Logistic-Normal Mixtures with Heterogeneous Components and High Dimensional Covariates

by

Yingchuan Wang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in the University of Michigan
2016

Doctoral Committee:

Professor Xuming He, Chair
Professor Bin Nan
Professor Kerby Shedden
Professor Naisyin Wang

2016

D E D I C A T I O N

I dedicate my dissertation work to my family. A special feeling of gratitude to my loving parents, Xianguang Wang and Liping Fu, and my grandparents, Quanbao Zhang and Liangyu Fu, whose words of encouragement and push for tenacity ring in my ears.

I also dedicate this dissertation to my many friends who have supported me throughout the process.

I dedicate this work and give special thanks to my wife Yiwei Zhang and my coming baby for all of the love.

A C K N O W L E D G M E N T S

Firstly, I would like to express my sincere gratitude to my advisor Prof. Xuming He for the continuous support of my Ph.D study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in my research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study.

I would also like to thank the rest of my thesis committee: Prof. Naisyin Wang, Prof. Kerby Shedden, and Prof. Bin Nan, for their insightful comments and encouragement, and for their constructive questions which incented me to widen my research from various perspectives.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

CHAPTER 1

Introduction

Finite mixture models are powerful statistical modeling tools to analyze the underlying data frame because of its flexible model structure and appealing interpretation. Numerous real data applications can be found in a variety of areas, such as economics, finance and clinical trails. The book by McLachlan and Peel (2000) has a very detailed summary of the finite mixture models when training data are multivariate. Jordan and Jacobs (1993) introduced the finite mixture of regression (FMR) models, which is an extension of the finite mixture models, under the term of the mixture of experts. Different from the finite mixture models, FMR relates a response variable Y to a set of baseline covariates $\boldsymbol{X} = (X_1, \cdots, X_p)$ to capture the heterogeneity for different subgroup populations. Jiang and Tanner (1999) showed that the FMR models are dense in the exponential family in the sense that any exponential density can be approximated by FMR models.

The logistic-normal mixtures is one of the FMR models. It allows both the mixing parameters and the mean parameters to depend on covariates. This is very distinct from many other FMR models, where the mixing parameters are treated as constants. Because of this unique feature, the logistic-normal mixtures can jointly model the subgroup membership and the regression component in each subgroup. Applications can be found at Wong and Li (2001), Muthen and Asparouhov (2009) and Muthen and Shedden (1999). Recently, Shen and He (2015) utilized the logistic-normal mixture model in testing the existence of subgroups with given covariates. In the paper, Shen and He (2015) proposed an EM test (the name EM is from the Expectation-Maximization algorithm) under the assumption that different subgroups have the same variance. However, the assumption of homogeneity in subgroups' variances does not generally hold in practice. If the equal subgroups' variances are insisted on, it is unclear whether the EM test will lose power and whether the model estimation will be biased. The purpose of Chapter 2 in

the dissertation is to relax the equal variance assumption in the logistic-normal mixture model. This additional flexibility in allowing unequal subgroup variances is highly valuable in practice, but brings technical challenges in the theoretical development. When the component variances are different, the unboundedness of the likelihood function renders the maximum likelihood estimators invalid. To overcome this difficulty, we propose to work with a penalized likelihood following the work of Chen (2008) for inference. We also propose a data-driven strategy to select the penalty parameter λ that maximizes the potential overall subgroup effect and provide the asymptotic theory for the penalized likelihood estimator and its associated EM tests.

It is unlikely that we always have affirmatory covariates to work with. Quite often, we have to consider numerous potential covariates and select important ones among them. There are some variable selection methods that are proposed for FMR models with constant mixing parameters and given number of components K. For example, when the dimension p of potential covariates is fixed, Khalili and Chen (2007) proposed a penalized maximum likelihood method based on the least absolute shrinkage and selection operator (LASSO) of Tibshirani (1996) and the smoothly clipped absolute deviation (SCAD) method of Fan and Li (2001), where the penalties depend on the scale of the regression parameters and the mixture structure. Khalili and Chen (2007) argued that, although the maximum likelihood estimators are consistent when p is fixed in theory, the empirical performances are very poor when the true models are sparse but the dimension p of potential covariates is moderately large. The poor empirical performances are in the sense that the maximum likelihood estimators are not stable and they do not help to identify the true non-zero coefficients. Sometimes, Akaike information criterion (AIC) and Bayes information criterion (BIC) are needed to perform model selection. Khalili and Chen (2007) showed that, however, their proposed penalized maximum likelihood method can select the true covariates almost surely, and in addition, they demonstrated in simulations that their method is computationally much more efficient than AIC and BIC. In many situations, however, the dimension of potential covariates p grows with n, where n is the sample size. In such cases, the standard asymptotic theory fails. The techniques in convex analysis, see Zhao and Yu (2006), Bickel, Ritov and Tsybakov (2009), Greenshtein et al. (2004), van de Geer (2008) , which are developed for large p small n scenario, cannot be applied to the non-convex log-likelihood of FMR models either. Städler, Bühlmann and van de Geer (2010) reparametrized the FMR models and considered a ℓ_1-penalized maximum likelihood estimators for the case of $p = o(e^n)$. They restricted the new parameters in a compact set and instead of focusing on the consistency of the parame-

ters, they showed that the Kullback-Leibler divergence between the estimated model and the true model goes to $zero$ with probability going to one. They also showed that the convergence rate is at the order of $\sqrt{\log n^3 \log(p \vee n)/n}$. van de Geer (2013) considered the same estimators as that in Städler, Bühlmann and van de Geer (2010), but improved the convergence rate to the order of $\sqrt{\log p/n}$ by developing the *Multivariate contraction theorem* and using generic chaining techniques. Note that the convergence rate for the ℓ_1 penalized maximum log-likelihood estimators of traditional linear regressions is at the rate of $\sqrt{\log p/n}$, see Lounici et al. (2008) and Zhang et al. (2009) for details. The log-likelihood function for linear regressions is convex and is in a simple quadratic form, whereas the FMR models have non-convex log-likelihoods, which are also in much more complicated forms than that of linear regressions, we do not expect the convergence rate for the ℓ_1 penalized maximum log-likelihood estimators of FMR to be better than that of linear regressions. In other words, $\sqrt{\log p/n}$ is the best convergence rate we could hope to get for FMR, and van de Geer (2013) showed that we indeed can achieve the optimal rates for FMR. Khalili and Lin (2013) also considered the large p small n cases, and showed that there exists a local penalized maximum likelihood estimator that can select the true covariates almost surely. However, their setting is rather restrictive; their p can be as large as $4n^{1/4} - 5$ at most.

For known number of components K, a common characteristic for the aforementioned results is that they only concerned the FMR models with constant mixing parameters. In Chapter 3, we consider logistic-normal mixtures, which allows both mixing parameters and component means to depend on covariates, in high dimensions. Because the mixing parameters may also depend on high dimensional covariates, and the logistic form renders them to behave very differently from the component means, the consistency theory in van de Geer (2013) cannot be directly applied to this situation. Furthermore, in numerical optimizations, the mixing parameters do not have explicit update formulas as that in Städler, Bühlmann and van de Geer (2010), which results lots of local minima in the optimization problem. In Chapter 3, we show that even with the logistic form, the convergence rate for the model as well as that for the ℓ_1 norm of all zero coefficients estimators can still achieve the optimal rates of $\sqrt{\log p/n}$. We also show that the non-zero coefficients estimates will converge to the true values asymptotically.

Note that the assumption of known number of components K is not guaranteed in practice, and we need a method to determine the number of components from data. Traditional AIC and BIC do not work in our setting, because the dimension of covariates is increasing with the sample size. Extended Bayesian Information Criteria (EBIC) which is developed by Chen and

Chen (2008) handles model selection in high dimensions, however, their work is under the linear regression framework. Corduneanu and Bishop (2001) proposed a variational bayesian method and very recently, Huang, Peng and Zhang (2013) proposed a penalized log-likelihood approach to select the number of components for mixture models, but these methods are all for a fixed dimension of covariates. In Chapter 3, we develop a selection criterion ($SCMM$) for selecting the number of components for logistic-normal mixture models with high dimensional covariates. We show that $SCMM$ choose the true K with probability going to one.

The rest of the dissertation is organized as follows. We study the penalized maximum likelihood estimator and the penalized EM test for logistic-normal mixture models with a fixed number of covariates in Chapter 2. The proposed method and the associated theory, simulations and real data applications of variable selection with high dimensional covariates for logistic normal mixtures will be discussed in Chapter 3. We conclude the work with a brief discussion in Chapter 4 and the detailed proofs are in the Appendix.

CHAPTER 2

Subgroup Inferences for Logistic-Normal Mixtures with Heterogenous Components

Subgroup analysis is important in clinical trials and market segmentation. In recent years, how to extract unknown subgroups with distinct responses to a treatment has gained increasing popularity. Much of the early research in subgroup analysis has focused on pre-specified subgroups (Simon (2002), Song and Chi (2007), and Altstein et al. (2011) among others). Su et al. (2009) introduced an interaction tree procedure to obtain subpopulations with heterogeneous treatment effects across subpopulations. Foster et al. (2011) proposed the "Virtual Twins" method to identify a subgroup for the binary response in a randomized clinical trial. A parametric scoring system based on multiple covariates was presented in Cai et al. (2011) and Zhao et al. (2013), to help assign treatments to new patients. Lipkovich et al. (2011) and Lipkovich and Dmitrienkoa (2014) provided a recursive partitioning method for treatment assignments to patient subpopulations. Berger, Wang and Shen (2014) proposed a Bayesian method for subgroup analysis of multiple subgroups defined by a binary predictive variable. Kang, Janes and Huang (2014) relied on a novel boosting algorithm to choose an optimal treatment. Besides interaction models, methods based on mixture models are proposed, for instance Shen and He (2015). Horn et al. (2015) showed that regression mixture models can be effective in evaluating differential treatment effects.

A critical concern with various subgroup identification methods is that they tend to identify a subgroup even when no meaningful subgroup exists. Back in 2000, Sleight (2000) described subgroup analyses as "fun to look at, but don't believe them". Two recent articles, Shen and He (2015) and Fan, Lu and Song (2015) have advocated the use of hypothesis testing for the existence of subgroups.

In Shen and He (2015), a structured logistic-mixture model was proposed to jointly model the subgroup membership and the performance in each subgroup. An EM test is constructed to test the existence of the subgroup based on the model. However, they assume constant variability for different subgroups, which does not hold in general. In this chapter, we relax the equal variance assumption and propose a penalized maximum likelihood estimator and a penalized EM test.

The rest of Chapter 2 is organized as follows. We introduce the logistic-normal mixture models in Chapter 2.1. In Chapter 2.2, we first discuss the behavior of the unbounded log-likelihood function when we do not have equal variance assumption for subgroups. Then, we propose a penalized maximum likelihood estimator and show its consistency. We propose a penalized EM test based on the penalized maximum likelihood estimator in Chapter 2.3, and study its asymptotic property in Chapter 2.4. Chapter 2.5 will discuss the issues of tuning parameter selection for penalties and simulation results are reported in Chapter 2.6.

