Statistical Inference for Multiple Change Points and Implicit Network Structures

by

Zhibing He

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved June 2022 by the
Graduate Supervisory Committee:

Yunpeng Zhao, Co-Chair
Dan Cheng, Co-Chair
Hedibert Lopes
John Fricks
Ming-Hung Kao

ARIZONA STATE UNIVERSITY

August 2022

ABSTRACT

This dissertation contains two research projects: Multiple Change Point Detection in Linear Models and Statistical Inference for Implicit Network Structures.

In the first project, a new method to detect the number and locations of change points in piecewise linear models under stationary Gaussian noise is proposed. The method transforms the problem of detecting change points to the detection of local extrema by kernel smoothing and differentiating the data sequence. The change points are detected by computing the p-values for all local extrema using the derived peak height distributions of smooth Gaussian processes, and then applying the Benjamini-Hochberg procedure to identify significant local extrema. Theoretical results show that the method can guarantee asymptotic control of the False Discover Rate (FDR) and power consistency, as the length of the sequence, and the size of slope changes and jumps get large. In addition, compared to traditional methods for change point detection based on recursive segmentation, The proposed method tests the candidate local extrema only one time, achieving the smallest computational complexity. Numerical studies show that the properties on FDR control and power consistency are maintained in non-asymptotic cases.

In the second project, identifiability and estimation consistency under mild conditions in hub model are proved. Hub Model is a model-based approach, introduced by Zhao and Weko [76], to infer implicit network structuress from grouping behavior. The hub model assumes that each member of the group is brought together by a member of the group called the hub. This paper generalize the hub model by introducing a model component that allows hubless groups in which individual nodes spontaneously appear independent of any other individual. The new model bridges the gap between the hub model and the degenerate case of the mixture model – the Bernoulli product.

Furthermore, a penalized likelihood approach is proposed to estimate the set of hubs when it is unknown.

This work is dedicated to my wife, Dr. Yan Zhang, and my parents, without whose constant support this dissertation paper was not possible.

ACKNOWLEDGMENTS

I owe my gratitude to several people who have advised, supported, or inspired me during this work.

First and foremost, I would like to express my sincere gratitude to my co-advisors Dr. Yunpeng Zhao and Dr. Dan Cheng for the continuous support of my Ph.D study and related research. Their patience, yet willingness to challenge me to think about my work at a deeper level has been of great inspiration. Their guidance helped me in all the time of the research on network analysis and change point detection. I could not have imagined having better co-advisors for supporting me throughout my PhD study.

Besides my advisors, I would like to thank the rest of my dissertation committee: Dr. Hedibert Lopes, Dr. John Fricks and Dr. Ming-Hung Kao, for the insightful and valuable comments and encouragement, but also for the "hard" questions which inspired me to widen my research from various perspectives.

My sincere thanks also goes to Dr. Armin Schwartzman, Dr. Peter Bickel, Mr. Charles Weko and Dr. Jirui Wang who are the co-authors of the related research papers. Without their work and support, it would not be possible to conduct this research.

Last but not the least, I would like to thank the School of Mathematical and Statistical Sciences, Arizona State University, for providing me Teaching Assistant scholarship.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

The dissertation contains two research projects: Multiple Change Point Detection in Linear Models and Statistical Inference for Implicit Network Structures.

1.1 Multiple Change Point Detection in Linear Models

In this project, we consider a canonical univariate statistical model:

$$y(t) = \mu(t) + z(t), \quad t \in \mathbb{R}, \tag{1.1}$$

where $z(t)$ is correlated stationary Gaussian noise and $\mu(t)$ is a piecewise linear signal of the form

$$\mu(t) = c_j + k_j t, \quad t \in (v_{j-1}, v_j],$$

where $c_j, k_j \in \mathbb{R}$, $j = 1, 2, \ldots$ and $-\infty = v_0 < v_1 < v_2 < \cdots$. Assume the structures of $\mu(t)$ are different at neighboring v_j, i.e., $(c_j, k_j) \neq (c_{j+1}, k_{j+1})$, resulting in a continuous break or jump at v_j (see Fig 1.1). Such v_j is called a change point or structural break.

Change point detection is a fundamental and important problem in statistics and other related fields such as econometrics, genomics, climatology and medical imaging. It is broadly applied to different areas based on different types of signals. For example, the piecewise constant signals with jumps occur in the contexts of medical condition monitoring [46, 37, 14] and image analysis [52, 47]. The piecewise linear signals with continuous breaks are very common in climate change detection [65, 63] and human activity analysis [67, 53]. In addition, the piecewise linear signals with noncontinuous breaks are applied in the stock market monitoring [17, 24] and pitch recognition of sound [16, 49].

Based on different ways of linear structure changes, we define the following two types of structural breaks:

Definition 1. *A point v_j is called a Type* I *change point (structural break) if $c_j + k_j v_j = c_{j+1} + k_{j+1} v_j$ and $k_j \neq k_{j+1}$, and called a Type* II *change point if $c_j + k_j v_j \neq c_{j+1} + k_{j+1} v_j$ for $j \geq 1$, respectively.*

At Type I change points, signals are continuous while the slopes change at v_j (see Fig 1.1 (a)). Type II change points are essentially jumps (see Fig 1.1 (d)). Note that, an important special case of Type II change points is that $\mu(t)$ is piecewise constant, i.e., $k_j \equiv 0$ and $c_j \neq c_{j+1}$ for $j \geq 1$. Throughout this paper, we consider the following three scenarios of signal $\mu(t)$:

Scenario 1. The signal $\mu(t)$ contains only Type I change points (continuous breaks).

Scenario 2. The signal $\mu(t)$ contains only Type II change points (jumps).

Scenario 3. The signal $\mu(t)$ contains both Type I and Type II change points.

We are interested in detecting the number of change points and their locations simultaneously. We propose a new generic approach to the problem of detecting an unknown number of multiple structural breaks occurring at unknown locations. Our method of change point detection for the three scenarios above are illustrated in Fig 1.1. The key idea is that a change point in $\mu(t)$ will become a local extremum in the first or second derivative of the smoothed signal $\mu_\gamma(t)$. Specifically,

1. A Type I change point becomes a local extremum in the second derivative $\mu''_\gamma(t)$ (see figure 1.1 (a) and (c)).

2. A Type II change point becomes a local extremum in the first derivative $\mu'_\gamma(t)$ (see figure 1.1 (d) and (e)).

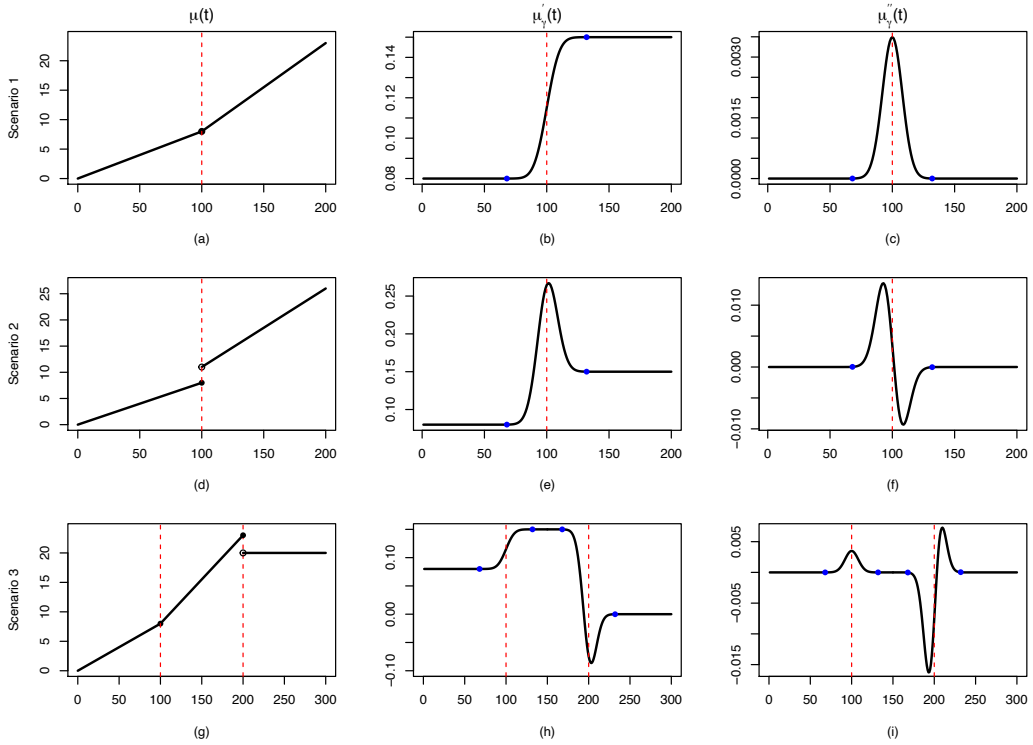

Figure 1.1: Illustration of change point detection. A change point in the piecewise linear signal $\mu(t)$ (left panel) becomes a local extremum in the first derivative of smoothed signal $\mu'_\gamma(t)$ (middle panel) or in the second derivative of smoothed signal $\mu''_\gamma(t)$ (right panel). The red dashed lines indicate the location of true change points. The blue points indicate the kernel smoothed signal regions. The top row shows a Type I change point v_j in $\mu(t)$ becomes a local extremum in $\mu''_\gamma(t)$ exactly at $t = v_j$. The middle row shows a Type II change point v_j in $\mu'_\gamma(t)$ becomes a local extremum in $\mu''_\gamma(t)$ around v_j. The bottom row reveals only Type II change point can generate a local extremum in $\mu'_\gamma(t)$, while both Type I and Type II change points can generate local extrema in $\mu''_\gamma(t)$.

Therefore, in scenario 1 and scenario 2, one can detect the change points by finding the local extrema in the second derivative $\mu''_\gamma(t)$ and first derivative $\mu'_\gamma(t)$ respectively. In scenario 3, note that a Type I change point does not generate any local extremum in the first derivative $\mu'_\gamma(t)$ (see Fig 1.1 (b)), thus one can first identify all Type II change points as local extrema in $\mu'_\gamma(t)$. On the other hand, since the local extrema in the second derivative $\mu''_\gamma(t)$ can be generated from both Type I and Type II change points (see Fig 1.1 (i)), one can detect all Type I change points as local extrema in $\mu''_\gamma(t)$ by removing those generated from Type II change points.

By focusing on the first and second derivatives of the smoothed signal $\mu_\gamma(t)$, the change point detection problem is transformed into a peak detection problem. Schwartzman *et al.* [61] and Cheng and Schwartzman [20] proposed a Smoothing and TEsting of Maxima/Minima (STEM) algorithm to find local maxima and minima of the derivative as candidate peaks. A multiple testing method is used to distinguish between the local extrema generated by change points and random noise. Furthermore, Cheng *et al.* [18] introduced differential STEM (dSTEM) method to detect the change points in data sequences modeled as piecewise constant signal-plus-noise.

Literature on change point detection contains a vast amount of research work of statistical inference, but most of it specially designed for the case of a single change point with unknown location. For example, with respect to the models with an unknown change point, Andrews [4] and Andrews *et al.* [5] proposed the comprehensive treatment and testing method for structural change. Perron [59] utilized unit root for a one-time change in the level or in the slope of the trend function in univariate time series. Bai [6] introduced the method of least squares to estimate a unknown shift point (change point) in a piecewise constant model. In recent years, the multiple change point detection problem has drawn an extensive interest, especially in terms of the multiple testing based methods. The literature on statistical inference

for multiple change points problems is diverse in different type of signals and different inferential questions. In the piecewise constant signal model, Yao and Au [72] studied the least squares estimators for the locations and the levels of the step function under known and unknown number of jumps. Lavielle [50] developed a penalized least squares method for estimating the number of change points and their locations. Another general approach is based on the idea of binary segmentation (BS) [68]. Hyun *et al.* [45] outlined similar post-selection tests for change point detection via Wild Binary Segmentation [29] and circular Binary Segmentation [57]. Hyun *et al.* [44] considered fused lasso [64] solutions for estimated change points, and test hypotheses of the equality of signal mean of either side of a given thus-detected change point. SMUCE [28, 58] estimated the number of change points as the minimum among all candidates fits for which the empirical residuals pass a certain multi-scale test at a given significance level. Li *et al.* [51] proposed an estimator, constructed similarly to SMUCE, which controls the FDR but with a generous definition of a true discovery. Hao *et al.* [38] and Cheng *et al.* [18] respectively showed FDR control for the SaRa and dSTEM estimator of multiple change point locations. For the general piecewise linear signal models, Bai and Perron [8] provided simultaneous asymptotic distribution results regarding the distance between the estimated change points and their true locations under the assumption of a known number of change points and the their minimum distance being $O(L)$. However, the distributional limits depend on the unknown magnitudes of parameter change, which are often difficult to estimate well. Baranowski *et al.* [10] proposed a narrowest-over-threshold (NOT) approach to estimate the number and locations of change points. NOT focus on the smallest local sections of the data on which the existence of a feature is suspected. Fryzlewicz [30] introduced a narrowest significance pursuit (NSP) algorithm for automatically detecting localised regions, each of which must contain a change point at a prescribed

global significance level. NSP works by fitting the postulated linear model over many regions of the data, using a certain multi-resolution sup-norm loss, and identifying the shortest interval on which the linearity is significantly violated. However, our proposed approach is unique in comparison with the existing literature in following ways.

1. Our method can estimate the number of change points and their locations simultaneously. Based on the distribution of local extrma in Gaussian process, the change points are identified as significant local extrema (in $\mu'_\gamma(t)$ or $\mu''_\gamma(t)$) under a global significant level. Moreover, the theoretical analysis in Section 2.2 guarantees that under mild conditions, our proposed method can truly control the false discovery rate (FDR) of detected change points under the significant level. Meanwhile, the power consistency is guaranteed.

2. Unlike the traditional change point detection methods, our approach can distinguish Type I and Type II change points, due to the fact that a Type I change point does not generate a local extremum in the first derivative $\mu'_\gamma(t)$, while a Type II change point will. This is especially important and useful for the data generated from Scenario 3 which is much more practical. One may only interest in the jumps (Type II) or continuous change points (Type I).

3. We assume that the noise $z(t)$ is a Gaussian process which allows the error terms to be correlated. The assumption of white noise in most of the change points literature is violated in practice [43]. Our method shows that change points methods can be devised for correlated noise, expanding the domain of their applicability.

4. Our proposed method can achieve lowest computational complexity. As we test the candidate peaks generated either by change points or random noise

only once, the computation of our method is the same as the number of the candidate peaks, which is much smaller than the data sequence length. Most of the traditional methods of change point detection require a computational cost of $O(L^2)$ ([28]). A few approaches under strict assumptions can achieve a linear computational cost ([48], [26],[29]).

1.2 Statistical Inference for Implicit Network Structures

In recent decades, network analysis has been applied in science and engineering fields including mathematics, physics, biology, computer science, social sciences and statistics (see [32, 34, 56] for reviews). Traditionally, statistical network analysis deals with parameter estimation of an observed network, i.e., an observed adjacency matrix. For example, community detection, a topic of broad interest, studies how to partition the node set of an observed network into cohesive overlapping or non-overlapping communities (see [1, 74] for recent reviews). Other well-studied statistical network models include the preferential attachment model [9], exponential random graph models [27, 60], latent space models [40, 39], and the graphon model [23, 31, 73].

In contrast to traditional statistical network analysis, this dissertation focuses on inferring a latent network structure. Specifically, we model data with the following format: each observation in the dataset is a subset of nodes that are observed simultaneously. An observation is called a *group* and a full dataset is called *grouped data*. [69] introduced this format using the toy example of a children's birthday party. In their simple example, children are treated as nodes and each party represents a group – i.e., a subset of children who attended the same party is a group. The reader is referred to [76, 70] for applications of such data to the social sciences and animal behavior.

The observed grouping behavior presumably results from a latent social structure that can be interpreted as a network structure of associated individuals [54]. The task is therefore to infer a latent network structure from grouped data. Existing methods mainly focus on ad-hoc descriptive approaches from the social sciences literature, such as the co-occurrence matrix [69] or the half weight index [15]. [76] propose the first model-based approach, called the *hub model*, which assumes that every observed group has a *hub* that brings together the other members of the group. When the hub nodes of grouped data are known, estimating the model parameters is a trivial task. In most research situations, hub nodes are unknown and need to be modeled as latent variables. Under this setup, estimating the model parameters becomes a more difficult task.

This project has three aims: first, to prove the identifiability of the canonical parameters and the asymptotic consistency for the estimators of those parameters *when hubs are unobserved*. The canonical parameters refer to the probabilities of being a hub node of a group and the probabilities of being included in a group formed by a particular hub node. The hub model is a restricted class from the family of finite mixtures of multivariate Bernoulli [76]. [36] showed that in general the parameters of finite mixture models of multivariate Bernoulli are not identifiable. [76] showed that the parameters are identifiable under two assumptions: the hub node of each group always appears in the group it forms and relationships are reciprocal. That is, the adjacency matrix is symmetric with diagonal entries as one. This paper considers identifiability when adjacency matrices are asymmetric. The model is therefore referred as to the *asymmetric hub model*. We prove that when the hub set (i.e., the set of possible hubs) contains at least one fewer member than the node set, the parameters are identifiable under mild conditions. The new setup is practical and less restrictive than the symmetry assumption. Moreover, allowing the hub set to be

smaller than the node set can reduce model complexity as pointed out by [70]. When proving the consistency of the estimators, we first prove the consistency of the hub estimates and then show that the estimators of model parameters are consistent as a corollary. Our proofs accommodate the most general setup in which the number of groups (i.e., sample size), the size of the node set, and the size of the hub set are all allowed to grow.

The second aim is to generalize the hub model to accommodate hubless groups and then prove identifiability and consistency of this generalized model. The classical hub model requires each group to have a hub. As observed in [70], when fitting the hub model to data, one sometimes has to choose an unnecessarily large hub set due to this requirement. For example, a node that appears infrequently in general but appears once as a singleton must be included in the hub set. To relax the *one-hub* restriction, we add a component to the hub model that allows hubless groups in which nodes appear independently. We call this additional component the *null component* and call the new model the *hub model with the null component*. The proofs of identifiability and consistency for the new model do not parallel the first set of proofs and are more challenging.

Since the new models assume the hub set is a subset of the nodes, this raises a natural question: how to estimate the hub set from data, which is the third aim. We formulate this problem as model selection for Bernoulli mixture models. We borrow the log penalty in Huang *et al.* [41], originally designed for Gaussian mixture models, to propose a penalized likelihood approach to select the hub set for the hub model with the null component. Instead of penalizing the mixing probability of every component as in Huang *et al.* [41], we modify the penalty function such that the probability of the null component is not penalized. The null component does not exist in the setup of Gaussian mixture models, but it creates a natural connection

between the hub model and a null model in our scenario. That is, when all other mixing probabilities are shrunken to zero, the model naturally degenerates to the model in which nodes appear independently in a group – in other words, each group is modeled by independent Bernoulli trials.

Chapter 2

MULTIPLE CHANGE POINT DETECTION IN LINEAR MODELS

2.1 Framework of Change Point Detection

2.1.1 The Kernel Smoothed Signal

We consider the following univariate statistical model:

$$y(t) = \mu(t) + z(t), \quad t \in \mathbb{R}, \tag{2.1}$$

where $z(t)$ is correlated stationary Gaussian noise and $\mu(t)$ is a piecewise linear signal of the form

$$\mu(t) = c_j + k_j t, \quad t \in (v_{j-1}, v_j],$$

where $c_j, k_j \in \mathbb{R}$, $j = 1, 2, \ldots$ and $-\infty = v_0 < v_1 < v_2 < \cdots$. The jump size a_j at v_j is defined as

$$a_j = c_{j+1} + k_{j+1} v_j - (c_j + k_j v_j) = (c_{j+1} - c_j) + (k_{j+1} - k_j) v_j, \ j \geq 1.$$

For the model (2.1), we assume $d = \inf_j (v_j - v_{j-1}) > 0$ so that the change points do not arbitrarily close to each other. In addition, we assume $k = \inf_j |k_{j+1} - k_j| > 0$ for the data sequence contains pure Type I change points so that the size of slope change does not become arbitrarily small. We assume $a = \inf_j |a_j| > 0$ for the data sequence contains pure Type II change points so that the size of jumps does not become arbitrarily small.

Let $w_\gamma(t)$ be the Gaussian kernel with compact support $[-c\gamma, c\gamma]$ and bandwidth γ, i.e.,

$$w_\gamma(t) = \frac{1}{\gamma} \phi(\frac{t}{\gamma}) \mathbb{1}\{-c\gamma \leq t \leq c\gamma\}.$$

Convolving the process (2.1) with the kernel $w_\gamma(t)$ results in a smoothed random process

$$y_\gamma(t) = w_\gamma(t) * y(t) = \int_{\mathbb{R}} w_\gamma(t-s)y(s)\,ds = \mu_\gamma(t) + z_\gamma(t), \tag{2.2}$$

where the smoothed signal and smoothed noise are defined respectively as

$$\mu_\gamma(t) = w_\gamma(t) * \mu(t) \quad \text{and} \quad z_\gamma(t) = w_\gamma(t) * z(t).$$

The smoothed noise $z_\gamma(t)$ is assumed to be a zero-mean and four-times differentiable stationary ergodic Gaussian process. To avoid the overlap of smoothing two neighboring change points, we assume $d = \inf_j(v_j - v_{j-1}) \geq 2c\gamma$.

2.1.2 Local Extrema for Derivatives of the Smoothed Signals

For a smooth function $f(t)$, denote by $f^{(\ell)}(t)$ its ℓ-th derivative, $\ell \geq 1$, and write by default $f'(t) = f^{(1)}(t)$ and $f''(t) = f^{(2)}(t)$ respectively. We have the following derivatives of the smoothed observed process (2.2),

$$y_\gamma^{(\ell)}(t) = w_\gamma^{(\ell)}(t) * y(t) = \int_{\mathbb{R}} w_\gamma^{(\ell)}(t-s)y(s)\,ds = \mu_\gamma^{(\ell)}(t) + z_\gamma^{(\ell)}(t), \tag{2.3}$$

where the derivatives of the smoothed signal and smoothed noise are respectively

$$\mu_\gamma^{(\ell)}(t) = w_\gamma^{(\ell)}(t) * \mu(t) \quad \text{and} \quad z_\gamma^{(\ell)}(t) = w_\gamma^{(\ell)}(t) * z(t), \quad \ell \geq 1.$$

Lemma 1. *For $\mu_\gamma(t)$ with support $(v_{j-1} + c\gamma, v_{j+1} - c\gamma)$, the first derivative is*

$$\mu'_\gamma(t) = \begin{cases} k_j[2\Phi(c) - 1], & t \in (v_{j-1} + c\gamma, v_j - c\gamma), \\ k_{j+1}[2\Phi(c) - 1], & t \in (v_j + c\gamma, v_{j+1} - c\gamma), \\ \frac{a_j}{\gamma}\phi(\frac{v_j - t}{\gamma}) + (k_j - k_{j+1})\Phi(\frac{v_j - t}{\gamma}) + (k_j + k_{j+1})\Phi(c) - k_j, & \text{otherwise.} \end{cases}$$

For $\mu_\gamma(t)$ with support $(v_{j-1} + c\gamma, v_{j+1} - c\gamma)$, the second derivative is

$$\mu''_\gamma(t) = \begin{cases} \frac{a_j(v_j - t) + (k_{j+1} - k_j)\gamma^2}{\gamma^3}\phi(\frac{v_j - t}{\gamma}), & t \in (v_j - c\gamma, v_j + c\gamma), \\ 0, & \text{otherwise.} \end{cases}$$

Remark 1. *$\mu_\gamma(t)$ is discontinuous at $v_j - c\gamma$ and $v_j + c\gamma$, where $\mu'_\gamma(t)$ and $\mu''_\gamma(t)$ have no definitions. Thus in our method, the locations, $t = v_j - c\gamma$ and $v_j + c\gamma$ will not be tested which does not affect the detection of the true change points.*

Lemma 2. *Define $q_j = \frac{k_{j+1}-k_j}{a_j}$ if $a_j \neq 0$. The local maximum/minimum of $\mu'_\gamma(t)$ with $t \in [v_j - c\gamma, v_j + c\gamma]$ is*

$$t = \begin{cases} \text{does not exist}, & a_j = 0, \quad (2.4a) \\ v_j + \gamma^2 q_j, & a_j \neq 0. \quad (2.4b) \end{cases}$$

The local maximum/minimum of $\mu''_\gamma(t)$ with $t \in (v_j - c\gamma, v_j + c\gamma)$ is

$$t = \begin{cases} v_j, & a_j = 0, \quad (2.5a) \\ v_j - \dfrac{\gamma^2 q_j \pm \gamma\sqrt{4 + q_j^2}}{2}, & a_j \neq 0. \quad (2.5b) \end{cases}$$

Proposition 1. *A Type* I *change point v_j in $\mu(t)$ becomes a local extremum (local maximum or local minimum) in the second derivative of the smoothed signal, $\mu''_\gamma(t)$, exactly at v_j (see* (2.5a)*); A Type* II *change point in $\mu(t)$ becomes a local extremum at $v_j + \gamma^2 q_j = v_j + o(1)$ in the first derivative, $\mu'_\gamma(t)$ (see* (2.4b)*).*

Remark 2. *A Type* I *change point does not generate any local extremum in the first derivative $\mu'_\gamma(t)$ (see* (2.4a)*). This key feature helps to identify the Type* II *change points in the mixture case (see Step 1 in Algorithm 3). Suppose q_j is a very small positive number, for the second derivative, $\mu''_\gamma(t)$, a Type* II *change point v_j will generate a pair of local maximum and local minimum at around $v_j - \gamma$ and $v_j + \gamma$ (see* (2.5b)*). This helps to remove the local extrema in the second derivative generated by Type* II *change points and thereafter to detect the Type* I *change points in the mixture case (see Step 2 in Algorithm 3).*

2.1.3 Main Ideas

Fig 1.1 shows the central idea of our method, that is to transform the problem of change point detection in piecewise linear signal to detection of local extrema in

the first or second derivatives of the smoothed signal. However, a local extremum is generated not only from change points in the signal, but from the random noise. Thus we need to use a multiple testing, based on the peak height distribution of $z'_\gamma(t)$ and $z''_\gamma(t)$ (see section 2.2.4), to identify the significant local extrema as true change points.

Toy examples in Fig 2.1 and Fig 2.2 illustrate the main ideas of detection of change points in a data sequence. Specifically, for a data sequence with piecewise linear signal in Scenario 1, a Type I change point will generate a peak (positive or negative) in the second derivative exactly at the same location (see (2.5a)). The peaks in $y''_\gamma(t)$ comes from both the six Type I change points and random noise. Then a multiple testing is applied to find the true change points. For the data sequence with signal in Scenario 2, a Type II change point can generate a peak in the first derivative around its location (see (2.4b)). All the six Type II change points are detected as significant peaks in the second derivative. For the data sequence with signal in Scenario 3, we first detect the Type II change points as the significant peaks in $y'_\gamma(t)$. Removing the peaks in $y''_\gamma(t)$ generated by Type II change points, the Type I change points are detected as the significant peaks in $y''_\gamma(t)$.

To detect the change points in piecewise linear model, we improve the method STEM (**S**moothing and **TE**sting of **M**axima/**M**inima) and propose mSTEM (**m**odified STEM) which consists of the following steps:

1. Differential kernel smoothing: to transform change points to local maxima or local minima (illustrated in Fig 1.1), and meanwhile increase the SNR.

2. Candidate peaks: to find local maxima and local minima of the first or second smoothed Gaussian process ($y'_\gamma(t)$ or $y''_\gamma(t)$).

3. P-values: to compute the p-value of each local maximum or local minimum under the null hypothesis of no change points in a local neighborhood.

4. Multiple testing: to apply a multiple testing procedure to the set of local maxima and local minima, a change point is claimed to be detected if the p-value (in Step 3) of its local maximum or local minumum is significant.

2.2 Multiple Change Point Detection for Linear Models

2.2.1 Type I *Change Point Detection*

Suppose we observe $y(t)$ with J Type I change points in a data sequence of length L centered at the origin, denoted by $U(L) = (-L/2, L/2)$.

Following the proposed *mSTEM* procedure of change point detection, we introduce the mSTEM algorithm for the detection of Type I change points.

Algorithm 1 (mSTEM algorithm for Type I break detection)**.**

1. Differential kernel smoothing*: Obtain the process* $y''_\gamma(t)$ *in* (2.3) *by convolution of* $y(t)$ *with the kernel derivative* $w''_\gamma(t)$*.*

2. Candidate peaks*: Find the set of local maxima and minima of* $y''_\gamma(t)$ *in* $U(L)$*, denoted by* $\tilde{T}_{\rm I} = \tilde{T}_{\rm I} \cup \tilde{T}_{\rm I}^-$*, where*

$$\tilde{T}_{\rm I}^+ = \left\{ t \in U(L) : y_\gamma^{(3)}(t) = 0,\ y_\gamma^{(4)}(t) < 0 \right\},$$
$$\tilde{T}_{\rm I}^- = \left\{ t \in U(L) : y_\gamma^{(3)}(t) = 0,\ y_\gamma^{(4)}(t) > 0 \right\}.$$

3. P-values*: For each* $t \in \tilde{T}_{\rm I}^+$*, compute the p-value* $p_{\rm I}(t)$ *for testing the (conditional) hypotheses*

$$\mathcal{H}_0(t) : \{\mu''_\gamma(s) = 0 \quad \textit{for all} \quad s \in (t-b, t+b)\} \quad \text{vs.}$$
$$\mathcal{H}_A(t) : \{\mu''_\gamma(s) > 0 \quad \textit{for some} \quad s \in (t-b, t+b)\};$$

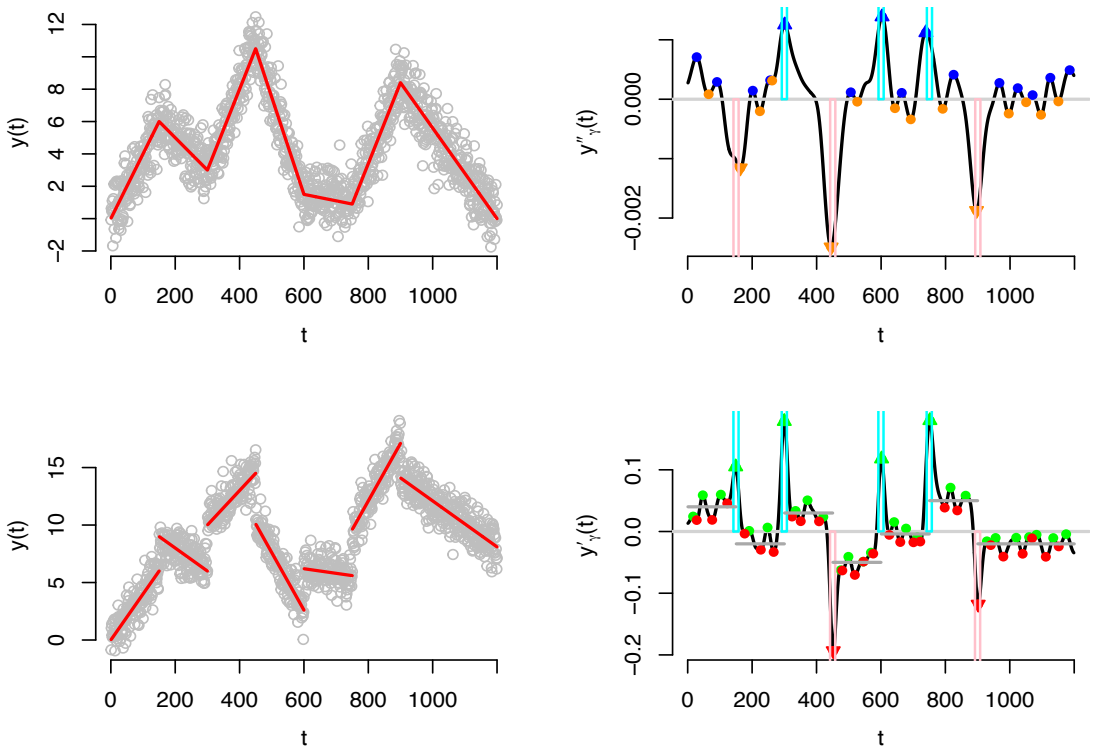

Figure 2.1: Procedure of Type I and Type II change point detection. The left panel shows the data sequences with six Type I change points and six Type II change points respectively, the red lines indicate the corresponding piecewise linear signals. The right panel shows the second and the first derivatives of the smoothed data $y_\gamma(t)$ ($\gamma = 8$) respectively. Local maxima and local minima of first derivatives are represented by green and red solid dots respectively. Local maxima and local minima of second derivatives are represented by blue and orange solid dots respectively. The solid triangles denote the significant local extrema under significant level $\alpha = 0.05$. The cyan and pink bars indicate the location tolerance intervals $(v_j - b, v_j + b)$ with $b = 5$ for the true change points. The deepgrey lines indicate the piecewise slopes of the piecewise linear signal (baselines for the testing of local extrema of $y'_\gamma(t)$).

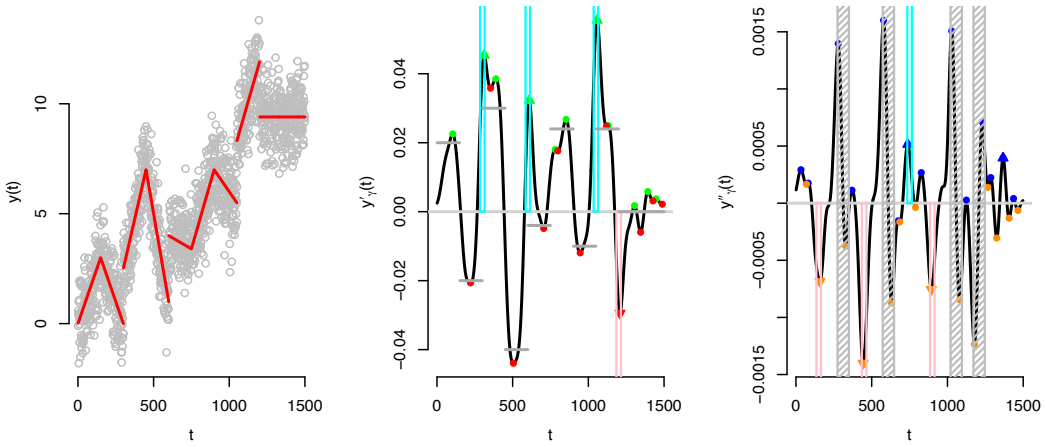

Figure 2.2: Procedure of mixture of Type I and Type II change point detection. Here, we use the same symbols and colors as in Fig 2.1. The left panel shows the data with four Type I and four Type II change points. The middle panel shows the first derivative of the data which can be used to detect Type II change points (see Step 1 in Algorithm 3). The right panel shows the second derivative of the data. Before detecting the Type I change points, we first remove the local extrema of $y''_\gamma(t)$ on the interval $[v_j - 2\gamma, v_j + 2\gamma]$ (represented by grey bars) avoiding the disturb of Type II change points. In this example, all the change points are detected and there is only one false discovery for Type I change points.

and for each $t \in \tilde{T}_{\mathrm{I}}^{-}$*, compute the p-value* $p_{\mathrm{I}}(t)$ *for testing the hypotheses*

$$\mathcal{H}_0(t) : \{\mu''_\gamma(s) = 0 \quad \textit{for all} \quad s \in (t-b, t+b)\} \quad \text{vs.}$$

$$\mathcal{H}_A(t) : \{\mu''_\gamma(s) < 0 \quad \textit{for some} \quad s \in (t-b, t+b)\},$$

where $b > 0$ *is an appropriate location tolerance.*

4. Multiple testing*: Apply a multiple testing procedure on the set of p-values* $\{p_{\mathrm{I}}(t), t \in \tilde{T}_{\mathrm{I}}\}$*, and declare significant all local extrema whose p-values are smaller than the significance threshold.*

P-values computation

The p-values in step 3 of the algorithm 1 is computed as

$$p_{\mathrm{I}}(t) = \begin{cases} F_{z''_\gamma}(y''_\gamma(t)), & t \in \tilde{T}_{\mathrm{I}}^{+}, \\ F_{z''_\gamma}(-y''_\gamma(t)), & t \in \tilde{T}_{\mathrm{I}}^{-}, \end{cases} \tag{2.6}$$

where $F_{z''_\gamma}(u)$ (see the chapter 2.2.4) denotes the right tail probability of $z''_\gamma(t)$ at the local maximum $t \in \tilde{T}_{\mathrm{I}}$, evaluated under the null model $\mu^{(3)}_\gamma(s) = 0, \forall s \in (t-b, t+b)$, that is,

$$F_{z''_\gamma}(u) = P\left(z''_\gamma(t) > u \middle|\ t \text{ is a local maximum of } z''_\gamma(t)\right). \tag{2.7}$$

The second line in (2.6) is obtained by

$$\begin{aligned} & P\left(z''_\gamma(t) < y''_\gamma(t) \middle|\ t \text{ is a local minimum of } z'_\gamma(t)\right) \\ =& P\left(-z''_\gamma(t) > -y''_\gamma(t) \middle|\ t \text{ is a local maximum of } -z''_\gamma(t)\right) \\ =& F_{z''_\gamma}(-y''_\gamma(t)), \end{aligned}$$

since $-z''_\gamma(t)$ and $z''_\gamma(t)$ have the same distribution.

