappa_0}}{v^2},\end{aligned}\tag{C.112}$$

which contradict the Equation (C.111), meaning for a given estimated threshold, we can

only have at most one true threshold that closes to it. Thus, we have $m_0 \leq \left|\hat{\mathcal{P}}\right| - 1$. By part (c) in Proposition 9, we have $m_0 \geq \left(\left|\hat{\mathcal{P}}\right| - 1\right)/2$, which implies $2m_0 \geq \left|\hat{\mathcal{P}}\right| - 1$. Then, by Proposition 10, we have $\mathbb{P}\left(\hat{m} = m_0\right) \to 1$. By part (a), (b) in Proposition 9, Equation (4.9) holds.

Appendix 3: Additional Simulation Results

Settings	Threshold(s)	Methods	Mean	Std	Selection Rate
Scenario 2 $\lvert r_1 - r_2 \rvert = 2.1$		dpTAR	3.90	0.67	0.69
	3.9	hdTAR	4.03	0.41	0.86
		TVAR	3.64	0.53	0.77
		dpTAR	5.96	0.48	0.69
	6	hdTAR	5.97	0.33	0.87
		TVAR	5.82	0.61	0.71
Scenario 2 $\lvert r_1 - r_2 \rvert = 2$		dpTAR	3.90	0.67	0.69
	4	hdTAR	4.16	0.41	0.83
		TVAR	3.82	0.56	0.76
		dpTAR	5.96	0.48	0.69
	6	hdTAR	6.04	0.40	0.91
		TVAR	5.75	0.62	0.72
Scenario 2 $\lvert r_1 - r_2 \rvert = 1.9$		dpTAR	4.07	0.63	0.60
	4.1	hdTAR	4.27	0.33	0.84
		TVAR	3.84	0.53	0.72
		dpTAR	6.03	0.44	0.60
	6	hdTAR	6.02	0.37	0.92
		TVAR	5.84	0.57	0.69
Scenario 2 $\lvert r_1 - r_2 \rvert = 1.8$		dpTAR	4.04	0.69	0.59
	4.2	hdTAR	4.34	0.25	0.84
		TVAR	3.92	0.54	0.75
		dpTAR	6.00	0.45	0.59
	6	hdTAR	6.02	0.33	0.89
		TVAR	5.85	0.58	0.68

Table C.1: Mean and standard deviation of estimated thresholds, the percentage of simulation runs where thresholds are correctly detected (selection rate) in Simulation Scenarios. If the estimated thresholds within one standard deviation of true threshold, we consider the estimated thresholds are correctly detected.

	Method	REE	SD(REE)	FPR	TPR
Scenario 2 $\lvert r_1 - r_2\rvert = 2.1$	dpTAR	0.48	0.09	0.22	0.74
	hdTAR	0.48	0.06	0.10	0.67
	TVAR	1.06	0.75	1.00	1.00
Scenario 2 $\lvert r_1 - r_2\rvert = 2$	dpTAR	0.48	0.09	0.22	0.74
	hdTAR	0.50	0.06	0.09	0.50
	TVAR	1.30	1.46	1.00	1.00
Scenario 2 $\lvert r_1 - r_2\rvert = 1.9$	dpTAR	0.47	0.10	0.22	0.74
	hdTAR	0.50	0.06	0.09	0.50
	TVAR	1.41	1.76	1.00	1.00
Scenario 2 $\lvert r_1 - r_2\rvert = 1.8$	dpTAR	0.47	0.09	0.23	0.74
	hdTAR	0.45	0.06	0.10	0.67
	TVAR	1.27	1.62	1.00	1.00

Table C.2: Results of parameter estimation for simulation scenarios. The table shows mean and standard deviation of relative estimation error (REE), true positive rate (TPR), and false positive rate (FPR) for estimated coefficients.

$\lvert r_1 - r_2\rvert = 2.1$	$\lvert r_1 - r_2\rvert = 2.0$	$\lvert r_1 - r_2\rvert = 1.9$	$\lvert r_1 - r_2\rvert = 1.8$
0.75	0.71	0.48	0.20

Table C.3: Results of selection rate for simulation Scenario 2. The table shows the rates of selecting z_t correctly.