2.1 Model

We consider the following logistic-normal mixture model that allows unequal variances in each component. For $i = 1, \ldots, n$,

$$\begin{aligned} Y_i &= \boldsymbol{Z}_i^T(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2\delta_i) + \varepsilon_i(\delta_i\sigma_1 + (1-\delta_i)\sigma_2), \\ P(\delta_i = 1|\boldsymbol{X}_i, \boldsymbol{Z}_i) &= \pi(\boldsymbol{X}_i^T\boldsymbol{\gamma}) \equiv \exp(\boldsymbol{X}_i^T\boldsymbol{\gamma})/(1+\exp(\boldsymbol{X}_i^T\boldsymbol{\gamma})), \\ P(\delta_i = 0|\boldsymbol{X}_i, \boldsymbol{Z}_i) &= 1 - P(\delta_i = 1|\boldsymbol{X}_i), \end{aligned} \tag{2.1}$$

where n is the sample size, $Y_i \in \mathbb{R}$ is the outcome, $\delta_i \in \{0, 1\}$ is the latent subgroup indicator, $\boldsymbol{Z}_i \in \mathbb{R}^{q_1}$ is the covariate associated with the subgroup mean, $\boldsymbol{X}_i \in \mathbb{R}^{q_2}$ is the baseline covariate associated with the group membership, $\boldsymbol{\beta}_1 \in \mathbb{R}^{q_1}, \boldsymbol{\beta}_2 \in \mathbb{R}^{q_1}, \boldsymbol{\gamma} \in \mathbb{R}^{q_2}$ are the corresponding coefficients, $\varepsilon_i \sim N(0,1)$ are independent of $\boldsymbol{Z}_i$, $\boldsymbol{X}_i$ and δ_i, and σ_1 and σ_2 are the standard derivations within each subgroup. The first element of $\boldsymbol{X}_i$ and the first component of $\boldsymbol{Z}_i$ are taken to be 1 to allow intercepts in the model, and the second element of $\boldsymbol{Z}_i$ is the treatment indicator. We can have overlapping variables in the random vectors of $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$.

In the two-component model, the overall model parameters are $\boldsymbol{\eta}^T = (\boldsymbol{\gamma}^T, \boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \sigma_1, \sigma_2)$. In this paper, we use $\boldsymbol{\theta}^T = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \sigma_1, \sigma_2)$ as the parameters except for $\boldsymbol{\gamma}$. We observe a random sample $\{\boldsymbol{W}_i = (Y_i, \boldsymbol{Z}_i^T, \boldsymbol{X}_i^T), i = 1, \ldots n\}$, but δ_i's are latent variables.

Remark. In the model formulation, we assume that the first nonzero component of $\boldsymbol{\beta}_2$ is positive, and in the case of $\boldsymbol{\beta}_2 = 0$ we assume that $\sigma_1 > \sigma_2$, to ensure parameter identifiability. The model becomes degenerate if $\boldsymbol{\beta}_2 = 0$ and $\sigma_1 = \sigma_2$. In our implementation, we identify the subgroups by taking the second component of $\boldsymbol{\beta}_2$ (the treatment effect difference) to be positive. We are not concerned with the special case where the treatment effect difference is zero, in which case the identification of subgroups is not practically important.

In the case of $\boldsymbol{X}_i = \boldsymbol{Z}_i$, we can think of the proposed model as a special case of the mixture-of-experts models (Jordan and Jacobs (1994)), which is well studied in machine learning. Unlike the mixture-of-experts models discussed in the literature, we have distinct and clear interpretations of the variables $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$. In particular, the covariates in $\boldsymbol{X}_i$ must be baseline measurements that are available prior to the treatment and can be used to predict subgroup membership for future subjects, while the covariates in $\boldsymbol{Z}_i$ include any variables relevant to the treatment effects within subgroups. For example, any treatment-related variables can be part of $\boldsymbol{Z}_i$, but not $\boldsymbol{X}_i$. Moreover, mixture-of-experts models are constructed to predict the response, and the response patterns in each component of the mixture model are not necessarily important or interpretable. The existence of meaningful subgroups with differential treatment effects is the focus of our work.

Another special case of the proposed model with $\boldsymbol{\gamma} = 0$ has been quite well studied in the literature; see, for instance, Goffinet et al. (1992), Chen et al. (2001), and Chen and Li (2009). In subgroup analysis, the case of $\boldsymbol{\gamma} = 0$ is rather uninteresting, because even if subgroups exist, no covariates are informative for predicting the subgroup membership. An important feature of the proposed model is to characterize subgroup membership given the baseline covariates $\boldsymbol{X}$.

2.2 Penalized Maximum Likelihood

To identify the existence of subgroup membership, without knowing the variances of potential subgroups are equivalent, we need to test $H_0 : \boldsymbol{\beta}_2 = 0, \sigma_1 = \sigma_2 \; vs. \; H_a = H_0^c$ for model (2.1). We hope the EM test, which is proposed by Shen and He (2015) for equal subgroups' variances scenario, could also be applied for unequal variances case. Unfortunately, the theory can not be worked out. Before we discussing the reason why EM test fails in unequal variances setting, we first briefly sketch its procedure.

In the EM test, Shen and He (2015) assumed $\sigma_1 = \sigma_2 = \sigma$ in model (2.1). In the first step, they get the maximum likelihood estimator of $(\boldsymbol{\beta}_1, \sigma)$ and calculate the likelihood under the null model ($\boldsymbol{\beta}_2 = 0$). In the second step, they initialize $\boldsymbol{\gamma}_0$ and get the maximum likelihood estimator of $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma)$ under the alternative model ($\boldsymbol{\beta}_2 \neq 0$). Using the EM algorithm, $\boldsymbol{\gamma}$ and $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma)$ can be updated finitely many times, and then, the likelihood under the alternative model is calculated based on these updated parameters. Finally, the test statistic is the difference between the log-likelihood under the null model and that under the alternative model.

This procedure fails in the unequal variances framework, because the different component variances make the log-likelihood of model (2.1) unbounded, for any given $\boldsymbol{\gamma}$ and any given sample size. The unbounded log-likelihood renders the maximum likelihood estimators in every step of EM test meaningless.

To see why the log-likelihood becomes unbounded, take a simple mixture normal model, with unequal variances as an example. Let $Y_1, \cdots, Y_n$ be i.i.d. from,

$$\pi N(\theta_1, \sigma_1^2) + (1-\pi) N(\theta_2, \sigma_2^2),$$

then the likelihood

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \left\{ \frac{1}{\sigma_1} \exp\{-\frac{-(Y_i - \theta_1)^2}{2\sigma_1^2}\} + \frac{1}{\sigma_2} \exp\{-\frac{-(Y_i - \theta_2)^2}{2\sigma_2^2}\} \right\}$$

goes to infinity by taking $\theta_1 = Y_1$ and letting σ_1 go to zero. The maximum likelihood estimator does not exist.

A simply way to overcome the difficulty of unboundedness of log-likelihood for unequal variance case is to impose a reasonable lower bound, say $a > 0$, for σ. Then, the maximum likelihood estimators exist for (σ_1, σ_2) over $[a, +\infty) \times [a, +\infty)$, and the EM test can be applied to this constrained parameter space. If the true values of σ_1, σ_2 do fall in this range, the estimation of the parameters would be consistent, and the EM test would follow the theoretical asymptotic distribution as given in **Theorem 2**. However, if the true values fall out of this range, the estimated parameters would be biased, and the EM test will be invalid. In practice it might be hard to specify an appropriate constant a. We will consider an alternative approach of penalized likelihood.

Let $f(Y|\boldsymbol{Z}, \boldsymbol{X}, \delta, \boldsymbol{\beta}, \sigma)$ denote the conditional density of Y given $(\boldsymbol{Z}, \boldsymbol{X}, \delta)$, then, the penal-

ized log-likelihood function is defined as

$$pl(\boldsymbol{\eta}; \boldsymbol{W}) = \sum_{i=1}^{n} \log[\sum_{j=0}^{1} f(Y_i|\boldsymbol{Z}_i, \boldsymbol{X}_i, \delta_i = j; \boldsymbol{\beta}_j, \sigma_j) P(\delta_i = j|\boldsymbol{X}_i; \boldsymbol{\gamma})] + p_n(\sigma_1) + p_n(\sigma_2), \quad (2.2)$$

where $p_n(\sigma_1)$ and $p_n(\sigma_2)$ are penalties on subgroups' variances. Consequently, the penalized maximum likelihood estimator is given by

$$\hat{\boldsymbol{\eta}} = \underset{\boldsymbol{\eta}}{\operatorname{argmax}}\, pl(\boldsymbol{\eta}; \boldsymbol{W}). \quad (2.3)$$

We consider data-dependent penalties and particularly, we use

$$p_n(\sigma) = -\lambda\big(\frac{S_n^2}{\sigma^2} + \log(\frac{\sigma^2}{S_n^2})\big) \quad (2.4)$$

for data analysis, where S_n^2 is the estimator for σ^2 from the equal variance model, and λ is a tuning parameter. Given any positive λ, $p_n(\sigma)$ achieves its maximum 0 at $\sigma^2 = S_n^2$, and goes to negative infinity as σ approaches zero or infinity.

The general conditions for the penalty are as follows.

C1. The penalty $p_n(\sigma) < 0$ almost surely.

C2. For any given constant C, for almost all sample $\omega \in \Omega$, there exists $n_0(\omega)$, such that when $n \geq n_0(\omega)$,

$$\inf\{p_n(\sigma)/[(\log n)^2 \log \sigma] : \, 0 < \sigma \leq (1/n)\} \geq C.$$

C3. If $\boldsymbol{\beta}_2 = 0$ and $\sigma_1 = \sigma_2 = \sigma_0$, we have $p_n(\sigma_0) = o(n)$ almost surely; otherwise (under the alternative model), $p_n(\sigma_1) = o(n)$ and $p_n(\sigma_2) = o(n)$ almost surely.

Remark. Condition C2 basically requires that the penalty should be small when σ is small, and Condition C3 requires that the penalty should not dominate the likelihood function evaluated at the true parameters. These two conditions together guarantee that the penalized likelihood does not attain its maximum when σ is near zero, and therefore, the estimator of σ stays away from zero. The conditions allow the penalties to depend on the data, which is quite useful in numerical

analysis in practice. We would discuss how to choose the tuning parameter λ for the penalty in equation (2.4).

For the variables $\boldsymbol{X}$ and $\boldsymbol{Z}$, we further impose the following conditions. If we partition the covariant vector $\boldsymbol{Z}$ into a discrete component $\boldsymbol{U}$ and a continuous component $\boldsymbol{V}$, that is, let $\boldsymbol{Z}^T = (1, \boldsymbol{U}^T, \boldsymbol{V}^T)$, where 1 corresponds to the intercept, $\boldsymbol{U}$ consists of only discrete variables, and $\boldsymbol{V}$ consists of only continuous variables. The separation of discrete and continuous covariates is to facilitate mathematical handling in the proofs of our asymptotic results. We assume

C4. The sample space of $\boldsymbol{U}$ is finite. For any unit vector $\boldsymbol{\alpha}$ of the same dimension as the vector $\boldsymbol{V}$, the conditional distribution of $\boldsymbol{V}^T\boldsymbol{\alpha}|\boldsymbol{U}$ is continuous, and the maximum of its density is uniformly bounded from above.