Error and Power definitions

We define that the *signal region* is $\mathbb{S}_1^b = \cup_{j=1}^{J}(v_j - b, v_j + b)$ and *null region* is $\mathbb{S}_0^b = U(L) \setminus \mathbb{S}_1^b$. For $u > 0$, let $\tilde{T}_{\mathrm{I}}(u) = \tilde{T}_{\mathrm{I}}^{+}(u) \cup \tilde{T}_{\mathrm{I}}^{-}(u)$, where

$$\begin{aligned} \tilde{T}_{\mathrm{I}}^{+}(u) &= \left\{t \in U(L) : y''_\gamma(t) > u,\ y^{(3)}_\gamma(t) = 0,\ y^{(4)}_\gamma(t) < 0\right\}, \\ \tilde{T}_{I}^{-}(u) &= \left\{t \in U(L) : y''_\gamma(t) < -u,\ y^{(3)}_\gamma(t) = 0,\ y^{(4)}_\gamma(t) > 0\right\}. \end{aligned}$$

The above equations indicate that $\tilde{T}_{\mathrm{I}}^{+}(u)$ and $\tilde{T}_{1,\gamma}^{-}(u)$ are respectively the set of local maxima of $y''_\gamma(t)$ above u and the set of local minima of $y''_\gamma(t)$ below $-u$. The number of totally and falsely detected change points at threshold u are defined

respectively as

$$R_{\mathrm{I}}(u) = \#\{t \in \tilde{T}_{\mathrm{I}}^{+}(u)\} + \#\{t \in \tilde{T}_{\mathrm{I}}^{-}(u)\},$$
$$V_{\mathrm{I}}(u;b) = \#\{t \in \tilde{T}_{\mathrm{I}}^{+}(u) \cap \mathbb{S}_0^b\} + \#\{t \in \tilde{T}_{\mathrm{I}}^{-}(u) \cap \mathbb{S}_0^b\}.$$

Both are defined as zero if $\tilde{T}_{\mathrm{I}}(u)$ is empty. The FDR at threshold u is defined as the expected proportion of falsely detected jumps

$$\mathrm{FDR_I(u;b)} = \mathrm{E}\left\{\frac{\mathrm{V_I(u;b)}}{\mathrm{R_I(u)} \vee 1}\right\}. \tag{2.8}$$

Note that definition of FDR for other type of change points can be defined similarly as (2.8).

Following the notation in Cheng and Schwartzman [20], define the *smoothed signal region* $\mathbb{S}_{1,\gamma}$ to be the support of $\mu'_\gamma(t)$ and *smoothed null region* $\mathbb{S}_{0,\gamma} = U(L) \setminus \mathbb{S}_{1,\gamma}$. We call the difference between the expanded signal support due to smoothing and the true signal support the *transition region* $\mathbb{T}_\gamma = \mathbb{S}_{1,\gamma} \setminus \mathbb{S}_1^b = \mathbb{S}_0^b \setminus \mathbb{S}_{0,\gamma}$.

Power

Denote by I^+ and I^- the collections of indices j corresponding to increasing and decreasing change points v_j, respectively. We define the power as the expected fraction of true discovered change points

$$\begin{aligned}\mathrm{Power_I}(u;b) &= \frac{1}{J}\sum_{j=1}^{J}\mathrm{Power_{I,}}_j(u;b) \\ &= \mathrm{E}\left[\frac{1}{J}\left(\sum_{j\in I^+}\mathbb{1}\left(\tilde{T}_{\mathrm{I}}^{+}(u) \cap (v_j - b, v_j + b) \neq \emptyset\right)\right.\right. \\ &\qquad \left.\left. + \sum_{j\in I^-}\mathbb{1}\left(\tilde{T}_{\mathrm{I}}^{-}(u) \cap (v_j - b, v_j + b) \neq \emptyset\right)\right)\right],\end{aligned} \tag{2.9}$$

where $\text{Power}_{j,\gamma}(u;b)$ is the probability of detecting change point v_j within a distance b,

$$\text{Power}_{\text{I},j}(u;b) = \begin{cases} \mathbb{P}\left(\tilde{T}_{\text{I}}^{+}(u) \cap (v_j - b, v_j + b) \neq \emptyset\right), & \text{if } j \in I^{+}, \\ \mathbb{P}\left(\tilde{T}_{\text{I}}^{-}(u) \cap (v_j - b, v_j + b) \neq \emptyset\right), & \text{if } j \in I^{-}. \end{cases} \tag{2.10}$$

Note that definition of Power for other type of change points can be defined similarly as (2.9). The indicator function in (2.9) ensures that only one significant local extremum is counted within a distance b of a change point, so power is not inflated. Note that when γ and u are fixed, $\text{Power}_{\text{I}}(u;b)$ and $\text{Power}_{j,\text{I}}(u;b)$ are increasing in b.

Asymptotic FDR control and Power consistency

Suppose the Benjamini-Hochberg (BH) procedure is applied in step 4 of dSTEM algorithm as follows. For a fixed $\alpha \in (0,1)$, let k be the largest index for which the ith smallest p-value is less than $i\alpha/\tilde{m}_\gamma$, where $\tilde{m}_\gamma$ is the number of local extrema of $y''_\gamma(t)$ on the smoothed signal region. Then the null hypothesis $\mathcal{H}_0(t)$ at $t \in \tilde{T}_{\text{I}}$ is rejected if

$$p_\gamma(t) < \frac{k\alpha}{\tilde{m}_\gamma} \iff \begin{cases} y'_\gamma(t) > \tilde{u}_{\text{BH}} = F^{-1}_{z''_\gamma}\left(\frac{k\alpha}{\tilde{m}_\gamma}\right) & \text{if } t \in \tilde{T}_{\text{I}}^{+}, \\ y'_\gamma(t) < -\tilde{u}_{\text{BH}} = -F^{-1}_{z''_\gamma}\left(\frac{k\alpha}{\tilde{m}_\gamma}\right) & \text{if } t \in \tilde{T}_{\text{I}}^{-}, \end{cases} \tag{2.11}$$

where $k\alpha/\tilde{m}_\gamma$ is defined as 1 if $\tilde{m}_\gamma = 0$. Since $\tilde{u}_{\text{BH}}$ is random, we define FDR in such BH procedure as

$$\text{FDR}_{\text{I,BH}}(b) = \text{E}\left\{\frac{V_\gamma(\tilde{u}_{\text{BH}};b)}{R_\gamma(\tilde{u}_{\text{BH}}) \vee 1}\right\}.$$

Similarly we can define the FDR in BH procedure for other type of change points.

To study the asymptotic theories of FDR and Power, we make the following assumptions:

(C1) The assumptions of §2.1.1 hold.

(C2) $L \to \infty$, $k = \inf_j |k_{j+1} - k_j| \to \infty$, and $k^2/\log L \to \infty$.

Let $\mathrm{E}[\tilde{m}_{z''_\gamma}(U(1))]$ and $\mathrm{E}[\tilde{m}_{z''_\gamma}(U(1), u)]$ be the expected number of local maxima and local maxima above level u of $z''_\gamma(t)$ on the unit interval $U(1) = (-1/2, 1/2)$, respectively. In particular, applying the Kac-Rice formula, we have the following explicit result [61],

$$\mathrm{E}[\tilde{m}_{z''_\gamma}(U(1))] = \frac{1}{2\pi}\sqrt{\frac{\mathrm{Var}(z^{(4)}(t))}{\mathrm{Var}(z^{(3)}(t))}}.$$

Theorem 1. *Under assumptions* $(C1)$ *and* $(C2)$*, for the data* $y(t)$ *containing pure Type* I *change points, if the number of change points* J *satisfies* $J/L \to A$ *as* $L \to \infty$*, then*

(i) suppose that Algorithm 1 is applied with a fixed threshold u*, we have*

$$\mathrm{FDR}_{\mathrm{I}}(u; b) \to \frac{E[\tilde{m}_{z''_\gamma}(U(1), u)](1 - 2c\gamma A)}{E[\tilde{m}_{z''_\gamma}(U(1), u)](1 - 2c\gamma A) + A}. \tag{2.12}$$

(ii) suppose that Algorithm 1 is applied with random threshold $\tilde{u}_{\mathrm{BH}}$*, we have*

$$\mathrm{FDR}_{\mathrm{I,BH}}(b) \to \alpha\frac{E[\tilde{m}_{z''_\gamma}(U(1))](1 - 2c\gamma A)}{E[\tilde{m}_{z''_\gamma}(U(1))](1 - 2c\gamma A) + A}. \tag{2.13}$$

Theorem 2. *Under conditions* $(C1)$ *and* $(C2)$*, for the data* $y(t)$ *containing only Type* I *change points, if the number of change points* J *satisfies* $J/L \to A$ *as* $L \to \infty$*, then*

(i) suppose that Algorithm 1 is applied with a fixed threshold u*, we have*

$$\mathrm{Power}_{\mathrm{I}}(u; b) = \frac{1}{J}\sum_{j=1}^{J} \mathrm{Power}_{\mathrm{I}}(j, u; b) \to 1. \tag{2.14}$$

(ii) suppose that Algorithm 1 is applied with the random threshold $\tilde{u}_{\mathrm{BH}}$*, we have*

$$\mathrm{Power}_{\mathrm{I,BH}}(b) = \frac{1}{J}\sum_{j=1}^{J} \mathrm{Power}_{\mathrm{I,BH}}(j; b) \to 1. \tag{2.15}$$

2.2.2 *Type* II *Change Point Detection*

As shown in Fig 1.1, for the detection of Type I change points, the piecewise slopes of $\mu''(t)$ are 0 everywhere except the smoothed signal region. However, for the detection of Type II change points, the piecewise slopes of $\mu'_\gamma(t)$ becomes piecewise constants (not necessary zeros), which are the baselines for the multiple testing (see Step 3 in Algorithm 2). Hence, we first need to estimate the piecewise slopes of the signals with Type II change points.

Piecewise slopes estimate

A basic idea of estimating the piecewise slopes is to cut the data sequence into segments in which there is no change points and thereafter the slopes are estimated by a linear regression. A Type II break is transformed into a pair of local maximum and minimum of $\mu''_\gamma(t)$ (see 2.5b), and the piecewise slope of $mu''_\gamma(t)$ are 0 everywhere except the smoothed signal region, which give us a hint for finding the segments.

(2.5b) shows the pairwise local maximum and local minimum for v_j are around at $v_j - \gamma$ and $v_j + \gamma$ when q_j is small. Note that $\mu''_\gamma(v_j - \gamma)$ and $\mu''_\gamma(v_j + \gamma)$ are not symmetric about 0 as shown in figure 2.1, thus it is possible that only one value is significant in the multiple testing of $y''_\gamma(t)$. Follow the algorithm 1 with a larger significant level such that we can get non-conservative estimators for the peaks in $y''_\gamma(t)$, say l_i for $i = 1, \ldots, K$, where K is the number of estimated peaks in $y''_\gamma(t)$. Since some of the breaks are detected as a pairwise local maxima and local minima, hence some of l_i are pairwise. But we can use the middle point of each pair as the estimator of a Type II change point.

We assume the distance between any two consecutive breaks is relatively large enough compared to γ. As the distance of the pairwised local maxima and local

minima is about 2γ, we define l_i and l_j are paired if

$$1.5\gamma \leq |l_i - l_j| \leq 2.5\gamma, \text{ for } i, j = 1, \ldots, K \text{ and } i \neq j. \tag{2.16}$$

Then we use the mean of paired l_i and l_j as the estimator of a Type II break. While if l_i, $l = 1, \ldots, K$ is identified as a single point (only local maximum or local minimum is significant in detection of $y''_\gamma(t)$), l_i will be the estimator of a Type II break. even though the distance between the true break and l_i is about γ, it does not affect the estimation of slopes if a robust regression method is applied. Finally, we can obtain the estimators for Type II breaks, say $\tilde{l}_i$ for $i = 1, \ldots, \tilde{K}$, where $\tilde{K}$ is the number of estimated breaks. Based on the estimators $\tilde{l}_i$, the whole data sequence can be divided into $\tilde{K} + 1$ segments. The slope in the jth segment, k_j, is estimated through a robust regression model [42] in the raw signal-plus-noise data over the interval $(\tilde{l}_{j-1} + 2\gamma, \tilde{l}_j - 2\gamma)$ for $j = 1, \ldots, \tilde{K}$, i.e.,

$$\hat{k}_j = \text{Robust Regression } (y(t) \sim t), \text{ for } t \in (\tilde{l}_{j-1} + 2\gamma, \tilde{l}_j - 2\gamma). \tag{2.17}$$

The following algorithm shows the procedure of estimating piecewise slopes.

Algorithm for estimating of piecewise slopes

1. Perform Algorithm 1 with a larger significant level to detect significant local extrema, denoted by l_i, for $i = 1, \ldots, K$.

2. Follow (2.16) to find paired local maxima and local minima.

3. Detect roughly Type II change points as $\tilde{l}_i$, for $i = 1, \ldots, \tilde{K}$, then the ith segment is defined as $(\tilde{l}_i, \tilde{l}_{i+1})$.

4. Estimate piecewise slopes by (2.17) in each segment.

Algorithm 2 (mSTEM algorithm for Type II break detection)**.**

1. Differential kernel smoothing*: Obtain the process $y'_\gamma(t)$ in* (2.3) *by convolution of $y(t)$ with the kernel derivative $w'_\gamma(t)$.*

2. Candidate peaks*: Find the set of local maxima and minima of $y'_\gamma(t)$ in $U(L)$, denoted by $\tilde{T}_{\text{II}} = \tilde{T}^+_{\text{II}} \cup \tilde{T}^-_{\text{II}}$, where*

$$\tilde{T}^+_{\text{II}} = \left\{t \in U(L) : y''_\gamma(t) = 0,\ y^{(3)}_\gamma(t) < 0\right\},$$
$$\tilde{T}^-_{\text{II}} = \left\{t \in U(L) : y''_\gamma(t) = 0,\ y^{(3)}_\gamma(t) > 0\right\}.$$

3. P-values*: For each $t \in \tilde{T}^+_{\text{II}}$, compute the p-value $p_{\text{II}}(t)$ for testing the (conditional) hypotheses*

$$\mathcal{H}_0(t) : \{\mu'_\gamma(s) - k(s) = 0 \quad \textit{for all} \quad s \in (t-b, t+b)\} \quad \text{vs.}$$
$$\mathcal{H}_A(t) : \{\mu'_\gamma(s) - k(s) > 0 \quad \textit{for some} \quad s \in (t-b, t+b)\},$$

where $k(s)$ is the estimated piecewise slope estimated by (2.17) *and $b > 0$ is an appropriate location tolerance. For each $t \in \tilde{T}^-_{\text{II}}$, compute the p-value $p_{\text{II}}(t)$ for testing the hypotheses*

$$\mathcal{H}_0(t) : \{\mu'_\gamma(s) - k(s) = 0 \quad \textit{for all} \quad s \in (t-b, t+b)\} \quad \text{vs.}$$
$$\mathcal{H}_A(t) : \{\mu'_\gamma(s) - k(s) < 0 \quad \textit{for some} \quad s \in (t-b, t+b)\},$$

4. Multiple testing*: Apply a multiple testing procedure on the set of p-values $\{p_{\text{II}}(t),\, t \in \tilde{T}_{\text{II}}\}$, and declare significant all local extrema whose p-values are smaller than the significance threshold.*

Asymptotic FDR control and Power consistency

For the proof of the asymptotic FDR and power consistency of data with Type II change points, we make the following assumption:

(C3) $a = \inf_j |a_j| \to \infty$ and $q = \sup_j |\frac{k_{j+1}-k_j}{a_j}| \to 0$ as $L \to \infty$.

Theorem 3. *Under conditions* $(C1)-(C3)$*, for the data* $y(t)$ *containing only Type* II *change points, if the number of change points* J *satisfies* $J/L \to A$ *as* $L \to \infty$*, then*

(i) for a fixed threshold u*, we have*

$$\mathrm{FDR}_{\mathrm{II}}(u;b) \to \frac{E[\tilde{m}_{z'_\gamma}(U(1),u)](1-2c\gamma A)}{E[\tilde{m}_{z'_\gamma}(U(1),u)](1-2c\gamma A)+A}. \tag{2.18}$$

(ii) for a random threshold $\tilde{u}_{\mathrm{BH}}$*, we have*

$$\mathrm{FDR}_{\mathrm{II,BH}}(b) \to \alpha\frac{E[\tilde{m}_{z'_\gamma}(U(1))](1-2c\gamma A)}{E[\tilde{m}_{z'_\gamma}(U(1))](1-2c\gamma A)+A}. \tag{2.19}$$

Theorem 4. *Under conditions* $(C1)-(C3)$*, for the data* $y(t)$ *containing only Type* I *change points, if the number of change points* J *satisfies* $J/L \to A$ *as* $L \to \infty$*, then*

(i) for a fixed threshold u*, we have*

$$\mathrm{Power}_{\mathrm{II}}(u;b) = \frac{1}{J}\sum_{j=1}^{J}\mathrm{Power}_{\mathrm{II}}(j,u;b) \to 1. \tag{2.20}$$

(ii) for a random threshold $\tilde{u}_{\mathrm{BH}}$*, we have*

$$\mathrm{Power}_{\mathrm{II,BH}}(b) = \frac{1}{J}\sum_{j=1}^{J}\mathrm{Power}_{\mathrm{II,BH}}(j;b) \to 1. \tag{2.21}$$

2.2.3 *Mixture of Type* I *and Type* II *Change Point Detection*

Type I change points can be detected through the peaks in the second derivative of $y_\gamma(t)$ (see algorithm 1), and Type II change points can be detected via finding significant peaks in $y'_\gamma(t)$ (see algorithm 2). However, it is very common that the real signal-plus-noise data contains both Type I and Type II change points. How to distinguish a Type I and Type II change point is the key problem. Our method for detecting change points of the signals in scenario 3 is based on the following features of $\mu'_\gamma(t)$ and $\mu''_\gamma(t)$:

1. A Type II change point will generate a peak in $\mu'_\gamma(t)$ (see Fig 1.1 (h)), but a Type I change point does not generate. Therefore, a detected change point in $y'_\gamma(t)$ can only be Type II.

2. Both Type I and Type II will generate peaks in $\mu''_\gamma(t)$ (see Fig 1.1 (i)).

Thus, our method is to detect all the Type II change points in $y'_\gamma(t)$ using algorithm 1, then we perform *mSTEM* procedure to find all the change points (includes both Type I and Type II), by removing the set of Type II change points in $y'_\gamma(t)$, we can obtain the Type I change points.

For the detection of mixture of Type I and Type II change points, the main idea is to detect Type II change points in the first derivatives (note that Type I change points do not generate local extrema of first derivative) and then detect Type I change points in the second derivatives by removing those peaks generated by Type II. The following algorithm shows the specific procedure of detecting Type I and II change points in the mixture case.

Algorithm 3 (mSTEM algorithm for mixture of Type I and II breaks detection)**.**

1. Estimate Type II breaks*: Perform the algorithm 2 to obtain the estimate of Type* II *breaks, say* $\mathcal{M}_{\text{II}} = \{\hat{v}_{\text{II},i}\}$ *for* $i = 1, 2, \ldots$. *We use a larger* γ *(compared to the signal in scenario 2) such that we can obtain better estimate for the piecewise slope, especially for the peaks in* $y''_\gamma(t)$ *generated by Type* I *breaks.*

2. Candidate Type I peaks*: Find the set of local maxima and minima of* $y''_\gamma(t)$ *in* $U(L)$*, denoted by* $\tilde{T}_{\text{I}} = \tilde{T}_{\text{I}}^+ \cup \tilde{T}_{\text{I}}^-$. *As* $\tilde{T}_{\text{I}}$ *contains the local extrema generated by both Type* I *and Type* II *breaks, to detect Type* I *breaks, it is necessary to remove the peaks generated by Type* II *breaks. Thus, the set of candidate peaks of Type*

I *breaks is defined as* $\tilde{T}_{\mathrm{I}\backslash\mathrm{II}} = \tilde{T}^{+}_{\mathrm{I}\backslash\mathrm{II}} \cup \tilde{T}^{-}_{\mathrm{I}\backslash\mathrm{II}}$*, where*

$$\tilde{T}^{+}_{\mathrm{I}\backslash\mathrm{II}} = \tilde{T}^{+}_{\mathrm{I}} \setminus \cup_{i=1}(\hat{v}_{\mathrm{II},i} - 2\gamma, \hat{v}_{\mathrm{II},i} + 2\gamma),$$
$$\tilde{T}^{-}_{\mathrm{I}\backslash\mathrm{II}} = \tilde{T}^{-}_{\mathrm{I}} \setminus \cup_{i=1}(\hat{v}_{\mathrm{II},i} - 2\gamma, \hat{v}_{\mathrm{II},i} + 2\gamma).$$

3. P-values*: For each* $t \in \tilde{T}^{+}_{\mathrm{I}\backslash\mathrm{II}}$*, compute the p-value* $p_{\mathrm{I}\backslash\mathrm{II}}(t)$ *for testing the (conditional) hypotheses*

$$\mathcal{H}_0(t) : \{\mu''_\gamma(s) = 0 \quad \textit{for all} \quad s \in (t-b, t+b)\} \quad \text{vs.}$$
$$\mathcal{H}_A(t) : \{\mu''_\gamma(s) > 0 \quad \textit{for some} \quad s \in (t-b, t+b)\};$$

and for each $t \in \tilde{T}^{-}_{\mathrm{I}\backslash\mathrm{II}}$*, compute the p-value* $p_{\mathrm{I}\backslash\mathrm{II}}(t)$ *for testing the (conditional) hypotheses*

$$\mathcal{H}_0(t) : \{\mu''_\gamma(s) = 0 \quad \textit{for all} \quad s \in (t-b, t+b)\} \quad \text{vs.}$$
$$\mathcal{H}_A(t) : \{\mu''_\gamma(s) < 0 \quad \textit{for some} \quad s \in (t-b, t+b)\},$$

where $b > 0$ *is an appropriate location tolerance.*

4. Multiple testing*: Apply a multiple testing procedure on the set of p-values* $\{p_{\mathrm{I}\backslash\mathrm{II}}(t),\, t \in \tilde{T}_{\mathrm{I}\backslash\mathrm{II}}\}$*, and declare significant all local extrema whose p-values are smaller than the significance threshold, then the set of Type* I *breaks can be obtained as* $\mathcal{M}_{\mathrm{I}} = \{\hat{v}_{\mathrm{I},i}\}$ *for* $i = 1, 2, \ldots$.

In the first step, Type I change points will generate some extra segments when estimating the piecewise slope, but it does not affect the estimate of Type II breaks as Type I breaks do not generate local maxima or local minima in $y'_\gamma(t)$.

Asymptotic FDR control and Power consistency

Theorem 5. *Under conditions* $(C1) - (C3)$*, assume* $y(t)$ *contains* J_1 *Type* I *change points and* J_2 *Type* II *change points with the constrains* $J_1/L \to A_1$ *and* $J_2/L \to A_2$ *as* $L \to \infty$.

(i) suppose that Algorithm 3 is applied with a fixed threshold u_1 and u_2 for Type I *and Type* II *change point detection respectively, then*

$$\limsup \mathrm{FDR}_{\mathrm{III}}(\mathrm{u}_1, \mathrm{u}_2; \mathrm{b}) \leq \frac{E[\tilde{m}_{z''_\gamma}(U(1), u_1)](1 - 2c\gamma A_1) + E[\tilde{m}_{z'_\gamma}(U(1), u_2)](1 - 4\gamma A_2)}{E[\tilde{m}_{z''_\gamma}(U(1), u_1)](1 - 2c\gamma A_1) + E[\tilde{m}_{z''_\gamma}(U(1), u_2)](1 - 4\gamma A_2) + A}, \tag{2.22}$$

where $A = A_1 + A_2$.

(ii) Suppose Algorithm 3 is applied with the random threshold $\tilde{u}_{\mathrm{BH}}$. *Then*

$$\limsup \mathrm{FDR}_{\mathrm{I,BH}}(\mathrm{b}) \leq \alpha. \tag{2.23}$$

$$\lim \mathrm{FDR}_{\mathrm{II,BH}}(\mathrm{b}) = \alpha \frac{\mathrm{E}[\tilde{\mathrm{m}}_{\mathrm{z}'_\gamma}(\mathrm{U}(1)](1 - 2\mathrm{c}\gamma\mathrm{A}_2)}{\mathrm{E}[\tilde{\mathrm{m}}_{\mathrm{z}'_\gamma}(\mathrm{U}(1))](1 - 2\mathrm{c}\gamma\mathrm{A}_2) + \mathrm{A}_2} \leq \alpha. \tag{2.24}$$

$$\limsup \mathrm{FDR}_{\mathrm{III,BH}}(\mathrm{b}) \leq \alpha. \tag{2.25}$$

Theorem 6. *Under conditions* $(C1) - (C3)$, *assume* $y(t)$ *contains* J_1 *Type* I *change points and* J_2 *Type* II *change points with the constrains* $J_1/L \to A_1$ *and* $J_2/L \to A_2$ *as* $L \to \infty$.

(i) suppose that Algorithm 3 is applied with a fixed threshold u_1 and u_2 for Type I *and Type* II *change point detection respectively, then*

$$\mathrm{Power}_{\mathrm{III}}(u_1, u_2; b) = \frac{1}{J} \sum_{j=1}^{J} \mathrm{Power}_{\mathrm{III}}(j, u_1, u_2; b) \to 1. \tag{2.26}$$

(ii) suppose that Algorithm 3 is applied with the random threshold $\tilde{u}_{\mathrm{BH}}$, *then*

$$\mathrm{Power}_{\mathrm{III,BH}}(b) = \frac{1}{J} \sum_{j=1}^{J} \mathrm{Power}_{\mathrm{III,BH}}(j; b) \to 1. \tag{2.27}$$

2.2.4 Gaussian Auto-correlation Model and Its Peak Height Distribution

Let $X(t)$ be a smoothed stationary Gaussian process with zero-mean and variance, σ^2. Define $\eta = \frac{\mathrm{Var(X')}}{\sqrt{\mathrm{Var(X)Var(X'')}}}$ (we omit t here because $X(t)$ is stationary), let $F_X(x)$ denote the right tail probability of $X(t)$ at the local maximim, that is,

$$F_X(x) = P(X > x \mid x \text{ is a local maximum of } X),$$

then

$$F_X(x) = 1 - \Phi\left(\frac{x}{\sigma\sqrt{1-\eta^2}}\right) + \sqrt{2\pi}\eta\phi\left(\frac{x}{\sqrt{\sigma}}\right)\Phi\left(\frac{\eta x}{\sqrt{\sigma}\sqrt{1-\eta^2}}\right). \tag{2.28}$$

Note that (2.28) is a general version of the peak height distribution in Schwartzman *et al.* [61].

We consider a simple example of $X(t)$. Let the noise $z(t)$ be

$$z(t) = \int_{\mathbb{R}} \frac{1}{\nu}\phi\left(\frac{t-s}{\nu}\right) dB(s), \qquad \nu > 0, \tag{2.29}$$

where ϕ is the standard Gaussian density, $dB(s)$ is Gaussian white noise ($z(t)$ is regarded by convention as Gaussian white noise when $\nu = 0$). Convolving with a Gaussian kernel $w_\gamma(t) = (1/\gamma)\phi(t/\gamma)$ with $\gamma > 0$ produces a zero-mean infinitely differentiable stationary ergodic Gaussian field

$$z_\gamma(t) = \int_{\mathbb{R}} w_\gamma(t-x)z(x)dx = \int_{\mathbb{R}} \frac{1}{\xi}\phi\left(\frac{t-s}{\xi}\right) dB(s), \quad \xi = \sqrt{\gamma^2+\nu^2}. \tag{2.30}$$

Lemma 3. *For the smoothed stationary Gaussian process $z_\gamma(t)$, defined in* (2.30)*, the variance of its derivatives are*

$$\mathrm{Var}(z'_\gamma(t)) = \frac{1}{4\sqrt{\pi}\xi^3}, \qquad \mathrm{Var}(z''_\gamma(t)) = \frac{3}{8\sqrt{\pi}\xi^5},$$
$$\mathrm{Var}(z^{(3)}_\gamma(t)) = \frac{15}{16\sqrt{\pi}\xi^7}, \qquad \mathrm{Var}(z^{(4)}_\gamma(t)) = \frac{105}{32\sqrt{\pi}\xi^9}.$$

Combining Lemma 3 and (2.28), we immediately have the following proposition:

Proposition 2. *Let $z_\gamma(t)$ be defined in* (2.30)*. The peak height distribution of $z'_\gamma(t)$ is*

$$F_{z'_\gamma}(x) = 1 - \Phi\left(\frac{x}{\sigma_1\sqrt{1-\eta_1^2}}\right) + \sqrt{2\pi}\eta_1\phi\left(\frac{x}{\sqrt{\sigma_1}}\right)\Phi\left(\frac{\eta_1 x}{\sqrt{\sigma_1}\sqrt{1-\eta_1^2}}\right), \tag{2.31}$$

where $\eta_1 = \frac{\sqrt{3}}{\sqrt{5}}$ *and* $\sigma_1^2 = \frac{1}{4\sqrt{\pi}\xi^3}$.

The peak height distribution of $z''_\gamma(t)$ *is*

$$F_{z''_\gamma}(x) = 1 - \Phi\left(\frac{x}{\sigma_2\sqrt{1-\eta_2^2}}\right) + \sqrt{2\pi}\eta_2\phi\left(\frac{x}{\sqrt{\sigma_2}}\right)\Phi\left(\frac{\eta_2 x}{\sqrt{\sigma_2}\sqrt{1-\eta_2^2}}\right), \tag{2.32}$$

where $\eta_2 = \frac{\sqrt{5}}{\sqrt{7}}$ *and* $\sigma_2^2 = \frac{3}{8\sqrt{\pi}\xi^5}$.

2.2.5 SNR

Smoothing the data can not only make it differentiable at v_j, but also can increase the SNR. Additionally, the asymptotic assumptions $(C3)$ and $(C4)$ are in fact to make SNR go to infinity.

Lemma 4. *For a Type* I *change point, the SNR at* v_j *is*

$$SNR_{\mathrm{I}}(v_j) = \frac{\mu''_\gamma(v_j)}{\sqrt{\mathrm{Var}(z''_\gamma(v_j))}} = \frac{\frac{k_{j+1}-k_j}{\sqrt{2\pi}\gamma}}{\sqrt{\frac{3}{8\sqrt{\pi}\gamma^5}}} = \frac{2\gamma^{3/2}}{\sqrt{3}\pi^{1/4}}(k_{j+1}-k_j). \tag{2.33}$$

For a Type II *change point, the SNR at* v_j *is*

$$SNR_{\mathrm{II}}(v_j) = \frac{\mu'_\gamma(v_j)}{\sqrt{\mathrm{Var}(z'_\gamma(v_j))}} = \frac{\frac{a_j}{\sqrt{2\pi}\gamma} + \frac{k_j+k_{j+1}}{2}}{\sqrt{\frac{1}{4\sqrt{\pi}\gamma^3}}} = \frac{\sqrt{2}a_j}{\pi^{1/4}}\sqrt{\gamma} + (k_j+k_{j+1})\pi^{1/4}\gamma^{3/2}. \tag{2.34}$$

Chapter 3

NUMERICAL STUDIES ON MULTIPLE CHANGE POINT DETECTION

3.1 Simulation Studies

3.1.1 Simulation Settings

In this section, we study the performance of our method for signals $\mu(t) = c_j + k_j t$, where $t = 1, \ldots, L$, $L = 15,000$, and the true change point locations are $v_j = jd$ for $j = 1, \ldots, \lfloor L/d \rfloor - 1$, and $d = 150$ is the distance between consecutive change points. In addition, the signals have 4 different scenarios: (1) piecewise linear mean with continuous change points (Type I); (2) piecewise constant mean with jumps (special case of Type II); (3) piecewise linear mean with discontinuous change points (Type II); (4) mixture of Type I and Type II change points. The noise is generated from a zero-mean stationary ergodic Gaussian process (see (2.29)). Note that the random error is white noise when $\nu = 0$, and is correlated when $\nu > 0$. The smoothing kernels are $w_\gamma(t) = (1/\gamma)\phi(t/\gamma)\mathbb{1}(t \in [-6\gamma, 6\gamma])$. The BH procedure was applied at FDR level $\alpha = 0.05$. Results were averaged over 2,000 replications.

3.1.2 Performance of Our Method

In this section, we further verify the properties of our method via numerical studies. Figure 3.1 and figure 3.2 show the excellent performance of our method for the 4 scenarios of signals. We see that as SNR increases FDR converges to its asymptotic limit which is under the FDR control level $\alpha = 0.05$, and Power converges to 1. The conditions (C2)-(C3) require SNR should be infinity, However, our examples show both FDR and Power have already converged when SNR is around 10.

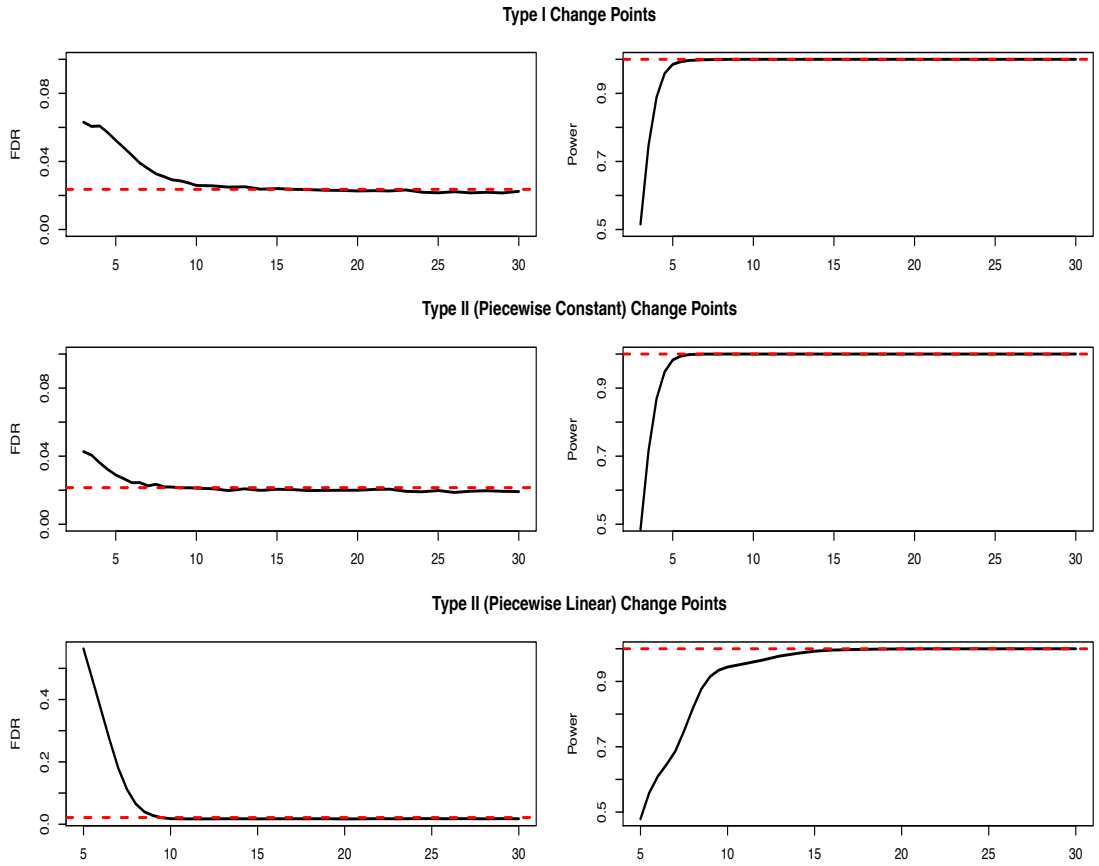

Figure 3.1: FDR and Power versus SNR for Type I and Type 2 change point detection. In this example, the kernel bandwidth $\gamma = 10$ and location tolerance $b = 10$. The red dashed lines indicate the theoretical limit of FDR and Power.