C5. The expectation $\mathbb{E}(||\boldsymbol{V}||_1 \mid \boldsymbol{U} = \boldsymbol{u}) < \infty$ uniformly in $\boldsymbol{u}$, where $||\cdot||_1$ is the L_1 norm.

The separation of $\boldsymbol{Z}$ is based on its own nature structure. It is also difficult to handle discrete variables and continuous variables simultaneously, because the probability density function for continuous variables and the probability mass function for discrete variables can not be unified to discuss in the proof; they have their own unique characters. More details will be presented in the Appendix.

With the above conditions, we are ready to state the consistency theorem of our proposed penalized maximum likelihood estimators. Note that the consistency of the penalized maximum likelihood estimators for unequal variances scenario plays the same role as the consistency of the maximum likelihood estimators for equal variances scenario. They are fundamental for constructing statistical tests for testing the existence of subgroups.

Theorem 1. *(Consistency of the penalized maximum likelihood estimator) Assume conditions C1-C5 hold, then*
(1) *under the null hypothesis that $\boldsymbol{\beta}_2 = 0$ and $\sigma_1 = \sigma_2 = \sigma_0$, if we fix any $\boldsymbol{\gamma}$ with $nonzero$ slope, the penalized maximum likelihood estimator of $\boldsymbol{\theta}^T = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \sigma_1, \sigma_2)$ from equation* (2.3) *is consistent,*
(2) *under the alternative hypothesis that $\boldsymbol{\beta}_2 \neq 0$ or $\sigma_1 \neq \sigma_2$, the penalized maximum likelihood estimator of $\boldsymbol{\eta}^T = (\boldsymbol{\gamma}^T, \boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \sigma_1, \sigma_2)$ defined in* (2.3) *is consistent.*

To illustrate the idea used in the proof for Theorem 1, we consider any sequence of positive

numbers σ_n and let

$$W_n(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n} 1(|Y_i - \boldsymbol{Z}_i^T\boldsymbol{\beta}| \leq |\sigma_n \log \sigma_n|),$$
$$A_n(C) = \{\sup_{\boldsymbol{\beta}\in R^{q_1}} W_n(\boldsymbol{\beta}) > C|\sigma_n \log \sigma_n|\},$$
$$B_n = \{\sup_{\boldsymbol{\beta}\in R^{q_1}} W_n(\boldsymbol{\beta}) > 4(\log n)^2/n\}.$$

With these quantities, we shall prove the following two statements:

S1. There exist $C_1, C_2 > 0$ such that for each given $\sigma_n \in (n^{-1}, \exp(-1))$, $P(A_n(C_1)) \leq C_2 n^{-2}$;

S2. There exists $C_3 > 0$ such that uniformly in $\sigma_n \in (0, n^{-1})$, $P(B_n) \leq C_3 n^{-2}$.

The log-likelihood of the normal mixture model becomes unbounded when some sample points are very close to one of the estimated component means and when the corresponding variance estimator goes to zero. Statements **S1** and **S2** above actually give an upper bound on the number of points that fall into such *trouble regions*. The upper bound is approximately limited to the order of $O((\log n)^2)$. The penalty that satisfies C1-C3 would ensure that the variance estimators will stay away from zero.

Note that **S1** and **S2** play the same role as **Lemma 1** of Chen et al.(2008) in a simpler setting. By **Lemma 2** of Chen et al. (2008), we can show that the number of sample points that fall within the range of $|\sigma \log \sigma|$ to either one of the estimated component means is in the order of $O((\log n)^2)$. As a consequence, we can show that the estimates of σ_1 and σ_2 will stay away from zero. Standard techniques in the large sample theory can then be applied to show the consistency of the maximum penalized likelihood estimators. Since proving Statement **S2** is essentially the same as proving **S1**, we only provide the details of the proof for **S1** in the supplementary file.

With **Theorem 1**, we are ready to propose the penalized EM test in Chapter 2.3.

2.3 Penalized EM test

Following the analog of the EM test proposed by Shen and He (2015), the procedure of the penalized EM test for testing the existence of subgroups with unequal variances is sketched as follows.

Firstly, we compute the penalized maximum likelihood estimates $\hat{\boldsymbol{\theta}}_0 = (\hat{\boldsymbol{\beta}}_1, \hat{\sigma})$ of $\boldsymbol{\beta}_1, \sigma$, under the null model ($\boldsymbol{\beta}_2 = 0, \sigma_1 = \sigma_2 = \sigma$). Then, initialize $\boldsymbol{\gamma}_1^{(0)}, \cdots, \boldsymbol{\gamma}_J^{(0)}$, where J is a

pre-specified integer. Typically, if $\boldsymbol{\gamma}$ is q_2 dimension and q_2 is small, we choose $J = 2^{q_2-1}$, and the set of initial values have positive and negative values in each coordinate to cover all the quadrants. For any fixed $j \in \{1, 2, \cdots, J\}$, we compute the penalized maximum likelihood estimates $\boldsymbol{\theta}_j^{(0)}$ of $\boldsymbol{\theta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1, \sigma_2)$ from (2.2) with fixed $\boldsymbol{\gamma}_j^{(0)}$. We then use the EM algorithm to update the parameters $\boldsymbol{\eta} = (\boldsymbol{\gamma}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1, \sigma_2)$ finitely many, say K, times, and let $\boldsymbol{\eta}_j^{(K)}$ be the final estimates. Then the penalized EM test statistic is defined by

$$pEM^{(K)} = \max\{pEM_j^{(K)} : j = 1, \ldots, J\}, \tag{2.5}$$

where

$$pEM_j^{(K)} = 2(pl(\boldsymbol{\eta}_j^{(K)}) - pl(\hat{\boldsymbol{\theta}}_0)), \tag{2.6}$$

in which $pl(\cdot)$ is defined in Equation (2.2). Note that $pl(\hat{\boldsymbol{\theta}}_0)$ is the penalized maximum log-likelihood under the single normal component model, where $\boldsymbol{\beta}_2 = 0$ and $\sigma_1 = \sigma_2$.

Next, we describe the specific form of $\boldsymbol{\eta}_j^{(k)}$ in each step of the EM algorithm in detail, as follows.

Let $\boldsymbol{\eta}^{(k)}$ denote the values at the $k-th$ step of the EM algorithm for (2.2). The objective function at the $k-th$ step is

$$\begin{aligned} Q(\boldsymbol{\eta}|\boldsymbol{\eta}^{(k)}) &= \textstyle\sum_{i=1}^n \mathbb{E}_{\delta_i|w_i,\boldsymbol{\eta}^{(k)}}\{I_{(\delta_i=1)} \log\big(\frac{\pi(\boldsymbol{X}_i^T\boldsymbol{\gamma})}{\sigma_1}\exp(-\frac{(Y_i-\boldsymbol{Z}_i^T(\boldsymbol{\beta}_1+\boldsymbol{\beta}_2))^2}{2\sigma_1^2})\big)+ \\ &\quad I_{(\delta_i=0)}\log\big(\frac{1-\pi(\boldsymbol{X}_i^T\boldsymbol{\gamma})}{\sigma_2}\exp(-\frac{(Y_i-\boldsymbol{Z}_i^T\boldsymbol{\beta}_2)^2}{2\sigma_2^2})\big) \\ &\quad +p_n(\sigma_1)+p_n(\sigma_2)\}, \end{aligned}$$

To evaluate it, the E step involves the calculation of

$$\begin{aligned} a_i^{(k)} &= P(\delta_i = 1|Y_i, \boldsymbol{Z}_i, \boldsymbol{X}_i; \eta^{(k)}) \\ &= f(Y_i|\delta_i = 1, \boldsymbol{Z}_i; \theta^{(k)})P(\delta_i = 1|\boldsymbol{X}_i; \boldsymbol{\gamma}^{(k)})/(f(Y_i|\delta_i = 1, \boldsymbol{Z}_i; \boldsymbol{\theta}^{(k)})P(\delta_i = 1|\boldsymbol{X}_i; \boldsymbol{\gamma}^{(k)}) \\ &\quad +f(Y_i|\delta_i = 0, \boldsymbol{Z}_i; \boldsymbol{\theta}^{(k)})P(\delta_i = 0|\boldsymbol{X}_i; \boldsymbol{\gamma}^{(k)})), \end{aligned}$$

and $b_i^{(k)} = 1 - a_i^{(k)}$, and the M step gives

$$\begin{aligned} \boldsymbol{\gamma}^{(k+1)} &= \text{argmax}_{\boldsymbol{\gamma}} \textstyle\sum_{i=1}^n a_i^{(k)} \log \pi(\boldsymbol{X}_i^T\boldsymbol{\gamma}) + b_i^{(k)} \log(1 - \pi(\boldsymbol{X}_i^T\boldsymbol{\gamma})); \\ (\boldsymbol{\beta}_{12}^{(k+1)}, \sigma_1^{(k+1)}) &= \text{argmax}_{\boldsymbol{\beta}_{12},\sigma} \textstyle\sum_{i=1}^n a_i^{(k)} \log(\frac{1}{\sigma}\exp(-(Y_i - \boldsymbol{Z}_i^T\boldsymbol{\beta}_{12})^2/(2\sigma^2))) + p_n(\sigma); \\ (\boldsymbol{\beta}_1^{(k+1)}, \sigma_2^{(k+1)}) &= \text{argmax}_{\boldsymbol{\beta}_1,\sigma} \textstyle\sum_{i=1}^n b_i^{(k)} \log(\frac{1}{\sigma}\exp(-(Y_i - \boldsymbol{Z}_i^T\boldsymbol{\beta}_1)^2/(2\sigma^2))) + p_n(\sigma), \end{aligned}$$

and $\boldsymbol{\beta}_2^{(k+1)} = \boldsymbol{\beta}_{12}^{(k+1)} - \boldsymbol{\beta}_1^{(k+1)}$.

In the M step, we use the Newton-Raphson method to update $\boldsymbol{\gamma}^{(k+1)}$, and the updating formula for $\boldsymbol{\beta}^{(k+1)}$ is a weighted least squares solution. For the particular penalty (2.4), the calculations for $\sigma_1^{(k+1)}$ and $\sigma_2^{(k+1)}$ given $\boldsymbol{\beta}_1^{(k+1)}$ and $\boldsymbol{\beta}_2^{(k+1)}$ are simply

$$\sigma_1^{(k+1)} = \Big(\frac{\sum_{i=1}^n a_i^{(k)}(Y_i - \boldsymbol{Z}_i^T(\boldsymbol{\beta}_1^{(k+1)} + \boldsymbol{\beta}_2^{(k+1)}))^2/2 + \lambda S_n^2}{\sum_{i=1}^n a_i^{(k)}/2 + \lambda}\Big)^{1/2},$$

and

$$\sigma_2^{(k+1)} = \Big(\frac{\sum_{i=1}^n b_i^{(k)}(Y_i - \boldsymbol{Z}_i^T\boldsymbol{\beta}_1^{(k+1)})^2/2 + \lambda S_n^2}{\sum_{i=1}^n b_i^{(k)}/2 + \lambda}\Big)^{1/2}.$$

We see that the two variances from the penalized likelihood are weighted sums of S_n^2 and the corresponding estimators without the penalty.