Choice of bandwidth γ

Fig 3.3 and Fig 3.4 show the results of FDR and Power as the kernel bandwidth γ increases. SNR is monotonically increasing with respect to γ, i.e., a larger γ indicates a larger SNR. Thus when γ is not large enough, the performance of FDR and Power get better as γ increases. However, a too large γ brings the issue of overlap, i.e., when the kernel region 12γ is longer than span of neighboring change points v_j and

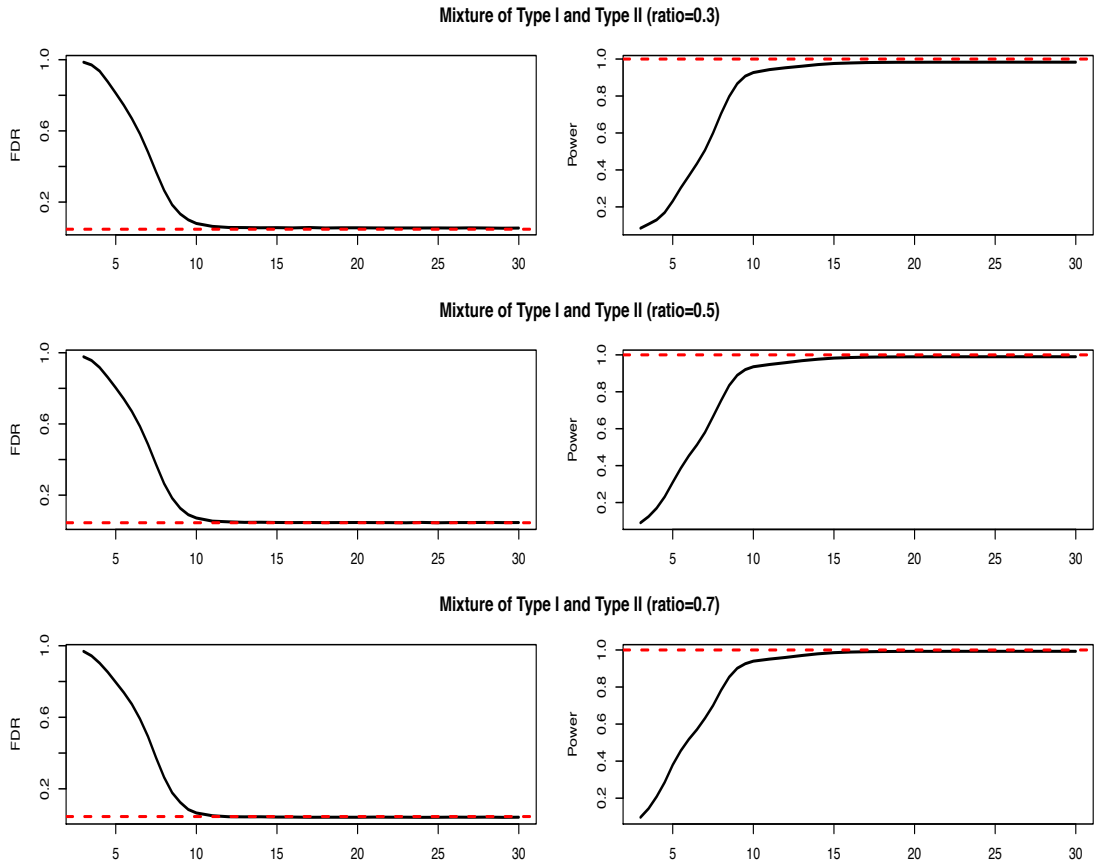

Figure 3.2: FDR and Power versus SNR for mixture of Type I and Type 2 change point detection. The ratio here is defined as the ratio of number of Type I change points to the numer of Type II (piecewise linear) change points. In this example, the kernel bandwidth $\gamma = 10$ and location tolerance $b = 10$. The red dashed lines indicate the theoretical limit of FDR and Power.

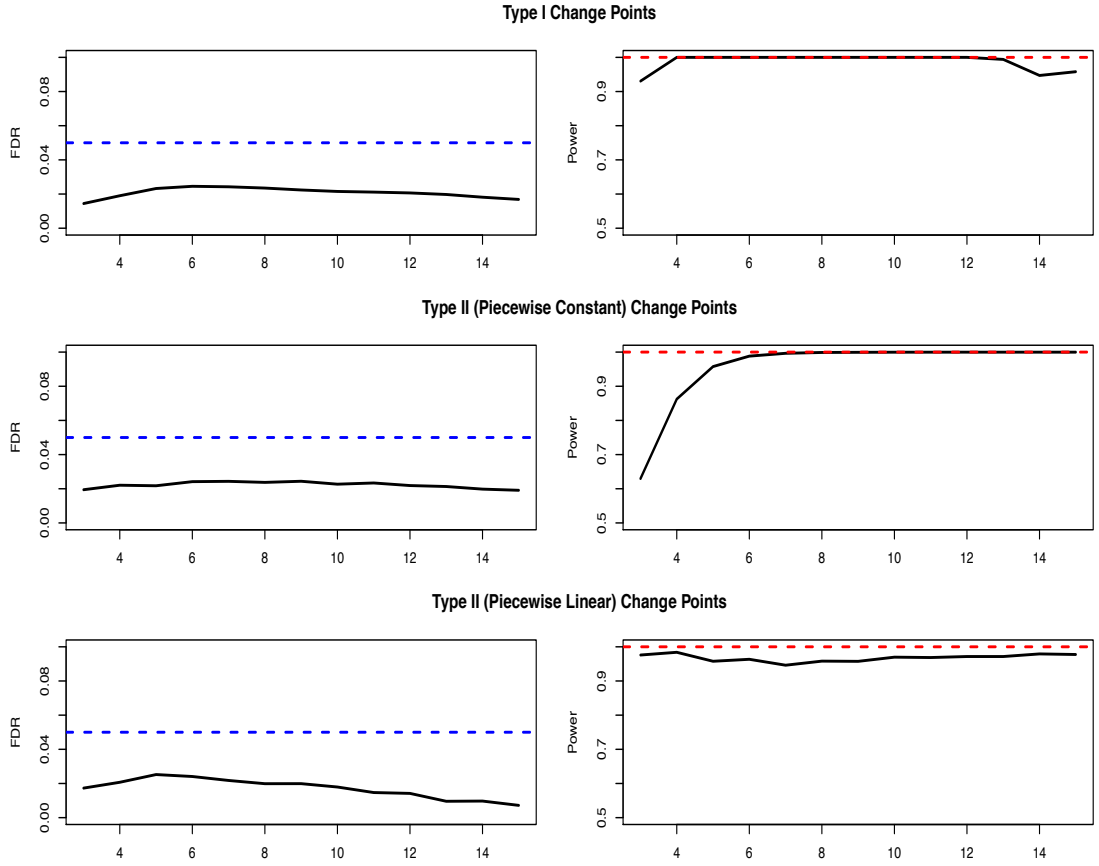

Figure 3.3: FDR and Power versus bandwidth γ for mixture of Type I and Type 2 change point detection. In this example, the kernel bandwidth γ varies from 3 to 15 and location tolerance $b = 10$. The blue dash lines indicate the FDR control level $\alpha = 0.05$ and the red dash lines indicate the theoretical limit of Power.

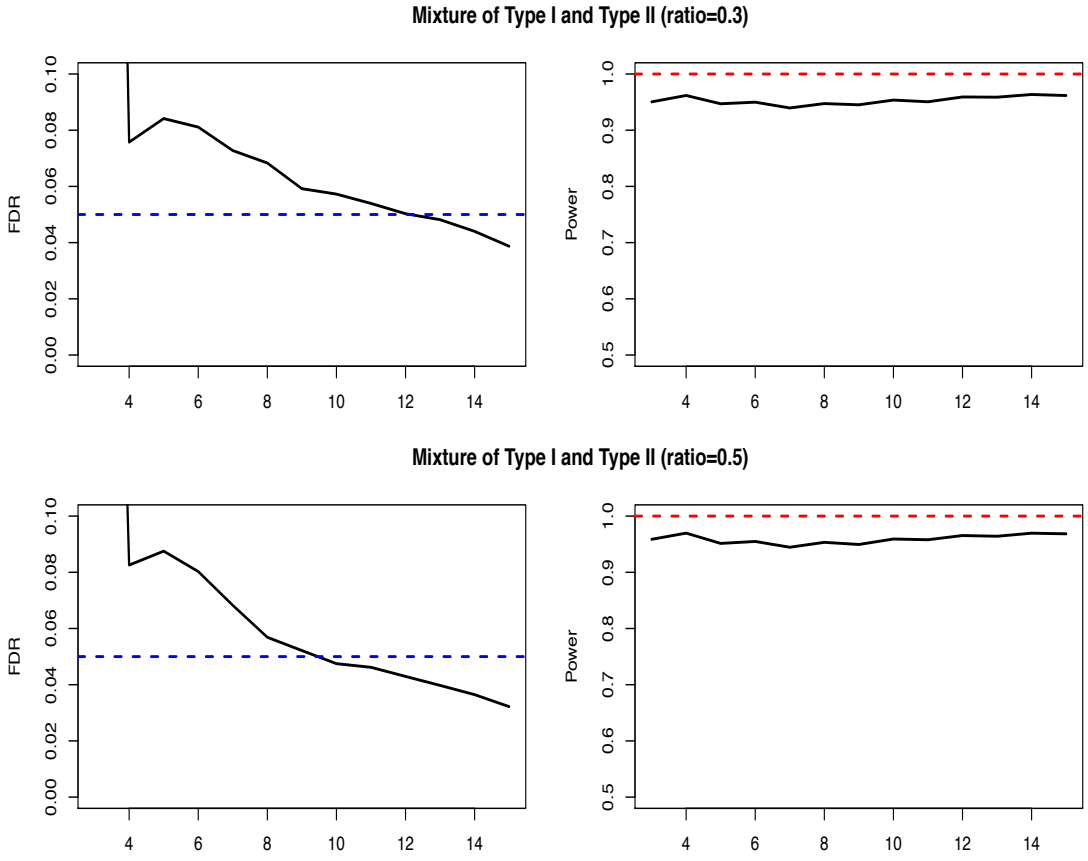

Figure 3.4: FDR and Power versus bandwidth γ for mixture of Type I and Type 2 change point detection. The ratio here is defined as the ratio of number of Type I change points to the number of Type II (piecewise linear) change points. In this example, the kernel bandwidth γ varies from 3 to 10 and location tolerance $b = 10$. The blue dash lines indicate the FDR control level $\alpha = 0.05$ and the red dash lines indicate the theoretical limit of Power.

v_{j+1}, then the kernel smoothing will affect not only the region $[v_j, v_{j+1})$, but also the regions $[v_{j-1}, v_j)$ and $[v_{j+1}, v_{j+2})$.

3.1.3 Comparison with Other Methods

For each scenario, we study the long-term data (the number of change points is large) and short-term data (only contains few change points). The long-term data is generated by just repeating the short-term data 10 times, thus the patterns of the long-term and short-term datas are the same. We compare our method with BP [7], NOT [10] and NSP [30] based on change point detection accuracy and computing time. As BP method takes too much time, we do not consider this method in the long-term data. To calculate FDR and Power, we let $b = \gamma$. $\hat{v}_j$ is defined as the estimation of its nearest true change point v_j, The distance $|\hat{v}_j - v_j|$ measures accuracy of detection. Table 3.1 shows NOT and our method have excellent performance in short-term data and our method is the fastest algorithm. In Table 3.2, we see only our method has the best performance smallest computing time while other two methods can not control the FDR and perform not as well as in short-term data.

For the long-term datasets, NOT and NSP are not a good choice for detection of change points. However, our method still keeps the excellent performance, even performs better in some cases.

3.2 Data Examples

In this section, we consider to apply our method to real applications and compare with NOT and NSP (B&P method was not included due to its weak performance and large computation). To have a better understanding of our method and comparation with other methods, a short-term dataset (has few change points) and a long-term dataset (has many change points) are used to evaluate the performance for

Table 3.1: Accuracy of estimation for change points in short-term data sequence

Signal Type	Method	Proportion of $\lvert\hat{v}_j - v_j\rvert$ within					FDR	Power	Time (s)
		$[0, \frac{1}{3}\gamma)$	$[\frac{1}{3}\gamma, \gamma)$	$[\gamma, 2\gamma)$	$[2\gamma, 4\gamma)$	$\geq 4\gamma$			
Type I	mSTEM	0.8400	0.1533	0.0217	0.0267	0.0483	0.0125	0.9933	0.1370
	NOT	0.9883	0.0117	0.0000	0.0017	0.0000	0.0572	1.0000	1.3550
	NSP	0.6183	0.1900	0.1617	0.0300	0.0000	0.0458	0.9999	3.2417
	B&P	0.4700	0.5083	0.0217	0.0000	0.0000	0.1632	0.9783	87.4562
Type II Piecewise Constant	mSTEM	0.9617	0.0383	0.0000	0.0367	0.0333	0.0227	1.0000	0.0290
	NOT	0.9833	0.0183	0.0017	0.0000	0.0017	0.0558	1.0000	0.2863
	NSP	0.6767	0.3133	0.2117	0.0867	0.0100	0.0517	0.9001	1.9430
	B&P	0.4117	0.0233	0.0467	0.2050	0.0000	0.1260	0.4350	71.0924
Type II Piecewise Linear	mSTEM	0.9983	0.0017	0.0000	0.0067	0.0233	0.0348	1.0000	0.0839
	NOT	0.8833	0.0933	0.0233	0.0000	0.0000	0.0727	0.9766	0.4524
	NSP	0.6967	0.2417	0.0333	0.0283	0.0000	0.0626	0.9384	2.8013
	B&P	0.3333	0.0000	0.3015	0.3333	0.0745	0.1633	0.3333	68.3345

Table 3.2: Accuracy of estimation for change points in long-term data sequence (with many change points)

Signal Type	Method	Proportion of $\lvert\hat{v}_j - v_j\rvert$ within					FDR	Power	Time (s)
		$[0, \frac{1}{3}\gamma)$	$[\frac{1}{3}\gamma, \gamma)$	$[\gamma, 2\gamma)$	$[2\gamma, 4\gamma)$	$\geq 4\gamma$			
Type I	mSTEM	0.7616	0.2241	0.0295	0.0244	0.0135	0.0127	0.9963	0.2469
	NOT	0.1358	0.1712	0.2475	0.5028	0.2369	0.0853	0.8732	112.5007
	NSP	0.3512	0.4607	0.1692	0.0086	0.0000	0.0792	0.8362	433.6193
Type II Piecewise Constant	mSTEM	0.9702	0.0000	0.000	0.0154	0.0161	0.01463	1.0000	0.1188
	NOT	0.8633	0.0037	0.007	0.0177	0.0090	0.0735	0.9164	48.4487
	NSP	0.6398	0.3202	0.030	0.0000	0.0000	0.08011	0.9228	213.0398
Type II Piecewise Linear	mSTEM	0.9899	0.0000	0.0000	0.0065	0.0133	0.0237	0.9992	1.0627
	NOT	0.8748	0.0002	0.0009	0.0032	0.0025	0.1217	0.8012	10.6064
	NSP	0.6059	0.3543	0.0298	0.0000	0.0000	0.1383	0.8255	400.6633

all methods. Moreover, our method can detect Type I and Type II change points simultaneously within one running, while NOT and NSP can not distinguish these two type of change points. For the NSP method, the change points are estimated as confidence intervals. To compare with other methods, we use the midpoints of the intervals as point estimate.

3.2.1 *Covid-19 Deaths in UK*

To compare the performance of our method with other methods for short-term data, we considered the same dataset in Fryzlewicz [30]. The dataset recorded the daily covid-19 associated deaths in UK from March 12, 2020 to July 23, 2020. In addition, same Anscombe transform (see Fryzlewicz [30]) was used to eliminate the weekly seasonality and make the data distribution closer to Gaussian with constant variance.

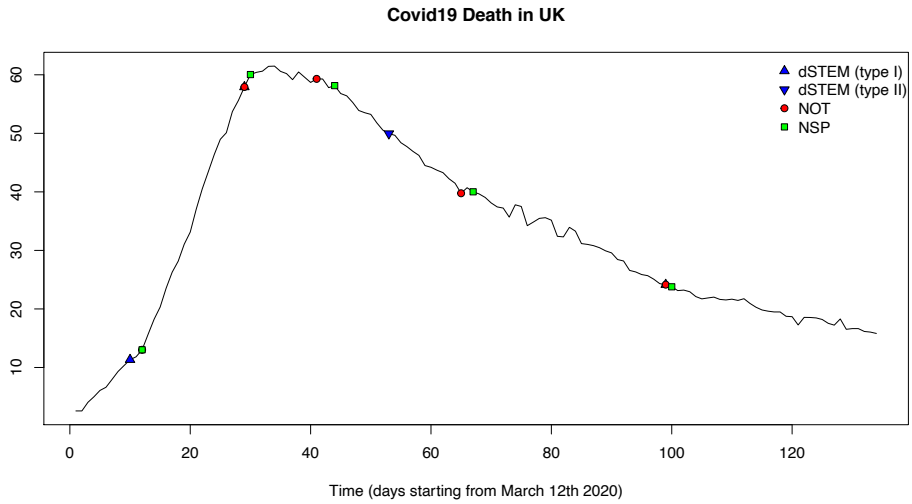

Figure 3.5: Covid-19 associated deaths in UK (March 12, 2020 – July 23, 2020). The change points estimate of our method are {10, 29, 53, 99}, {12, 29, 41, 65 99} for NOT and {12, 30, 44, 67, 100} for NSP.

Figure 3.5 shows the change points estimate of the three methods. Our method detected three Type I change points (10,29,99) and one Type II change point (55). According to the results of our method, we can split the time interval into five phases:
Phase I (1 – 10 days after March 12, 2020): Covid-19 virus began to spread out and the associated deaths increased slowly during this time period.
Phase II (11 – 29): The growth rate of this phase is much larger than that of phase

I. The rapid expanding number of infected population resulted in explosive growth in deaths which attain the peak in this phase.

Phase III (30 – 53): The covid-19 pandemic got controlled and the deaths began to decrease slowly. The shut down policies of pubic areas which was deployed in phase II began to take effect.

Phase IV (54 – 99): The deaths decreased faster than that in phase III due to hysteresis of restrictions and shut down policies.

Phase V (100 – 134): The pandemic was completely under control and the deaths decreased stably and slowly. Moreover, the deaths was controlled under a low level.

NOT and NSP have very similar estimate results and both detected five change points. Compared to our method, the main difference is NOT and NSP fitted phase II and phase III by three stages: The deaths kept its highest level in the first stage (30 – 44), then decreased rapidly in the second stage (45 – 67), and in the last stage the deaths decreased slowly. Our purpose is not to judge which method gave the correct or best results, all of the three methods' results make sense and have reasonable interpretation.

3.2.2 Stock Price of Host Hotel & Resorts

For a long-term time series, we studied the daily stock price (close) of Host Hotel & Resorts, Inc. (HST) from the period from January 1,2018 to November 5, 2021. Host Hotel & Resorts, Inc. is the world's largest lodging and real estate investment trust (REIT). Our interest in the company is because of its leading positions in the industry and eventful history during the last four years (since 2018) in hotel industry.

The historical data for HST stock price is available at `https://finance.yahoo.com`. Our method was applied to detect the change points and then compared with other two methods (NOT and NSP).

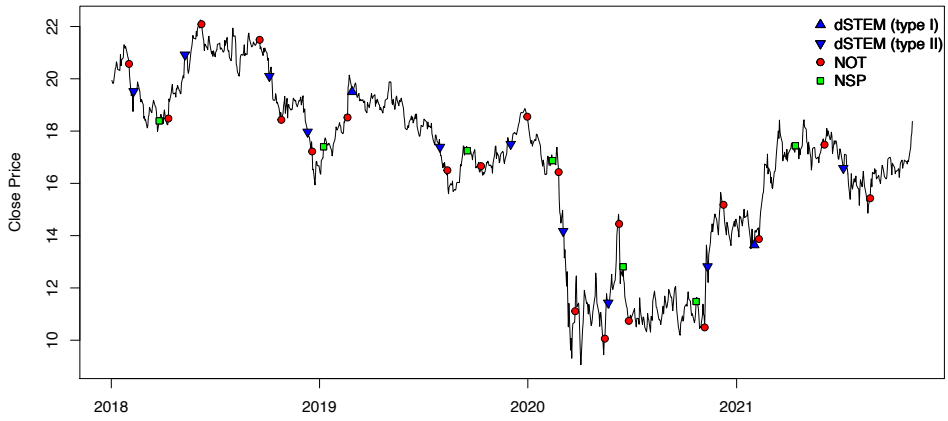

Figure 3.6: HST daily stock price (January 1, 2018 – November 5, 2020).

From Figure 3.6, we see NOT is very sensitive to variations in the dataset. It tended to detect more local peaks which might result from noise and thus will give a large FDR. Conversely, NSP tended to detect only few change points which might miss some true change points and result in small TPR. But the results of our method can be interpreted reasonably. For example, the change points in 2018 – 2019 were consistent with its large-cap stock S&P 500 due to the trade war between USA and China. And the change points after 2020 were consistent with the timeline of Covid-19 outbreak.

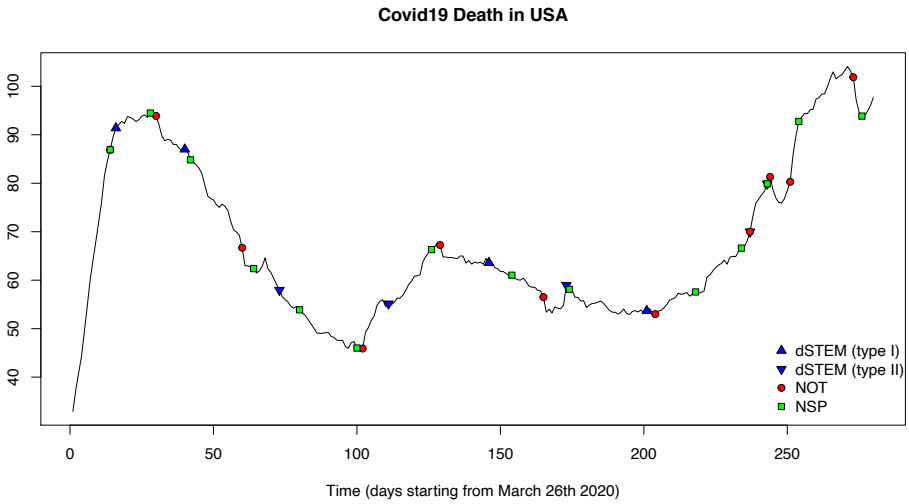

Figure 3.7: Covid-19 associated deaths in USA (March 26, 2020 – December 31, 2021).

Chapter 4

SUMMARY AND DISCUSSION

In this project, we combined both local maxima and minima of the derivative as candidate peaks, and then applied a multiple testing procedure to find a uniform threshold (in absolute value) for detecting all change points. This approach is sensible when the distributions (number and height) of true increasing and decreasing change points are about the same. Alternatively, different thresholds for detecting increasing and decreasing change points could be found by applying separate multiple testing procedures to the sets of candidate local maxima and local minima. While we applied the BH algorithm to control FDR, in principle other multiple testing procedures may be used to control other error rates.

A natural and important question is how to choose the smoothing bandwidth γ. We can see that either a small γ (if the noise is highly autocorrelated) or a relatively large γ (if the noise is less autocorrelated) is preferred in order to increase power, but only to the extent that the smoothed signal regions have little overlap and that detected change points are not displaced by more than the desired tolerance b (recall that the value of b is not used in the dSTEM algorithm itself, but it may be determined by the needs of the specific scientific application). Considering the Gaussian kernel to have an effective support of $\pm c\gamma$, a good value of γ may be about $\min(b, d/(2c))$, where d is the separation between change points. For example, if we consider the Gaussian kernel to have an effective support of $\pm 4\gamma$ and the separation between change points is $d = 100$, we may choose γ to be no larger than $\gamma = 10$. Since the location of the change points is unknown, a more precise optimization of γ may require an iterative procedure. Moreover, if some change points are close together

and others are far apart, an adaptive bandwidth may be preferable. We leave these as problems for future research.

We have assumed stationary Gaussian noise in our model for simplicity. The stationarity assumption allowed us to use an explicit formula for the height distribution of local extrema [19, 21]. However, in many applications, nonstationary noise is more realistic. To our knowledge, there are no existing methods for change point detection with nonstationary noise. Our work here provides a promising approach to solving this problem. To compute p-values, the height distribution of local extrema for smooth nonstationary Gaussian processes can be computed explicitly as long as the covariance function of the process is known [19, 21]. Otherwise, p-values could be approximated using the approximate overshoot distribution. However, challenges include how to estimate the covariance function and to prove FDR control and power consistency in this setting.

Chapter 5

TECHNICAL DETAILS OF MULTIPLE CHANGE POINT DETECTION

5.1 Proofs in Chapter 2.1

Proof of Lemma 1. For $t \in (v_j - c\gamma, v_j + c\gamma)$,

$$\begin{aligned}\mu_\gamma(t) = w_\gamma(t) * \mu(t) &= \int_{t-c\gamma}^{t+c\gamma} w_\gamma(t-s)\mu(s)ds \\ &= \int_{t-c\gamma}^{v_j} \frac{1}{\gamma}\phi(\frac{t-s}{\gamma})(c_j + k_j s)ds + \int_{v_j}^{t+c\gamma} \frac{1}{\gamma}\phi(\frac{t-s}{\gamma})(c_{j+1} + k_{j+1}s)ds \\ &= [c_j + k_j t - (c_{j+1} + k_{j+1}t)]\Phi(\frac{v_j - t}{\gamma}) + [c_j + k_j t + (c_{j+1} + k_{j+1}t)]\Phi(c) \\ &\quad - (c_j + k_j t) + (k_j - k_{j+1})\gamma\phi(c) + (k_{j+1} - k_j)\gamma\phi(\frac{v_j - t}{\gamma}).\end{aligned} \tag{5.1}$$

For $t \in (v_j + c\gamma, v_{j+1} - c\gamma)$,

$$\begin{aligned}\mu_\gamma(t) = w_\gamma(t) * \mu(t) &= \int_{t-c\gamma}^{t+c\gamma} \frac{1}{\gamma}\phi(\frac{t-s}{\gamma})(c_{j+1} + k_{j+1}s)ds \\ &= (c_{j+1} + k_{j+1}t)[2\Phi(c) - 1].\end{aligned} \tag{5.2}$$

For $t \in (v_{j-1} + c\gamma, v_j - c\gamma)$,

$$\begin{aligned}\mu_\gamma(t) = w_\gamma(t) * \mu(t) &= \int_{t-c\gamma}^{t+c\gamma} \frac{1}{\gamma}\phi(\frac{t-s}{\gamma})(c_j + k_j s)ds \\ &= (c_j + k_j t)[2\Phi(c) - 1].\end{aligned} \tag{5.3}$$

Note that $\mu_\gamma(t)$ is noncontinuous at $v_j - c\gamma$ and $v_j + c\gamma$. Take the first and second derivatives of $\mu_\gamma(t)$ in (5.1), (5.2) and (5.3) respectively, we have

$$\mu'_\gamma(t) = \begin{cases} k_j[2\Phi(c) - 1] & t \in (v_{j-1} + c\gamma, v_j - c\gamma), \\ k_{j+1}[2\Phi(c) - 1] & t \in (v_j + c\gamma, v_{j+1} - c\gamma), \\ \frac{a_j}{\gamma}\phi(\frac{v_j - t}{\gamma}) + (k_j - k_{j+1})\Phi(\frac{v_j - t}{\gamma}) + (k_j + k_{j+1})\Phi(c) - k_j & \text{otherwise.} \end{cases}$$

And

$$\mu''_\gamma(t) = \begin{cases} \frac{a_j(v_j-t)+(k_{j+1}-k_j)\gamma^2}{\gamma^3}\phi(\frac{v_j-t}{\gamma}) & t \in (v_j - c\gamma, v_j + c\gamma), \\ 0 & \text{otherwise.} \end{cases}$$

□

Proof of Lemma 2. To find the local maxima/minima in $\mu'_\gamma(t)$, by letting $\mu''_\gamma(t) = 0$, we have

$$t = \begin{cases} \text{does not exist} & a_j = 0, \\ v_j + \gamma^2 q_j & a_j \neq 0. \end{cases}$$

To find the local maxima/minima in $\mu''_\gamma(t)$, by letting $\mu'''_\gamma(t) = 0$, we have

$$t = \begin{cases} v_j & a_j = 0, \\ v_j - \frac{\gamma^2 q_j \pm \gamma\sqrt{4+q_j^2}}{2} & a_j \neq 0. \end{cases}$$

□

5.2 Peak Height Distribution for $z'_\gamma(t)$ and $z''_\gamma(t)$

Proof of Lemma 3. Note that, due to the stationarity of $z_\gamma(t)$, for simplicity, we only consider the case of $t = 0$. We can compute the variances of $z_\gamma^{(d)}(t)$, $d = 0, \ldots, 4$ as follows.

$$\begin{aligned} \mathrm{Var}(z'_\gamma(0)) &= \int_{\mathbb{R}} \frac{s^2}{\xi^6}\phi^2\left(\frac{s}{\xi}\right) ds \\ &= \frac{1}{\xi^6}\frac{\xi}{2\sqrt{\pi}}\int_{\mathbb{R}} s^2 \frac{\sqrt{2}}{\xi}\phi\left(\frac{s}{\xi/\sqrt{2}}\right) ds \\ &= \frac{1}{\xi^6}\frac{\xi}{2\sqrt{\pi}}\frac{\xi^2}{2} = \frac{1}{4\sqrt{\pi}\xi^3}. \end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(z_\gamma''(0)) &= \int_{\mathbb{R}} \frac{1}{\xi^6}\phi^2\left(\frac{s}{\xi}\right) ds - 2\int_{\mathbb{R}} \frac{s^2}{\xi^8}\phi^2\left(\frac{s}{\xi}\right) ds + \int_{\mathbb{R}} \frac{s^4}{\xi^{10}}\phi^2\left(\frac{s}{\xi}\right) ds \\
&= \frac{1}{\xi^4}\frac{1}{2\sqrt{\pi}\xi} - \frac{2}{\xi^2}\frac{1}{4\sqrt{\pi}\xi^3} + \frac{1}{\xi^{10}}\frac{\xi}{2\sqrt{\pi}}\int_{\mathbb{R}} s^4 \frac{\xi}{\sqrt{2}}\phi\left(\frac{s}{\xi/\sqrt{2}}\right) ds \\
&= \frac{1}{\xi^{10}}\frac{\xi}{2\sqrt{\pi}}\frac{3\xi^4}{4} = \frac{3}{8\sqrt{\pi}\xi^5}.
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(z_\gamma'''(0)) &= 9\int_{\mathbb{R}} \frac{s^2}{\xi^{10}}\phi^2\left(\frac{s}{\xi}\right) ds - 6\int_{\mathbb{R}} \frac{s^4}{\xi^{12}}\phi^2\left(\frac{s}{\xi}\right) ds + \int_{\mathbb{R}} \frac{s^6}{\xi^{14}}\phi^2\left(\frac{s}{\xi}\right) ds \\
&= \frac{9}{\xi^4}\frac{1}{4\sqrt{\pi}\xi^3} - \frac{6}{\xi^2}\frac{3}{8\sqrt{\pi}\xi^5} + \frac{1}{\xi^{14}}\frac{\xi}{2\sqrt{\pi}}\int_{\mathbb{R}} s^6 \frac{\sqrt{2}}{\xi}\phi\left(\frac{s}{\xi/\sqrt{2}}\right) ds \\
&= \frac{1}{\xi^{14}}\frac{\xi}{2\sqrt{\pi}}\frac{15\xi^6}{8} = \frac{15}{16\sqrt{\pi}\xi^7}.
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(z_\gamma^{(4)}(0)) &= 9\int_{\mathbb{R}} \frac{1}{\xi^{10}}\phi^2\left(\frac{s}{\xi}\right) ds - 36\int_{\mathbb{R}} \frac{s^2}{\xi^{12}}\phi^2\left(\frac{s}{\xi}\right) ds + 42\int_{\mathbb{R}} \frac{s^4}{\xi^{14}}\phi^2\left(\frac{s}{\xi}\right) ds \\
&\quad - 12\int_{\mathbb{R}} \frac{s^6}{\xi^{16}}\phi^2\left(\frac{s}{\xi}\right) ds + \int_{\mathbb{R}} \frac{s^8}{\xi^{18}}\phi^2\left(\frac{s}{\xi}\right) ds \\
&= \frac{1}{\xi^9}\left(\frac{9}{2\sqrt{\pi}} - 36\times\frac{1}{4\sqrt{\pi}} + 42\times\frac{3}{8\sqrt{\pi}} - 12\times\frac{15}{16\sqrt{\pi}} + \frac{1}{2\sqrt{\pi}}\times\frac{105}{16}\right) \\
&= \frac{105}{32\sqrt{\pi}\xi^9}.
\end{aligned}$$

□

5.3 FDR Control and Power Consistency for Type I Change Points

5.3.1 Supporting Results for FDR Control and Power Consistency

To prove the FDR control and power consistent of our method, we need the theoretical results in this section, in which we borrow the notations defined in Schwartzman *et al.* [61].

Lemma A1. *Assume that there exist a universal δ such that $I_j^{\mathrm{mode}} := \{t \in U(L) : |t - v_j| \le \delta \ll \gamma\} \subset S_j$ for all j. Suppose that $q = \sup_j |q_j|$ is sufficiently small, then*

(1) $M_\gamma = \inf M_{j,\gamma} > 0$ *where* $M_{j,\gamma} = \frac{\mu'_\gamma(\tau_{j,\gamma}) - k(\tau_{j,\gamma})}{a_j}$.

(2) $C_\gamma = \inf_j C_{j,\gamma} > 0$ *and* $D_\gamma = \inf_j D_{j,\gamma} > 0$, *where* $C_{j,\gamma} = \frac{1}{|a_j|} \inf_{I_j^{\text{side}}} |\mu''_\gamma(t)|$, $I_j^{\text{side}} = S_{j,\gamma} \setminus I_j^{\text{mode}}$, *and* $D_{j,\gamma} = \frac{1}{|a_j|} \inf_{I_j^{\text{mode}}} |\mu'''_\gamma(t)|$.

Proof.

$$\begin{aligned} M_{j,\gamma} &= \frac{\mu'_\gamma(\tau_{j,\gamma}) - k(\tau_{j,\gamma})}{a_j} \\ &= \frac{1}{a_j}[\frac{a_j}{\gamma}\phi(\frac{v_j - t}{\gamma}) - (k_{j+1} - k_j)\Phi(\frac{v_j - t}{\gamma}) + k_{j+1}] - \frac{k(\tau_{j,\gamma})}{a_j} \\ &\geq \frac{1}{\gamma}\phi(\frac{v_j - t}{\gamma}) + q_j[\mathbb{1}_{q_j \leq 0} - \Phi(\frac{v_j - t}{\gamma})]. \end{aligned}$$

As q is sufficiently small and $t \in S_{j,\gamma} = (v_j - c\gamma, v_j + c\gamma)$, thus there exist a very small universal $\varepsilon_1 > 0$ such that $M_{j,\gamma} \geq \frac{\phi(c)}{\gamma} - \varepsilon_1 > 0$. Therefore, $M_\gamma = \inf M_{j,\gamma} \geq \frac{\phi(c)}{\gamma} - \varepsilon_1 > 0$.

$$\begin{aligned} C_{j,\gamma} &= \frac{1}{|a_j|} \inf_{I_j^{\text{side}}} |\mu''_\gamma(t)| \\ &= \inf_{I_j^{\text{side}}} |\frac{(v_j - t) + q_j\gamma^2}{\gamma^3}|\phi(\frac{v_j - t}{\gamma}). \end{aligned}$$

As $t \in I_j^{\text{side}} = (v_j - c\gamma, v_j - \delta) \cup (v_j + \delta, v_j + c\gamma)$, then $\delta < |v_j - t| < c\gamma$, thus there exist a universal $\varepsilon_2 > 0$ and $\varepsilon_2 < \delta$ such that $C_{j,\gamma} \geq \frac{\delta - \varepsilon_2}{\gamma^3}\phi(c)\frac{1}{2\Phi(c)-1} > 0$. Therefore, $C_\gamma = \inf_j C_{j,\gamma} > 0$.