2.4 Distribution of the Penalized EM Test Statistic

To evaluate the limiting distribution of the proposed pEM test statistic, we first define the *Fisher information matrix* for $\boldsymbol{\theta}$ given $\boldsymbol{\gamma}$ based on the penalized likelihood:

$$I_{\boldsymbol{\gamma}}^*(\boldsymbol{\theta}) = -\mathbb{E}[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\frac{pl(\boldsymbol{\eta})}{n}].$$

By direct calculations, for a given $\boldsymbol{\gamma}$, under the null hypothesis of $\boldsymbol{\beta}_2 = 0, \sigma_1 = \sigma_2 = \sigma$,

$$I_{\boldsymbol{\gamma}}^*(\boldsymbol{\theta}) = \frac{1}{\sigma^2}\begin{pmatrix} I_1 & 0_{2\times 2} \\ 0_{2\times 2} & I_2 \end{pmatrix}. \tag{2.7}$$

where

$$I_1 = \begin{pmatrix} \mathbb{E}\big(\boldsymbol{Z}\boldsymbol{Z}^T\big) & \mathbb{E}\big(\pi(\boldsymbol{X}^T\boldsymbol{\gamma})\boldsymbol{Z}\boldsymbol{Z}^T\big) \\ \mathbb{E}\big(\pi(\boldsymbol{X}^T\boldsymbol{\gamma})\boldsymbol{Z}\boldsymbol{Z}^T\big) & \mathbb{E}\big(\pi^2(\boldsymbol{X}^T\boldsymbol{\gamma})\boldsymbol{Z}\boldsymbol{Z}^T\big) \end{pmatrix}$$

and

$$I_2 = \begin{pmatrix} 2\mathbb{E}\big(\pi^2(\boldsymbol{X}^T\boldsymbol{\gamma}) - \sigma^2\frac{\mathbb{E}\left(p_n''(\sigma)\right)}{n}\big) & 2\mathbb{E}\big(\pi(\boldsymbol{X}^T\boldsymbol{\gamma})(1-\pi(\boldsymbol{X}^T\boldsymbol{\gamma}))\big) \\ 2\mathbb{E}\big(\pi(\boldsymbol{X}^T\boldsymbol{\gamma})(1-\pi(\boldsymbol{X}^T\boldsymbol{\gamma}))\big) & 2\mathbb{E}\big((1-\pi(\boldsymbol{X}^T\boldsymbol{\gamma}))^2\big) - \sigma^2\frac{\mathbb{E}\left(p_n''(\sigma)\right)}{n} \end{pmatrix}.$$

The above matrix is positive definite if the variable vectors $\boldsymbol{X}$ and $\boldsymbol{Z}$ are non-degenerate, and $\mathbb{E}\big(p_n''(\sigma)\big) < 0$.

To ensure that the penalty effect vanishes asymptotically, we further impose the following condition:

C6. Under the null model of ($\boldsymbol{\beta}_2 = 0, \sigma_1 = \sigma_2 = \sigma > 0$), we have $\mathbb{E}p_n''(\sigma) < 0$, $\mathbb{E}\big(p_n''(\sigma)\big) = o_p(n)$, and $p_n'(\sigma) = o_p(\sqrt{n})$.

Under C6, we have,

$$I_{\boldsymbol{\gamma}}^*(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}}\frac{pl(\boldsymbol{\eta})}{n}\left(\frac{\partial}{\partial \boldsymbol{\theta}}\frac{pl(\boldsymbol{\eta})}{n}\right)^T\right] + o_p(1),$$

and $I_{\boldsymbol{\gamma}}^*(\boldsymbol{\theta})$ works just like the usual Fisher information matrix for deriving the limiting distribution of the likelihood ratio statistic in a standard problem.

To be more specific, given $\boldsymbol{\gamma}$ with nonzero slope, let $\hat{\boldsymbol{\theta}}_n = \text{argmax}_{\boldsymbol{\theta}}\, pl(\boldsymbol{\eta})$ and $\hat{\boldsymbol{\theta}}_0 = \text{argmax}_{\boldsymbol{\theta}\in H_0}\, pl(\boldsymbol{\eta})$, then we have a quadratic approximation of the penalized likelihood ratio statistic $T^*(\boldsymbol{\gamma})$ and

$$T^*(\boldsymbol{\gamma}) = 2(pl(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\gamma}) - pl(\hat{\boldsymbol{\theta}}_0, \boldsymbol{\gamma})) = ||\frac{1}{\sqrt{n}}\psi^*(Y_i, \boldsymbol{Z}_i, \boldsymbol{X}_i; \boldsymbol{\gamma})||^2 + o_p(1), \tag{2.8}$$

where $\psi^*(Y_i, \boldsymbol{Z}_i, \boldsymbol{X}_i; \boldsymbol{\gamma}) = (\psi(Y_i, \boldsymbol{Z}_i, \boldsymbol{X}_i; \boldsymbol{\gamma})^T, \psi_0(Y_i, \boldsymbol{Z}_i, \boldsymbol{X}_i; \boldsymbol{\gamma}))$, in which

$$\boldsymbol{\psi}(Y_i, \boldsymbol{Z}_i, \boldsymbol{X}_i; \boldsymbol{\gamma}) = \frac{1}{\sigma_0}\boldsymbol{D}(\boldsymbol{\gamma})^{-1/2}\{\pi(\boldsymbol{X}_i^T\boldsymbol{\gamma})\boldsymbol{I}_{q_2} - \boldsymbol{B}(\boldsymbol{\gamma})\boldsymbol{A}^{-1}\}(Y_i - \boldsymbol{Z}_i^T\boldsymbol{\beta}_0)\boldsymbol{Z}_i, \tag{2.9}$$

and

$$\begin{aligned}\psi_0(Y_i, \boldsymbol{Z}_i, \boldsymbol{X}_i; \boldsymbol{\gamma}) \;=\; & \big(2[\mathbb{E}\big(\pi^2(\boldsymbol{X}^T\boldsymbol{\gamma})\big) - (\mathbb{E}\big(\pi(\boldsymbol{X}^T\boldsymbol{\gamma})\big))^2]\big)^{-1/2}(\mathbb{E}\big(\pi(\boldsymbol{X}^T\boldsymbol{\gamma})\big) - \pi(\boldsymbol{X}^T\boldsymbol{\gamma})) \\ & (\tfrac{(Y_i - \boldsymbol{Z}_i^T\boldsymbol{\beta}_1)^2}{\sigma^2} - 1),\end{aligned}$$

with $A = \mathbb{E}\big(\boldsymbol{Z}\boldsymbol{Z}^T\big); B(\boldsymbol{\gamma}) = \mathbb{E}\big(\pi(\boldsymbol{X}^T\boldsymbol{\gamma})\boldsymbol{Z}\boldsymbol{Z}^T\big), C(\boldsymbol{\gamma}) = \mathbb{E}\big(\pi^2(\boldsymbol{X}^T\boldsymbol{\gamma})\boldsymbol{Z}\boldsymbol{Z}^T\big)$; and $D(\boldsymbol{\gamma}) = C(\boldsymbol{\gamma}) - B(\boldsymbol{\gamma})A^{-1}B(\boldsymbol{\gamma})$. Direct calculations show that both $\psi(Y_i, \boldsymbol{Z}_i, \boldsymbol{X}_i; \boldsymbol{\gamma})$ and $\psi_0(Y_i, \boldsymbol{Z}_i, \boldsymbol{X}_i; \boldsymbol{\gamma})$ have mean zero, and the covariance matrix of ψ^* is I_{q_1+1}. Therefore, $T^*(\boldsymbol{\gamma})$ has a χ^2 limiting distribution with the degrees of freedom $q_1 + 1$. Note that here we have not updated the estimates through the EM iterations. In other words, $T^*(\boldsymbol{\gamma}) = pEM^{(0)}$ with only one starting value, $\boldsymbol{\gamma}$.

Following similar arguments to those used for Theorem 1 of Shen and He (2015), we see that

the representation in (2.8) holds uniformly in $\boldsymbol{\gamma} \in \tilde{\Gamma}$, where $\boldsymbol{\gamma} \in \tilde{\Gamma}$ is a prespecified compact set for $\boldsymbol{\gamma}$. We have the following theorem.

Theorem 2. *Under the null hypothesis and Assumptions C1-C6, for any finite integers $J > 0$ and $K \geq 0$, the penalized EM test statistic $pEM^{(K)}$ converges in distribution as $n \to \infty$. Specially, for $J = 1$ and $K = 0$, the limiting distribution is $\chi^2_{q_1+1}$ where q_1 is the dimension of $\boldsymbol{\beta}_2$.*

If the null hypothesis H_0 is rejected, the model parameter estimator is consistent from the penalized likelihood due to Theorem 1. Furthermore, from (2.8) it follows that the bootstrap method can be used to compute the p-value from the pEM test. The limiting distribution of the test statistic under the null hypothesis is not a simple chi-square distribution when $J > 1$ and $K \geq 1$, and moreover, the convergence to the limiting distribution is very slow for the test statistic even without covariates (Goffinet et al. (1992)). Therefore, we recommend to use the bootstrap method for computing the p-values.

2.5 Choice of the Tuning Parameter

The choice of the tuning parameter λ is practically important. Our conditions allow data-dependent penalties. For the specific penalty in (2.4), we show that Conditions C1-C3 and C6 are satisfied with any choice of λ in the interval

$$[n^{-2+c_1}(\log n)^3,\ n^{c_2}], \tag{2.10}$$

where $c_1, c_2 \in (0, 1)$ are any constants.

C1 and C3 are satisfied under the null model by noting that $c_2 \in (0, 1)$. Also note that $n^{c_2} = o(n)$ and $n^{c_2-1/2} = o(\sqrt{n})$ imply $\mathbb{E}\big(p''_n(\sigma_0)\big) = -4\lambda/\sigma_0^2 = o(n)$ and $p'_n(\sigma_0) = 2(n^{-1/2}\lambda)\{n^{1/2}(S_n^2 - \sigma_0^2)\}/\sigma_0^3 = 2(n^{-1/2}\lambda)O_p(1) = o_p(\sqrt{n})$, henceforth C6 is satisfied.

For C2, note that, under the null hypothesis, $S_n^2 \to \sigma_0^2$ almost surely. Write $S_n^2 = \sigma_0^2 + \epsilon_n$, where $\epsilon_n \to 0$ almost surely. Then for any $\sigma \in (0, 1/n)$ and sufficiently large n, we have

$$\begin{gathered} p_n(\sigma)\{(\log n)^2 \log \sigma\}^{-1} = \\ -\lambda\{(\sigma_0^2 + \epsilon_n)/\sigma^2 + \log(\sigma^2/(\sigma_0^2 + \epsilon_n))\}\{(\log n)^2 \log \sigma\}^{-1} \\ \geq (-\lambda/2)(\sigma_0^2/\sigma^2)\{(\log n)^2 \log \sigma\}^{-1}. \end{gathered} \tag{2.11}$$

Let $f_n(\sigma) = (-\lambda/2)(\sigma_0^2/\sigma^2)\{(\log n)^2 \log \sigma\}^{-1}$, then

$$\inf\{p_n(\sigma)\{(\log n)^2 \log \sigma\}^{-1} : 0 < \sigma \leq (1/n)\} \geq \inf\{f_n(\sigma) : 0 < \sigma \leq (1/n)\}.$$

Because $f_n(\sigma)$ is decreasing in $\sigma \in (0, n^{-1})$ for large n, we have

$$\inf\{p_n(\sigma)\{(\log n)^2 \log \sigma\}^{-1} : 0 < \sigma \leq (1/n)\} \geq f_n(n^{-1}) = O(\lambda n^2 (\log n)^{-3}).$$

By the choice of λ, for sufficiently large n, $\inf\{p_n(\sigma)\{(\log n)^2 \log \sigma\}^{-1} : 0 < \sigma \leq (1/n)\} \geq O(n^{c_1}) > C$, for any given constant C. Then, C2 is satisfied under the null hypothesis. The same results can be obtained under the alternative hypothesis due to the fact that S_n^2 is almost surely bounded.

2.6 Simulations

In this chapter, we report the performance of the proposed methods through Monte Carlo simulations. We compare the parameter estimates from the structured logistic-normal mixture model with equal and unequal variances, and the performance of the proposed pEM test versus the EM test of Shen and He (2015). We use $q_2 = 2$ and other parameters are given in details below. The bootstrap method is used to compute the p values of the tests for the empirical studies. The simulation results are part of the joint research that appeared in Shen, Wang and He (2016).

2.6.1 Estimation

We start with an evaluation of the parameter estimates under the two-component model, that is, when the mixture model parameters are all well defined. Data as random samples of sizes $n = 400$ are generated from

$$Y_i = \beta_{11} + \beta_{12} T_i + \beta_{13} Z_i + (\beta_{21} + \beta_{22} T_i + \beta_{23} Z_i)\delta_i + \varepsilon_{1i}\delta_i + \varepsilon_{2i}(1 - \delta_i),$$

$$P(\delta_i = 1 | X_i) = \pi(\gamma_{11} + \gamma_{12} X_i),$$

for $i = 1, \ldots, n$, where $\varepsilon_{1i} \sim N(0, \sigma_1^2)$ and $\varepsilon_{2i} \sim N(0, \sigma_2^2)$, independent of X_i, Z_i and T_i. We generate $X_i = Z_i$ from Uniform $(0, 4)$, and $T_i \in \{0, 1\}$ used to mimic the treatment indicator is

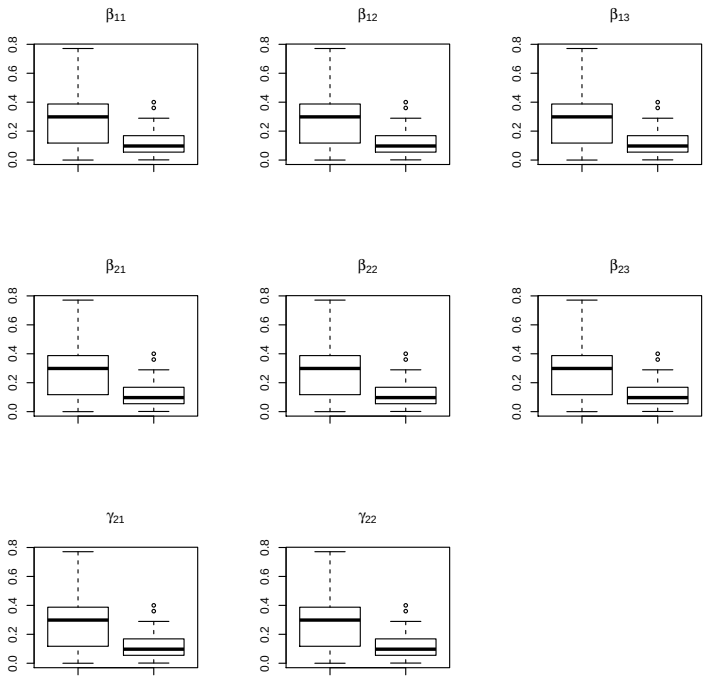

Figure 2.1: The boxplots of the absolute biases in 100 experiments discussed in Chapter 2.6.1. In each sub-panel, the left and the right boxes are for the estimates under the equal and the unequal variance models, respectively.

generated from the Bernoulli distribution with $P(T_i = 1) = 0.5$. We fix $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12}, \beta_{13}) = (2, 0, 2), \boldsymbol{\beta}_2 = (\beta_{21}, \beta_{22}, \beta_{23}) = (1, 2, 0), \boldsymbol{\gamma} = (\gamma_{11}, \gamma_{12}) = (2, -1), \sigma_1 = 1.5$ and $\sigma_2 = 0.5$. In the computations, we adopt the constraint that $\beta_{22} > 0$ to guarantee the uniqueness of the parameters. We show in Figure 2.1 the boxplots of the absolute bias of the parameter estimates based on 100 data sets. Not surprisingly, the estimates from the equal variance model have larger biases than those from the unequal variances model, so it is helpful to take the heterogeneity in the variances into consideration.

2.6.2 Type I errors

To evaluate the validity of the pEM test, we generate data from Model (2.1) with $q_1 = 3, q_2 = 2$, $\boldsymbol{\beta}_1 = (1, 0, 2)^T$, $\boldsymbol{\beta}_2 = (0, 0, 0)^T$, $\boldsymbol{Z} = (1, t, x)^T$, $\boldsymbol{X} = (1, x)^T$, where t resembles a treatment in-

dicator distributed as Bernoulli(0.5), x is independent of t with the distribution $N(-1, 1)$, and the error ε is white noise $N(0, 0.5^2)$. The pEM test uses the initial values $\mathbf{\Gamma} = \{(1, -2)^T, (1, 2)^T\}$. The resulting type I errors at $n = 60$ and 100 are summarized in Table 2.1, from which we can see the type I errors are quite close to the nominal levels for $K = 0, 3$, and 9, even for relatively small sample sizes. The tuning parameter is set as $\lambda = 1$ here, but the results are similar for other choices of λ.

Table 2.1: Type I errors of the pEM tests with bootstrap approximations in 1000 data sets with standard errors in the parenthesis, with $\lambda = 1$.

n	Nominal level α	$pEM^{(0)}$	$pEM^{(3)}$	$pEM^{(9)}$
n=60	0.01	0.012(0.003)	0.011(0.003)	0.011(0.003)
	0.05	0.055(0.007)	0.055(0.007)	0.050(0.007)
	0.10	0.102(0.010)	0.103(0.010)	0.106(0.010)
n=100	0.01	0.010(0.003)	0.011(0.003)	0.010(0.003)
	0.05	0.049(0.007)	0.051(0.007)	0.050(0.007)
	0.10	0.102(0.010)	0.099(0.009)	0.104(0.010)

2.6.3 Power Comparison

We use the same model and the same pEM test as in the previous subsection, except that $\boldsymbol{\beta}_2 = (1, a, b)^T$, $\boldsymbol{\gamma} = (1, 1)^T$ for some non-negative values of a and b to be given in the tables and for different sets of σ values. In particular, we consider $(\sigma_1 = 0.5, \sigma_2 = 0.5), (\sigma_1 = 0.4, \sigma_2 = 0.6), (\sigma_1 = 0.5, \sigma_2 = 1.0)$ and $(\sigma_1 = 0.5, \sigma_2 = 1.5)$ in Table 2.2 to represent different levels of heterogeneity. The power is obtained from the EM or pEM test under the equal or unequal variance model. We only show the comparisons at the iterations times $K = 9$ as this is our recommended choice. When the two component variances are close, the EM test based on the equal variance assumption is slightly more powerful, but when the two σ's differ with their ratios equal to 2 and 3, the pEM test under the unequal variance model is significantly more powerful.

Table 2.2: Power (%) of the (penalized) EM tests at the 5% level. The (penalized) EM test uses $\mathbf{\Gamma} = \{(1,2)^T, (1,-2)^T\}$, with $K = 9$ iterations. The parameters of Model (2.1) are $\boldsymbol{\beta}_1 = (1,0,2)^T$, $\boldsymbol{\beta}_2 = (1,a,b)^T$, $\boldsymbol{\gamma} = (1,1)^T$, and the tuning parameter is $\lambda = 1.0$.

n	a	b	$pEM^{(9)}$	$EM^{(9)}$	$pEM^{(9)}$	$EM^{(9)}$
			($\sigma_1 = 0.5$, $\sigma_2 = 0.5$)		($\sigma_1 = 0.4$, $\sigma_2 = 0.6$)	
60	0.5	1	71.2	77.8	73.4	73.6
60	0.5	0	35.6	36.0	42.2	37.6
60	1.0	1	85.2	87.8	86.6	87.8
60	1.0	0	81.4	84.8	82.8	82.0
100	0.5	1	92.0	96.8	92.8	94.8
100	0.5	0	57.8	54.8	74.6	49.6
100	1.0	1	96.8	99.4	97.8	98.8
100	1.0	0	95.8	97.6	97.2	96.0
			($\sigma_1 = 0.5$, $\sigma_2 = 1.0$)		($\sigma_1 = 0.5$, $\sigma_2 = 1.5$)	
60	0.5	1	49.0	38.4	53.8	31.0
60	0.5	0	36.8	27.2	55.0	40.8
60	1.0	1	63.4	47.2	63.8	39.6
60	1.0	0	63.0	44.8	70.8	47.8
100	0.5	1	77.6	60.6	81.2	42.0
100	0.5	0	65.8	34.2	81.8	51.2
100	1.0	1	87.6	75.4	86.6	51.8
100	1.0	0	89.8	58.6	90.0	58.6

CHAPTER 3

Logistic-Normal Mixture Model with High Dimensional Covariates

When the number of potential covariates is large, we have to perform variable selection. This type of problem falls into the area of the so-called variable selection with high dimensional covariates. There are quite a few methods, such as LASSO of Tibshirani (1996) and SCAD of Fan and Li (2001), that have been proposed for linear regression, quantile regression, logistic regression and etc, in which the objective log-likelihood functions are convex. However, finite mixture regression (FMR) models have non-convex log-likelihood, which makes variable selection in high dimensions substantially difficult. As far as we know, only Städler, Bühlmann and van de Geer (2010) and van de Geer (2013) considered the variable selection for FMR models for the case of $p = o(e^n)$, where p is the dimension of the potential covariates and n is the sample size. Khalili and Lin (2013) also studied the problem, but their setting is very restrictive; the p can be as large as $4n^{1/4} - 5$ at most, which is not really a high dimensional problem. Other than that, there is no discussion about FMR models in large p and small n scenario in literature.