In addition,

$$\begin{aligned} D_{j,\gamma} &= \frac{1}{|a_j|} \inf_{I_j^{\text{mode}}} |\mu'''_\gamma(t)| \\ &= \inf_{I_j^{\text{mode}}} |\frac{(v_j - t)^2 + q_j\gamma^2 - \gamma^2}{\gamma^5}|\phi(\frac{v_j - t}{\gamma}). \end{aligned}$$

As $0 < \delta \ll \gamma$ and $t \in I_j^{\text{mode}} = [v_j - \delta, v_j + \delta]$, then $|v_j - t| \leq \delta$, thus there exist a very small universal $\varepsilon_3 > 0$ such that $D_{j,\gamma} \geq \frac{\gamma^2 - \delta^2 - \varepsilon_3}{\gamma^5}\phi(\frac{\delta}{\gamma})\frac{1}{2\Phi(c)-1} > 0$. Therefore, $D_\gamma = \inf_j D_{j,\gamma} > 0$. □

Lemma A2. *Suppose that $q = \sup_j |q_j|$ is sufficiently small,*

then $\tau_{j,\gamma} = v_j + \gamma^2 q_j \in I_j^{\text{mode}}$ for all j. Let

$$\tau_{j,\gamma} = \begin{cases} \tau_{j,\gamma}^+ & \text{if } \mu_\gamma'(\tau_{j,\gamma}) \text{ is a local maximum} \\ \tau_{j,\gamma}^- & \text{if } \mu_\gamma'(\tau_{j,\gamma}) \text{ is a local minimum,} \end{cases} \quad I_j^{\text{mode}} = \begin{cases} I_j^{\text{mode}+} & \text{if } \tau_{j,\gamma} = \tau_{j,\gamma}^+ \\ I_j^{\text{mode}-} & \text{if } \tau_{j,\gamma} = \tau_{j,\gamma}^-. \end{cases}$$

Define $\sigma_1 = sd(z_\gamma'(t))$, $\sigma_2 = sd(z_\gamma''(t))$ and $\sigma_3 = sd(z_\gamma'''(t))$, then for any threshold u,

(1)

$$P(\#\{t \in \tilde{T} \cap I_j^{\text{side}}\} = 0) \geq 1 - \exp(-\frac{a_j^2 C_{j,\gamma}^2}{2\sigma_2^2}).$$

(2)

$$P(\#\{t \in \tilde{T} \cap I_j^{\text{mode}}\} = 1) \geq -1 + 2\Phi(\frac{|a_j| C_{j,\gamma}}{\sigma_2}) - 2\exp(-\frac{a_j^2 D_{j,\gamma}^2}{2\sigma_3^2}).$$

(3)

$$\begin{aligned}
&P(\#\{t \in \tilde{T}_1^+ \cap I_j^{\text{mode}+} : y_\gamma'(t) - k(t) > u\} = 1) \geq 1 - \exp(-\frac{a_j^2 D_{j,\gamma}^2}{2\sigma_3^2}) - \Phi(\frac{u - |a_j| M_{j,\gamma}}{\sigma_1}), \\
&P(\#\{t \in \tilde{T}_1^- \cap I_j^{\text{mode}-} : y_\gamma'(t) - k(t) < -u\} = 1) \geq 1 - \exp(-\frac{a_j^2 D_{j,\gamma}^2}{2\sigma_3^2}) + \Phi(\frac{u - |a_j| M_{j,\gamma}}{\sigma_1}), \\
&P(\#\{t \in \tilde{T}_1^+ \cap I_j^{\text{mode}-} : y_\gamma'(t) - k(t) > u\} = 0) \geq 1 - \exp(-\frac{a_j^2 D_{j,\gamma}^2}{2\sigma_3^2}), \\
&P(\#\{t \in \tilde{T}_1^- \cap I_j^{\text{mode}+} : y_\gamma'(t) - k(t) < -u\} = 0) \geq 1 - \exp(-\frac{a_j^2 D_{j,\gamma}^2}{2\sigma_3^2}).
\end{aligned}$$

Proof. (1). As the probability that there are no local extrema of $y_\gamma'(t)$ in I_j^{side} is greater than the probability that $|y_\gamma''(t)| > 0$ for all $t \in I_j^{\text{side}}$, thus

$$\begin{aligned}
P(\#\{t \in \tilde{T} \cap I_j^{\text{side}}\} = 0) &\geq P(\inf_{I_j^{\text{side}}} |y_\gamma''(t)| > 0) \\
&\geq P(\sup_{I_j^{\text{side}}} |z_\gamma''(t)| < \inf_{I_j^{\text{side}}} |\mu_\gamma''(t)|) \\
&= 1 - P(\sup_{I_j^{\text{side}}} |z_\gamma''(t)| > \inf_{I_j^{\text{side}}} |\mu_\gamma''(t)|) \\
&\geq 1 - \exp(-\frac{a_j^2 C_{j,\gamma}^2}{2\sigma_2^2}).
\end{aligned} \tag{5.4}$$

The last line follows from Borell-TIS ineuqality.

(2). The probability that $y'_\gamma(t)$ has no local maxima in I_j^{mode} is less than the probability that $y''_\gamma(v_j - \delta) \leq 0$ or $y''_\gamma(v_j + \delta) \geq 0$. Thus the probability of no local maxima in $I_j^{\text{mode}+}$ is bounded above by

$$\begin{aligned}
P(\#\{t \in \tilde{T}^+ \cap I_j^{\text{mode}+}\} = 0) &\leq P(y''_\gamma(v_j - \delta) \leq 0 \cup y''_\gamma(v_j + \delta) \geq 0) \\
&\leq P(y''_\gamma(v_j - \delta) \leq 0) + P(y''_\gamma(v_j + \delta) \geq 0) \\
&= \Phi(-\frac{\mu''_\gamma(v_j - \delta)}{\sigma_2}) + \Phi(\frac{\mu''_\gamma(v_j + \delta)}{\sigma_2}) \\
&= 1 - \Phi(\frac{\mu''_\gamma(v_j - \delta)}{\sigma_2}) + 1 - \Phi(-\frac{\mu''_\gamma(v_j + \delta)}{\sigma_2}) \\
&\leq 2 - 2\Phi(\frac{|a_j|C_{j,\gamma}}{\sigma_2}).
\end{aligned}$$

The last line holds because $y''_\gamma(t) \sim N(\mu''_\gamma(t), \sigma_2^2)$ and for $t \in I_j^{\text{mode}+}$, $\mu''_\gamma(v_j - \delta) > |a_j|C_{j,\gamma} > 0$ and $-\mu''_\gamma(v_j + \delta) > |a_j|C_{j,\gamma} > 0$. Similarly, $P(\#\{t \in \tilde{T}^- \cap I_j^{\text{mode}-}\} = 0) \leq 2 - 2\Phi(\frac{|a_j|C_{j,\gamma}}{\sigma_2})$.

The probability that $y'_\gamma(t)$ has no local minima in $I_j^{\text{mode}+}$ is greater than the probability that $y'''_\gamma(t) < 0$ for all $t \in I_j^{\text{mode}+}$. Thus,

$$\begin{aligned}
P(\#\{t \in \tilde{T}^- \cap I_j^{\text{mode}+}\} = 0) &\geq P(\sup_{I_j^{\text{mode}+}} y'''_\gamma(t) < 0) \\
&\geq P(\sup_{I_j^{\text{mode}+}} z'''_\gamma < -\sup_{I_j^{\text{mode}+}} \mu'''_\gamma(t)) \\
&= 1 - P(\sup_{I_j^{\text{mode}+}} z'''_\gamma(t) \geq \inf_{I_j^{\text{mode}+}} -\mu'''_\gamma(t)) \\
&\geq 1 - P(|\sup_{I_j^{\text{mode}+}} z'''_\gamma(t)| \geq \inf_{I_j^{\text{mode}+}} -\mu'''_\gamma(t)) \\
&\geq 1 - \exp(-\frac{a_j^2 D_{j,\gamma}^2}{2\sigma_3^2}).
\end{aligned}$$

The last line holds because $\mu'''_\gamma(t)) < 0$ for all $t \in I_j^{\text{mode}+}$. Similarly, $P(\#\{t \in \tilde{T}^+ \cap I_j^{\text{mode}-}\} = 0) \geq 1 - \exp(-\frac{a_j^2 D_{j,\gamma}^2}{2\sigma_3^2})$.

On the other hand, the probability that $y'_\gamma(t)$ has at least two local maxima in I_j^{mode} is less than the probability that $y'''_\gamma(t) > 0$ for some $t \in I_j^{\text{mode}}$ and $y'''_\gamma(t) < 0$ for some other $t \in I_j^{\text{mode}}$. Thus

$$
\begin{aligned}
&P(\#\{t \in \tilde{T}^+ \cap I_j^{\text{mode}}\} \geq 2) \\
&\leq P(\sup_{I_j^{\text{mode}}} y'''_\gamma(t) > 0 \cap \inf_{I_j^{\text{mode}}} y'''_\gamma(t) < 0) \\
&\leq P(\sup_{I_j^{\text{mode}+}} y'''_\gamma(t) > 0) \wedge P(\inf_{I_j^{\text{mode}-}} y'''_\gamma(t) < 0) \\
&\leq P(\sup_{I_j^{\text{mode}+}} z'''_\gamma(t) > \inf_{I_j^{\text{mode}+}} -\mu'''_\gamma(t)) \wedge P(\sup_{I_j^{\text{mode}-}} z'''_\gamma(t) > \inf_{I_j^{\text{mode}-}} \mu'''_\gamma(t)) \\
&\leq \exp(-\frac{a_j^2 D_{j,\gamma}^2}{2\sigma_3^2}).
\end{aligned}
$$

The last line holds because $\mu'''_\gamma(t) < 0$ for all $t \in I_j^{\text{mode}+}$ and $\mu'''_\gamma(t) > 0$ for all $t \in I_j^{\text{mode}-}$. Similarly, $P(\#\{t \in \tilde{T}^- \cap I_j^{\text{mode}}\} \geq 2) \leq \exp(-\frac{a_j^2 D_{j,\gamma}^2}{2\sigma_3^2})$. Therefore,

$$
P(\#\{t \in \tilde{T} \cap I_j^{\text{mode}}\} \geq 2) \leq \exp(-\frac{a_j^2 D_{j,\gamma}^2}{2\sigma_3^2}).
$$

The probability that $y'_\gamma(t)$ has only one local maximum in $I_j^{\text{mode}+}$ is calculated as

$$
\begin{aligned}
&P(\#\{t \in \tilde{T}^+ \cap I_j^{\text{mode}+}\} = 1) \\
&= 1 - P(\#\{t \in \tilde{T}^+ \cap I_j^{\text{mode}+}\} = 0) - P(\#\{t \in \tilde{T}^+ \cap I_j^{\text{mode}+}\} \geq 2) \\
&\geq -1 + 2\Phi(\frac{|a_j| C_{j,\gamma}}{\sigma_2}) - \exp(-\frac{a_j^2 D_{j,\gamma}^2}{2\sigma_3^2}).
\end{aligned}
$$

Similarly, $P(\#\{t \in \tilde{T}^- \cap I_j^{\text{mode}-}\} = 1) \geq -1 + 2\Phi(\frac{|a_j| C_{j,\gamma}}{\sigma_2}) - \exp(-\frac{a_j^2 D_{j,\gamma}^2}{2\sigma_3^2})$.

The probability that $y'_\gamma(t)$ has only one local extreme in $I_j^{\text{mode}+}$ is greater than the probability that $y'_\gamma(t)$ has only one local maximum and zero local minimum.

$$\begin{aligned}
&P(\#\{t \in \tilde{T} \cap I_j^{\text{mode}+}\} = 1) \\
&\geq P(\#\{t \in \tilde{T}^+ \cap I_j^{\text{mode}+}\} = 1 \cap \#\{t \in \tilde{T}^- \cap I_j^{\text{mode}+}\} = 0) \\
&\geq P(\#\{t \in \tilde{T}^+ \cap I_j^{\text{mode}+}\} = 1) + P(\#\{t \in \tilde{T}^- \cap I_j^{\text{mode}+}\} = 0) - 1 \\
&\geq -1 + 2\Phi(\frac{|a_j|C_{j,\gamma}}{\sigma_2}) - 2\exp(-\frac{a_j^2 D_{j,\gamma}^2}{2\sigma_3^2}).
\end{aligned}$$

Similarly, $P(\#\{t \in \tilde{T} \cap I_j^{\text{mode}-}\} = 1) \geq -1 + 2\Phi(\frac{|a_j|C_{j,\gamma}}{\sigma_2}) - 2\exp(-\frac{a_j^2 D_{j,\gamma}^2}{2\sigma_3^2})$. Therefore,

$$P(\#\{t \in \tilde{T} \cap I_j^{\text{mode}}\} = 1) \geq -1 + 2\Phi(\frac{|a_j|C_{j,\gamma}}{\sigma_2}) - 2\exp(-\frac{a_j^2 D_{j,\gamma}^2}{2\sigma_3^2}).$$

(3). The probability that at least two local maxima of $y'_\gamma(t)$ in I_j^{mode} is exceed $u + k(t)$ is less than the probability that $y'_\gamma(t)$ has at least two maxima in I_j^{mode}.

$$\begin{aligned}
&P(\#\{t \in \tilde{T}^+ \cap I_j^{\text{mode}} : y'_\gamma(t) - k(t) > u\} \geq 2) \\
&\leq P(\#\{t \in \tilde{T}^+ \cap I_j^{\text{mode}}\} \geq 2) \\
&\leq \exp(-\frac{a_j^2 D_{j,\gamma}^2}{2\sigma_3^2}).
\end{aligned}$$

Similarly, $P(\#\{t \in \tilde{T}^- \cap I_j^{\text{mode}} : y'_\gamma(t) - k(t) < -u\} \geq 2) \leq \exp(-\frac{a_j^2 D_{j,\gamma}^2}{2\sigma_3^2})$.

On the other hand, the probability that no local maxima of $y'_\gamma(t)$ in I_j^{mode} is exceed $u + k(t)$ is less than the probability that $y'_\gamma(t) - k(t)$ is below u anywhere in I_j^{mode}, that is,

$$\begin{aligned}
&P(\#\{t \in \tilde{T}^+ \cap I_j^{\text{mode}+} : y'_\gamma(t) - k(t) > u\} = 0) \\
&\leq P(y'_\gamma(t) - k(t) \leq u \text{ for } \forall t \in I_j^{\text{mode}+}) \\
&\leq \Phi(\frac{u + k(\tau_{j,\gamma}^+) - \mu'_\gamma(\tau_{j,\gamma}^+)}{\sigma_1}) \\
&= \Phi(\frac{u - |a_j|M_{j,\gamma}}{\sigma_1}).
\end{aligned}$$

The last line holds because $a_j M_{j,\gamma} = \mu'_\gamma(\tau^+_{j,\gamma}) - k(\tau^+_{j,\gamma}) > 0$ as $a_j > 0$ for $\mu'_\gamma(\tau_{j,\gamma})$ is the local maximum. Similarly, $P(\#\{t \in \tilde{T}^- \cap I_j^{\text{mode}-} : y'_\gamma(t) - k(t) < -u\} = 0) \leq \Phi(\frac{u-|a_j|M_{j,\gamma}}{\sigma_1})$.

Therefore, the probability that only one local maximum of $y'_\gamma(t)$ in $I_j^{\text{mode}+}$ is exceed $u + k(t)$ is

$$\begin{aligned} & P(\#\{t \in \tilde{T}^+ \cap I_j^{\text{mode}+} : y'_\gamma(t) - k(t) > u\} = 1) \\ &= 1 - P(\#\{t \in \tilde{T}^+ \cap I_j^{\text{mode}+} : y'_\gamma(t) - k(t) > u\} = 0) \\ &\quad - P(\#\{t \in \tilde{T}^+ \cap I_j^{\text{mode}+} : y'_\gamma(t) - k(t) > u\} \geq 2) \\ &\geq 1 - \Phi(\frac{\mu - |a_j|M_{j,\gamma}}{\sigma_1}) - \exp(-\frac{a_j^2 D_{j,\gamma}^2}{2\sigma_3^2}). \end{aligned} \tag{5.5}$$

Similarly, $P(\#\{t \in \tilde{T}^- \cap I_j^{\text{mode}-} : y'_\gamma(t) - k(t) < -u\} = 1) \geq 1 - \Phi(\frac{\mu-|a_j|M_{j,\gamma}}{\sigma_1}) - \exp(-\frac{a_j^2 D_{j,\gamma}^2}{2\sigma_3^2})$.

The probability that no local maxima of $y'_\gamma(t)$ in $I_j^{\text{mode}-}$ is exceed $u+k(t)$ is greater than the probability that $y'_\gamma(t)$ has no local maxima in $I_j^{\text{mode}-}$, that is,

$$\begin{aligned} P(\#\{t \in \tilde{T}^+ \cap I_j^{\text{mode}-} : y'_\gamma(t) - k(t) > u\} = 0) &\geq P(\#\{t \in \tilde{T}^+ \cap I_j^{\text{mode}-}\} = 0) \\ &\geq 1 - \exp(-\frac{a_j^2 D_{j,\gamma}^2}{2\sigma_3^2}). \end{aligned} \tag{5.6}$$

Similarly, $P(\#\{t \in \tilde{T}^- \cap I_j^{\text{mode}+} : y'_\gamma(t) - k(t) < -u\} = 0) \geq 1 - \exp(-\frac{a_j^2 D_{j,\gamma}^2}{2\sigma_3^2})$. □

Lemma A3. *Let $W_{1,\gamma} = \#\{t \in \tilde{T}_1 \cap \mathbb{S}_{1,\gamma}\}$ be the number of local extrema in the set $\mathbb{S}_{1,\gamma}$ and $W_{1,\gamma}(u) = W^+_{1,\gamma}(u) + W^-_{1,\gamma}(u)$, where $\tilde{W}^+_{1,\gamma}(u) = \#\{t \in \tilde{T}_1^+ \cap \mathbb{S}_{1,\gamma} : y_\gamma(t) - k(t) > u\}$ and $\tilde{W}^-_{1,\gamma}(u) = \#\{t \in \tilde{T}_1^- \cap \mathbb{S}_{1,\gamma} : y_\gamma(t) - k(t) < -u\}$. Under conditions (C1)–(C3), there exists some $\eta > 0$ such that*

(1)

$$P(\#\{t \in \tilde{T}_1 \cap \mathbb{T}_{1,\gamma}\} \geq 1) = o(\exp(-\eta a^2)).$$

(2)

$$P(W_{1,\gamma} = J) = P(\#\{t \in \tilde{T}_1 \cap \mathbb{S}_{1,\gamma}\} = J) = 1 - o(\exp(-\eta a^2)).$$

(3)

$$P(W_{1,\gamma}(u) = J) = P(\tilde{m}^+_{1,\gamma}(u) + \tilde{m}^-_{1,\gamma}(u) = J) = 1 - o(\exp(-\eta a^2)).$$

(4)

$$W_{1,\gamma}/L = A_1 + o_p(1).$$

(5)

$$W_{1,\gamma}(u)/W_{1,\gamma} = 1 + o_p(1).$$

Proof. (1). As the transition region for peak j, $T_{j,\gamma} = S_{j,\gamma} \setminus S_j$, is a subset of I_j^{side}, thus $\mathbb{T}_\gamma = \cup_{j=1}^J T_{j,\gamma} \subset \cup_{j=1}^J I_j^{\text{side}}$. Then for some $\eta > 0$,

$$\begin{aligned}
P(\#\{\tilde{T} \cap \mathbb{T}_\gamma \geq 1) &\leq P(\#\{t \in \tilde{T} \cap \cup_{j=1}^J I_j^{\text{side}}\} \geq 1) \\
&= P(\cup_{j=1}^J \#\{t \in \tilde{T} \cap I_j^{\text{side}}\} \geq 1) \\
&\leq \sum_{j=1}^J [1 - P(\#\{t \in \tilde{T} \cap I_j^{\text{side}}\} = 0)] \\
&\leq \sum_{j=1}^J \exp(-\frac{a_j^2 C_{j,\gamma}^2}{2\sigma_2^2}) \leq \sum_{j=1}^J \exp(-\frac{a^2 C_\gamma^2}{2\sigma_2^2}) \\
&= \frac{J}{L} L \exp(-\frac{a^2 C_\gamma^2}{2\sigma_2^2}) = o(\exp(-\eta a^2)).
\end{aligned} \tag{5.7}$$

The last line holds as long as $\eta > \frac{C_\gamma^2}{2\sigma_2^2}$,

$$L \exp(-\frac{a^2 C_\gamma^2}{2\sigma_2^2}) = \exp\{a^2(\frac{\log L}{a^2} - \frac{C_\gamma^2}{2\sigma_2^2})\} = O(\exp(-\frac{a^2 C_\gamma^2}{2\sigma_2^2})) = o(\exp(-\eta a^2)).$$

(2).

$$
\begin{aligned}
&P(\#\{t \in \tilde{T} \cap \mathbb{S}_{1,\gamma} = J) \geq P[\cap_{j=1}^{J}(\#\{t \in \tilde{T} \cap I_j^{\text{mode}}\} = 1 \cap \#\{t \in \tilde{T} \cap I_j^{\text{side}}\} = 0)] \\
&\geq 1 - \sum_{j=1}^{J}[1 - P(\#\{t \in \tilde{T} \cap I_j^{\text{mode}}\} = 1 \cap \#\{t \in \tilde{T} \cap I_j^{\text{side}}\} = 0)] \\
&\geq 1 - \sum_{j=1}^{J}[2 - P(\#\{t \in \tilde{T} \cap I_j^{\text{mode}}\} = 1) - P(\#\{t \in \tilde{T} \cap I_j^{\text{side}}\} = 0)] \\
&= 1 - \sum_{j=1}^{J}[2 - 2\Phi(\frac{|a_j|C_{j,\gamma}}{\sigma_2}) + \exp(-\frac{a_j^2 D_{j,\gamma}^2}{2\sigma_3^2}) + \exp(-\frac{a_j^2 C_{j,\gamma}^2}{2\sigma_2^2})] \\
&\geq 1 - \frac{J}{L}\{2L[1 - \Phi(\frac{aC_\gamma}{\sigma_2})] + L\exp(-\frac{a^2 C_\gamma^2}{2\sigma_2^2}) + L\exp(-\frac{a^2 D_\gamma^2}{2\sigma_3^2})\} \\
&= 1 - o(\exp(-\eta a^2)).
\end{aligned}
\tag{5.8}
$$

As $L[1-\Phi(Ka)] \leq L\phi(Ka)/(Ka)$ for any $K > 0$, the last line holds as long as $\eta > \frac{C_\gamma^2}{2\sigma_2^2}$ and $\eta > \frac{D_\gamma^2}{2\sigma_3^2}$.

(3). Define the index set of breaks as $\mathbb{J} = \{1, 2, \ldots, J\}$. Let $\mathbb{J}^+ = \{j \in \mathbb{J} : \mu'_\gamma(\tau_{j,\gamma})$ is a local maximum$\}$ and $\mathbb{J}^- = \{j \in \mathbb{J} : \mu'_\gamma(\tau_{j,\gamma})$ is a local minimum$\}$. Note that $\mathbb{J} = \mathbb{J}^+ \cup \mathbb{J}^-$. Let

$$
\begin{aligned}
B_{j,0}:&\ \#\{t \in \tilde{T} \cap I_j^{\text{mode}}\} = 0, \\
B_{j,1}:&\ \#\{t \in \tilde{T}^+ \cap I_j^{\text{mode}+} : y'_\gamma(t) - k(t) > u\} = 1, \\
B_{j,2}:&\ \#\{t \in \tilde{T}^- \cap I_j^{\text{mode}+} : y'_\gamma(t) - k(t) < -u\} = 0, \\
B_{j,3}:&\ \#\{t \in \tilde{T}^+ \cap I_j^{\text{mode}-} : y'_\gamma(t) - k(t) > u\} = 0, \\
B_{j,4}:&\ \#\{t \in \tilde{T}^- \cap I_j^{\text{mode}-} : y'_\gamma(t) - k(t) < -u\} = 1.
\end{aligned}
$$

$$
\begin{aligned}
P(\tilde{m}_{1,\gamma}(u)=J) &\geq P\{[\cap_{j\in\mathbb{J}^+}(B_{j,1}\cap B_{j,2}\cap B_{j,0})]\cap[\cap_{j\in\mathbb{J}^-}(B_{j,3}\cap B_{j,4}\cap B_{j,0})]\}\\
&\geq P(\cap_{j\in\mathbb{J}^+}(B_{j,1}\cap B_{j,2}\cap B_{j,0}))+P(\cap_{j\in\mathbb{J}^-}(B_{j,3}\cap B_{j,4}\cap B_{j,0}))-1\\
&\geq 1-\sum_{j\in\mathbb{J}^+}[1-P(B_{j,1}\cap B_{j,2}\cap B_{j,0})]-\sum_{j\in\mathbb{J}^-}[1-P(B_{j,3}\cap B_{j,4}\cap B_{j,0})]\\
&\geq 1-\sum_{j\in\mathbb{J}^+}[\exp(-\frac{a_j^2C_{j,\gamma}^2}{2\sigma_2^2})+2\exp(-\frac{a_j^2D_{j,\gamma}^2}{2\sigma_3^2})+\Phi(\frac{u-|a_j|M_{j,\gamma}}{\sigma_1})]\\
&\quad+1-\sum_{j\in\mathbb{J}^-}[\exp(-\frac{a_j^2C_{j,\gamma}^2}{2\sigma_2^2})+2\exp(-\frac{a_j^2D_{j,\gamma}^2}{2\sigma_3^2})+\Phi(\frac{u-|a_j|M_{j,\gamma}}{\sigma_1})]-1\\
&=1-\sum_{j\in\mathbb{J}}[\exp(-\frac{a_j^2C_{j,\gamma}^2}{2\sigma_2^2})+2\exp(-\frac{a_j^2D_{j,\gamma}^2}{2\sigma_3^2})+1-\Phi(\frac{-u+|a_j|M_{j,\gamma}}{\sigma_1})]\\
&=1-o(\exp(-\eta a^2)).
\end{aligned}
$$

The last line holds for any $\eta > \frac{C_\gamma^2}{2\sigma_2^2}$ and $\eta > \frac{D_\gamma^2}{2\sigma_3^2}$. □

As $\mu'_\gamma(t)$ is piecewise constant, the detection of type I breaks is the same as the detection of change points in piecewise constant signal ([18]). For the proofs of the FDR control and power consistency for type I breaks, one can see the arguments in [18].

5.4 FDR Control and Power Consistency for Type II Change Points

In this section, we consider the case that the linear model contains only the type II change points. For simplicity, we omit the subscript II for all the related notations.

5.4.1 FDR Control

Proof of Theorem 5. Let $V(u) = \#\{t \in \tilde{T}(u) \cap \mathbb{N}\}$ and $W(u) = \#\{t \in \tilde{T}(u) \cap \mathbb{S}\}$, by the definition of FDR and Lemma 12 in [61], for any fixed u,

$$
\mathrm{FDR}(u) = E[\frac{V(u)}{V(u)+W(u)}] = E[\frac{V(u)/L}{V(u)/L+W(u)/L}]. \tag{5.9}
$$

Notice that

$$\begin{aligned} P(V_\gamma(u) = V(u)) &= 1 - P(V_\gamma(u) \neq V(u)) \\ &= 1 - P(\#\{t \in \mathbb{T}_\gamma(u) \neq 0\}) \\ &= 1 - o(\exp(-\eta a^2)) \to 1. \end{aligned}$$

Similarly, $W(u) = \tilde{W}_\gamma(u) + o_p(1)$. Then we have

$$\begin{aligned} \frac{V(u)}{L} &= \frac{V(u)}{V_{1,\gamma}(u)} \frac{V_\gamma(u)}{L} = (1 + 2c\gamma A) E[\tilde{m}_\gamma(U(1), u)] + o_p(1), \\ \frac{W(u)}{L} &= \frac{W_1(u)}{W_\gamma(u)} \frac{W_\gamma(u)}{L} = A + o_p(1). \end{aligned}$$

Hence,

$$\frac{V(u)/L}{V(u)/L + W(u)/L} = \frac{(1 + 2c\gamma A) E[\tilde{m}_\gamma(U(1), u)]}{A + (1 + 2c\gamma A) E[\tilde{m}_\gamma(U(1), u)]} + o_p(1) \leq 1.$$

By the Dominated Convergence Theorem (DCT),

$$\begin{aligned} \lim E[\frac{V(u)/L}{V(u)/L + W(u)/L}] &= E[\lim \frac{V(u)/L}{V(u)/L + W(u)/L}] \\ &= \frac{(1 + 2c\gamma A) E[\tilde{m}_\gamma(U(1), u)]}{A + (1 + 2c\gamma A) E[\tilde{m}_\gamma(U(1), u)]}. \end{aligned}$$

That is, $\lim_{a,L\to\infty} \mathrm{FDR}(u) = \frac{(1+2c\gamma A)E[\tilde{m}_\gamma(U(1),u)]}{A+(1+2c\gamma A)E[\tilde{m}_\gamma(U(1),u)]}$, completing the part (i) in Theorem 3.

Let $\tilde{G}(u) = \#\{t \in \tilde{T}(u)\}/\#\{t \in \tilde{T}\}$ be the empirical marginal right cdf of $y_\gamma(t)$ given $t \in \tilde{T}$. Then the BH threshold $\tilde{u}_{\mathrm{BH}}$ satisfies $\alpha\tilde{G}(\tilde{u}_{\mathrm{BH}}) = k\alpha/\tilde{m} = F_\gamma(\tilde{u}_{\mathrm{BH}})$, so $\tilde{u}_{\mathrm{BH}}$ is the largest u that solves the equation

$$\alpha\tilde{G}(u) = F_{z''_\gamma}(u). \tag{5.10}$$

We first find the limit of $\tilde{G}(u)$.

$$\tilde{G}(u) = \frac{V_{1,\gamma}(u) + W_\gamma(u)}{V_\gamma + W_\gamma} = \frac{V_\gamma(u)}{V_\gamma} \frac{V_\gamma}{V_\gamma + W_\gamma} + \frac{W_\gamma(u)}{W_\gamma} \frac{W_\gamma}{V_\gamma + W_\gamma} \tag{5.11}$$

By Lemma 8 in [61],

$$\frac{V_\gamma(u)/L}{V_\gamma/L} \xrightarrow{P} \frac{E[V_\gamma(u)]}{E[V_\gamma]} = F_{z''_\gamma}(u).$$

In addition,

$$\frac{V_\gamma}{V_\gamma + W_\gamma} = \frac{V_\gamma/L}{V_\gamma/L + W_\gamma/L} \xrightarrow{P} \frac{E[\tilde{m}_\gamma(U(1))](1-2c\gamma A)}{E[\tilde{m}_\gamma(U(1))](1-2c\gamma A)+A},$$

and

$$\frac{W_\gamma}{\tilde{V}_\gamma + W_\gamma} = \frac{W_\gamma/L}{V_\gamma/L + W_{1,\gamma}/L} \xrightarrow{P} \frac{A}{E[\tilde{m}_{1,\gamma}(U(1))](1-2c\gamma A)+A}.$$

Combined with the part (4) in Lemma A3, we obtain

$$\tilde{G}(u) \xrightarrow{P} \frac{F_\gamma(u)E[\tilde{m}_\gamma(U(1))](1-2c\gamma A)+A}{E[\tilde{m}_\gamma(U(1))](1-2c\gamma A)+A}.$$

Plugging $\tilde{G}_1(u)$ by its limit in (5.10) and solving for u gives the deterministic solution

$$F_\gamma(u^*_{\mathrm{BH}}) = \frac{\alpha A}{A + E[\tilde{m}_{1,\gamma}(U(1))](1-2c\gamma A)(1-\alpha)}. \tag{5.12}$$

Note that $\tilde{u}_{\mathrm{BH}}$ is the solution of $\alpha\tilde{G}(u) = F_\gamma(u)$ and u^*_{BH} is the solution of $\lim \alpha\tilde{G}(u) = F_\gamma(u)$, as $F_\gamma^{-1}(\cdot)$ is monotonic, $\tilde{u}_{\mathrm{BH}} \xrightarrow{P} u^*_{\mathrm{BH}}$, i.e., for any $\delta > 0$, $P(|\tilde{u}_{\mathrm{BH}} - u^*_{\mathrm{BH}}| \le \delta) = 1$.

For the random threshold $\tilde{u}_{\mathrm{BH}}$,

$$\begin{aligned}
&\mathrm{FDR}(\tilde{u}_{\mathrm{BH}}) = E[\frac{V(\tilde{u}_{\mathrm{BH}})}{V(\tilde{u}_{\mathrm{BH}}) + W(\tilde{u}_{\mathrm{BH}})}] \\
&= E[\frac{V(\tilde{u}_{\mathrm{BH}})}{V(\tilde{u}_{\mathrm{BH}}) + W(\tilde{u}_{\mathrm{BH}})}\mathbb{1}\left(|\tilde{u}_{\mathrm{BH}} - u^*_{\mathrm{BH}}| \le \delta\right)] \\
&\quad + E[\frac{V(\tilde{u}_{\mathrm{BH}})}{V(\tilde{u}_{\mathrm{BH}}) + W(\tilde{u}_{\mathrm{BH}})}\mathbb{1}\left(|\tilde{u}_{\mathrm{BH}} - u^*_{\mathrm{BH}}| > \delta\right)] \\
&= E[\frac{V(\tilde{u}_{\mathrm{BH}})}{V(\tilde{u}_{\mathrm{BH}}) + W(\tilde{u}_{\mathrm{BH}})}\mathbb{1}\left(|\tilde{u}_{\mathrm{BH}} - u^*_{\mathrm{BH}}| \le \delta\right)] + o(1).
\end{aligned}$$

For $\frac{V(\tilde{u}_{\mathrm{BH}})}{V(\tilde{u}_{\mathrm{BH}})+W(\tilde{u}_{\mathrm{BH}})}\mathbb{1}\left(|\tilde{u}_{\mathrm{BH}} - u^*_{\mathrm{BH}}| \le \delta\right)$, we have a lower bound:

$$\begin{aligned}
&\frac{V(\tilde{u}_{\mathrm{BH}})}{V(\tilde{u}_{\mathrm{BH}}) + W(\tilde{u}_{\mathrm{BH}})}\mathbb{1}\left(|\tilde{u}_{\mathrm{BH}} - u^*_{\mathrm{BH}}| \le \delta\right) \ge \\
&\quad \frac{V(u^*_{\mathrm{BH}} + \delta)}{V(u^*_{\mathrm{BH}} - \delta) + W(u^*_{\mathrm{BH}} - \delta)}\mathbb{1}\left(|\tilde{u}_{\mathrm{BH}} - u^*_{\mathrm{BH}}| \le \delta\right),
\end{aligned}$$

and an upper bound:

$$\frac{V(\tilde{u}_{\mathrm{BH}})}{V(\tilde{u}_{\mathrm{BH}})+W(\tilde{u}_{\mathrm{BH}})}\mathbb{1}\left(|\tilde{u}_{\mathrm{BH}}-u^*_{\mathrm{BH}}|\le\delta\right)\le\frac{V(u^*_{\mathrm{BH}}-\delta)}{V(u^*_{\mathrm{BH}}+\delta)+W(u^*_{\mathrm{BH}}+\delta)}.$$

As

$$V(u^*_{\mathrm{BH}}+\delta)/L\to(1+2c\gamma A)E[\tilde{m}_{0,\gamma}(U(1),u^*_{\mathrm{BH}}+\delta)],$$
$$V(u^*_{\mathrm{BH}}-\delta)/L\to(1+2c\gamma A)E[\tilde{m}_{0,\gamma}(U(1),u^*_{\mathrm{BH}}-\delta)],$$

by Kac-Rice formula, $E[\tilde{m}_{0,\gamma}(U(1),u]$ is continuous, thus $E[\tilde{m}_{0,\gamma}(U(1),u^*_{\mathrm{BH}}+\delta)]=E[\tilde{m}_{0,\gamma}(U(1),u^*_{\mathrm{BH}}-\delta)]\to E[\tilde{m}_{0,\gamma}(U(1),u^*_{\mathrm{BH}})]$.

As

$$\begin{aligned}
&E[\frac{V(u^*_{\mathrm{BH}}+\delta)}{V(u^*_{\mathrm{BH}}-\delta)+W(u^*_{\mathrm{BH}}-\delta)}\mathbb{1}\left(|\tilde{u}_{\mathrm{BH}}-u^*_{\mathrm{BH}}|\le\delta\right)]\\
&=E[\frac{V(u^*_{\mathrm{BH}}+\delta)}{V(u^*_{\mathrm{BH}}-\delta)+W(u^*_{\mathrm{BH}}-\delta)}]P(|\tilde{u}_{\mathrm{BH}}-u^*_{\mathrm{BH}}|\le\delta)\\
&=E[\frac{V(u^*_{\mathrm{BH}}+\delta)}{V(u^*_{\mathrm{BH}}-\delta)+W(u^*_{\mathrm{BH}}-\delta)}]\\
&\to E[\frac{V(u^*_{\mathrm{BH}})}{V(u^*_{\mathrm{BH}})+W(u^*_{\mathrm{BH}})}],
\end{aligned}$$

then by DCT,

$$\begin{aligned}
\lim\mathrm{FDR}(\tilde{u}_{\mathrm{BH}})&=\lim E[\frac{V(\tilde{u}_{\mathrm{BH}})}{V(\tilde{u}_{\mathrm{BH}})+W(\tilde{u}_{\mathrm{BH}})}]\\
&=E[\lim\frac{V(\tilde{u}_{\mathrm{BH}})}{V(\tilde{u}_{\mathrm{BH}})+W(\tilde{u}_{\mathrm{BH}})}]=E[\frac{V(u^*_{\mathrm{BH}})}{V(u^*_{\mathrm{BH}})+W(u^*_{\mathrm{BH}})}].
\end{aligned}$$

On the other hand,

$$W(u^*_{\mathrm{BH}}+\delta)/L=A+o_p(1),$$
$$W(u^*_{\mathrm{BH}}-\delta)/L=A+o_p(1).$$

Similar as the arguments in part (i), we have

$$\begin{aligned}\frac{V(u^*_{\mathrm{BH}})}{V(u^*_{\mathrm{BH}})+W(u^*_{\mathrm{BH}})} &= \frac{V(u^*_{\mathrm{BH}})/L}{V(u^*_{\mathrm{BH}})/L+W(u^*_{\mathrm{BH}})/L}\\ &= \frac{\tilde{m}_{0,\gamma}(u^*_{\mathrm{BH}})/L}{\tilde{m}_{0,\gamma}(u^*_{\mathrm{BH}})/L+\tilde{m}_{1,\gamma}(u^*_{\mathrm{BH}})/L}\\ &\to \frac{E[\tilde{m}_{0,\gamma}(u^*_{\mathrm{BH}})]/L}{E[\tilde{m}_{0,\gamma}(u^*_{\mathrm{BH}})]/L+A}\\ &= \frac{F_\gamma(u^*_{\mathrm{BH}})E[\tilde{m}_{0,\gamma}(U(1))](1-2c\gamma A)}{F_\gamma(u^*_{\mathrm{BH}})E[\tilde{m}_{0,\gamma}(U(1))](1-2c\gamma A)+A}\\ &= \alpha\frac{E[\tilde{m}_{0,\gamma}(U(1)](1-2c\gamma A)}{E[\tilde{m}_{0,\gamma}(U(1))](1-2c\gamma A)+A}.\end{aligned} \tag{5.13}$$

Therefore, by DCT,

$$\lim \mathrm{FDR}(\tilde{u}_{\mathrm{BH}}) \to E[\lim \frac{V(u^*_{\mathrm{BH}})}{V(u^*_{\mathrm{BH}})+W(u^*_{\mathrm{BH}})}] = \alpha\frac{E[\tilde{m}_{0,\gamma}(U(1)](1-2c\gamma A)}{E[\tilde{m}_{0,\gamma}(U(1))](1-2c\gamma A)+A}.$$

□

5.4.2 Power Consistency

Proof of Theorem 6. By Lemma A2, for any fixed u,

$$\begin{aligned}\mathrm{Power}_{j,\gamma}(u) &= P(\#\{t\in\tilde{T}(u)\cap S_j\}\geq 1) \geq P(\#\{t\in\tilde{T}(u)\cap I_j^{\mathrm{mode}}\}\geq 1)\\ &\geq 1-\exp(-\frac{a_j^2D_{j,\gamma}^2}{2\sigma^3})+\Phi(\frac{u-|a_j|M_{j,\gamma}}{\sigma_1}) \to 1.\end{aligned}$$

Therefore,

$$\mathrm{Power}_\gamma(u) = \frac{1}{J}\sum_{j=1}^{J}\mathrm{Power}_{j,\gamma}(u) \to 1.$$

For the random threshold $\tilde{u}_{\mathrm{BH}}$ and arbitrary $\delta>0$, we have

$$\begin{aligned}&P(\#\{t\in\tilde{T}(\tilde{u}_{\mathrm{BH}})\cap S_j\}\geq 1)\\ &= P(\#\{t\in\tilde{T}(\tilde{u}_{\mathrm{BH}})\cap S_j\}\geq 1, |\tilde{u}_{\mathrm{BH}}-u^*_{\mathrm{BH}}|\leq\delta)\\ &\quad + P(\#\{t\in\tilde{T}(\tilde{u}_{\mathrm{BH}})\cap S_j\}\geq 1, |\tilde{u}_{\mathrm{BH}}-u^*_{\mathrm{BH}}|>\delta)\end{aligned} \tag{5.14}$$

As

$$P(\#\{t\in\tilde{T}(\tilde{u}_{\mathrm{BH}})\cap S_j\}\geq 1, |\tilde{u}_{\mathrm{BH}}-u^*_{\mathrm{BH}}|>\delta) \leq P(|\tilde{u}_{\mathrm{BH}}-u^*_{\mathrm{BH}}|>\delta) = 0,$$

then

$$
\begin{aligned}
&P(\#\{t \in \tilde{T}(\tilde{u}_{\mathrm{BH}}) \cap S_j\} \geq 1) \\
&= P(\#\{t \in \tilde{T}(\tilde{u}_{\mathrm{BH}}) \cap S_j\} \geq 1, |\tilde{u}_{\mathrm{BH}} - u^*_{\mathrm{BH}}| \leq \delta) \\
&\geq P(\#\{t \in \tilde{T}(u^*_{\mathrm{BH}} + \delta) \cap S_j\} \geq 1, |\tilde{u}_{\mathrm{BH}} - u^*_{\mathrm{BH}}| \leq \delta) \\
&= P(\#\{t \in \tilde{T}(u^*_{\mathrm{BH}} + \delta) \cap S_j\} \geq 1).
\end{aligned}
$$

The last line holds because $P(\#\{t \in \tilde{T}(u^*_{\mathrm{BH}} + \delta) \cap S_j\} \geq 1, |\tilde{u}_{\mathrm{BH}} - u^*_{\mathrm{BH}}| > \delta) = 0$.

On the other hand, similarly

$$
\begin{aligned}
&P(\#\{t \in \tilde{T}(\tilde{u}_{\mathrm{BH}}) \cap S_j\} \geq 1) \\
&\leq P(\#\{t \in \tilde{T}(u^*_{\mathrm{BH}} - \delta) \cap S_j\} \geq 1, |\tilde{u}_{\mathrm{BH}} - u^*_{\mathrm{BH}}| \leq \delta) \\
&= P(\#\{t \in \tilde{T}(u^*_{\mathrm{BH}} - \delta) \cap S_j\} \geq 1).
\end{aligned}
$$

As $\delta > 0$ is arbitrary, let $\delta \to 0$, by the part (3) in Lemma A3,

$$
\begin{aligned}
P(\#\{t \in \tilde{T}(u^*_{\mathrm{BH}} + \delta) \cap S_j\} \geq 1) &\to 1, \\
P(\#\{t \in \tilde{T}(u^*_{\mathrm{BH}} - \delta) \cap S_j\} \geq 1) &\to 1.
\end{aligned}
$$

Therefore,

$$\mathrm{Power}_{j,\gamma}(\tilde{u}_{\mathrm{BH}}) \to 1.$$

Then

$$\mathrm{Power}_{\gamma}(\tilde{u}_{\mathrm{BH}}) = \frac{1}{J}\sum_{j=1}^{J} \mathrm{Power}_{j,\gamma}(\tilde{u}_{\mathrm{BH}}) \to 1.$$

□

Chapter 6

STATISTICAL INFERENCE FOR IMPLICIT NETWORK STRUCTURES USING HUB MODELS

6.1 Hub Model and Variants

6.1.1 Model Setup

First, we review the grouped data structure and then propose a modified version of the hub model, called the *asymmetric hub model*. For a set of n individuals, $V = \{1, \ldots, n\}$, we observe T subsets, called *groups*.

In this paper, groups are treated as a random sample of size T with each group being an observation. Each group is represented by an n length row vector $G^{(t)}$, where

$$G_i^{(t)} = \begin{cases} 1 & \text{if node } i \text{ appears in group } t, \\ 0 & \text{otherwise,} \end{cases}$$

for $i = 1, \ldots, n$ and $t = 1, \ldots, T$. The full dataset is a $T \times n$ matrix $\boldsymbol{G}$ with $G^{(t)}$ being its rows.

Let V_0 be the set of all nodes which can serve as a hub and let $n_L = |V_0|$. We refer to V_0 as the *hub set* and call the nodes in this set *hub set member*. In contrast to the setup in [76] where the hub set contains all nodes, we assume that the hub set contains fewer members than the whole set of nodes, i.e., $n_L < n$. We assume in this section that V_0 is known and consider the problem of estimating V_0 in Section 6.2. For simplicity of notation, we further assume $V_0 = \{1, \ldots, n_L\}$ in this section. We refer to nodes from $n_L + 1$ to n as *followers*. Given this notation, the true hub of $G^{(t)}$ is represented by $z_*^{(t)}$ which takes on values from $1, \ldots, n_L$.

Under the hub model, each group $G^{(t)}$ is independently generated by the following two-step process:

1. The hub is sampled from a multinomial trial with parameter $\rho = (\rho_1, \ldots, \rho_{n_L})$, i.e., $\mathbb{P}(z_*^{(t)} = i) = \rho_i$, with $\sum_{i=1}^{n_L} \rho_i = 1$.

2. Given the hub node i, each node j appears in the group independently with probability A_{ij}, i.e., $\mathbb{P}(G_j^{(t)} = 1 | z_*^{(t)} = i) = A_{ij}$.

Note that multiple hub set members may appear in the same group although only one of them will be the hub of that group.

A key assumption from [76] which we adopt in this paper is that a hub node must appear in any group that it forms (i.e., $A_{ii} \equiv 1$, for $i = 1, \ldots, n_L$). The parameters for the hub model are thus

$$\rho = (\rho_1, \ldots, \rho_{n_L}),$$
$$A_{n_L \times n} = \begin{pmatrix} 1 & A_{12} & \cdots & A_{1,n_L} & A_{1,n_L+1} & \cdots & A_{1,n} \\ A_{21} & 1 & \cdots & A_{2,n_L} & A_{2,n_L+1} & \cdots & A_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ A_{n_L,1} & A_{n_L,2} & \ldots & 1 & A_{n_L,n_L+1} & \cdots & A_{n_L,n} \end{pmatrix}.$$

As in [76], we interpret A_{ij} as the strength of the relationship between node i and j. We differ from [76] in that A is a non-square matrix and A_{ij} is not necessarily equal to A_{ji}. The setting in this article is more natural. Social relationships are usually non-reciprocal and in most organizations there are members who do not have the authority or willingness to initiate groups.

We begin with the case where both $\boldsymbol{G}$ and $z_* = (z_*^{(1)}, \ldots, z_*^{(T)})$ are observed. The likelihood function is

$$\mathbb{P}(\boldsymbol{G}, z_* | A, \rho) = \prod_{t=1}^{T} \prod_{i=1}^{n_L} \prod_{j=1}^{n} \left[A_{ij}^{G_j^{(t)}} (1 - A_{ij})^{(1 - G_j^{(t)})} \right]^{1(z_*^{(t)} = i)} \prod_{i=1}^{n_L} \rho_i^{1(z_*^{(t)} = i)},$$

where $1(\cdot)$ is the indicator function. With both $\boldsymbol{G}$ and z_* being observed, it is straightforward to estimate A and ρ by their respective maximum likelihood estimators (MLEs):

$$\hat{A}_{ij}^{z_*} = \frac{\sum_t G_j^{(t)} 1(z_*^{(t)} = i)}{\sum_t 1(z_*^{(t)} = i)}, \quad i = 1, \ldots, n_L, j = 1, \ldots, n,$$
$$\hat{\rho}_i^{z_*} = \frac{\sum_t 1(z_*^{(t)} = i)}{T}, \quad i = 1, \ldots, n_L.$$

When the hub node of each group is latent, i.e., when z_* is unobserved, the estimation problem becomes challenging. Integrating out z_*, the marginal likelihood of $\boldsymbol{G}$ is

$$\mathbb{P}(\boldsymbol{G}|A, \rho) = \prod_{t=1}^{T} \sum_{i=1}^{n_L} \rho_i \prod_{j=1}^{n} A_{ij}^{G_j^{(t)}} (1 - A_{ij})^{1 - G_j^{(t)}}, \tag{6.1}$$

which has the form of a Bernoulli mixture model. Hereafter the term hub model refers to the case where z_* is unobserved, unless otherwise specified.

Although less stringent than the original symmetric hub model, the asymmetric hub model has a significant limitation: it cannot naturally transition to a null model. In general, a null model generates data that match the basic features of the observed data, but which is otherwise a random process without structured patterns. In other words, a null model is the degenerate case of the model class being studied. The null model for grouped data, naturally, generates each group by independent Bernoulli trials. That is, if the grouping behavior is not governed by a network structure then every node is assumed to appear independently in a group. The likelihood of $G^{(t)}$ under the null model is

$$\mathbb{P}(G^{(t)}) = \prod_{j=1}^{n} \pi_j^{G_j^{(t)}} (1 - \pi_j)^{1 - G_j^{(t)}},$$

where π_j is the probability that node j appears in a group.

The asymmetric hub model needs generalization to accommodate the null model because if there is only one component in (6.1), say, node i is the only hub set member,

the likelihood of $G^{(t)}$ becomes

$$\mathbb{P}(G^{(t)}) = \prod_{j=1}^{n} A_{ij}^{G_j^{(t)}} (1 - A_{ij})^{1-G_j^{(t)}},$$

which is not a proper null model because the assumption $A_{ii} \equiv 1$ forces node i to appear in every group.

To allow the hub model to degenerate to the null model, we add the null component. This null component allows groups without hubs where nodes independently appear in such groups. We call this model the *hub model with the null component*. We use $z_*^{(t)} = 0$ to represent a hubless group.

The parameters for the hub model with the null component are $\rho = (\rho_0, \rho_1, \ldots, \rho_{n_L})$, $A_{(n_L+1)\times n} = [A_{ij}]_{i=0,1,\ldots,n_L, j=1,\ldots,n}$. Here the row indices of A start from 0, i.e., $A_{0j} \equiv \pi_j$ for $j = 1, \ldots, n$. We will use A_{0j} and π_j interchangeably below. As before we assume $A_{ii} \equiv 1$ for $i = 1, \ldots, n_L$. The marginal likelihood of $\boldsymbol{G}$ under the new model is

$$\mathbb{P}(\boldsymbol{G}|A, \rho) = \prod_{t=1}^{T} \sum_{i=0}^{n_L} \rho_i \prod_{j=1}^{n} A_{ij}^{G_j^{(t)}} (1 - A_{ij})^{1-G_j^{(t)}}. \tag{6.2}$$

The above model degenerates to the null model when $\rho_0 = 1$. For simplicity of notation, we use the same notation such as ρ and A for both the hub model with and without the null component when the meaning is clear from context.

The new model has an advantage in data analysis in addition to the theoretical benefit. Grouped data usually contain a number of tiny groups such as singletons and doubletons. When fitting the asymmetric hub model to such a dataset, one sometimes has to include these nodes in the hub set due to the one-hub restriction. Doing so may result in an unnecessarily large hub set (see Section 4 in the Supplemental Materials). In the hub model with the null component, these small groups can be treated as hubless groups and the corresponding nodes may be removed from the hub set. Therefore, the model complexity is significantly reduced.

6.1.2 Model Identifiability

Before considering estimation of ρ and A under (6.1) and (6.2), we need to establish the identifiability of parameters ρ and A. [76] proved model identifiability under the symmetry condition. We seek a new set of identifiability conditions as the new models do not assume symmetry of A.

To precisely define identifiability, let $\mathcal{P}$ be the parameter space of the hub model with the null component, where $\mathcal{P} = \{(\rho, A)|0 < \rho_i < 1, i = 0, \ldots, n_L; A_{ii} = 1, i = 1, \ldots, n_L; 0 \leq A_{ij} \leq 1, i = 0, \ldots, n_L, j = 1, \ldots, n, i \neq j\}$. The parameter space of the hub model without the null component is similar except that the index i always begins with 1. Let $\boldsymbol{g} = (g_j^{(t)})_{t=1,\ldots,T, j=1,\ldots,n}$ be any realization of $\boldsymbol{G}$ under the hub model.

Definition 2. *The parameters (ρ, A) within the parameter space $\mathcal{P}$ are identifiable (under the hub model with or without the null component) if the following holds:*

$$\forall \boldsymbol{g}, \forall (\tilde{\rho}, \tilde{A}) \in \mathbb{P}(\boldsymbol{G} = \boldsymbol{g}|\rho, A) = \mathbb{P}(\boldsymbol{G} = \boldsymbol{g}|\tilde{\rho}, \tilde{A}) \iff (\rho, A) = (\tilde{\rho}, \tilde{A}).$$

We define identifiability in the strictest sense and the above definition does not allow label swapping of latent classes. In cluster analysis label swapping refers to the fact that nodes can be successfully partitioned into latent classes, but individual classes cannot be uniquely identified. For example, community detection may correctly partition voters into communities based on their political preferences, but cannot identify which political party each community prefers. This is not an issue in the hub model due to the constraint $A_{ii} = 1$. In addition, note that we only need to consider identifiability for the distribution of a single observation, i.e., $T = 1$ because the data are independently and identically distributed. Let g be a realization of a single observation hereafter.

We now give the identifiability result for the asymmetric hub model.

Theorem 7. *The parameters* (ρ, A) *of the asymmetric hub model are identifiable under the following conditions:*

1. $A_{ij} < 1$, *for* $i = 1, \ldots, n_L, j = 1, \ldots, n, i \neq j$;
2. *for all* $i = 1, \ldots, n_L$, $i' = 1, \ldots, n_L, i \neq i'$, *the vectors* $(A_{i,n_L+1}, A_{i,n_L+2}, \ldots, A_{i,n})$ *and* $(A_{i',n_L+1}, A_{i',n_L+2}, \ldots, A_{i',n})$ *are not identical.*

Condition (ii) implies that for any pair of nodes in the hub set, there exists a follower with different probability of being included in groups formed by the two hubs, respectively. All proofs are given in the Supplementary Materials.

Identifiability under the model with the null component is more difficult to prove than the case of the asymmetric hub model due to the extra null component in the model. In particular, there is no constraint such as $\pi_i = 1$ on parameters of the null component. The conditions for identifiability in the following theorem are; however, as natural as those in Theorem 7.

Theorem 8. *The parameters* (ρ, A) *of the hub model with the null component are identifiable under conditions (i) and (ii) in Theorem 7 (index* i *begins with 0 in (i)), and*

3. *for any* $i = 1, \ldots, n_L$, *the vectors* $(A_{i,n_L+1}, A_{i,n_L+2}, \ldots, A_{i,n})$ *and* $(\pi_{n_L+1}, \pi_{n_L+2}, \ldots, \pi_n)$ *are different by at least two entries.*

Condition (iii) adds the requirement that for any hub i, there exist two followers which each has different probabilities of appearing in a group led by hub i than of appearing in a hubless group. This condition implies that there should exist at least two more nodes in the node set than in the hub set. This condition is natural if one compares it to condition (ii), as both imply that there exists at least one more column than rows in A.

6.1.3 Consistency of the Maximum Profile Likelihood Estimator

We consider the asymptotic consistency for the hub model in the most general setting. That is, we allow the number of groups (T), the size of the node set (n), and the size of the hub set (n_L) to grow. As mentioned in Section 1, we reformulate the problem as a clustering problem where a cluster is defined as the groups formed by the same hub node. We borrow the techniques from the community detection literature to prove the consistency of class labels, i.e., the consistency of hub labels. The consistency of parameter estimation then holds as a corollary. Note that n is necessarily to go to infinity for proving the consistency of hub labels because when n is fixed, the posterior probability of the hub label of a group given the data cannot concentrate on a single node. If one is only interested in the consistency of parameter estimation, it is possible to allow n fixed. The problem degenerates to the classical case, that is, estimating a non-growing number of parameters, and the classical theory of MLE is expected to be applicable.

We first consider the asymmetric hub model without the null component. Let $z = (z^{(t)})_{t=1,\dots,T}$ be an assignment of hub labels. Given z, the log-likelihood of the full dataset $\boldsymbol{G}$ is

$$L_G(A|z) = \sum_{t=1}^{T}\sum_{j=1}^{n} G_j^{(t)} \log A_{z^{(t)},j} + (1 - G_j^{(t)}) \log(1 - A_{z^{(t)},j}). \tag{6.3}$$

For $i = 1, \dots, n_L$, let $t_i = \sum_t 1(z^{(t)} = i)$ be the number of groups with hub i. Given z, the MLE of A is

$$\hat{A}^z_{ij} = \frac{\sum_t G_j^{(t)} 1(z^{(t)} = i)}{t_i}, \text{ for } t_i > 0.$$

If $t_i = 0$, define $\hat{A}^z_{ij} = 0$. We will omit the upper index z when it is clear from the context. Plugging $\hat{A}_{ij}$ back into (6.3), we obtain the profile log-likelihood

$$L_G(z) = \max_A L_G(A|z) = \sum_t \sum_j G_j^{(t)} \log \hat{A}_{z^{(t)},j} + (1 - G_j^{(t)}) \log(1 - \hat{A}_{z^{(t)},j}).$$

Furthermore, let

$$\hat{z} = \operatorname*{argmax}_z L_G(z).$$

The framework of profile likelihoods are adopted from the community detection literature [11, 22], where z is treated as an unknown parameter and we search for the z that optimizes the profile likelihood.

Recall that z_* is the true class assignment. We will treat z_* as a random vector to maintain continuity with the previous sub-section.

Let $P_j^{(t)} = \mathbb{P}(G_j^{(t)} = 1|z_*^{(t)}) = A_{z_*^{(t)},j}$. Then by replacing $G_j^{(t)}$ by $P_j^{(t)}$, we obtain a "population version" of $L_G(z)$:

$$L_P(z) = \sum_t \sum_j P_j^{(t)} \log \bar{A}_{z^{(t)},j} + (1 - P_j^{(t)}) \log(1 - \bar{A}_{z^{(t)},j}),$$

where

$$\bar{A}_{ij} = \frac{\sum_t P_j^{(t)} 1(z^{(t)} = i)}{t_i}, \text{ for } t_i > 0. \tag{6.4}$$

Otherwise, define $\bar{A}_{ij} = 0$. Let $T_e = \sum_t 1(z_*^{(t)} \neq \hat{z}^{(t)})$ be the number of groups with incorrect hub labels. As discussed previously, we do not allow label swapping in the definition of T_e. Our aim is to prove

$$T_e/T = o_p(1), \quad \text{as } n_L \to \infty, n \to \infty, T \to \infty.$$

We make the following assumptions throughout the proof of consistency under the asymmetric hub model:

H_1: $Tc_{\min}/n_L \leq t_{i*} \leq Tc_{\max}/n_L$ for $i = 1, \ldots, n_L$, where $t_{i*} = \sum_t 1(z_*^{(t)} = i)$ and $c_{\min}$ and $c_{\max}$ are constants.

H_2: $A_{ij} = s_{ij}d$ for $i = 1, \ldots, n_L, j = 1, \ldots, n$ and $i \neq j$ where s_{ij} are unknown constants satisfying $0 < s_{\min} \leq s_{ij} \leq s_{\max} < \infty$ while d goes to zero as n goes to infinity.

H_3: There exists a set $V_i \subset \{n_L + 1, \ldots, n\}$ for $i = 1, \ldots, n_L$ with [1] $|V_i| \geq vn/n_L$ such that $\tau = \min_{i,i'=1,\ldots,n_L, i \neq i', j \in V_i} (s_{ij} - s_{i'j})$ is bounded away from 0.

H_4: $A_{ii'} \leq c_0/n_L$ for $i = 1, \ldots, n_L$, $i' = 1, \ldots, n_L$, $i \neq i'$, where c_0 is a positive constant.

H_1 ensures that no hub set members appear too infrequently. The assumption in fact automatically holds with high probability if $(n_L^2 \log n_L)/T = o(1)$, which can be proved by applying Hoeffding's inequality. Here we directly assume the condition for simplicity. H_2 allows the expected density of A to shrink as n grows, which is a common setup in the community literature. H_3 implies that for every hub set member there exists a set of nodes that are more likely to join groups initiated by this particular hub set member than others. The size of this set is influenced by v and the magnitude of this preference is influenced by d (since $A_{ij} = ds_{ij}$). The decay rates of d and v, as well as the growth rates of n_L, n and T, will be specified in the following consistency results. H_4 is a technical assumption that prevents label swapping from influencing the consistency results.

Now we state a lemma that T_e/T is bounded by $L_P(z_*) - L_P(\hat{z})$. That is, z_* is a *well-separated* point of maximum of L_P. The reader is referred to Section 5.2 in [66] for the classical case of this concept.

[1] $|\cdot|$ is the cardinality of a set.

Lemma 5. *Under H_1 – H_4, for some positive constant δ,*

$$\mathbb{P}\left(\frac{\delta n_L}{dvnT}(L_P(z_*) - L_P(\hat{z})) \geq \frac{T_e}{T}\right) \to 1, \quad \text{as } n_L \to \infty, n \to \infty, T \to \infty.$$

We consider the most general setup in which n_L, n, and T all go to infinity in the main text. For the easier case of n_L being fixed, we give the corresponding results (Theorem 3′ and 4′ for the asymmetric hub model and Theorem 5′ and 6′ for the hub model with the null component) in the Supplementary Materials. Based on Lemma 5, we establish label consistency:

Theorem 9. *Under H_1 – H_4, if $n_L^2 \log T/(dTv) = o(1)$, $(\log d)^2 n_L^2 \log n_L/(dnv^2) = o(1)$, and $(\log T)^2 n_L^2 \log n_L/(dnv^2) = o(1)$, then*

$$T_e/T = o_p(1), \quad \text{as } n_L \to \infty, n \to \infty, T \to \infty.$$

The next result addresses the consistency for parameter estimation of A, which is based upon a faster decay rate of T_e/T than Theorem 9 (see the proof of Theorem 10 in the Supplemental Materials for details).

Theorem 10. *Under H_1 – H_4, if $n_L \log n/T = o(1)$, $n_L^3 \log T/(dTv) = o(1)$, $(\log d)^2 n_L^4 \log n_L/(dnv^2) = o(1)$, and $(\log T)^2 n_L^4 \log n_L/(dnv^2) = o(1)$, then*

$$\max_{i \in \{1,\ldots,n_L\}, j \in \{1,\ldots,n\}} \left|\hat{A}_{ij}^{\hat{z}} - A_{ij}\right| = o_p(1), \quad \text{as } n_L \to \infty, n \to \infty, T \to \infty.$$

We now establish the consistency for the hub model with the null component. The proofs are more challenging due to the extra null component. We make the following assumptions throughout the proofs, parallel to H_1 – H_4:

H_1^*: $Tc_{\min}/n_L \leq t_{i*} \leq Tc_{\max}/n_L$ for $i = 0, \ldots, n_L$, where $t_{i*} = \sum_t 1(z_*^{(t)} = i)$ and $c_{\min}$ and $c_{\max}$ are constants.

H_2^*: $A_{ij} = s_{ij}d$ for $i = 0, \ldots, n_L, j = 1, \ldots, n$ and $i \neq j$ where s_{ij} are unknown constants satisfying $0 < s_{\min} \leq s_{ij} \leq s_{\max} < \infty$ while d goes to zero as n goes to infinity.

H_3^*: There exists a set $V_i \subset \{n_L + 1, \ldots, n\}$ for $i = 1, \ldots, n_L$ with $|V_i| \geq vn/n_L$ such that $\tau = \min_{i=1\ldots,n_L,i'=0,\ldots,n_L,i\neq i',j\in V_i} (s_{ij} - s_{i'j})$ is bounded away from 0.

H_4^*: $A_{ii'} \leq c_0/n_L$ for $i = 0, \ldots, n_L$, $i' = 1, \ldots, n_L$, $i \neq i'$, where c_0 is a positive constant.

The main difference between the two sets of assumptions is on the range of the indices. For example, index i is from 0 to n_L in H_1^*. In particular, t_{0*} is the true number of hubless groups. Index i starts from 1 in H_3^* because we only define the set V_i for each hub set member i but not for the hubless case.

We need a result on the separation of $L_P(z_*)$ from $L_P(\hat{z})$ which is similar to Lemma 5. However, the technique in the original proof cannot be directly applied to the new model. A key step in the proof of Lemma 5 relies on the fact that we can obtain a non-zero lower bound for the number of correctly classified groups with node i as the hub node in the asymmetric hub model. Specifically, let $t_{ii'} = \sum_t 1(z_*^{(t)} = i, \hat{z}^{(t)} = i')$ for $i = 0, \ldots, n_L$, $i' = 0, \ldots, n_L$. Thus, t_{ii} is the number of correctly classified groups where node i is the hub node. For the asymmetric hub model, we obtain a lower bound for t_{ii}/t_{i*} $(i = 1, \ldots, n_L)$ from the fact that a node cannot be labeled as the hub of a particular group if the node does not appear in the group. This is due to the assumption $A_{ii} \equiv 1$ for $i = 1, \ldots, n_L$. For the hub model with the null component, the lower bound for t_{ii}/t_{i*} cannot be proved by the same technique because all groups can be classified as hubless groups without violating the assumption $A_{ii} \equiv 1$.

We take a different path in the proof to overcome this issue and other technical difficulties due to the null component. We first bound t_{i0}/t_{i*} for $i = 1, \ldots, n$.

Lemma 6. *Under H_1^* – H_4^*, if $n_L^4 \log T/(dTv) = o(1)$, $(\log d)^2 n_L^6 \log n_L/(dnv^2) = o(1)$ and $(\log T)^2 n_L^6 \log n_L/(dnv^2) = o(1)$, then for all $\eta > 0$,*

$$\frac{t_{i0}}{t_{i*}} \leq \eta, \quad i = 1, \ldots, n_L,$$

with probability approaching 1.

Based on the result in Lemma 6, we establish the label consistency for the hub model with the null component.

Theorem 11. *Under the conditions of Lemma 6,*

$$\frac{T_e}{T} = o_p(1), \quad \text{as } n_L \to \infty, n \to \infty, T \to \infty.$$

We conclude this section by the result on consistency for parameter estimation of A under the hub model with the null component.

Theorem 12. *Under H_1^* – H_4^*, if $n_L \log n/T = o(1)$, $n_L^5 \log T/(dTv) = o(1)$, $(\log d)^2 n_L^8 \log n_L/(dnv^2) = o(1)$ and $(\log T)^2 n_L^8 \log n_L/(dnv^2) = o(1)$, then*

$$\max_{i \in \{0,\ldots,n_L\}, j \in \{1,\ldots,n\}} \left| \hat{A}_{ij}^{\hat{z}} - A_{ij} \right| = o_p(1), \quad \text{as } n_L \to \infty, n \to \infty, T \to \infty.$$

6.2 Hub Model with the Null Component and Unknown Hub Set

6.2.1 Model Setup

The asymmetric hub model (with or without the null component) assumes that the hub set is a subset of the nodes. The previous section addressed the estimation problem when the hub set is known, but in practice, the hub set is usually not known a priori. In this section, we study the selection of the hub set under the hub model with the null component.

Recall that V_0 denotes the hub set with $|V_0| = n_L$. In the following, we no longer assume $V_0 = \{1, \ldots, n_L\}$ and the goal is to estimate V_0. We begin with a

known *potential hub set*, denoted by $\bar{V}_0$, which is subset containing all nodes that can potentially serve as hub set members. One might assume that the ideal $\bar{V}_0$ would be the same as the entire node set V; however, to prove identifiability of parameters when the hub set is unknown (see Theorem $S1$ in the Supplemental Materials), we require the potential hub set $\bar{V}_0$ to be smaller than V. In practice, this means we have prior knowledge that certain nodes do not play an important role in group formation and are therefore not included in the hub set. Let $M = |\bar{V}_0|$ with $n_L < M < n$. Without loss of generality, assume $\bar{V}_0 = \{1, \ldots, M\}$.

The data generation mechanism is the same as the hub model with the null component. The parameters are $\rho = (\rho_0, \rho_1, \ldots, \rho_M)$, $A_{(M+1)\times n} = [A_{ij}]_{i=0,1,\ldots,M, j=1,\ldots,n}$. For $i = 1, \ldots, M$, $\rho_i = 0$ if $i \notin V_0$. The corresponding $\{A_{ij}\}_{j=1,\ldots,n}$ therefore do not play a role in the model and will not be estimated. If all $\rho_i = 0$, $i = 1, \ldots, M$, the model degenerates to the null model in which nodes appear independently in all groups. The marginal likelihood of $\boldsymbol{G}$ is

$$\mathbb{P}(\boldsymbol{G}|A, \rho) = \prod_{t=1}^{T} \sum_{i=0}^{M} \rho_i \prod_{j=1}^{n} A_{ij}^{G_j^{(t)}} (1 - A_{ij})^{1 - G_j^{(t)}}.$$

6.2.2 Penalized Likelihood

We propose to maximize the following penalized log-likelihood function to estimate V_0:

$$L(A, \rho) - T\lambda \sum_{i=1}^{M} [\log(\epsilon + \rho_i) - \log \epsilon], \tag{6.5}$$

$$\text{subject to } \rho_i \geq 0,\ i = 0, 1, \ldots, M,\ \sum_{i=0}^{M} \rho_i = 1,$$

where

$$L(A, \rho) = \log \mathbb{P}(\boldsymbol{G}|A, \rho) = \sum_{t=1}^{T} \log \left[\sum_{i=1}^{M} \rho_i \prod_{j=1}^{n} A_{ij}^{G_j^{(t)}} (1 - A_{ij})^{1 - G_j^{(t)}} \right].$$

λ is the tuning parameter which controls the penalty on the mixing weights. ϵ is a small positive number. We use $\epsilon = 10^{-8}$ in all numerical studies. The estimated hub set V_0 includes node i ($i = 1, \ldots, M$) if and only if $\hat{\rho}_i \neq 0$ in the maximizer of (6.5).

The penalty function in (6.5) was inspired by a similar penalty function proposed by Huang *et al.* [41] for selecting the number of components in Gaussian mixture models. However, our penalty function has a subtle but substantial difference: the hub node index m in the penalty function begins with 1 instead of 0 – that is, we do not penalize the coefficient of the null component ρ_0. The model is therefore penalized toward the null model, i.e., the independent Bernoulli model, when λ is sufficiently large. The penalty function uses $\log(\epsilon + \rho_i)$ instead of $\log \rho_i$ as in [41], because $\log(\epsilon + \rho_i)$ will not go to infinity when ρ_i goes to zero, which makes it possible for $\hat{\rho}_i$ to reach exactly zero.