Different from the FMR models with constant mixing parameters studied by Städler, Bühlmann and van de Geer (2010), van de Geer (2013) and Khalili and Lin (2013), we study the high dimensional variable selection method for logistic-normal mixture, which is a very special model in the family of FMR that allows the mixing parameters to depend on possibly high dimensional covariates. Besides, we note that the variable selection methods in high dimensions for FMR models discussed in literature rely on the assumption of known number of components. One may use BIC to select the number of components empirically, but BIC is only proved to work for exponential families with fixed number of covariates; there is no theoretical guarantee for BIC to work for FMR models with high dimensional covariates. We propose a new selection criterion

($SCMM$) for selecting the number of components for FMR models in high dimensions. We show $SCMM$ is consistent and we examine its empirical performance via simulations.

The rest of Chapter 3 is organized as follows. We introduce the logistic-normal mixture model with high dimensional covariates and a reparameterization in Chapter 3.1. Investigation of the variable selection methods with known number of components will be given in Chapter 3.2. In Chapter 3.3, we propose a new selection criterion $SCMM$ for selecting the number of components in high dimensional setting and show its consistency. We evaluate the performance of the proposed methods via simulations in Chapter 3.4, and we apply the methods to a real data example in Chapter 3.5. Note that the notations in this chapter are independent of those in Chapter 2.

3.1 Logistic Normal Mixtures

3.1.1 Model setup

Suppose that $\boldsymbol{X}_i \in R^p$, $i = 1, 2, \cdots, n$ are random or fixed covariates which take values in space $\mathcal{X}$. Assume $\mathcal{X}$ is bounded and $\{Y_i\}_{i=1}^n$ are independent conditional on $\{\boldsymbol{X}_i\}_{i=1}^n$, and

$$Y_i|\boldsymbol{X}_i = \boldsymbol{x}_i \sim \sum_{k=1}^{K} \pi(\boldsymbol{x}_i^T\boldsymbol{\gamma}_k)\frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_k)^2}{2\sigma_k^2}\right), \tag{3.1}$$

where $\pi(\boldsymbol{x}_i^T\boldsymbol{\gamma}_k) \equiv \exp(\boldsymbol{x}_i^T\boldsymbol{\gamma}_k)/(\sum_{k=1}^K \exp(\boldsymbol{x}_i^T\boldsymbol{\gamma}_k))$ with $\boldsymbol{\gamma}_1 = \mathbf{0}$.

The model (3.1) is the logistic normal mixtures with K components, where in each component, there are two p dimensional parameter vectors in both mixing and mean parts. We consider the scenario that $p \gg n$, and study the performance of the ℓ_1 penalized maximum log-likelihood estimator. Note that we assume K is known in chapter 3.1 and chapter 3.2.

3.1.2 Reparametrization

Section 3.1 of Städler, Bühlmann and van de Geer (2010) has a detailed discussion of the reparametrization. But to make the dissertation complete, we will briefly illustrate the necessity of reparametrizing the model.

Consider a Gaussian linear model,

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon, \tag{3.2}$$

where $\mathbf{Y}$ is a $n \times 1$ response vector, and $\mathbf{X}$ is a $n \times p$ design matrix. β is the unknown coefficients with dimensions $p \times 1$, and ϵ is the error vector that follows $N(0, \sigma^2 I_n)$. When $p \gg n$, the ℓ_1 norm penalized estimator, called LASSO, aims to find a sparse solution for β. It is defined as

$$\hat{\beta}_\lambda = \underset{\beta}{\operatorname{argmin}} \left\{ ||\mathbf{Y} - \mathbf{X}\beta||_2^2 + \lambda ||\beta||_1 \right\}, \tag{3.3}$$

where $||\cdot||_1$, $||\cdot||_2$ are the ℓ_1 and l_2 norm respectively, and λ is the tuning parameter.

Note that the log-likelihood function for the Gaussian linear model can be calculated as (without considering the constants)

$$l(\mathbf{Y}|\mathbf{X}, \beta, \sigma) = -n \log \sigma - ||\mathbf{Y} - \mathbf{X}\beta||_2^2 / 2\sigma^2.$$

Therefore the LASSO estimator in equation (3.3) is essentially a penalized log-likelihood based method, but without accounting for σ. There are two reasons for it to do so, one is that for a certain range of selections of the tuning parameters λ, the estimates of β by considering the full penalized log-likelihood method would be the same as the estimates of LASSO; the other reason is that σ for Gaussian linear model (3.2) is a nuisance parameter that people usually do not care.

However, there is no way to avoid the discussion of σ for mixture models. Mathematically, we can not separate β and σ completely in the log-likelihood function, and besides that, a good estimator of σ is usually very crucial for heterogeneous regression models as discussed in Chen, Tan and Zhang (2008) and McLachlan et al. (2004). Hence, it is inevitable to consider the optimization of the full penalized log-likelihood function for FMR models.

For Gaussian linear model, the full penalized maximum log-likelihood estimator is defined as

$$\hat{\boldsymbol{\eta}}_\lambda := (\hat{\beta}_\lambda, \hat{\sigma}_\lambda) = \underset{\beta, \lambda}{\operatorname{argmin}} \left\{ \log \sigma + ||\mathbf{Y} - \mathbf{X}\beta||_2^2 / (2n\sigma^2) + \lambda ||\beta||_1 \right\}. \tag{3.4}$$

There are two main drawbacks of the above estimator. Firstly, the estimator in (3.4) is not equivariant, in the sense that different scales of $\boldsymbol{Y}$ and $\boldsymbol{X}$ yield different estimates. Secondly and

more importantly, the objective function in equation (3.4) is non-convex. Many fast algorithms, e.g. pathwise coordinate optimization of Friedman et al. (2007), that are developed for convex optimizations such as LASSO, can no longer be applied.

Städler, Bühlmann and van de Geer (2010) resolved the above two issues by penalizing β and σ simultaneously and reparametrizing the model. Specifically, they first consider the penalized log-likelihood function as

$$\log \sigma + ||\mathbf{Y} - \mathbf{X}\beta||_2^2/(2n\sigma^2) + \lambda(||\beta||_1/\sigma), \tag{3.5}$$

and then, they reparameterize

$$\phi = \beta/\sigma,\ \rho = \sigma^{-1}.$$

The optimazation problem is then defined as

$$\hat{\boldsymbol{\theta}}_\lambda := (\hat{\phi}_\lambda, \hat{\rho}_\lambda) = \underset{\phi,\rho}{\operatorname{argmin}} \left\{ -\log \rho + \frac{1}{2n}||\rho\mathbf{Y} - \mathbf{X}\phi||_2^2 + \lambda(||\phi||_1) \right\}. \tag{3.6}$$

Note that the estimators in (3.6) are equivariant, and the objective function in (3.6) is now convex in ρ and ϕ.

We extend the idea of reparametrization to our model (3.1), and consider the ℓ_1 norm penalized maximum log-likelihood estimator. Let

$$\phi = \beta/\sigma,\ \rho = \sigma^{-1},$$

then

$$Y_i|\boldsymbol{X}_i = \boldsymbol{x}_i \sim \sum_{k=1}^{K} \pi(\boldsymbol{x}_i^T\boldsymbol{\gamma}_k)\frac{\rho_k}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\rho_k y_i - \boldsymbol{x}_i^T\boldsymbol{\phi}_k)^2\right). \tag{3.7}$$

Then, the ℓ_1 norm penalized estimator is

$$\begin{aligned}\hat{\boldsymbol{\theta}}_\lambda = \operatorname{argmin}_{\boldsymbol{\theta}\in\tilde{\boldsymbol{\Theta}}} & -\tfrac{1}{n}\textstyle\sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi(\boldsymbol{x}_i^T\boldsymbol{\gamma}_k)\frac{\rho_k}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\rho_k y_i - \boldsymbol{x}_i^T\boldsymbol{\phi}_k)^2\right)\right) \\ & +\lambda_1 \textstyle\sum_{k=1}^{K} ||\boldsymbol{\phi}_k||_1 + \lambda_2 \sum_{k=1}^{K} ||\boldsymbol{\gamma}_k||_1,\end{aligned} \tag{3.8}$$

where we use $\boldsymbol{\theta}$ to denote all the parameters, i.e., $\boldsymbol{\theta} = (\boldsymbol{\gamma}_2, \cdots, \boldsymbol{\gamma}_K, \boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_K, \boldsymbol{\eta})$, and $\boldsymbol{\eta} = (\rho_1, \cdots, \rho_K)$. λ_1, λ_2 are tuning parameters, $\tilde{\boldsymbol{\Theta}}$ is the parameter space, and we will discuss its regularizations in Chapter 3.2.

Note that with the reparametrization, the objective function in (3.8) is still non-convex. However, in each optimization step of the EM algorithm to solve the optimization problem, the log and the sum interchange and hence we have a convex objective function in every update step. Although the reparametrization does not alleviate the difficulty in studying the theoretical properties of the ℓ_1 penalized maximum log-likelihood estimator, it does help a lot in numeric data analysis in the sense that it avoids lots of local minima solutions.

In the next chapter, we will discuss the properties of $\hat{\boldsymbol{\theta}}_\lambda$ defined in (3.8) as well as the necessary conditions to ensure consistency.

3.2 Variable Selection for Known K

We consider the ℓ_1 penalized estimator defined in (3.8) for logistic normal mixtures in high dimensions. Because of the non-convexity, we need certain conditions on the parameter space to control the behavior of the log-likelihood function. Note that the constants below vary line by line.

3.2.1 Conditions

We assume the covariate space $\mathcal{X}$ is bounded, and

$$\tilde{\boldsymbol{\Theta}} \subset \boldsymbol{\Theta} := \left\{ \boldsymbol{\theta}; ||\log \boldsymbol{\eta}||_\infty \leq \tilde{K}, ||\boldsymbol{\theta} - \boldsymbol{\theta}_0||_1 \leq M \right\}, \tag{3.9}$$

where $\boldsymbol{\theta}_0$ is the true parameter and $\tilde{K}$ and M are fixed constants.