Maximizing the Lagrangian form of the penalized log-likelihood function (6.5) is equivalent to maximizing $L(A, \rho)$ under the following constraints

$$\rho_i \geq 0,\ i = 0, 1, \ldots, M,\ \sum_{i=0}^{M} \rho_i = 1,\ \sum_{i=1}^{M} [\log(\epsilon + \rho_i) - \log \epsilon] \leq t.$$

To show how the constraints can result in sparse solutions, we consider a toy model containing only two nodes, both of which are potential hub set members, that is, $M = 2$. The constraints become

$$\begin{aligned} &\rho_1 \geq 0,\ \rho_2 \geq 0,\ \rho_1 + \rho_2 \leq 1, \\ &\log(1 + \frac{\rho_1}{\epsilon}) + \log(1 + \frac{\rho_2}{\epsilon}) \leq t. \end{aligned} \tag{6.6}$$

Figure 6.1 shows the feasible regions of the log penalties for $t = 3, 4, 5$ and $\epsilon = 0.01$, where the crosses mark the intersection of $\log(1 + \rho_1/\epsilon) + \log(1 + \rho_2/\epsilon) = t$ and the axes, and the dashed line indicates $\rho_1 + \rho_2 = 1$. For $t = 3$ and 4, $\hat{\rho}_1$ (resp. $\hat{\rho}_2$) can potentially reach 0 with $\hat{\rho}_2$ (resp. $\hat{\rho}_1$) being non-zero, indicated by the cross markers

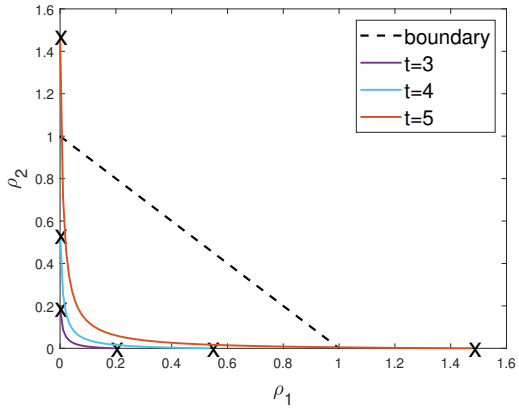

Figure 6.1: Feasible regions of the log penalty with different values of t.

within the region defined by (6.6). For $t = 5$ (corresponding to a smaller λ), this cannot happen because $\log(1 + \rho_1/\epsilon) + \log(1 + \rho_2/\epsilon) = 5$ intersects with the axes outside of the region defined by (6.6).

6.2.3 Algorithm

We propose a modified expectation-maximization (EM) algorithm for optimizing (6.5).

Modified EM Algorithm

Iteratively update $\hat{A}$ and $\hat{z}$ by the following E-step and M-step until convergence. Define $h_{ti} = \mathbb{P}(z^{(t)} = i|\boldsymbol{G}, A)$ for $t = 1, \ldots, T$ and $i = 0, \ldots, M$.

E-step: Given $\hat{A}$ and $\hat{\rho}$,

$$\hat{h}_{ti} = \frac{\hat{\rho}_i \mathbb{P}(G^{(t)}|z^{(t)} = i, \hat{A})}{\sum_{i=0}^{M} \hat{\rho}_i \mathbb{P}(G^{(t)}|z^{(t)} = i, \hat{A})}, \quad \text{for } i = 0, \ldots, M.$$

M-step: For i such that $\hat{\rho}_i \neq 0$, given $\hat{h}_{ti}$,

$$\hat{A}_{ij} = \frac{\sum_{t=1}^{T} \hat{h}_{ti} G_j^{(t)}}{\sum_{t=1}^{T} \hat{h}_{ti}}, \text{ for } j = 1, \ldots, n.$$

Update $\hat{\rho}$ by solving the following optimization problem:

$$\hat{\rho} = \underset{\rho}{\operatorname{argmax}}\, L(\hat{A}, \rho) - T\lambda \sum_{i=1}^{M} \log(\epsilon + \rho_i), \tag{6.7}$$

$$\text{subject to } \rho_i \geq 0,\ i = 0, \ldots, M,\ \sum_{i=0}^{M} \rho_i = 1.$$

The only difference between modified EM and the standard EM algorithm is the update of $\hat{\rho}$ in the M-step. In the standard EM algorithm for the likelihood without the penalty term, $\hat{\rho}_i$ has a closed-form solution, that is, $\hat{\rho}_i = \sum_{t=1}^{T} \hat{h}_{ti}/T,\ i = 0, \ldots, M$. By contrast, (6.7) is a non-linear optimization problem with inequality constraints, which we use a numerical technique – the augmented Lagrange multiplier ([33]) method to solve the problem. In addition, since (6.5) is a non-convex optimization problem, we use multiple different initial values (20 random initial values are used in this paper) to help guard against local maxima.

Chapter 7

NUMERICAL STUDIES ON HUB MODELS

7.1 Numerical Studies

7.1.1 Numerical Studies When the Hub Set is Known

In this sub-section, we examine the performance of the estimators for the asymmetric hub model and the hub model with the null component when the hub set is known, under varying n_L, n and T. Hub set selection will be considered in the next sub-section. The parameters are estimated by the standard EM algorithm and the estimated hub labels are determined according to the largest posterior probabilities.

For the asymmetric hub model, let ρ_i be generated independently from $U(0,1)$ and renormalize ρ_i such that $\sum_{i=1}^{n_L} \rho_i = 1$. Let the size of the node set, n, be 100 or 500. We partition the follower set $\{n_L+1, \ldots, n\}$ into n_L non-overlapping sets $V_1, \ldots, V_{n_L}$. Each set V_i is the set of followers with a preference for hub set member i over other hub set members. As in Theorem 5, we assume different ranges of probabilities of joining a group for followers that prefer a specific hub set member than for followers which do not prefer that member. Specifically, for $j \in V_i$, the parameters A_{ij} are generated independently from $U(0.2, 0.4)$, and for $j \notin V_i$, the parameters A_{ij} are generated independently from $U(0, 0.2)$. The numerical results for sparser A will be given in Section 4 of the Supplemental Materials. For clarification, we will not use prior information about how A was generated in the estimating procedure. That is, we still treat A as unknown fixed parameters in the estimation. We generate these probabilities from uniform distributions for the sole purpose of adding more variations to the parameter setup. In each setup, we consider four different sample

sizes, $T = 500, 1000, 1500$ and 2000, and two different values of the size of hub set, $n_L = 10$ and 20.

For the hub model with the null component, let the probability of hubless groups $\rho_0 = 0.2$, and let ρ_i be generated independently from $U(0,1)$ and renormalize ρ_i such that $\sum_{i=1}^{n_L} \rho_i = 0.8$ for $i = 1, \ldots, n_L$. For a hubless group, each node will independently join the group with probability $\pi_j \equiv 0.05$ for $j = 1, \ldots, n$. The setups on n_L, n, $\{V_1, \ldots, V_{n_L}\}$, A, n_L and T are identical to the asymmetric hub model case.

Table 7.1: Asymmetric hub model results. Mis-labels: the fraction of groups with incorrect hub labels. RMSE($\hat{A}_{ij}$): average RMSEs when the hub labels are unknown. RMSE*: average RMSEs when the hub labels are known.

$n_L = 10$	$n = 100$			$n = 500$		
	Mis-labels	RMSE($\hat{A}_{ij}$)	RMSE*	Mis-labels	RMSE($\hat{A}_{ij}$)	RMSE*
$T = 500$	0.0479	0.0501	0.0475	0.0011	0.0483	0.0483
$T = 1000$	0.0335	0.0344	0.0332	0.0000	0.0337	0.0337
$T = 1500$	0.0295	0.0280	0.0272	0.0000	0.0274	0.0274
$T = 2000$	0.0262	0.0243	0.0236	0.0000	0.0235	0.0235
$n_L = 20$	$n = 100$			$n = 500$		
	Mis-labels	RMSE($\hat{A}_{ij}$)	RMSE*	Mis-labels	RMSE($\hat{A}_{ij}$)	RMSE*
$T = 500$	0.2396	0.0791	0.0662	0.0605	0.0686	0.0673
$T = 1000$	0.1528	0.0548	0.0463	0.0096	0.0466	0.0463
$T = 1500$	0.1186	0.0433	0.0375	0.0029	0.0380	0.0379
$T = 2000$	0.0998	0.0366	0.0325	0.0013	0.0328	0.0328

Table 7.1 and 7.2 show the performance of the estimators for the asymmetric hub model and the hub model with the null component, respectively. The first measure of

Table 7.2: Hub model with null component results. Mis-labels: the fraction of groups with incorrect hub labels. RMSE($\hat{A}_{ij}$): average RMSEs when the hub labels are unknown. RMSE*: average RMSEs when the hub labels are known.

$n_L = 10$	$n = 100$			$n = 500$		
	Mis-labels	RMSE($\hat{A}_{ij}$)	RMSE*	Mis-labels	RMSE($\hat{A}_{ij}$)	RMSE*
$T = 500$	0.0842	0.0542	0.0511	0.0058	0.0516	0.0516
$T = 1000$	0.0595	0.0376	0.0357	0.0006	0.0362	0.0362
$T = 1500$	0.0512	0.0308	0.0294	0.0001	0.0292	0.0292
$T = 2000$	0.0489	0.0264	0.0253	0.0001	0.0253	0.0253
$n_L = 20$	$n = 100$			$n = 500$		
	Mis-labels	RMSE($\hat{A}_{ij}$)	RMSE*	Mis-labels	RMSE($\hat{A}_{ij}$)	RMSE*
$T = 500$	0.3206	0.0839	0.0734	0.1146	0.0732	0.0719
$T = 1000$	0.2102	0.0607	0.0506	0.0229	0.0510	0.0509
$T = 1500$	0.1598	0.0488	0.0411	0.0076	0.0418	0.0416
$T = 2000$	0.1419	0.0414	0.0355	0.0022	0.0359	0.0359

performance we are interested in is the proportion of mislabeled groups, T_e/T. As the proportion of mislabeled groups approaches zero, we expect the parameter estimates to approach the accuracy achievable if the hub nodes are known. The second measure of performance is the RMSE($\hat{A}_{ij}$). As a reference point, we also provide the RMSE achieved when we treat the hub nodes as known, RMSE*. All results are averaged by 1000 replicates.

From the tables, the estimators for the asymmetric hub model generally outperform those for the hub model with the null component as the latter is a more complex model. The patterns within the two tables are, however, similar. First, the perfor-

mance becomes better as the sample size T grows, which is in line with common sense in statistics. Second, the performance becomes worse as n_L grows because n_L is the number of components in the mixture model, and thus a larger n_L indicates a more complex model. Third, the effect of n is more complicated: the RMSE* for the case that hub labels are known slightly increases as n grows because the model contains more parameters. What we are interested in is the case where hub labels are unknown, and this is what our theoretical studies focused on. In this case, the RMSE($\hat{A}_{ij}$) significantly improves as n grows. This is because the clustered pattern becomes clearer as the number of followers increases, which is in line with the label consistency results.

7.1.2 Numerical Results for Hub Set Selection

We study the performance of hub set selection by the penalized log-likelihood (6.5), which is optimized by the modified EM algorithm (Algorithm 1). We use the same settings as the hub model with the null component in the previous sub-section. The only difference is we need to specify the potential hub set $\bar{V}_0 = \{1, \ldots, M\}$: we consider $M = 80$ for $n = 100$ and $M = 80, 200$ and 300 for $n = 500$. In each setup, AIC and BIC are used to select the tunning parameter, λ. Let $\widehat{V}_0$ be the estimate of V_0. The performance of hub set selection is evaluated by the true positive rate (TPR) and the false positive rate (FPR), where

$$\text{TPR} = \frac{\sum_{i=1}^{M} 1(i \in V_0, i \in \widehat{V}_0)}{n_L}, \ \text{FPR} = \frac{\sum_{i=1}^{M} 1(i \notin V_0, i \in \widehat{V}_0)}{M - n_L}.$$

Table 7.3 shows the TPR and FPR for hub set selection under various settings. The patterns in the table with respect to n_L, n and T are similar to Table 7.1 and 7.2. That is, the performance of hub set selection is better for smaller n_L, larger n, and/or larger T. Among all settings, the model with $n_L = 10, T = 2000$ and $n = 500$

Table 7.3: TPR and FPR for hub set selection.

n_L	T	Criteria	$n = 100$		$n = 500$					
			$M = 80$		$M = 80$		$M = 200$		$M = 300$	
			TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
10	1000	AIC	0.6438	0.0719	0.9460	0.0128	0.7338	0.0081	0.6986	0.0128
		BIC	0.5787	0.0283	0.9381	0.0127	0.6831	0.0042	0.6472	0.0081
20	1000	AIC	0.5140	0.1410	0.6972	0.0249	0.4831	0.0229	0.4780	0.0370
		BIC	0.5100	0.1350	0.6859	0.0239	0.4494	0.0132	0.4673	0.0318
10	2000	AIC	0.8613	0.0187	0.9909	0.0010	0.9130	0.0018	0.8585	0.0015
		BIC	0.7675	0.0043	0.9883	0.0005	0.8956	0.0007	0.8400	0.0004
20	2000	AIC	0.6560	0.1050	0.8551	0.0074	0.6770	0.0155	0.6250	0.0140
		BIC	0.4438	0.0344	0.7884	0.0034	0.5848	0.0058	0.5519	0.0056

is the simplest for hub set selection purpose, which has the largest TPR and smallest FPR with λ selected by either AIC or BIC. Furthermore, the selection performance becomes worse as M grows because a larger M corresponds to a larger potential hub set and hence a larger candidate set of models.

7.1.3 Additional Simulation Results

To further study the performance of the estimates under the setting of sparse A, we introduce a scale factor α to control the density of A. Specifically, $A_{ij} \sim U(0.2\alpha, 0.4\alpha)$ for $j \in V_i$ and $A_{ij} \sim U(0, 0.2\alpha)$ for $j \notin V_i$, where $\alpha = 0.1, 0.2, \ldots, 1$. We study how the ratios of the RMSEs when the hub labels are unknown to those when the hub labels are known i.e., $\text{RMSE}(\hat{A}_{ij})/\text{RMSE}^*$, change with the degree of sparsity. We present the results for the case when $n = 100$. Other simulation settings are the same with those in Section 4.1.

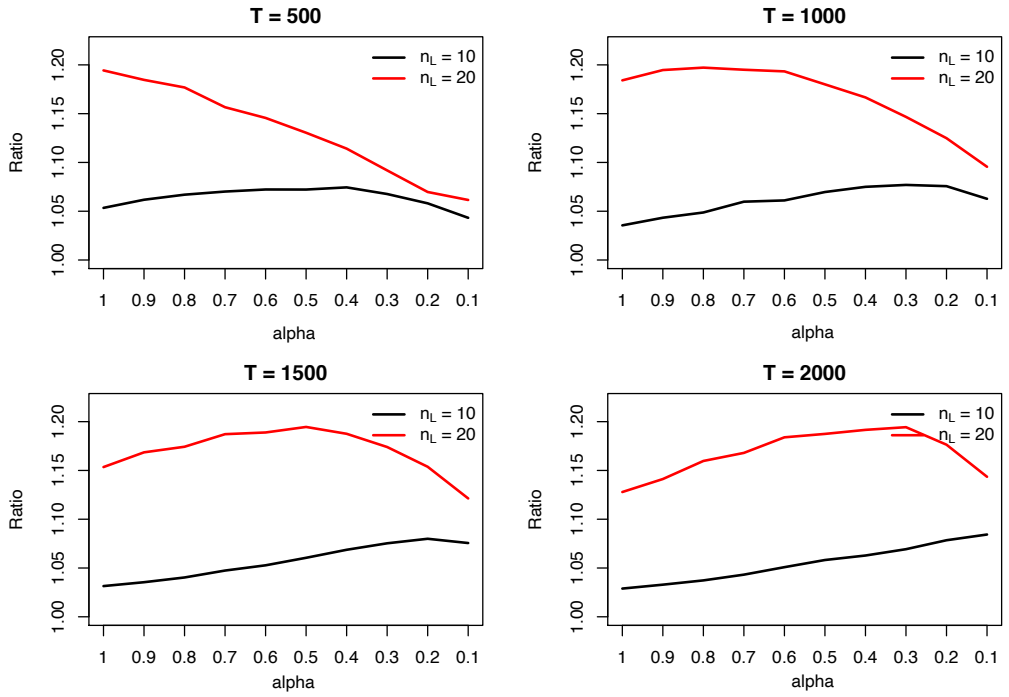

Figure 7.1: The asymmetric hub model results. The ratio is RMSE($\hat{A}_{ij}$)/RMSE*.

Figure 7.1 and 7.2 show the results of ratio versus α for the asymmetric hub model and the hub model with the null component, respectively. The overall ratio typically first increases and then decreases as α decreases. This implies that estimators for both models perform well when A is dense, and then the problem becomes more difficult for the estimator with unknown hub labels as A becomes sparse, but eventually when A becomes too sparse, the matrix A cannot be well estimated even for the case of known hub labels (i.e., the baseline).

Moreover, Figure 7.1 and 7.2 show that the turning point, i.e., the maximizer of the ratio, comes earlier when A is more difficult to estimate, which corresponds to the cases with larger n_L, smaller T, and the hub model with the null component. The turning point corresponds to the α value that gives the largest gap between the RMSE for the estimator with unknown hub labels and the baseline, and when the

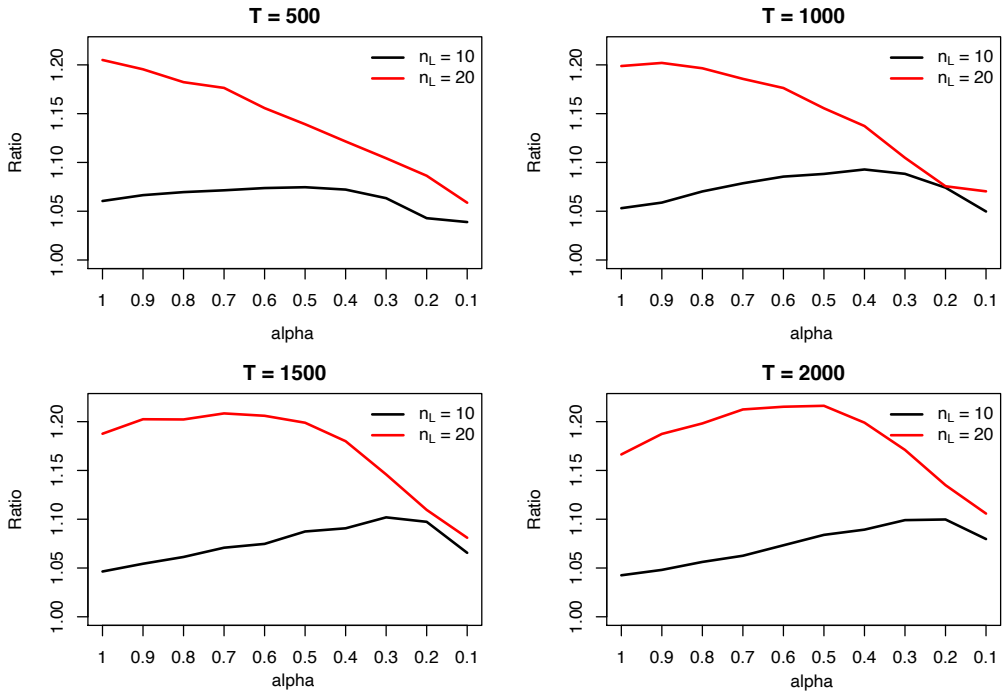

Figure 7.2: The hub model with the null component results. The ratio is RMSE($\hat{A}_{ij}$)/RMSE*.

settings become more difficult, the estimator with unknown hub labels starts to face challenges on a denser graph.

7.1.4 Analysis of Extended Bakery Data

We apply the hub model with the null component to the extended bakery dataset (`http://wiki.csc.calpoly.edu/datasets/wiki/ExtendedBakery`) to find the hub items and relationships among all the items. The dataset is a collection of purchases in a chain of bakery stores. The stores provide 50 items including 40 bakery goods (1-40) and 10 drinks (41-50). The goods can be divided into five categories: cakes (1-10), tarts (11-20), cookies (21-30) and pastries (31-40). Each purchase contains a collection of items bought together.

The extended bakery data was used as a benchmark dataset to test certain machine learning methods. For example, [2] used association rule mining to extract the hidden relationships of items and [55] applied a multinomial logit (MNL) model to address the problem of collaboratively learning representations of the users and the items in recommendation systems.

In our experiment, we use the 5,000 receipts in the dataset. Since drinks are typically purchased as affiliated items of food, we use the 40 bakery goods as the potential hub set, i.e., $\bar{V}_0 = \{1, \ldots, 40\}$. We use $\lambda = 0.025, 0.030, \ldots, 0.045$ to estimate the hub set.

Table 7.4: Estimated hub set for extended bakery data

λ	Selected hub nodes								
0.025	1	4	5	6	12	13	25	29	33
0.030	1	4	5	15	23	29	33		
0.035	5	15	23	29	34				
0.040	15	16	23	29	34				
0.045	15	23	29	34					

Table 7.4 shows the estimated hub sets. As λ increases, nodes are removed gradually from the hub set. According to the BIC criteria, the optimal λ is 0.045, at which the estimated hub set contains v_{15}, v_{23}, v_{29} and v_{34}, where v_{15} is tart, v_{23} and v_{29} are cookies, and v_{34} is pastry.

In addition, if the data was fitted by the hub model without the null component, then the entire node set has to be used as the hub set. In fact, each of the 50 items was purchased individually for at least once, and therefore must serve as a hub if the hubless groups are not assumed. When the hub model with the null component

is used, the corresponding items may be removed from the hub set, which greatly reduces the model complexity.

7.2 Analysis of Passerine Data

We apply the hub model with the null component to analyze a dataset on grouping behavior of passerines [62]. The dataset includes 63 color-marked passerines in Australia for daily observations, which are 2 scarlet robins (Petroica boodang), 13 striated thornbills (Acanthiza lineata), 26 buff-rumped thornbills (Acanthiza reguloides), 14 yellow-rumped thornbills (Acanthiza chrysorrhoa), 4 speckled warblers (Chthonicola sagittatus), 2 white-throated treecreepers (Cormobates leucophaea), one white-eared honeyeater (Lichenostomous leucotis), and one unkown bird. A group is defined as individuals observed together in a flock, and in total there are 109 groups, i.e., $T = 109$. Species information is summarized in Table 7.5.

Table 7.5: Summary of passerine species

Species	Binomial Nomenclature	Number	Label
scarlet robin	Petroica boodang	2	$v_1 - v_2$
striated thornbill	Acanthiza lineata	13	$v_3 - v_{15}$
buff-rumped thornbill	Acanthiza reguloides	26	$v_{16} - v_{41}$
yellow-rumped thornbill	Acanthiza chrysorrhoa	14	$v_{42} - v_{55}$
speckled warbler	Chthonicola sagittatus	4	$v_{56} - v_{59}$
white-throated treecreeper	Cormobates leucophaea	2	$v_{60} - v_{61}$
white-eared honeyeater	Lichenostomus leucotis	1	v_{62}
unknown	unknown	1	v_{63}

In the following analysis, we set the potential hub set $\bar{V}_0$ with $M = 55$ as the collection of birds in the first four species (Table 7.5) and the other eight birds belonging to small-scale species as followers [1] . Table 7.6 shows the estimated hub set under

[1]Nodes v_1 and v_2 appear frequently so we include them in the poential hub set

Table 7.6: Estimated hub set for passerine data

λ	v_7	v_9	v_{10}	v_{20}	v_{30}	v_{33}	v_{37}	v_{42}	v_{46}
0.045	■	■	■	■	■	■	■	■	■
0.050		■	■	■		■		■	
0.055		■			■			■	
0.060		■			■			■	
0.065									

various λ values where a grey block indicates that a node is included in the hub set. As λ increases, nodes are removed gradually from the hub set and at $\lambda = 0.065$, the hub model degenerates to the null model where the hub set is empty. The BIC selects $\lambda = 0.055$, where the estimated hub set includes v_9, v_{30} and v_{42}, each belonging to one of the three large-scale species.

In addition, we bootstrap 1,000 samples from the original data to evaluate the stability of the proposed hub set selection method. Specifically, we perform our method on each bootstrapped sample under λ from 0.045 to 0.065 and compute the proportion of each node being selected as a hub set member. Table 7.7 demonstrates the stability of the proposed method: the majority of the birds are not selected as a hub set member in any bootstrap sample, and v_9, v_{30} and v_{42}, the three birds identified from the original data dominate in the selection proportions across the bootstrapped samples.

Table 7.7: Selection proportion from bootstrap

λ	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}	v_{11}	v_{12}	v_{13}	v_{14}	v_{15}
0.045	0	0	0	0.045	0	0	0.81	0	0.995	0.870	0	0	0	0	0
0.050	0	0	0	0.050	0	0	0	0	1	0.600	0	0	0	0	0
0.055	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
0.060	0	0	0	0.005	0	0	0	0	0.965	0	0	0	0	0	0
0.065	0	0	0	0	0	0	0	0	0	0.005	0	0	0	0	0
λ	v_{16}	v_{17}	v_{18}	v_{19}	v_{20}	v_{21}	v_{22}	v_{23}	v_{24}	v_{25}	v_{26}	v_{27}	v_{28}	v_{29}	v_{30}
0.045	0.01	0	0	0	0.810	0	0	0	0.010	0.095	0.115	0.025	0	0	0.890
0.050	0	0	0	0	0.600	0	0	0	0.015	0.075	0.100	0.015	0	0	0.625
0.055	0	0	0	0	0.005	0	0	0	0	0.005	0	0.005	0	0	0.945
0.060	0	0	0	0	0.025	0	0	0	0	0	0.015	0.005	0	0	0.830
0.065	0	0	0	0	0.010	0	0	0	0	0	0.005	0	0	0	0.015
λ	v_{31}	v_{32}	v_{33}	v_{34}	v_{35}	v_{36}	v_{37}	v_{38}	v_{39}	v_{40}	v_{41}	v_{42}	v_{43}	v_{44}	v_{45}
0.045	0	0	0.825	0	0	0	0.830	0	0	0	0	0.965	0	0.005	0
0.050	0	0	0.625	0	0	0	0.105	0	0	0	0	0.935	0	0	0
0.055	0	0	0.010	0	0	0	0.015	0	0	0	0	0.985	0	0	0
0.060	0	0	0.040	0	0	0	0.020	0	0	0	0	0.910	0	0	0
0.065	0	0	0.045	0	0	0	0.050	0	0	0	0	0.080	0	0	0
λ	v_{46}	v_{47}	v_{48}	v_{49}	v_{50}	v_{51}	v_{52}	v_{53}	v_{54}	v_{55}					
0.045	0.845	0	0	0	0	0	0	0	0	0					
0.050	0.235	0	0	0	0	0	0	0	0	0					
0.055	0	0	0	0	0	0	0	0	0	0					
0.060	0	0	0	0	0	0	0	0	0	0					
0.065	0	0	0	0	0	0	0	0	0	0					

Chapter 8

CONCLUSION AND DISCUSSION

In this project, we studied the theoretical properties of the hub model and its variants from the perspective of Bernoulli mixture models. The contributions of the paper are four-fold. First, we proved the model identifiability of the hub model. Bernoulli mixture models are a notoriously difficult model to prove identifiability on, especially under mild conditions. Second, we proved the label consistency and estimation consistency of the hub model. Third, we generalized the hub model by adding the null component that allows nodes to independently appear in hubless groups. The new model can naturally degenerate to the null model – the Bernoulli product. We also proved identifiability and consistency of the newly proposed model. Finally, we proposed a penalized likelihood method to select the hub set, which estimates not only the size of the hub set, n_L, but also which nodes belong to the set. The new method can handle data with no prior knowledge of the hub set and hence greatly expands the domain of the applicability of the hub model.

A natural constraint from [76] that we apply in this paper is $A_{ii} = 1$ $(i = 1, \ldots, n_L)$, which turns out to be a key condition for ensuring model identifiability and avoiding the label swapping issue in the proof of consistency. On the other hand, this constraint prevents the asymmetric hub model from naturally degenerating to the null model because one node always appear in every group when there is only one component in the hub model, which motivated adding the null component to the model.

We consider the profile likelihood estimator in the proofs of consistency. The marginal likelihood MLE could also be studied using a different framework. [12] and

[13] proved the consistency of the marginal likelihood MLE under the block models for undirected and directed networks, respectively. Their approach is to first prove the consistency of the MLE under the complete data likelihood and to further show that the marginal likelihood is asymptotically equivalent to the complete data likelihood, which implies the consistency of the MLE under the marginal likelihood. We plan to extend the above framework to the hub model for future works. Moreover, we plan to study the model selection consistency of the proposed hub set selection method, especially when n_L, n and T are all allowed to grow. What we would also like to explore is to go beyond the independence assumption and to develop theories and model selection methodologies for correlated or temporally dependent groups [75].

Finally, we briefly review other work on Bernoulli mixture models. [36] first showed that finite mixtures of Bernoulli products are not identifiable. [3] introduced and studied the concept of generic identifiability, which means that the set of non-identifiable parameters has Lebesgue measure zero. Identifiability under another class of mixture Bernoulli models has been recently studied [71, 35]. This class of models, for example, the DINA (Deterministic Input, Noisy "And" gate) model, has applications in psychological and educational research. The motivation, the model setup, and the proof techniques presented in this paper are all different from previous research, and the result of neither implies the other.

Chapter 9

TECHNICAL DETAILS OF STATISTICAL INFERENCE FOR IMPLICIT NETWORK STRUCTURES

9.1 Proofs in Chapter 6.1.2

Proof of Theorem 7. Let $(\tilde{\rho}, \tilde{A}) \in \mathcal{P}$ be a set of parameters such that $\mathbb{P}(g|\rho, A) = \mathbb{P}(g|\tilde{\rho}, \tilde{A})$ for all g. For all $i = 1, \ldots, n_L$, $k = n_L + 1, \ldots, n$, consider the probability that only i appears under parameterizations (ρ, A) and $(\tilde{\rho}, \tilde{A})$, respectively

$$\tilde{\rho}_i(1 - \tilde{A}_{ik}) \prod_{j=1,\ldots,n, j\neq i, j\neq k} (1 - \tilde{A}_{ij}) = \rho_i(1 - A_{ik}) \prod_{j=1,\ldots,n, j\neq i, j\neq k} (1 - A_{ij}),$$

and the probability that only i and k appear

$$\tilde{\rho}_i\tilde{A}_{ik} \prod_{j=1,\ldots,n, j\neq i, j\neq k} (1 - \tilde{A}_{ij}) = \rho_i A_{ik} \prod_{j=1,\ldots,n, j\neq i, j\neq k} (1 - A_{ij}).$$

As $A_{ij} < 1$ in condition (i), dividing the second equation by the first, we obtain $\tilde{A}_{ik}/(1-\tilde{A}_{ik}) = A_{ik}/(1-A_{ik})$ and hence $\tilde{A}_{ik} = A_{ik}$ for $i = 1, \ldots, n_L$, $k = n_L+1, \ldots, n$.

For any $i = 1, \ldots, n_L$, $i' = 1, \ldots, n_L$, $i \neq i'$, suppose that k is the follower such that $A_{ik} \neq A_{i'k}$. Consider the probability that only i and i' appear

$$\begin{aligned}
&\tilde{\rho}_i\tilde{A}_{ii'}(1 - \tilde{A}_{ik}) \prod_{j\neq i, j\neq i', j\neq k} (1 - \tilde{A}_{ij}) + \tilde{\rho}_{i'}\tilde{A}_{i'i}(1 - \tilde{A}_{i'k}) \prod_{j=1,\ldots,n, j\neq i, j\neq i', j\neq k} (1 - \tilde{A}_{i'j}) \\
=&\rho_i A_{ii'}(1 - A_{ik}) \prod_{j\neq i, j\neq i', j\neq k} (1 - A_{ij}) + \rho_{i'} A_{i'i}(1 - A_{i'k}) \prod_{j=1,\ldots,n, j\neq i, j\neq i', j\neq k} (1 - A_{i'j}),
\end{aligned}$$

and the probability that i, i' and k appear

$$\begin{aligned}
&\tilde{\rho}_i\tilde{A}_{ii'}\tilde{A}_{ik} \prod_{j=1,\ldots,n, j\neq i, j\neq i', j\neq k} (1 - \tilde{A}_{ij}) + \tilde{\rho}_{i'}\tilde{A}_{i'i}\tilde{A}_{i'k} \prod_{j=1,\ldots,n, j\neq i, j\neq i', j\neq k} (1 - \tilde{A}_{i'j}) \\
=&\rho_i A_{ii'}A_{ik} \prod_{j=1,\ldots,n, j\neq i, j\neq i', j\neq k} (1 - A_{ij}) + \rho_{i'} A_{i'i}A_{i'k} \prod_{j=1,\ldots,n, j\neq i, j\neq i', j\neq k} (1 - A_{i'j}).
\end{aligned}$$

As $\tilde{A}_{ik} = A_{ik}$ for $i = 1, \ldots, n_L$, $k = n_L + 1, \ldots, n$, the above two equations become

$$\begin{aligned}
&\tilde{\rho}_i \tilde{A}_{ii'}(1 - A_{ik}) \prod_{j \neq i, j \neq i', j \neq k} (1 - \tilde{A}_{ij}) + \tilde{\rho}_{i'} \tilde{A}_{i'i}(1 - A_{i'k}) \prod_{j=1,\ldots,n, j \neq i, j \neq i', j \neq k} (1 - \tilde{A}_{i'j}) \\
=&\rho_i A_{ii'}(1 - A_{ik}) \prod_{j \neq i, j \neq i', j \neq k} (1 - A_{ij}) + \rho_{i'} A_{i'i}(1 - A_{i'k}) \prod_{j=1,\ldots,n, j \neq i, j \neq i', j \neq k} (1 - A_{i'j}),
\end{aligned} \tag{9.1}$$

$$\begin{aligned}
&\tilde{\rho}_i \tilde{A}_{ii'} A_{ik} \prod_{j \neq i, j \neq i', j \neq k} (1 - \tilde{A}_{ij}) + \tilde{\rho}_{i'} \tilde{A}_{i'i} A_{i'k} \prod_{j=1,\ldots,n, j \neq i, j \neq i', j \neq k} (1 - \tilde{A}_{i'j}) \\
=&\rho_i A_{ii'} A_{ik} \prod_{j \neq i, j \neq i', j \neq k} (1 - A_{ij}) + \rho_{i'} A_{i'i} A_{i'k} \prod_{j=1,\ldots,n, j \neq i, j \neq i', j \neq k} (1 - A_{i'j}).
\end{aligned} \tag{9.2}$$

(9.1) and (9.2) can be viewed as a system of linear equations with unknown variables

$$\tilde{\rho}_i \tilde{A}_{ii'} \prod_{j=1,\ldots,n, j \neq i, j \neq i', j \neq k} (1 - \tilde{A}_{ij}),$$

and

$$\tilde{\rho}_{i'} \tilde{A}_{i'i} \prod_{j=1,\ldots,n, j \neq i, j \neq i', j \neq k} (1 - \tilde{A}_{i'j}).$$

By condition (ii), as $A_{ik} \neq A_{i'k}$, the system has full rank and hence has one and only one solution:

$$\begin{aligned}
\tilde{\rho}_i \tilde{A}_{ii'} \prod_{j=1,\ldots,n, j \neq i, j \neq i', j \neq k} (1 - \tilde{A}_{ij}) &= \rho_i A_{ii'} \prod_{j=1,\ldots,n, j \neq i, j \neq i', j \neq k} (1 - A_{ij}), \\
\tilde{\rho}_{i'} \tilde{A}_{i'i} \prod_{j=1,\ldots,n, j \neq i, j \neq i', j \neq k} (1 - \tilde{A}_{ij}) &= \rho_{i'} A_{i'i} \prod_{j=1,\ldots,n, j \neq i, j \neq i', j \neq k} (1 - A_{i'j}).
\end{aligned} \tag{9.3}$$

Combining (9.3) with

$$\tilde{\rho}_i (1 - \tilde{A}_{ii'}) \prod_{j=1,\ldots,n, j \neq i, j \neq i', j \neq k} (1 - \tilde{A}_{ij}) = \rho_i (1 - A_{ii'}) \prod_{j=1,\ldots,n, j \neq i, j \neq i', j \neq k} (1 - A_{ij}),$$

we obtain $\tilde{A}_{ii'} = A_{ii'}$ for $i = 1, \ldots, n_L, i' = 1, \ldots, n_L$ by a similar argument to that at the beginning of the proof. It follows immediately that $\tilde{\rho}_i = \rho_i$ for $i = 1, \ldots, n_L$. □

Remark 3. *Neither conditions in Theorem 1 can be removed. That is, if either condition is removed, then there exists* $(\rho, A) \in \mathcal{P}$ *such that* (ρ, A) *is not identifiable. In fact,*

$$\rho = (1/2, 1/2),\ A = \begin{pmatrix} 1 & 1/2 & 0 \\ 1 & 1 & 1/2 \end{pmatrix}$$

and

$$\rho = (1/4, 3/4),\ A = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1/3 \end{pmatrix}$$

give the same probability distribution, which implies condition (i) – that is, $A_{ij} < 1$*, for* $i = 1, \ldots, n_L, j = 1, \ldots, n, i \neq j$*, is necessary.*

Moreover,

$$\rho = (1/2, 1/2),\ A = \begin{pmatrix} 1 & 1/2 & 1/2 \\ 1/2 & 1 & 1/2 \end{pmatrix}$$

and

$$\rho = (1/4, 3/4),\ A = \begin{pmatrix} 1 & 0 & 1/2 \\ 2/3 & 1 & 1/2 \end{pmatrix}$$

give the same probability distribution, which implies condition (ii) – that is, for all $i = 1, \ldots, n_L$*,* $i' = 1, \ldots, n_L, i \neq i'$*, there exists* $k \in \{n_L + 1, \ldots, n\}$ *such that* $A_{ik} \neq A_{i'k}$*, is necessary.*

Proof of Theorem 8. Let $(\tilde{\rho}, \tilde{A}) \in \mathcal{P}$ be a set of parameters of the hub model with the null component such that $\mathbb{P}(g|\rho, A) = \mathbb{P}(g|\tilde{\rho}, \tilde{A})$ for all g. Consider the probability that no one appears:

$$\tilde{\rho}_0 \prod_{j=1}^{n} (1 - \tilde{\pi}_j) = \rho_0 \prod_{j=1}^{n} (1 - \pi_j).$$

For $k = n_L + 1, \ldots, n$, consider the probability that only k appears:

$$\tilde{\rho}_0 \tilde{\pi}_k \prod_{j=1,\ldots,n, j\neq k} (1 - \tilde{\pi}_j) = \rho_0 \pi_k \prod_{j=1,\ldots,n, j\neq k} (1 - \pi_j).$$

From the above equations, we obtain

$$\begin{aligned} &\tilde{\pi}_k = \pi_k, \quad k = n_L + 1, \ldots, n, \\ &\tilde{\rho}_0 \prod_{j=1}^{n_L} (1 - \tilde{\pi}_j) = \rho_0 \prod_{j=1}^{n_L} (1 - \pi_j). \end{aligned} \tag{9.4}$$

By condition (iii), for $i = 1, \ldots, n_L$, let k and k' be the nodes from $\{n_L + 1, \ldots, n\}$ such that $\pi_k \neq A_{ik}$ and $\pi_{k'} \neq A_{ik'}$.

Consider the probability that i appears but no other nodes from $\{1, \ldots, n_L\}$ appears (the rest do not matter)

$$\begin{aligned} &\tilde{\rho}_0 \tilde{\pi}_i \prod_{j=1,\ldots,n_L, j\neq i} (1 - \tilde{\pi}_j) + \tilde{\rho}_i \prod_{j=1,\ldots,n_L, j\neq i} (1 - \tilde{A}_{ij}) \\ =&\rho_0 \pi_i \prod_{j=1,\ldots,n_L, j\neq i} (1 - \pi_j) + \rho_i \prod_{j=1,\ldots,n_L, j\neq i} (1 - A_{ij}); \end{aligned} \tag{9.5}$$

the probability that i and k appear but no other nodes from $\{1, \ldots, n_L\}$ appears (the rest do not matter)

$$\begin{aligned} &\tilde{\rho}_0 \tilde{\pi}_i \prod_{j=1,\ldots,n_L, j\neq i} (1 - \tilde{\pi}_j) \pi_k + \tilde{\rho}_i \prod_{j=1,\ldots,n_L, j\neq i} (1 - \tilde{A}_{ij}) \tilde{A}_{ik} \\ =&\rho_0 \pi_i \prod_{j=1,\ldots,n_L, j\neq i} (1 - \pi_j) \pi_k + \rho_i \prod_{j=1,\ldots,n_L, j\neq i} (1 - A_{ij}) A_{ik}; \end{aligned} \tag{9.6}$$

the probability that i and k' appear but no other nodes from $\{1, \ldots, n_L\}$ appears (the rest do not matter)

$$\begin{aligned} &\tilde{\rho}_0 \tilde{\pi}_i \prod_{j=1,\ldots,n_L, j\neq i} (1 - \tilde{\pi}_j) \pi_{k'} + \tilde{\rho}_i \prod_{j=1,\ldots,n_L, j\neq i} (1 - \tilde{A}_{ij}) \tilde{A}_{ik'} \\ =&\rho_0 \pi_i \prod_{j=1,\ldots,n_L, j\neq i} (1 - \pi_j) \pi_{k'} + \rho_i \prod_{j=1,\ldots,n_L, j\neq i} (1 - A_{ij}) A_{ik'}; \end{aligned} \tag{9.7}$$

and the probability that i, k and k' appear but no other nodes from $\{1, \ldots, n_L\}$ appears (the rest do not matter)

$$\begin{aligned}
&\tilde{\rho}_0\tilde{\pi}_i \prod_{j=1,\ldots,n_L, j\neq i} (1-\tilde{\pi}_j)\pi_k\pi_{k'} + \tilde{\rho}_i \prod_{l=1,\ldots,n_L, j\neq i} (1-\tilde{A}_{ij})\tilde{A}_{ik}\tilde{A}_{ik'} \\
=&\rho_0\pi_i \prod_{j=1,\ldots,n_L, j\neq i} (1-\pi_j)\pi_k\pi_{k'} + \rho_i \prod_{l=1,\ldots,n_L, j\neq i} (1-A_{ij})A_{ik}A_{ik'}. \qquad (9.8)
\end{aligned}$$

Note that the above equations are not probabilities of a single realization g but are sums of multiple $\mathbb{P}(g)$. Moreover, we put $\pi_k, \pi_{k'}$ instead of $\tilde{\pi}_k, \tilde{\pi}_{k'}$ on the LHS of the equations, since we have proved $\tilde{\pi}_k = \pi_k, k = n_L + 1, \ldots, n$.

Let

$$\begin{aligned}
x &= \rho_0\pi_i \prod_{j=1,\ldots,n_L, j\neq i} (1-\pi_j), \\
\tilde{x} &= \tilde{\rho}_0\tilde{\pi}_i \prod_{j=1,\ldots,n_L, j\neq i} (1-\tilde{\pi}_j), \\
y &= \rho_i \prod_{j=1,\ldots,n_L, j\neq i} (1-A_{ij}), \\
\tilde{y} &= \tilde{\rho}_i \prod_{l=1,\ldots,n_L, j\neq i} (1-\tilde{A}_{ij}).
\end{aligned}$$

Then (9.5), (9.6) (9.7) and (9.8) become

$$\begin{aligned}
\tilde{x} + \tilde{y} &= x + y, \\
\tilde{x}\pi_k + \tilde{y}\tilde{A}_{ik} &= x\pi_k + yA_{ik}, \\
\tilde{x}\pi_{k'} + \tilde{y}\tilde{A}_{ik'} &= x\pi_{k'} + yA_{ik'}, \\
\tilde{x}\pi_k\pi_{k'} + \tilde{y}\tilde{A}_{ik}\tilde{A}_{ik'} &= x\pi_k\pi_{k'} + yA_{ik}A_{ik'}.
\end{aligned}$$

Plugging $\tilde{x} - x = y - \tilde{y}$ into the last three equations, we obtain

$$\tilde{y}\tilde{A}_{ik} = \tilde{y}\pi_k + y(A_{ik} - \pi_k), \qquad (9.9)$$

$$\tilde{y}\tilde{A}_{ik'} = \tilde{y}\pi_{k'} + y(A_{ik'} - \pi_{k'}), \qquad (9.10)$$

$$y\pi_k\pi_{k'} + \tilde{y}\tilde{A}_{ik}\tilde{A}_{ik'} = \tilde{y}\pi_k\pi_{k'} + yA_{ik}A_{ik'}. \qquad (9.11)$$

Multiplying (9.11) by $\tilde{y}$, and plugging the right hand sides of (9.9) and (9.10) into the resulting equation, we obtain

$$\begin{aligned}
&y\tilde{y}\pi_k\pi_{k'} + \tilde{y}^2\pi_k\pi_{k'} + \tilde{y}\pi_k y(A_{ik'} - \pi_{k'}) + \tilde{y}\pi_{k'} y(A_{ik} - \pi_k) + y^2(A_{ik} - \pi_k)(A_{ik'} - \pi_{k'}) \\
&= \tilde{y}^2\pi_k\pi_{k'} + y\tilde{y}A_{ik}A_{ik'}, \\
\Rightarrow &y(A_{ik} - \pi_k)(A_{ik'} - \pi_{k'}) = \tilde{y}(A_{ik} - \pi_k)(A_{ik'} - \pi_{k'}).
\end{aligned}$$

Therefore, $\tilde{y} = y$ since $\pi_k \neq A_{ik}$ and $\pi_{k'} \neq A_{ik'}$. It follows that $\tilde{x} = x$, i.e.,

$$\tilde{\rho}_0\tilde{\pi}_i \prod_{j=1,\ldots,n_L, j\neq i} (1 - \tilde{\pi}_j) = \rho_0\pi_i \prod_{j=1,\ldots,n_L, j\neq i} (1 - \pi_j), \quad i = 1, \ldots, n_L.$$

Combining the above equation with (9.4), we obtain

$$\tilde{\pi}_i = \pi_i, \quad i = 1, \ldots, n_L,$$

$$\tilde{\rho}_0 = \rho_0.$$

Note that $\mathbb{P}(g) = \mathbb{P}(g|z = 0)\mathbb{P}(z = 0) + \mathbb{P}(g|z \neq 0)\mathbb{P}(z \neq 0)$. So far we have proved parameters of $\mathbb{P}(g|z = 0)$ and $\mathbb{P}(z = 0)$ are identifiable. We only need to prove the identifiability of $\mathbb{P}(g|z \neq 0)$, which is the case of the asymmetric hub model and has been proved by Theorem 1. □

Remark 4. *No conditions in Theorem 2 can be removed. Here we only give a counterexample when condition (iii), that is, for any $i = 1, \ldots, n_L$, there exist $k \in \{n_L + 1, \ldots, n\}$ and $k' \in \{n_L + 1, \ldots, n\}$ such that $\pi_k \neq A_{ik}$ and $\pi_{k'} \neq A_{ik'}$, is not satisfied since the other two are similar to the case of Theorem 1. In fact,*

$$\rho = (1/2, 1/2), \ A = \begin{pmatrix} 1/2 & 0 & 1/2 \\ 1 & 1/2 & 1/2 \end{pmatrix}$$

and

$$\rho = (1/4, 3/4), \ A = \begin{pmatrix} 0 & 0 & 1/2 \\ 1 & 1/3 & 1/2 \end{pmatrix}$$

give the same probability distribution.

9.2 Proofs in Chapter 6.1.3

We start by recalling notations defined in the main text. Recall that z_* is the true label assignment, z is an arbitrary label assignment, and $\hat{z}$ is the maximum profile likelihood estimator. Furthermore, $t_{i*} = \sum_t 1(z_*^{(t)} = i)$, and $t_i = \sum_t 1(z^{(t)} = i), t_{ii'} = \sum_t 1(z_*^{(t)} = i, \hat{z}^{(t)} = i')$.

Proof of Lemma 5. We first prove a fact: under H_1 and H_4, for $0 < \delta_1 < e^{-c_0}$,

$$\mathbb{P}\left(\bigcup_{i=1}^{n_L}\left\{\frac{t_{ii}}{t_{i*}} \le \delta_1\right\}\right) \to 0.$$

Note that $\hat{z}$ must be feasible (the estimated hub must appear in the group as we assume $A_{ii} \equiv 1$), we have

$$\begin{aligned}
&\mathbb{P}\left(\frac{t_{ii}}{t_{i*}} \le \delta_1 \middle| z_*\right) \\
\le &\mathbb{P}\left(\frac{1}{t_{i*}}\sum_{t=1}^{T} 1(z_*^{(t)} = i) \prod_{k\in\{1,\dots,n_L\},k\neq i} (1 - G_k^{(t)}) \le \delta_1 \middle| z_*\right). \qquad (9.12)
\end{aligned}$$

Now since

$$\mathbb{E}\left[\prod_{k\in\{1,\dots,n_L\},k\neq i} (1 - G_k^{(t)}) \middle| z_*^{(t)} = i\right] = \prod_{k\in\{1,\dots,n_L\},k\neq i} (1 - A_{ik}) \ge (1 - c_0/n_L)^{n_L} \ge e^{-c_0},$$

by Hoeffding's inequality,

$$\begin{aligned}
(9.12) \le &\mathbb{P}\left(\frac{1}{t_{i*}}\sum_{t=1}^{T} 1(z_*^{(t)} = i)\left[\prod_{k\in\{1,\dots,n_L\},k\neq i} (1 - G_k^{(t)}) - \prod_{k\in\{1,\dots,n_L\},k\neq i} (1 - A_{ik})\right] \le \delta_1 - e^{-c_0} \middle| z_*\right) \\
&\le \exp\{-2t_{i*}(e^{-c_0} - \delta_1)^2\}.
\end{aligned}$$

Hence

$$
\begin{aligned}
&\mathbb{P}\left(\bigcup_{i=1}^{n_L}\left\{\frac{t_{ii}}{t_{i*}} \le \delta_1\right\}\middle| z_*\right) \\
=&\mathbb{P}\left(\bigcup_{i=1}^{n_L}\left\{\frac{t_{ii}}{t_{i*}} \le \delta_1\right\}, \{t_{i*} \ge c_{\min}T/n_L, \text{for all } i\}\middle| z_*\right) \\
&+\mathbb{P}\left(\bigcup_{i=1}^{n_L}\left\{\frac{t_{ii}}{t_{i*}} \le \delta_1\right\}, \{t_{i*} < c_{\min}T/n_L, \text{for some } i\}\middle| z_*\right) \\
\le&\sum_{i=1}^{n_L}\mathbb{P}\left(\frac{t_{ii}}{t_{i*}} \le \delta_1\middle| z_*\right)1(t_{i*} \ge c_{\min}T/n_L) \\
&+1(t_{i*} < c_{\min}T/n_L, \text{for some } i) \\
\le&n_L \exp\{-2c_{\min}T/(n_L)(e^{-c_0}-\delta_1)^2\} + 1(t_{i*} < c_{\min}T/n_L, \text{for some } i).
\end{aligned}
$$

It follows that

$$
\begin{aligned}
&\mathbb{P}\left(\bigcup_{i=1}^{n_L}\left\{\frac{t_{ii}}{t_{i*}} \le \delta_1\right\}\right) \\
=&\mathbb{E}_{z_*}\left[\mathbb{P}\left(\bigcup_{i=1}^{n_L}\left\{\frac{t_{ii}}{t_{i*}} \le \delta_1\right\}\middle| z_*\right)\right] \\
\le&n_L \exp\{-2c_{\min}T/(n_L)(e^{-c_0}-\delta_1)^2\} + \mathbb{P}(t_{i*} < c_{\min}T/n_L, \text{for some } i) \to 0.
\end{aligned}
$$

Therefore, $\frac{t_{ii}}{t_{i*}} \ge \delta_1$ for $i = 1, \ldots, n_L$ with probability approaching 1.

Let $\mathcal{E} = \{\frac{t_{ii}}{t_{i*}} \ge \delta_1 \text{ and } t_{i*} \ge c_{\min}T/n_L, i = 1, \ldots, n_L\}$. We have shown $\mathbb{P}(\mathcal{E}) \to 1$. The inequalities below are proved within the set $\mathcal{E}$, and thus hold with probability approaching 1.

For $i = 1, \ldots, n_L$, $k = 1, \ldots, n_L$, $k \ne i$,

$$
\frac{t_{ik}}{t_k} = \frac{t_{ik}}{\sum_{k'=1}^{n_L} t_{k'k}} \le \frac{t_{ik}}{t_{ik}+t_{kk}} = \frac{t_{ik}/t_{i*}}{t_{ik}/t_{i*} + t_{kk}/t_{k*}\cdot t_{k*}/t_{i*}} \le \frac{1}{1+\delta_1 \cdot c_{\min}/c_{\max}} = \delta_2 < 1.
$$

Under H_2 and H_3, $\min_{i,i'=1,\ldots,n_L, i\neq i', j\in V_i} A_{ij} - A_{i'j} = \tau d$, where τ is bounded away from 0. Now we give a lower bound for $A_{ij} - \bar{A}_{kj}$ for $j \in V_i$ and $k \neq i$,

$$\begin{aligned} A_{ij} - \bar{A}_{kj} &= \frac{\sum_t (A_{ij} - P_j^{(t)}) 1(\hat{z}^{(t)} = k)}{t_k} \\ &= \frac{\sum_{k'=1}^{n_L} (A_{ij} - A_{k'j}) t_{k'k}}{t_k} \\ &\geq \frac{\tau d \sum_{k'\neq i} t_{k'k}}{t_k} \geq \tau(1-\delta_2)d. \end{aligned} \tag{9.13}$$

Next, we show the following fact: if $p = \rho_1 d, q = \rho_2 d$ where $\rho_1 > \rho_2$ are fixed positive numbers, then there exists $\delta_3 > 0$ such that $\mathrm{KL}(p,q) \geq \delta_3 d$.

$$\begin{aligned} \mathrm{KL}(p,q) &= p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q} \\ &= p\log\frac{p}{q} + \log(1 - \frac{p-q}{1-q}) + p\log\frac{1-p}{1-q} \\ &= -p\log\frac{q}{p} + \frac{p-q}{1-q} + o(d) + \rho_1 d\, o(1) \\ &\geq -p\log\frac{q}{p} + (q-p) + o(d) \\ &= p\left[\frac{q-p}{p} - \log\left(1 + \frac{q-p}{p}\right)\right] + o(d) \\ &\geq \delta_3 d. \end{aligned}$$

The last line holds for sufficiently small δ_3 because $\frac{q-p}{p} - \log(1 + \frac{q-p}{p}) = c_{\rho_1,\rho_2} > 0$ where $\frac{q-p}{p} \in (-1, 0)$ and c_{ρ_1,ρ_2} is a constant depending on ρ_1 and ρ_2.

As $\bar{A}_{kj} = \frac{\sum_t P_j^{(t)} 1(z^{(t)}=j)}{t_i} = [\sum_t A_{z_*^{(t)},j} 1(z^{(t)}=j)]/t_i \asymp d$, combining the above fact and (9.13), we have

$$\begin{aligned}
L_P(z_*) - L_P(\hat{z}) &= \sum_t \sum_j \mathrm{KL}(P_j^{(t)}, \bar{A}_{\hat{z}^{(t)},j}) \\
&\geq \sum_{i=1}^{n_L} \sum_{k\neq i} \sum_{t: z_*^{(t)}=i, \hat{z}^{(t)}=k} \sum_{j\in V_i} \mathrm{KL}(A_{ij}, \bar{A}_{\hat{z}^{(t)},j}) \\
&\geq \sum_{i=1}^{n_L} \sum_{k\neq i} \sum_{t: z_*^{(t)}=i, \hat{z}^{(t)}=k} \sum_{j\in V_i} \tau(1-\delta_2)\delta_3 d \\
&\geq \tau(1-\delta_2)\delta_3 dvnT_e/n_L.
\end{aligned}$$

Letting $\delta = 1/[\tau(1-\delta_2)\delta_3]$,

$$\frac{\delta n_L}{dvnT}(L_P(z_*) - L_P(\hat{z})) \geq \frac{T_e}{T},$$

with probability approaching 1. □

To prove Theorem 10, we need the following lemma.

Lemma S1.

$$\begin{aligned}
&\mathbb{P}(\max_z |L_G(z) - L_P(z)| \geq 2\eta) \leq \\
&\quad n_L^T (T/n_L + 1)^{n_L n} e^{-\eta} + 2n_L^T \exp\left\{ -\frac{\eta^2/4}{\sum_t \sum_j (\log \bar{A}_{ij})^2 \mathrm{Var}(G_j^{(t)}) + \max_{ij} |\log \bar{A}_{ij}| \eta/6} \right\} \\
&\quad + 2n_L^T \exp\left\{ -\frac{\eta^2/4}{\sum_{ij:\bar{A}_{ij}<1} \left((\log(1-\bar{A}_{ij}))^2 \sum_{t:z^{(t)}=i} \mathrm{Var}(G_j^{(t)}) \right) + \max_{ij:\bar{A}_{ij}<1} |\log(1-\bar{A}_{ij})| \eta/6} \right\}.
\end{aligned}$$

Proof of Lemma S1.

$$\begin{aligned}
L_G(z) - L_P(z) = &\left(\sum_{i=1}^{n_L} t_i \sum_j \hat{A}_{ij} \log \hat{A}_{ij} + (1 - \hat{A}_{ij}) \log(1 - \hat{A}_{ij}) \right) \\
&- \left(\sum_{i=1}^{n_L} t_i \sum_j \hat{A}_{ij} \log \bar{A}_{ij} + (1 - \hat{A}_{ij}) \log(1 - \bar{A}_{ij}) \right) \\
&+ \left(\sum_{i=1}^{n_L} t_i \sum_j \hat{A}_{ij} \log \bar{A}_{ij} + (1 - \hat{A}_{ij}) \log(1 - \bar{A}_{ij}) \right) \\
&- \left(\sum_{i=1}^{n_L} t_i \sum_j \bar{A}_{ij} \log \bar{A}_{ij} + (1 - \bar{A}_{ij}) \log(1 - \bar{A}_{ij}) \right) \\
= &\sum_{i=1}^{n_L} t_i \sum_j D(\hat{A}_{ij} | \bar{A}_{ij}) + B_{n_L,n,T}.
\end{aligned}$$

To bound $\sum_{i=1}^{n_L} t_i \sum_j D(\hat{A}_{ij}|\bar{A}_{ij})$, we adopt the approach in [22], which is based on a heterogeneous Chernoff bound in [25]. Let ν be any realization of $\hat{A}$.

$$\mathbb{P}(\hat{A}_{ij} = \nu_{ij} | z_*) \leq e^{-t_i D(\nu_{ij}|\bar{A}_{ij})}.$$

By the independence of $\hat{A}_{ij}$ conditional on z_*,

$$\mathbb{P}(\hat{A} = \nu | z_*) \leq \exp\left\{ - \sum_{i=1}^{n_L} \sum_j t_i D(\nu_{ij}|\bar{A}_{ij}) \right\}.$$

Let $\hat{\mathcal{A}}$ be the range of $\hat{A}$ for a fixed z. Then $|\hat{\mathcal{A}}| \leq \prod_{i=1}^{n_L} (t_i + 1)^n \leq \prod_{i=1}^{n_L} (t_i + 1)^n \leq (T/n_L + 1)^{n_L n}$, as $\hat{A}_{ij}$ can only take values from $0/t_i, 1/t_i, \ldots, t_i/t_i$.

For all $\eta > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^{n_L} \sum_j t_i D(\hat{A}_{ij}|\bar{A}_{ij}) \geq \eta \middle| z_* \right) \\
=&\sum_{\nu \in \hat{\mathcal{A}}} \mathbb{P}\left(\hat{A} = \nu, \sum_{i=1}^{n_L} \sum_j t_i D(\nu_{ij}|\bar{A}_{ij}) \geq \eta \middle| z_* \right) \\
\leq&\sum_{\nu \in \hat{\mathcal{A}}} \exp\left\{ - \sum_{i=1}^{n_L} \sum_j t_i D(\nu_{ij}|\bar{A}_{ij}) \right\} 1\left\{ - \sum_{i=1}^{n_L} \sum_j t_i D(\nu_{ij}|\bar{A}_{ij}) \leq -\eta \right\} \\
\leq&\sum_{\nu \in \hat{\mathcal{A}}} e^{-\eta} \leq |\hat{\mathcal{A}}| e^{-\eta} \leq (T/n_L + 1)^{n_L n} e^{-\eta},
\end{aligned}$$

and then

$$\mathbb{P}\left(\max_z \sum_{i=1}^{n_L}\sum_j t_i D(\hat{A}|\bar{A}_{ij}) \geq \eta\right) \leq n_L^T (T/n_L+1)^{n_L n} e^{-\eta}. \tag{9.14}$$

Next, we bound $B_{n_L,n,T}$. Let $B_{n_L,n,T} = B_{1,n_L,n,T} + B_{2,n_L,n,T}$, where

$$B_{1,n_L,n,T} = \sum_i \left(\sum_j \sum_{t:z^{(t)}=i} (G_j^{(t)} - P_j^{(t)}) \log \bar{A}_{ij}\right),$$

$$B_{2,n_L,n,T} = \sum_i \left(\sum_j \sum_{t:z^{(t)}=i} (G_j^{(t)} - P_j^{(t)}) \log(1-\bar{A}_{ij})\right).$$

As $\left|(G_j^{(t)} - P_j^{(t)}) \log \bar{A}_{ij}\right| \leq |\log \bar{A}_{ij}|$, by Bernstein's inequality, we have

$$\mathbb{P}(|B_{1,n_L,n,T}| \geq \eta/2) \leq 2\exp\left\{-\frac{\eta^2/4}{\sum_t \sum_j (\log \bar{A}_{ij})^2 \mathrm{Var}(G_j^{(t)}) + \max_{ij} |\log \bar{A}_{ij}|\eta/6}\right\},$$

$$\mathbb{P}(\max_z |B_{1,n_L,n,T}| \geq \eta/2) \leq 2n_L^T \exp\left\{-\frac{\eta^2/4}{\sum_t \sum_j (\log \bar{A}_{ij})^2 \mathrm{Var}(G_j^{(t)}) + \max_{ij} |\log \bar{A}_{ij}|\eta/6}\right\}.$$

In addition, if $\bar{A}_{ij} = 1$, $\sum_{t:z^{(t)}=i}(G_j^{(t)} - P_j^{(t)}) \equiv 0$, which implies the term $\sum_{t:z^{(t)}=i}(G_j^{(t)} - P_j^{(t)}) \log(1-\bar{A}_{ij})$ in $B_{2,n_L,n,T}$ can be dropped. As $\left|(G_j^{(t)} - P_j^{(t)}) \log(1-\bar{A}_{ij})\right| \leq |\log(1-\bar{A}_{ij})|$, by Bernstein's inequality,

$$\mathbb{P}(|B_{2,n_L,n,T}| \geq \eta/2)$$
$$\leq 2\exp\left\{-\frac{\eta^2/4}{\sum_{ij:\bar{A}_{ij}<1}\left((\log(1-\bar{A}_{ij}))^2 \sum_{t:z^{(t)}=i} \mathrm{Var}(G_j^{(t)})\right) + \max_{ij:\bar{A}_{ij}<1} |\log(1-\bar{A}_{ij})|\eta/6}\right\},$$

$$\mathbb{P}(\max_z |B_{2,n_L,n,T}| \geq \eta/2)$$
$$\leq 2n_L^T \exp\left\{-\frac{\eta^2/4}{\sum_{ij:\bar{A}_{ij}<1}\left((\log(1-\bar{A}_{ij}))^2 \sum_{t:z^{(t)}=i} \mathrm{Var}(G_j^{(t)})\right) + \max_{ij:\bar{A}_{ij}<1} |\log(1-\bar{A}_{ij})|\eta/6}\right\}.$$

Finally, combining (9.14), (9.2) and (9.2), we obtain

$$\mathbb{P}(\max_z |L_G(z) - L_P(z)| \geq 2\eta)$$
$$\leq \mathbb{P}\left(\max_z \sum_{i=1}^{n_L}\sum_j t_i D(\hat{A}|\bar{A}_{ij}) \geq \eta\right) + \mathbb{P}(\max_z |B_{1,n_L,n,T}| \geq \eta/2) + \mathbb{P}(\max_z |B_{2,n_L,n,T}| \geq \eta/2)$$
$$\leq n_L^T (T/n_L+1)^{n_L n} e^{-\eta} + 2n_L^T \exp\left\{-\frac{\eta^2/4}{\sum_t \sum_j (\log \bar{A}_{ij})^2 \mathrm{Var}(G_j^{(t)}) + \max_{ij} |\log \bar{A}_{ij}|\eta/6}\right\}$$
$$+ 2n_L^T \exp\left\{-\frac{\eta^2/4}{\sum_{ij:\bar{A}_{ij}<1}\left((\log(1-\bar{A}_{ij}))^2 \sum_{t:z^{(t)}=i} \mathrm{Var}(G_j^{(t)})\right) + \max_{ij:\bar{A}_{ij}<1} |\log(1-\bar{A}_{ij})|\eta/6}\right\}.$$

□

Proof of Theorem 3. First we show the following fact: under $H_1 - H_4$, if $n_L^2 \log T/(dTv) \to 0$, $(\log d)^2 n_L^2 \log n_L/(dnv^2) \to 0$ and $(\log T)^2 n_L^2 \log n_L/(dnv^2) \to 0$, then

$$\max_z \frac{n_L}{dvnT} |L_P(z) - L_G(z)| = o_p(1), \quad \text{as } n_L \to \infty, n \to \infty, T \to \infty. \tag{9.15}$$

Letting $\eta = dvnT\epsilon/n_L$, the LHS in Lemma S1 becomes $\mathbb{P}(\max_z \frac{n_L}{dvnT}|L_G(z) - L_P(z)| \geq 2\epsilon)$. To prove the above fact, we need to show each term in the RHS of Lemma S1 goes to 0.

For the first term, it is easy to check that if $n_L \log n_L/(dvn) \to 0$ and $n_L^2 \log T/(dvT) \to 0$, then

$$n_L^T (T/n_L)^{n_L n} e^{-\frac{dvnT\epsilon}{n_L}} \to 0.$$

Under H_2, $A_{ij} \asymp d$ and $|\log \bar{A}_{ij}| = O(|\log d|)$ for $i \neq j$. We can therefore find a constant C_1 such that

$$\mathbb{P}(|B_{1,n_L,n,T}| \geq dvnT\epsilon/(2n_L)) \leq 2\exp\left\{-\frac{d^2v^2n^2T^2\epsilon^2/(4n_L^2)}{C_1^2 Tn(\log d)^2 d + C_1|\log d| dvnT\epsilon/(6n_L)}\right\},$$

and

$$\mathbb{P}(\max_z |B_{1,n_L,n,T}| \geq dvnT\epsilon/(2n_L)) \leq 2n_L^T \exp\left\{-\frac{d^2v^2n^2T^2\epsilon^2/(4n_L^2)}{C_1^2 Tn(\log d)^2 d + C_1|\log d| dvnT\epsilon/(6n_L)}\right\}.$$

Then if $(\log d)^2 n_L^2 \log n_L/(dnv^2) \to 0$,

$$\mathbb{P}(\max_z |B_{1,n_L,n,T}| \geq dvnT\epsilon/(2n_L)) \to 0.$$

For the third term, when $\bar{A}_{ij} < 1$, we have

$$\bar{A}_{ij} \leq \frac{(t_i - 1) + P_j^{(t)}}{t_i},$$
$$1 - \bar{A}_{ij} \geq \frac{1 - P_j^{(t)}}{t_i} \geq \frac{1 - P_j^{(t)}}{T},$$

which imply $|\log(1-\bar{A}_{ij})| \le C_2 \log T$ for some constant $C_2 > 0$. Therefore,

$$\mathrm{P}\left(\max_z |B_{2,n_L,n,T}| \ge dvnT\epsilon/(2n_L)\right) \le 2n_L^T \exp\left\{-\frac{d^2v^2n^2T^2\epsilon^2/(4n_L^2)}{C_2^2(\log T)^2Tnd + C_2(\log T)dnvT\epsilon/(6n_L)}\right\}.$$

Furthermore, if $(\log T)^2 n_L^2 \log n_L/(dnv^2) \to 0$,

$$\mathbb{P}(\max_z |B_{2,n_L,n,T}| \ge dvnT\epsilon/(2n_L)) \to 0.$$

Combining the inequalities of the above three terms, we have proved (9.15).

Finally, for all $\epsilon > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\frac{T_e}{T} \ge \epsilon\right) \\
=&\mathbb{P}\left(\frac{T_e}{T} \ge \epsilon, \frac{\delta n_L}{dvnT}(L_P(z_*) - L_P(\hat{z})) \ge \frac{T_e}{T}\right) \\
&+ \mathbb{P}\left(\frac{T_e}{T} \ge \epsilon, \frac{\delta n_L}{dvnT}(L_P(z_*) - L_P(\hat{z})) < \frac{T_e}{T}\right) \\
=&\mathbb{P}\left(\frac{\delta n_L}{dvnT}(L_P(z_*) - L_P(\hat{z})) \ge \epsilon\right) + o(1) \quad \text{(by Lemma 1)} \\
=&\mathbb{P}\left(\frac{\delta n_L}{dvnT}\left[(L_P(z_*) - L_G(z_*)) + (L_G(z_*) - L_G(\hat{z})) + (L_G(\hat{z}) - L_P(\hat{z}))\right] \ge \epsilon\right) + o(1) \\
\le&\mathbb{P}\left(\frac{\delta n_L}{dvnT}\left(|L_P(z_*) - L_G(z_*)| + |L_G(\hat{z}) - L_P(\hat{z})|\right) \ge \epsilon\right) + o(1) \\
\to& 0.
\end{aligned}$$

□

We now give the result of label consistency for fixed n_L. We make the following assumptions similar to H_1 – H_4.

H_1': $c_{\min} T \le t_{i*} \le c_{\max} T$ for $i = 1, \ldots, n_L$.

H_2': $A_{ij} = s_{ij} d$ for $i = 1, \ldots, n_L$,$j = 1, \ldots, n$ and $i \ne j$ where s_{ij} are unknown constants satisfying $0 < s_{\min} \le s_{ij} \le s_{\max} < \infty$ while d goes to 0 as n goes to infinity.

H_3': There exists a set $V_i \subset \{n_L+1,\ldots,n\}$ for $i=1,\ldots,n_L$ with $|V_i| \geq vn$ such that $\tau = \min_{i,i'=1,\ldots,n_L, i\neq i', j\in V_i}(s_{ij}-s_{i'j})$ is bounded away from 0.

H_4': $A_{ii'}$ is bounded away from 1 for $i=1,\ldots,n_L$,$i'=1,\ldots,n_L$ and $i\neq i'$.

Theorem 3′. *Under* $H_1'-H_4'$, *if* $\log T/(dTv)=o(1)$, $(\log d)^2/(dnv^2)=o(1)$ *and* $(\log T)^2/(dnv^2)=o(1)$, *then*

$$T_e/T = o_p(1), \quad \text{as } n\to\infty, T\to\infty.$$

We omit all the proofs for fixed n_L because they are trivial corollaries of the results for growing n_L.

Proof of Theorem 4. First we show the following fact: under the conditions in Theorem 4,

$$n_L T_e/T = o_p(1), \text{ as } n_L\to\infty, n\to\infty, T\to\infty.$$

According to the proof in Theorem 3, we need

$$\mathbb{P}\left(\frac{\delta n_L^2}{dvnT}\left(|L_P(z_*)-L_G(z_*)|+|L_G(\hat{z})-L_P(\hat{z})|\right)\geq\epsilon\right)\to 0,$$

which holds if we can show

$$\max_z \frac{n_L^2}{dvnT}|L_G(z)-L_P(z)| = o_p(1).$$

As in the proof of Lemma S1, this holds by letting $\eta = dvnT\epsilon/n_L^2$.