Define

$$\sup_{\boldsymbol{x}\in\mathcal{X}} ||\boldsymbol{\phi}^T\boldsymbol{x}||_\infty = \sup_{\boldsymbol{x}\in\mathcal{X}} \max_{1\leq k\leq K} |\boldsymbol{\phi}_k^T\boldsymbol{x}|, \quad \sup_{\boldsymbol{x}\in\mathcal{X}} ||\boldsymbol{\gamma}^T\boldsymbol{x}||_\infty = \sup_{\boldsymbol{x}\in\mathcal{X}} \max_{1\leq k\leq K} |\boldsymbol{\gamma}_k^T\boldsymbol{x}|.$$

Then for any $\boldsymbol{\theta} \in \tilde{\boldsymbol{\Theta}}$, we have

$$\sup_{\boldsymbol{x}\in\mathcal{X}} ||\boldsymbol{\phi}^T\boldsymbol{x}||_\infty \leq C \text{ and } \sup_{\boldsymbol{x}\in\mathcal{X}} ||\boldsymbol{\gamma}^T\boldsymbol{x}||_\infty \leq C \tag{3.10}$$

for some constant C. This is because the boundedness of $\mathcal{X}$ and

$$|\boldsymbol{\phi}^T\boldsymbol{x}| \leq |(\boldsymbol{\phi}-\boldsymbol{\phi}_0)^T\boldsymbol{x}| + |\boldsymbol{\phi}_0^T\boldsymbol{x}| \leq ||\boldsymbol{x}||_\infty(||\boldsymbol{\phi}-\boldsymbol{\phi}_0||_1 + ||\boldsymbol{\phi}_0||_1).$$

We also define

$$\psi(\boldsymbol{x}) = (\pi(\boldsymbol{x}^T\boldsymbol{\gamma}_1), \cdots, \pi(\boldsymbol{x}^T\boldsymbol{\gamma}_K), \boldsymbol{x}^T\boldsymbol{\phi}_1, \cdots, \boldsymbol{x}^T\boldsymbol{\phi}_K, \rho_1, \cdots, \rho_K).$$

Note that although $\psi(\boldsymbol{x})$ depends on $\boldsymbol{x}$ and $\boldsymbol{\theta}$, it has a fixed 3K dimension that is independent of n and p. Let $\boldsymbol{\psi}$ be the space for the collections of $\psi(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{\theta} \in \tilde{\boldsymbol{\Theta}}$, and we write $\psi_0(\boldsymbol{x})$ as the vector evaluated at the true parameter $\boldsymbol{\theta}_0$.

Let f_ψ denote the conditional density of Y given $\boldsymbol{x}$ through ψ, and let $l_\psi = \log f_\psi$, which is the log-density. Define $s_\psi = \partial l_\psi/\partial\psi$ to be the score function, and define the Fisher information as

$$I(\psi) = \int s_\psi s_\psi^T f_\psi d\mu,$$

where μ is the dominating measure of f_ψ.

Let the Kullback-Leibler information be denoted as

$$\varepsilon(\psi|\psi_0) = -\int \log\left[\frac{f_\psi}{f_{\psi_0}}\right] f_{\psi_0} d\mu.$$

The following 5 conditions are needed in this chapter, but note that Conditions 1, 2, 3 and 5 are met automatically for the logistic-normal mixture model in (3.1). See Städler, Bühlmann and van de Geer (2010) for details.

C1. It holds that

$$\sup_{\psi\in\boldsymbol{\psi}} \max_{(j_1,j_2,j_3)\in\{1,\cdots,3K\}^3} \left|\frac{\partial^3}{\partial\psi_{j_1}\partial\psi_{j_2}\partial\psi_{j_3}} l_\psi(\cdot)\right| \leq G_2(\cdot),$$

where

$$\sup_{\boldsymbol{x}\in\mathcal{X}} \int G_2(y) f_{\psi_0}(y|\boldsymbol{x}) d\mu(y) < \infty.$$

For a matrix A, let $\Lambda_{\min}(A)$ be its smallest eigenvalue.

C2. For all $\boldsymbol{x} \in \mathcal{X}$, the Fisher information matrix $I(\psi_0(\boldsymbol{x}))$ is positive definite, and,

$$\Lambda_{\min} = \inf_{\boldsymbol{x} \in \boldsymbol{X}} \Lambda_{\min}(I(\psi_0(\boldsymbol{x}))) > 0.$$

C3. For any $\epsilon > 0$, there exists an $\alpha_\epsilon > 0$ such that

$$\inf_{\boldsymbol{x} \in \mathcal{X}} \inf_{\psi \in \boldsymbol{\psi}, ||\psi - \psi_0(\boldsymbol{x})||_2 > \epsilon} \epsilon(\psi | \psi_0(\boldsymbol{x})) \geq \alpha_\epsilon.$$

Let the active set, i.e., the set of non-zero coefficients of $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_K) \in R^{Kp}$ be

$$S_1 = \{(k, j); \boldsymbol{\phi}_{k,j} \neq 0\}, \text{ and } s_1 = |S_1|,$$

and the active set of $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_K) \in R^{Kp}$ be

$$S_2 = \{(k, j); \boldsymbol{\gamma}_{k,j} \neq 0\}, \text{ and } s_2 = |S_2|.$$

C4. There exists a constant $\kappa \geq 1$ such that, for any $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_K) \in R^{Kp}$ satisfies

$$||\boldsymbol{\alpha}_{S_i^c}||_1 \leq 6||\boldsymbol{\alpha}_{S_i}||_1,$$

then

$$||\boldsymbol{\alpha}_{S_i}||_2^2 \leq \kappa^2 \sum_{k=1}^{K} \boldsymbol{\alpha}_k^T \Sigma_n \boldsymbol{\alpha}_k, \ i = 1 \text{ and } 2,$$

where $\Sigma_n = \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^T / n$.

C5.

$$\sup_{\boldsymbol{x} \in \mathcal{X}, \boldsymbol{\theta} \in \tilde{\boldsymbol{\Theta}}} ||s_{\psi(\boldsymbol{x})}(Y)||_\infty \leq G_1(Y),$$

where by direction calculations, $G_1(Y) = K \exp(2\tilde{K})(|Y|^2 + |Y| + 1)$.

Condition $C4$ is also called a restricted eigenvalue condition. Basically, it requires the active covariates and the non-active covariates not to be strongly correlated. With Conditions 1, 2 and 3, and by **Lemma 1** of Städler, Bühlmann and van de Geer (2010), we have, for any $\boldsymbol{x} \in \mathcal{X}$,

there exists a constant $c_0 \geq 1$, such that

$$||\psi(\boldsymbol{x}) - \psi_0(\boldsymbol{x})||_2^2 \leq c_0^2 \varepsilon(\psi|\psi_0(\boldsymbol{x})). \tag{3.11}$$

With these regularizations, we are ready to state our theorems.

3.2.2 Consistency Results

Let $\lambda_0 = (\log p/n)^{1/2}$, and

$$L_{\boldsymbol{\theta}}(\boldsymbol{x}, Y) = \log\left(\sum_{k=1}^{K} \pi(\boldsymbol{x}^T\boldsymbol{\gamma}_k)\frac{\rho_k}{\sqrt{2\pi}}\exp\left(-\tfrac{1}{2}(\rho_k Y - \boldsymbol{x}^T\boldsymbol{\phi}_k)^2\right)\right),$$
$$V_n(\boldsymbol{\theta}) = \tfrac{1}{n}\sum_{i=1}^{n}[L_{\boldsymbol{\theta}}(\boldsymbol{x}_i, Y_i) - E_{\boldsymbol{\theta}_0}\{L_{\boldsymbol{\theta}}(\boldsymbol{x}_i, Y_i)\}].$$

For any constant T, define the event

$$\mathcal{E}(T) = \left\{\sup_{\boldsymbol{\theta}\in\tilde{\Theta}} \frac{|V_n(\boldsymbol{\theta}) - V_n(\boldsymbol{\theta}_0)|}{(||\boldsymbol{\phi} - \boldsymbol{\phi}_0||_1 + ||\boldsymbol{\gamma} - \boldsymbol{\gamma}_0||_1 + ||\boldsymbol{\eta} - \boldsymbol{\eta}_0||_2) \vee \lambda_0} \leq T\lambda_0\right\}.$$

Denote the empirical average Kullback-Leibler information by

$$\bar{\varepsilon}(\hat{\theta}|\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\varepsilon(\hat{\theta}(\boldsymbol{x}_i)|\theta_0(\boldsymbol{x}_i)).$$

Then, we have following theorems.

Theorem 3. *Under the logistic-normal mixture model, if $p = o(\exp(n))$, then there exists a constant T, such that as $n \to \infty$, we have $P(\mathcal{E}(T)) \to 1$.*

Theorem 4. *Under Condition C4, for $\lambda_1 \geq 5T\lambda_0, \lambda_2 \geq 5T\lambda_0$, we have, on $\mathcal{E}(T)$,*

$$\bar{\varepsilon}(\hat{\theta}|\theta_0) + 2(\lambda_1 - T\lambda_0)||\hat{\boldsymbol{\phi}}_{S_1^c}||_1 + 2(\lambda_2 - T\lambda_0)||\hat{\boldsymbol{\gamma}}_{S_2^c}||_1$$
$$\leq C(\lambda_1 \vee \lambda_2 + T\lambda_0)^2(s_1 \vee s_2),$$

$$(\lambda_1 \wedge \lambda_2 + T\lambda_0)(||(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)_{S_1}||_1 + ||(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)_{S_2}||_1) \leq C(\lambda_1 \vee \lambda_2 + T\lambda_0)^2.$$

Remark. **Theorem 3** and **Theorem 4** together imply that with probability going to 1, we have an uniform upper bound for the empirical average Kullback-Leibler information as well

as the ℓ_1 norm of the estimated zero coefficients. Moreover, the estimated non-zero coefficients converge to the true values eventually. If we take $\lambda_1 = O(\lambda_0)$ and $\lambda_2 = O(\lambda_0)$, then the upper bound in **Theorem 4** is at the order of λ_0, which is $(\log p/n)^{1/2}$, and this is smaller than $\sqrt{\log n^3 \log(p \vee n)/n}$ of Städler, Bühlmann and van de Geer (2010). Because we have a theoretical upper bound for the estimated zero coefficients, in practice, we may use $\lambda_0 = (\log p/n)^{1/2}$ as a threshold to perform variable selection in order to get a sparse but still accurate model. A theorem for consistent variable selection with a beta-min condition are given as follows.

C6. **Beta-min** condition: There exists a large enough constant C, such that

$$\min_{j \in S_1} \phi_j \geq C\sqrt{\log p/n} \text{ and } \min_{j \in S_2} \gamma_j \geq C\sqrt{\log p/n}.$$

Define $\hat{S}_1 = \left\{(k,j); \hat{\phi}_{k,j} > C\sqrt{\log p/n}\right\}$ and $\hat{S}_2 = \left\{(k,j) : \hat{\gamma}_{k,j} > C\sqrt{\log p/n}\right\}$ to be the estimated activate sets for $\boldsymbol{\phi}$ and $\boldsymbol{\gamma}$ respectively. Then, if we properly choose some constant C with the threshold value $C\sqrt{\log(p)/n}$, we have

Theorem 5. *Under Conditions C4 and C6, as $n \to \infty$,*

$$P(\hat{S}_1 = S_1 \textit{ and } \hat{S}_2 = S_2) \to 1.$$

3.3 Selection for K

Chapter 3.2 provides the consistency theorems for variable selection assuming known number of components K. In practice, we often need to estimate K from data. We propose a procedure and a selection criterion called $SCMM$ in this chapter that can help to determine the true number of components K_0 consistently. With the estimated K_0, we can then utilize the ℓ_1 penalized likelihood approach discussed in Chapter 3.2 to perform consistent variable selection. We describe the procedure and $SCMM$ as follows.

Suppose we would like to identify the true underlying K_0 from the set $\mathcal{B} = \{1, 2, 3, \cdots, B\}$, where B is any given finite number. We may choose B large enough so that $K_0 \in \mathcal{B}$. For any $b \in \mathcal{B}$, we first optimize (3.8) with $K = b$ and denote the selected variable set as $\hat{s}(b)$. We then calculate $SCMM(\hat{s}(b))$, where $SCMM(\hat{s}(b))$ is the selection criterion $SCMM$ evaluating at the model b with the variable set $\hat{s}(b)$. Our estimated $\hat{K}_0$ is the b that minimizes $SCMM(\hat{s}(b))$.