Then we bound $\left|\hat{A}_{ij}^{\hat{z}}-\hat{A}_{ij}^{z_*}\right|$:

$$|\hat{A}_{ij}^{\hat{z}}-\hat{A}_{ij}^{z_*}| = \left|\frac{\sum_t G_j^{(t)}1(\hat{z}^{(t)}=i)}{t_i}-\frac{\sum_t G_j^{(t)}1(z_*^{(t)}=i)}{t_{i*}}\right|$$

$$\leq \left|\frac{\sum_t G_j^{(t)}1(\hat{z}^{(t)}=i)}{t_i}-\frac{\sum_t G_j^{(t)}1(\hat{z}^{(t)}=i)}{t_{i*}}\right| + \left|\frac{\sum_t G_j^{(t)}1(\hat{z}^{(t)}=i)}{t_{i*}}-\frac{\sum_t G_j^{(t)}1(z_*^{(t)}=i)}{t_{i*}}\right|$$

$$\leq \left|\frac{t_{i*}-t_i}{t_{i*}}\right| + \frac{\sum_t\left|1(\hat{z}^{(t)}=i)-1(z_*^{(t)}=i)\right|}{t_{i*}} \leq \delta n_L T_e/T,$$

where δ is a constant. The last line holds by H_1'.

Furthermore,

$$\begin{aligned}
&\mathbb{P}\left(\max_{ij}\left|\hat{A}_{ij}^{\hat{z}}-A_{ij}\right|\geq\epsilon\right)\\
\leq&\mathbb{P}\left(\max_{ij}\left|\hat{A}_{ij}^{\hat{z}}-\hat{A}_{ij}^{z_*}\right|\geq\epsilon/2\right)+P\left(\max_{ij}\left|\hat{A}_{ij}^{z_*}-A_{ij}\right|\geq\epsilon/2\right)\\
\leq&\mathbb{P}\left(\delta n_L T_e/T\geq\epsilon\right)+P\left(\max_{ij}\left|\hat{A}_{ij}^{z_*}-A_{ij}\right|\geq\epsilon/2\right).
\end{aligned}$$

The second term vanishes by Hoeffding's inequality: for all $\epsilon>0$,

$$\begin{aligned}
&\mathbb{P}\left(\left|\hat{A}_{ij}^{z_*}-A_{ij}\right|\geq\epsilon/2\middle|\, z_*\right)\\
=&\mathbb{P}\left(\left|\sum_t 1(z_*^{(t)}=i)(G_j^{(t)}-A_{ij})\right|\geq\epsilon t_{i*}/2\middle|\, z_*\right)\\
\leq&2\exp\{-\epsilon^2 t_{i*}/2\}.
\end{aligned}$$

Therefore, if $n_L\log n/T\to 0$,

$$\begin{aligned}
&\mathbb{P}\left(\max_{ij}\left|\hat{A}_{ij}^{z_*}-A_{ij}\right|\geq\epsilon/2\right)\\
\leq&2nn_L\exp\{-\epsilon^2 c_{\min}T/(2n_L)\}+\mathbb{P}(t_{i*}<c_{\min}T/n_L,\text{for some } i)\to 0.
\end{aligned}$$

□

The following theorem is on estimation consistency for fixed n.

Theorem 4′. *Under $H_1'-H_4'$, if $\log n/T=o(1)$, $\log T/(dTv)=o(1)$, $(\log d)^2/(dnv^2)=o(1)$ and $(\log T)^2/(dnv^2)=o(1)$, then*

$$\max_{i\in\{1,\ldots,n_L\},j\in\{1,\ldots,n\}}\left|\hat{A}_{ij}^{z_*}-A_{ij}\right|=o_p(1),\quad\text{as } n\to\infty,T\to\infty.$$

Finally, we give the simplest version of the estimation consistency result, which only considers the rates of n and T but treats n_L, d, and v as fixed.

Theorem 4″. *Under $H_1' - H_4'$, for fixed d and v, if $\log n/T = o(1)$ and $(\log T)^2/n = o(1)$, then*

$$\max_{i\in\{1,\ldots,n_L\},j\in\{1,\ldots,n\}} \left|\hat{A}_{ij}^{z_*} - A_{ij}\right| = o_p(1), \quad \text{as } n\to\infty, T\to\infty.$$

The first condition means n can grow faster than T as long as $\log n/T \to 0$. Such a condition is common in the literature of high-dimensional statistics. The second condition is more of a technical one: for proving the label consistency, we need an upper bound of the growth rate of T due to the concentration bound in Lemma S1.

Proof of Lemma 2. By the proof of Lemma 1, there exists $\delta_1 > 0$ such that

$$t_{ii} + t_{i0} \geq \delta_1 t_{i*}, \quad i = 1,\ldots,n_L, \tag{9.16}$$

$$t_{00} \geq \delta_1 t_{0*}, \tag{9.17}$$

with probability approaching 1.

Therefore [1] , for $i = 1,\ldots,n_L$, $j \in V_i$,

$$\begin{aligned} A_{ij} - \bar{A}_{0j} &= \frac{\sum_t (A_{ij} - P_j^{(t)}) 1(\hat{z}^{(t)} = 0)}{t_0} \\ &= \frac{\sum_{k=0}^{n_L} (A_{ij} - A_{kj}) t_{k0}}{t_0} \\ &\geq \frac{(A_{ij} - A_{0j}) t_{00}}{t_0} \geq \tau d \frac{t_{00}}{T} \geq \tau d \frac{t_{00}}{(n_L+1) t_{0*}/c_{\min}} \geq \frac{\tau d c_{\min} \delta_1}{n_L}. \end{aligned}$$

[1] Some inequalities below hold with probability approaching 1. We omit this sentence occasionally.

Using the same argument in Lemma 1, it follows that

$$\begin{aligned} L_P(z_*) - L_P(\hat{z}) &= \sum_t \sum_j \mathrm{KL}(P_j^{(t)}, \bar{A}_{\hat{z}^{(t)},j}) \\ &\geq \max_{i=1,\ldots,n_L} \sum_{t:z_*^{(t)}=i,\hat{z}^{(t)}=0} \sum_{j\in V_i} \mathrm{KL}(A_{ij}, \bar{A}_{0j}) \\ &\geq \max_{i=1,\ldots,n_L} \frac{\tau d c_{\min}\delta_1\delta_3}{n_L}\frac{vn}{n_L}t_{i0} \\ &\geq \max_{i=1,\ldots,n_L} \frac{\tau d c_{\min}\delta_1\delta_3}{n_L}\frac{vn}{n_L}\frac{t_{i0}}{t_{i*}}\frac{c_{\min}T}{n_L} \\ &\geq \max_{i=1,\ldots,n_L} \tau\epsilon\frac{dvnT}{n_L^3}\frac{t_{i0}}{t_{i*}}, \end{aligned} \tag{9.18}$$

where ϵ is a positive constant and τ is bounded away from 0.

Next, we show the following fact: under the conditions in Lemma 2,

$$\max_z \frac{n_L^3}{dvnT}|L_G(z) - L_P(z)| = o_p(1).$$

As in the proofs of Lemma S1 and Theorem 3, the above statement holds by letting $\eta = dvnT\epsilon/n_L^3$. Combining (9.18) and the above fact, by the same argument in Theorem 3, we have

$$\mathbb{P}\left(\max_{i=1,\ldots,n_L}\frac{t_{i0}}{t_{i*}} \leq \eta\right) \to 1. \tag{9.19}$$

□

Proof of Theorem 5. Due to (9.16) and (9.19), there exists $\delta_2 > 0$ such that

$$t_{ii} \geq \delta_2 t_{i*} \quad \text{for } i = 0, \ldots, n_L,$$

with probability approaching 1. By the same argument in Lemma 1,

$$\begin{aligned} L_P(z_*) - L_P(\hat{z}) &= \sum_{t=1}^{T}\sum_{j=1}^{n} \mathrm{KL}(P_j^{(t)}, \bar{A}_{\hat{z}^{(t)},j}) \\ &\geq \sum_{i=1}^{n_L} \sum_{0\leq k\leq n_L, k\neq i} \sum_{t:z_*^{(t)}=i,\hat{z}^{(t)}=k} \sum_{j\in V_i} \mathrm{KL}(A_{ij}, \bar{A}_{kj}) \\ &\geq \frac{vn}{n_L}\sum_{i=1}^{n_L} \sum_{0\leq k\leq n_L, k\neq i} t_{ik}\tau(1-\delta_2)\delta_3 d, \end{aligned}$$

which implies that there exists $\delta > 0$ such that with probability approaching 1,

$$\frac{\delta n_L}{dvnT}(L_P(z_*) - L_P(\hat{z})) \geq \sum_{i=1}^{n_L} \sum_{0 \leq k \leq n_L, k \neq i} \frac{t_{ik}}{T}. \tag{9.20}$$

By the same argument in Theorem 3, this further implies

$$\sum_{i=1}^{n_L} \sum_{0 \leq k \leq n_L, k \neq i} \frac{t_{ik}}{T} = o_p(1), \quad \text{as } n_L \to \infty, n \to \infty, T \to \infty, \tag{9.21}$$

if $n_L^2 \log T/(dvT) = o(1)$, $n_L^2 (\log T)^2 \log n_L/(dnv^2) = o(1)$ and
$n_L^2 (\log d)^2 \log n_L/(dnv^2) = o(1)$.

As in the proof of Theorem 4,

$$\sum_{i=1}^{n_L} \sum_{0 \leq k \leq n_L, k \neq i} (n_L + 1)\frac{t_{ik}}{T} = o_p(1), \quad \text{as } n_L \to \infty, n \to \infty, T \to \infty, \tag{9.22}$$

if $n_L^3 \log T/(dvT) = o(1)$, $n_L^4 (\log T)^2 \log n_L/(dnv^2) = o(1)$ and
$n_L^4 (\log d)^2 \log n_L/(dnv^2) = o(1)$.

Now we bound t_{0i}, $i = 1, \ldots, n_L$. From (9.22), $\sum_{1 \leq k \leq n_L, k \neq i} t_{ki} = o_p(T/(n_L + 1))$. And from $\delta_2 T c_{\min}/(n_L + 1) \leq \delta_2 t_{i*} \leq t_{ii}$, $\sum_{1 \leq k \leq n_L, k \neq i} t_{ki} \leq t_{ii}$, with probability approaching 1. Moreover, from (9.17), $t_{0i} \leq (1 - \delta_1) t_{0*}$.

Therefore, there exists $\delta_4 > 0$ such that for $i = 1, \ldots, n_L$, $j \in V_i$,

$$\begin{aligned}
A_{ij} - \bar{A}_{ij} &= \frac{\sum_t (A_{ij} - P_j^{(t)}) 1(\hat{z}^{(t)} = i)}{t_i} \\
&\geq \frac{(A_{ij} - A_{0j}) t_{0i}}{t_i} \\
&\geq \frac{\tau d t_{0i}}{t_{0i} + t_{ii} + \sum_{1 \leq k \leq n_L, k \neq i} t_{ki}} \\
&\geq \frac{\tau d t_{0i}}{(1 - \delta_1) t_{0*} + 2 t_{ii}} \\
&\geq \frac{\tau d t_{0i}}{(1 - \delta_1) t_{0*} + 2 t_{i*}} \geq \frac{\tau d n_L t_{0i}}{\delta_4 T}.
\end{aligned}$$

It follows that

$$\begin{aligned} L_P(z_*) - L_P(\hat{z}) &\geq \max_{i=1,\ldots,n_L} \sum_{t:z_*^{(t)}=i,\hat{z}^{(t)}=i} \sum_{j\in V_i} \mathrm{KL}(A_{ij}, \bar{A}_{ij}) \\ &\geq \max_{i=1,\ldots,n_L} \frac{\tau d n_L t_{0i} \delta_3}{\delta_4 T} \frac{vn}{n_L+1} t_{ii} \\ &\geq \max_{i=1,\ldots,n_L} \frac{d}{\delta_4} \frac{n_L t_{0i}}{T} \frac{vn}{n_L+1} \tau \delta_2 \delta_3 t_{i*} \\ &\geq \max_{i=1,\ldots,n_L} \frac{d}{\delta_4} \frac{n_L t_{0i}}{T} \frac{vn}{n_L+1} \tau \delta_2 \delta_3 T \frac{c_{\min}}{n_L+1} \\ &\geq \max_{i=1,\ldots,n_L} \frac{dvnT}{n_L^2} \frac{n_L t_{0i}}{T} \delta, \end{aligned} \tag{9.23}$$

where $\delta = \tau\delta_2\delta_3 c_{min}/\delta_4$ is positive constant.

By using the same argument in Theorem 3,

$$\max_{i=1,\ldots,n_L} \frac{n_L t_{0i}}{T} = o_p(1), \tag{9.24}$$

if $n_L^4 \log T/(dvT) = o(1)$, $(\log T)^2 n_L^6 \log n_L/(dnv^2) = o(1)$ and $n_L^6 (\log d)^2 \log n_L/(dnv^2) = o(1)$. It follows that

$$\sum_{i=1}^{n_L} \frac{t_{0i}}{T} = o_p(1).$$

Combining (9.21) and (9.24),

$$\frac{T_e}{T} = o_p(1), \quad \text{as } n \to \infty, T \to \infty.$$

□

For label consistency under the hub model with the null component with fixed n_L, we make the following assumptions:

$H_1^{*\prime}$: $Tc_{\min}/n_L \leq t_{i*} \leq Tc_{\max}/n_L$ for $i = 0, \ldots, n_L$.

$H_2^{*\prime}$: $A_{ij} = s_{ij} d$ for $i = 0, \ldots, n_L$,$j = 1, \ldots, n$ and $i \neq j$ where s_{ij} are unknown constants satisfying $0 < s_{\min} \leq s_{ij} \leq s_{\max} < \infty$ while d goes to 0 as n goes to infinity.

$H_3^{*\prime}$: There exists a set $V_i \subset \{n_L+1,\ldots,n\}$ for $i=1,\ldots,n_L$ with $|V_i| \geq vn$ such that $\tau = \min_{i=1,\ldots,n_L, i'=0,\ldots,n_L, i\neq i', j\in V_i}(s_{ij}-s_{i'j})$ is bounded away from 0.

$H_4^{*\prime}$: $A_{ii'}$ is bounded away from 1 for $i=0,\ldots,n_L$,$i'=1,\ldots,n_L$ and $i\neq i'$.

Theorem 5′. *Under $H_1^{*\prime}-H_4^{*\prime}$, if $\log T/(dTv)=o(1)$, $(\log d)^2/(dnv^2)=o(1)$ and $(\log T)^2/(dnv^2)=o(1)$, then*

$$T_e/T = o_p(1), \quad \text{as } n\to\infty, T\to\infty.$$

Proof of Theorem 6. By the same argument in Theorem 4, it is sufficient to show

$$\frac{(n_L+1)T_e}{T} = o_p(1), \quad \text{as } n_L\to\infty, n\to\infty, T\to\infty. \tag{9.25}$$

From (9.22), we have shown

$$\sum_{i=1}^{n_L}\sum_{0\leq k\leq n_L, k\neq i}\frac{(n_L+1)t_{ik}}{T} = o_p(1), \quad \text{as } n_L\to\infty, n\to\infty, T\to\infty.$$

From (9.23), there exists $\delta'>0$ such that

$$L_P(z_*) - L_P(\hat{z}) \geq \max_{i=1,\ldots,n_L}\frac{dvnT}{\delta' n_L^3}\frac{n_L(n_L+1)t_{0i}}{T},$$

which further implies

$$\max_{i=1,\ldots,n_L}\frac{n_L(n_L+1)t_{0i}}{T} = o_p(1),$$

if $n_L^5\log T/(dTv)=o(1)$, $(\log d)^2 n_L^8\log n_L/(dnv^2)=o(1)$ and $(\log T)^2 n_L^8\log n_L/(dnv^2)=o(1)$.

It follows that

$$\sum_{i=1}^{n_L}\frac{(n_L+1)t_{0i}}{T} = o_p(1).$$

Eq. (9.25) is therefore proved and so is the theorem. □

Finally, we give the result for estimation consistency under the hub model with the null component with fixed n_L:

Theorem 6′. *Under $H_1^{*\prime} - H_4^{*\prime}$, if $\log n/T = o(1)$, $\log T/(dTv) = o(1)$, $\log T/(dTv) = o(1)$, $(\log d)^2/(dnv^2) = o(1)$ and $(\log T)^2/(dnv^2) = o(1)$, then*

$$\max_{i\in\{0,\ldots,n_L\},j\in\{1,\ldots,n_L\}} |\hat{A}^{\hat{\tilde{z}}}_{ij} - A_{ij}| = o_p(1), \quad \text{as } n\to\infty, T\to\infty.$$

9.3 Identifiability Under Hub Model with the Null Component and Unknown Hub Set

We give a new identifiability result for the hub model with the null component and unknown hub set. Recall that V_0 is the true hub set with $|V_0| = n_L$. Let $\tilde{V}_0$ be another potential hub set with the corresponding parameters $(\tilde{\rho}, \tilde{A}) \in \mathcal{P}$ such that $\mathbb{P}(g|\rho, A) = \mathbb{P}(g|\tilde{\rho}, \tilde{A})$.

Theorem S1. *The parameters (ρ, A) of the hub model with the null component and unknown hub set are identifiable under the following conditions:*

(i′) $A_{ij} < 1$ *for* $i \in V_0 \cup \{0\}$ *and* $\tilde{A}_{ij} < 1$ *for* $i \in \tilde{V}_0 \cup \{0\}, j = 1, \ldots, n, j \neq i$;

(ii′) *for all* $i \in V_0, i' \in V_0, i \neq i'$, *there exists* $k \in V \setminus V_0$ *such that* $A_{ik} \neq A_{i'k}$;

(iii′) *for all* $i \in V_0$, *there exist* $k, k' \in V \setminus V_0$ *and* $k \neq k'$ *such that* $\pi_k \neq A_{ik}$ *and* $\pi_{k'} \neq A_{ik'}$;

(iv′) *there exists* $k \notin V_0 \cup \tilde{V}_0$ *such that for any* $i \in V_0$, $\pi_k \neq A_{ik}$, *and for any* $l \in \tilde{V}_0$, $\tilde{\pi}_k \neq \tilde{A}_{lk}$.

Conditions (i') - (iii') are identical to those in Theorem 1 and Theorem 2. Condition (iv') requires there exists at least one node that can only play a role as a follower.

Proof of Theorem S1. Theorem 2 shows when $V_0 = \tilde{V}_0$, the parameters in the hub model with null component are identifiable. Therefore, we only need to show $V_0 = \tilde{V}_0$ if $\mathbb{P}(g|\rho, A) = \mathbb{P}(g|\tilde{\rho}, \tilde{A})$ for all g.

Suppose there exist $(\tilde{\rho}, \tilde{A}) \neq (\rho, A)$ such that $\mathbb{P}(g|\rho, A) = \mathbb{P}(g|\tilde{\rho}, \tilde{A})$ for any g. Let $B_1 = \tilde{V}_0 \setminus V_0$ and $B_2 = V \setminus (V_0 \cup \tilde{V}_0)$. First, we consider the probability that no node appears

$$\rho_0 \prod_{j=1}^{n}(1 - A_{0j}) = \tilde{\rho}_0 \prod_{j=1}^{n}(1 - \tilde{A}_{0j}), \tag{9.26}$$

and the probability that only $k \in B_2$ appears,

$$\rho_0 A_{0k} \prod_{j\neq k}^{n}(1 - A_{0j}) = \tilde{\rho}_0 \tilde{A}_{0k} \prod_{j\neq k}^{n}(1 - \tilde{A}_{0j}). \tag{9.27}$$

Dividing (9.27) by (9.26), since $A_{0k} < 1$, we have $A_{0k} = \tilde{A}_{0k}$ for any $k \in B_2$.

Next we show that $B_1 = \tilde{V}_0 \setminus V_0 = \emptyset$. Suppose $B_1 \neq \emptyset$. By condition (iv'), for any $i \in B_1$, there exists a $k \in B_2$ such that $\tilde{A}_{0k} \neq \tilde{A}_{ik}$. Consider the probability that only i appears,

$$\tilde{\rho}_0 \tilde{A}_{0i} \prod_{j=1,\ldots,n,j\neq i}(1-\tilde{A}_{0j}) + \tilde{\rho}_i \prod_{j=1,\ldots,n,j\neq i}(1-\tilde{A}_{ij}) = \rho_0 A_{0i}(1-A_{0k}) \prod_{j=1,\ldots,n,j\notin\{i,k\}}(1-A_{0j}), \tag{9.28}$$

and the probability that only i and k appear

$$\tilde{\rho}_0 \tilde{A}_{0i} A_{0k} \prod_{j=1,\ldots,n,j\notin\{i,k\}}(1 - \tilde{A}_{0j}) + \tilde{\rho}_i \tilde{A}_{ik} \prod_{j\notin\{i,k\}}(1 - \tilde{A}_{ij}) = \rho_0 A_{0i} A_{0k} \prod_{j\notin\{i,k\}}(1 - A_{0j}). \tag{9.29}$$

Let

$$\tilde{x} = \tilde{\rho}_0 \tilde{A}_{0i} \prod_{j=1,\ldots,n,j\notin\{i,k\}}(1 - \tilde{A}_{0j}),$$

$$\tilde{y} = \tilde{\rho}_i \prod_{j=1,\ldots,n,j\notin\{i,k\}}(1 - \tilde{A}_{ij}).$$

Then (9.28) and (9.29) can be viewed as a system of linear equations with unknown variables $\tilde{x}$ and $\tilde{y}$:

$$\begin{pmatrix} A_{0k} & \tilde{A}_{ik} \\ 1-A_{0k} & 1-\tilde{A}_{ik} \end{pmatrix} \begin{pmatrix} \tilde{x} \\ \tilde{y} \end{pmatrix} = \begin{pmatrix} \rho_0 A_{0i} A_{0k} \prod_{j=1,\ldots,n, j\notin\{i,k\}} (1-A_{0j}) \\ \rho_0 A_{0i}(1-A_{0k}) \prod_{j=1,\ldots,n, j\notin\{i,k\}} (1-A_{0j}) \end{pmatrix}.$$

Since $A_{0k} = \tilde{A}_{0k} \neq \tilde{A}_{ik}$, the system is full rank and hence has a unique solution:

$$\tilde{\rho}_0 \tilde{A}_{0i} \prod_{j=1,\ldots,n, j\notin\{i,k\}} (1-\tilde{A}_{0j}) = \rho_0 A_{0i} \prod_{j=1,\ldots,n, j\notin\{i,k\}} (1-A_{0j}),$$

$$\tilde{\rho}_i \prod_{j=1,\ldots,n, j\notin\{i,k\}} (1-\tilde{A}_{ij}) = 0.$$

Combining with (9.26), we have

$$\tilde{\rho}_0(1-\tilde{A}_{0i}) \prod_{j=1,\ldots,n, j\notin\{i,k\}} (1-\tilde{A}_{0j}) = \rho_0(1-A_{0i}) \prod_{j=1,\ldots,n, j\notin\{i,k\}} (1-A_{0j}).$$

As $\tilde{A}_{ij} < 1$, $A_{0i} = \tilde{A}_{0i}$ for any $i \in B_1 \subset \tilde{V}_0$ and $\tilde{\rho}_i = 0$, which contradicts the assumption that $0 < \tilde{\rho}_i < 1$ for any $i \in \tilde{V}_0$. Therefore, $\tilde{V}_0 \setminus V_0 = \emptyset$ implies that $\tilde{V}_0$ does not contain any redundant component.

By the same argument, we obtain $A_{0i} = \tilde{A}_{0i}$ for any $i \in V_0 \setminus \tilde{V}_0$ and $\rho_i = 0$, which contradicts the assumption $0 < \rho_i < 1$ for $i \in V_0$. Therefore, $V_0 \setminus \tilde{V}_0 = \emptyset$. Hence, $V_0 = \tilde{V}_0$. By Theorem 2, we have $(\tilde{\rho}, \tilde{A}) = (\rho, A)$. □

REFERENCES

[1] Abbe, E., "Community detection and stochastic block models: recent developments", The Journal of Machine Learning Research **18**, 1, 6446–6531 (2017).

[2] Agarwal, A. and N. Nanavati, "Association rule mining using hybrid ga-pso for multi-objective optimisation", in "2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)", pp. 1–7 (IEEE, 2016).

[3] Allman, E. S., C. Matias and J. A. Rhodes, "Identifiability of parameters in latent structure models with many observed variables", The Annals of Statistics **37**, 6A, 3099–3132 (2009).

[4] Andrews, D. W., "Tests for parameter instability and structural change with unknown change point", Econometrica: Journal of the Econometric Society pp. 821–856 (1993).

[5] Andrews, D. W., I. Lee and W. Ploberger, "Optimal changepoint tests for normal linear regression", Journal of Econometrics **70**, 1, 9–38 (1996).

[6] Bai, J., "Least squares estimation of a shift in linear processes", Journal of Time Series Analysis **15**, 5, 453–472 (1994).

[7] Bai, J., "Estimation of a change point in multiple regression models", Review of Economics and Statistics **79**, 4, 551–563 (1997).

[8] Bai, J. and P. Perron, "Estimating and testing linear models with multiple structural changes", Econometrica **66**, 1, 47–78 (1998).

[9] Barabási, A.-L. and R. Albert, "Emergence of scaling in random networks", Science **286**, 509–512 (1999).

[10] Baranowski, R., Y. Chen and P. Fryzlewicz, "Narrowest-over-threshold detection of multiple change points and change-point-like features", Journal of the Royal Statistical Society: Series B (Statistical Methodology) **81**, 3, 649–672 (2019).

[11] Bickel, P. J. and A. Chen, "A nonparametric view of network models and Newman-Girvan and other modularities", Proc. Natl. Acad. Sci. USA **106**, 21068–21073 (2009).

[12] Bickel, P. J., D. Choi, X. Chang and H. Zhang, "Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels", ArXiv:1207.0865 (2012).

[13] Brault, V., C. Keribin and M. Mariadassou, "Consistency and asymptotic normality of latent block model estimators", Electronic journal of statistics **14**, 1, 1234–1268 (2020).

[14] Bulut, A., A. K. Singh, P. Shin, T. Fountain, H. Jasso, L. Yan and A. Elgamal, "Real-time nondestructive structural health monitoring using support vector machines and wavelets", in "Advanced Sensor Technologies for Nondestructive Evaluation and Structural Health Monitoring", vol. 5770, pp. 180–189 (SPIE, 2005).

[15] Cairns, S. J. and S. J. Schwager, "A comparison of association indices", Animal Behavior **35** (1987).

[16] Camacho, A. and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music", The Journal of the Acoustical Society of America **124**, 3, 1638–1652 (2008).

[17] Chang, P.-C., C.-Y. Fan and C.-H. Liu, "Integrating a piecewise linear representation method and a neural network model for stock trading points prediction", IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) **39**, 1, 80–92 (2008).

[18] Cheng, D., Z. He and A. Schwartzman, "Multiple testing of local extrema for detection of change points", Electronic Journal of Statistics **14**, 2, 3705–3729 (2020).

[19] Cheng, D. and A. Schwartzman, "Distribution of the height of local maxima of gaussian random fields", Extremes **18**, 213–240 (2015).

[20] Cheng, D. and A. Schwartzman, "Multiple testing of local maxima for detection of peaks in random fields", The Annals of Statistics **45**, 2, 529–556 (2017).

[21] Cheng, D. and A. Schwartzman, "Expected number and height distribution of critical points of smooth isotropic gaussian random fields", Bernoulli **24**, 4B, 3422–3446 (2018).

[22] Choi, D. S., P. J. Wolfe and E. M. Airoldi, "Stochastic blockmodels with growing number of classes", Biometrika **99**, 273–284 (2012).

[23] Diaconis, P. and S. Janson, "Graph limits and exchangeable random graphs", arXiv preprint arXiv:0712.2749 (2007).

[24] Ding, Y., X. Yang, A. J. Kavs and J. Li, "A novel piecewise linear segmentation for time series", in "2010 The 2nd International Conference on Computer and Automation Engineering (ICCAE)", vol. 4, pp. 52–55 (IEEE, 2010).

[25] Dubhashi, D. P. and A. Panconesi, *Concentration of measure for the analysis of randomized algorithms* (Cambridge University Press, 2009).

[26] Efron, B. and N. R. Zhang, "False discovery rates and copy number variation", Biometrika **98**, 2, 251–271 (2011).

[27] Frank, O. and D. Strauss, "Markov graphs", Journal of the American Statistical Asscociation **81**, 832–842 (1986).

[28] Frick, K., A. Munk and H. Sieling, "Multiscale change point inference", Journal of the Royal Statistical Society: Series B (Statistical Methodology) **76**, 3, 495–580 (2014).

[29] Fryzlewicz, P., "Wild binary segmentation for multiple change-point detection", The Annals of Statistics **42**, 6, 2243–2281 (2014).

[30] Fryzlewicz, P., "Narrowest significance pursuit: inference for multiple change-points in linear models", arXiv preprint arXiv:2009.05431 (2020).

[31] Gao, C., Y. Lu, H. H. Zhou *et al.*, "Rate-optimal graphon estimation", The Annals of Statistics **43**, 6, 2624–2652 (2015).

[32] Getoor, L. and C. P. Diehl, "Link mining: A survey", ACM SIGKDD Explorations Newsletter **7**, 2, 3–12 (2005).

[33] Ghalanos, A. and S. Theussl, "Rsolnp: general non-linear optimization using augmented lagrange multiplier method", R package version **1** (2012).

[34] Goldenberg, A., A. X. Zheng, S. E. Fienberg and E. M. Airoldi, "A survey of statistical network models", Foundations and Trends in Machine Learning **2**, 129–233 (2010).

[35] Gu, Y. and G. Xu, "The sufficient and necessary condition for the identifiability and estimability of the dina model", Psychometrika **84**, 2, 468–483 (2019).

[36] Gyllenberg, M., T. Koski, E. Reilink and M. Verlaan, "Non-uniqueness in probabilistic numerical identification of bacteria", Journal of Applied Probability **31**, 2, 542–548 (1994).

[37] Hann, C. E., I. Singh-Levett, B. L. Deam, J. B. Mander and J. G. Chase, "Real-time system identification of a nonlinear four-story steel frame structureâĂŤapplication to structural health monitoring", IEEE Sensors Journal **9**, 11, 1339–1346 (2009).

[38] Hao, N., Y. S. Niu and H. Zhang, "Multiple change-point detection via a screening and ranking algorithm", Statistica Sinica **23**, 4, 1553 (2013).

[39] Hoff, P. D., "Modeling homophily and stochastic equivalence in symmetric relational data", in "Advances in Neural Information Processing Systems", vol. 19 (MIT Press, Cambridge, MA, 2007).

[40] Hoff, P. D., A. E. Raftery and M. S. Handcock, "Latent space approaches to social network analysis", Journal of the American Statistical Asscociation **97**, 1090–1098 (2002).

[41] Huang, T., H. Peng and K. Zhang, "Model selection for gaussian mixture models", arXiv preprint arXiv:1301.3558 (2013).

[42] Huber, P. J., *Robust statistics*, vol. 523 (John Wiley & Sons, 2004).

[43] Hung, Y., Y. Wang, V. Zarnitsyna, C. Zhu and C. J. Wu, "Hidden markov models with applications in cell adhesion experiments", Journal of the American Statistical Association **108**, 504, 1469–1479 (2013).

[44] Hyun, S., M. G'Sell and R. J. Tibshirani, "Exact post-selection inference for the generalized lasso path", Electronic Journal of Statistics **12**, 1, 1053–1097 (2018).

[45] Hyun, S., K. Z. Lin, M. G'Sell and R. J. Tibshirani, "Post-selection inference for changepoint detection algorithms with application to copy number variation data", Biometrics **77**, 3, 1037–1049 (2021).

[46] Khan, F., A. Ghaffar, N. Khan and S. H. Cho, "An overview of signal processing techniques for remote health monitoring using impulse radio uwb transceiver", Sensors **20**, 9, 2479 (2020).

[47] Kharinov, M., "Image segmentation using optimal and hierarchical piecewise-constant approximations", Pattern recognition and image analysis **24**, 3, 409–417 (2014).

[48] Killick, R., 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).

[49] Klapuri, A. P., "Automatic music transcription as we know it today", Journal of New Music Research **33**, 3, 269–282 (2004).

[50] Lavielle, M., "Using penalized contrasts for the change-point problem", Signal processing **85**, 8, 1501–1510 (2005).

[51] Li, H., A. Munk and H. Sieling, "Fdr-control in multiscale change-point segmentation", Electronic Journal of Statistics **10**, 1, 918–959 (2016).

[52] Liao, S. X. and M. Pawlak, "On image analysis by moments", IEEE Transactions on Pattern analysis and machine intelligence **18**, 3, 254–266 (1996).

[53] Lu, J., X. Zheng, Q. Z. Sheng, Z. Hussain, J. Wang and W. Zhou, "Mfe-har: multiscale feature engineering for human activity recognition using wearable sensors", in "Proceedings of the 16th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services", pp. 180–189 (2019).

[54] Moreno, J. L., *Who Shall Survive? A New Approach to the Problem of Human Interactions* (Nervous and Mental Disease Publishing Co., 1934).

[55] Negahban, S., S. Oh, K. K. Thekumparampil and J. Xu, "Learning from comparisons and choices", The Journal of Machine Learning Research **19**, 1, 1478–1572 (2018).

[56] Newman, M. E. J., *Networks: An introduction* (Oxford University Press, 2010).

[57] Olshen, A. B., E. Venkatraman, R. Lucito and M. Wigler, "Circular binary segmentation for the analysis of array-based dna copy number data", Biostatistics **5**, 4, 557–572 (2004).

[58] Pein, F., H. Sieling and A. Munk, "Heterogeneous change point inference", Journal of the Royal Statistical Society: Series B (Statistical Methodology) **79**, 4, 1207–1227 (2017).

[59] Perron, P., "The great crash, the oil price shock, and the unit root hypothesis", Econometrica: journal of the Econometric Society pp. 1361–1401 (1989).

[60] Robins, G., P. Pattison, Y. Kalish and D. Lusher, "An introduction to exponential random graph (p*) models for social networks", Social networks **29**, 2, 173–191 (2007).

[61] Schwartzman, A., Y. Gavrilov and R. J. Adler, "Multiple testing of local maxima for detection of peaks in 1d", Annals of statistics **39**, 6, 3290 (2011).

[62] Shizuka, D. and D. R. Farine, "Measuring the robustness of network community structure using assortativity", Animal behaviour **112**, 237–246 (2016).

[63] Tebaldi, C. and D. Lobell, "Towards probabilistic projections of climate change impacts on global crop yields", Geophysical Research Letters **35**, 8 (2008).

[64] Tibshirani, R., M. Saunders, S. Rosset, J. Zhu and K. Knight, "Sparsity and smoothness via the fused lasso", Journal of the Royal Statistical Society: Series B (Statistical Methodology) **67**, 1, 91–108 (2005).

[65] Tomé, A. and P. Miranda, "Piecewise linear fitting and trend changing points of climate parameters", Geophysical Research Letters **31**, 2 (2004).

[66] Van der Vaart, A. W., *Asymptotic statistics*, vol. 3 (Cambridge university press, 2000).

[67] Van Laerhoven, K., E. Berlin and B. Schiele, "Enabling efficient time series analysis for wearable activity data", in "2009 International Conference on Machine Learning and Applications", pp. 392–397 (IEEE, 2009).

[68] Vostrikova, L. Y., "Detecting "disorder" in multidimensional random processes", in "Doklady Akademii Nauk", vol. 259, pp. 270–274 (Russian Academy of Sciences, 1981).

[69] Wasserman, S. and C. Faust, *Social Network Analysis: Methods and Applications* (Cambridge University Press, 1994).

[70] Weko, C. and Y. Zhao, "Penalized component hub models", Social Networks **49**, 27–36 (2017).

[71] Xu, G. *et al.*, "Identifiability of restricted latent class models with binary responses", The Annals of Statistics **45**, 2, 675–707 (2017).

[72] Yao, Y.-C. and S.-T. Au, "Least-squares estimation of a step function", Sankhyā: The Indian Journal of Statistics, Series A pp. 370–381 (1989).

[73] Zhang, Y., E. Levina and J. Zhu, "Estimating network edge probabilities by neighbourhood smoothing", Biometrika **104**, 4, 771–783 (2017).

[74] Zhao, Y., "A survey on theoretical advances of community detection in networks", Wiley Interdisciplinary Reviews: Computational Statistics **9**, 5 (2017).

[75] Zhao, Y., "Network inference from temporal-dependent grouped observations", arXiv preprint arXiv:1808.08478 (2018).

[76] Zhao, Y. and C. Weko, "Network inference from grouped observations using hub models", Statistica Sinica **29**, 1, 225–244 (2019).