Specifically,

$$\hat{K}_0 := \underset{b \in \mathcal{B}}{\operatorname{argmin}}\ SCMM(\hat{s}(b)), \tag{3.12}$$

where the selection criterion $SCMM(\hat{s}(b))$ is defined as

$$\begin{gathered} SCMM(\hat{s}(b)) := \\ -\frac{1}{n}\sum_{i=1}^{n} \log\left(\sum_{k=1}^{b} \pi(\boldsymbol{x}_{\hat{s}(b),i}^T \hat{\boldsymbol{\gamma}}_k) \frac{\hat{\rho}_k}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\hat{\rho}_k y_i - \boldsymbol{x}_{\hat{s}(b),i}^T \hat{\boldsymbol{\phi}}_k)^2\right)\right) \\ + b|\hat{s}(b)| n^{0.5+\delta_1} \log p/n, \end{gathered} \tag{3.13}$$

where $\{\hat{\boldsymbol{\gamma}}_k,\ \hat{\rho}_k,\ \hat{\boldsymbol{\phi}}_k\}_{k=1}^b$ are the maximum likelihood estimates by fitting the b-component logistic normal mixture model (model b) with the variable set $\hat{s}(b)$, and $|\hat{s}(b)|$ is the size of the variable set $\hat{s}(b)$ and δ_1 is any small positive constant. The need for a positive δ_1 is for the asymptotic consistency result that we prove, but in practice we use $\delta_1 = 0$, which works fine.

$SCMM$ is essentially the sum of the negative log-likelihood and the penalty on the model size. Let $\hat{\boldsymbol{\theta}} = \{\hat{\boldsymbol{\gamma}}_k,\ \hat{\rho}_k,\ \hat{\boldsymbol{\phi}}_k\}_{k=1}^b$ and write $l^b(\hat{s}(b), \hat{\boldsymbol{\theta}})$ as the log-likelihood evaluated at $\hat{\boldsymbol{\theta}}$ for model b. Also define the penalty function for any model b (b-component) with any variable set $\hat{s}(b)$ as $p_n(\hat{s}(b)) = b|\hat{s}(b)| n^{0.5+\delta_1} \log p/n$. Then we can write $SCMM(\hat{s}(b))$ from (3.13) in short,

$$SCMM(\hat{s}(b)) := -l^b(\hat{s}(b), \hat{\boldsymbol{\theta}}) + p_n(\hat{s}(b)). \tag{3.14}$$

Remark. The exact model size for model b with variable set $\hat{s}(b)$ is $(2b-1)|\hat{s}(b)| + b$, which is the sum of $b|\hat{s}(b)|$ for $\boldsymbol{\phi}$ and $(b-1)|\hat{s}(b)|$ for $\boldsymbol{\gamma}$ and b for $\boldsymbol{\rho}$, whereas we use $b|\hat{s}(b)|$ in the penalty function to roughly captures the model size. Since we compare models between different b so as to select the true K_0, the asymptotic behavior for $SMCC$ is only affected by the order of $n^{0.5+\delta_1} \log p/n$. Hence, using either the exact model size $(2b-1)|\hat{s}(b)| + b$ or the rough model size $b|\hat{s}(b)|$ in the penalty function will provide the same asymptotic property for $SMCC$. We use the rough model size for notational convenience. Note that if we select models within the same b, we need to use the exact model size. However, our goal is to select K, then there is no difference between using the exact model size and using the rough model size in the penalty function.

Briefly speaking, we propose a two-step procedure for selecting K_0. For each model b, the first step is to use the ℓ_1 penalized likelihood approach to select a small subset of variables $\hat{s}(b)$. The second step is to calculate $SCMM(\hat{s}(b))$, which is the sum of the negative $\log$ likelihood

evaluated at $\hat{s}(b)$ and the penalty on the model size. Because the size of $\hat{s}(b)$ is small, we hope the log-likelihood $l^b(\hat{s}(b), \hat{\boldsymbol{\theta}})$ can capture the model specific characteristics. Note that although $|\hat{s}(b)|$ is small, it is data dependent. The traditional AIC and BIC may not work in this setting. In order to take account for all $C_p^{\hat{s}(b)} = O(p^{\hat{s}(b)})$ possible subsets of variables with size $|\hat{s}(b)|$, we need a larger penalty in (3.13) than what is used in BIC. We see from the simulations in the next chapter that BIC tends to choose overfitted models, which suggests the penalty in BIC is not heavy enough.

Let $S_0 = \{(k, j) : \boldsymbol{\phi}_{k,j} \neq 0 \text{ or } \boldsymbol{\gamma}_{k,j} \neq 0\}$ be the true activate variable set. The following theorem provides a key property of SCMM, which is quite essential for showing its consistency.

Theorem 6. *Suppose $K_0 \in \mathcal{B} = \{1, 2, 3, \cdots, B\}$ and $p = o(n^c)$ for some c. For any fixed constant s, assume*

$$\sup_{\{b<K_0, |M|\leq s\} \bigcup \{b>K_0, S_0 \not\subset M, |M|\leq s\}} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} E(l^b(M, \boldsymbol{\theta})) < E(l^{K_0}(S_0, \boldsymbol{\theta}_0)). \tag{3.15}$$

Then as $n \to \infty$, we have

$$P\left(SCMM(S_0(K_0)) < \inf_{b\neq K_0, |M|\leq s} SCMM(M(b))\right) \to 1. \tag{3.16}$$

Remark. We recall the notations in equation (3.15) that $l^b(M, \boldsymbol{\theta})$ is the log-likelihood evaluated at the model b (b-component) with the variable set M at the parameters $\boldsymbol{\theta}$, and $SCMM(M(b))$ is the $SCMM$ evaluated at the model b with the variable set M. Therefore, $l^{K_0}(S_0, \boldsymbol{\theta}_0)$ and $SCMM(S_0(K_0))$ are the log-likelihood and the SCMM evaluated at the true model, respectively. Equation (3.15) is a necessary but mild assumption. Basically, it requires the maximum of the expected log-likelihood to be obtained only at the true model and in addition, there is a non-zero gap between the maximum and the expected log-likelihood evaluated at any other model which differs from the truth. If the gap can be infinitely close to $zero$ as p grows, we can not distinguish the truth from other models in terms of likelihood.

Theorem 7. *(Consistency of SCMM)*
Suppose $K_0 \in \mathcal{B} = \{1, 2, 3, \cdots, B\}$ and $p = o(n^c)$ for some c. Also assume conditions $C4$ and $C6$ and equation (3.15)*. Then as $n \to \infty$, we have $P(\hat{K}_0 = K_0) \to 1$.*

Proof: This is a direct result from **Theorem 5** and **Theorem 6**. To see this, let $\hat{S}_0$ be the

estimated active variable set given K_0 and define the set

$$A_n := \{\omega : \hat{S}_0 = S_0\}.$$

Also define the set

$$B_n := \{\omega : SCMM(S_0(K_0)) < \inf_{b \neq K_0, |M| \leq s} SCMM(M(b))\}.$$

By the definition of A_n and B_n, for any $\omega \in A_n \bigcap B_n$, $K_0 = \text{argmin}_{b \in \mathcal{B}}\ SCMM(\hat{s}(b))$. Hence, $A_n \bigcap B_n \subset \{\omega : \hat{K}_0 = K_0\}$. By **Theorem 5** and **Theorem 6**, $P(A_n) \to 1$ and $P(B_n) \to 1$. Therefore,

$$P(\hat{K}_0 = K_0) \geq P(A_n \bigcap B_n) \to 1.$$

3.4 Simulations

Chapter 3.2 considered the ℓ_1 penalized likelihood approach to perform variable selection with given K for logistic normal mixture models, whereas Chapter 3.3 discussed how to select K from data. In this chapter, we investigate the empirical performance of our proposed methods.

3.4.1 Variable selection with given K

In this chapter, we consider 2 examples for the purpose of demonstration that our proposed ℓ_1 norm penalized maximum likelihood estimator works well in terms of variable selection for high dimensional logistic normal mixture models with given K. Our main focus is on variable selection, that is, we want to see if the proposed estimator can select the true non-zero coefficients consistently while the number of the selected variables is small. After selection, we could estimate the parameters by re-fitting the model with the selected variables. The classical asymptotic theory of MLE guarantees consistent estimation if the selected model is correct.

Let T be the true model, and $\tilde{T}$ be the estimated model. Also let $T \subset \tilde{T}$ denote the case that the estimated model includes all the variables in the true model. We consider the sparsity of the estimated model by looking at its model size relative to the true model size. Specifically, we

define

$$\text{Relative Sparsity (RS)} = \frac{\#\text{of selected variables}}{\#\text{of variables in the true model}}.$$

Hence, if the true model contains 3 variables, then RS=2 means the estimated model selects 6 out of p variables.

To optimize (3.8) numerically, we use the generalized EM algorithm described in Städler, Bühlmann and van de Geer (2010). The idea is that in the M-step of the EM algorithm, instead of obtaining the minimizer of $\boldsymbol{\phi}, \boldsymbol{\gamma}, \rho$ simultaneously, we perform coordinate-wise updates. Note that in the M-step, there is no explicit update formula for $\boldsymbol{\gamma}$, and we approximate it by applying the Newton-Raphson method. Due to the non-convexity of (3.8), we try multiple starting values to try to find the global minimum. Based on **Theorem 5**, we also investigate the performance of a thresholding method in variable selection.

The empirical performances depend on the signal-to-noise ratio (SNR). We use the definition from Städler, Bühlmann and van de Geer (2010),

$$SNR = \frac{Var(Y)}{Var(Y|\boldsymbol{\theta} = \mathbf{0})}.$$

By setting all the coefficients $\boldsymbol{\theta} = 0$, $Var(Y|\boldsymbol{\theta} = \mathbf{0})$ denotes the variance purely from noise.

Example 1

We consider 4 logistic normal mixture models with 2 mixture components: M_1, M_2, M_3 and M_4. Models M_1 and M_2 have independent baseline covariates $\boldsymbol{X}$, whereas models M_3 and M_4 have correlated baseline covariates $\boldsymbol{X}$. We use the notation $corr_{l,m}$ to denote the correlation between covariates l and m. In all cases, $\boldsymbol{X}$ are simulated from a 200-variate Gaussian distribution, i.e., $p = 200$, however, only 3 of them are included in the true model. We consider small variances for mixture components in models M_1 and M_3, which produce high SNR. Models M_2 and M_4 have large variances for the components and hence have low SNR. The ℓ_1 normed penalized estimator is obtained from samples of size $n = 200$. The particular values for parameters $\boldsymbol{\rho}, \boldsymbol{\phi}, \boldsymbol{\gamma}$ are specified in Table 3.1, and SNR is calculated via Monte Carlo simulations.

We consider 100 realizations for each model. For each realization, we try 25 random starting values. To maintain the sparse nature of the true model, we require each starting value contains at most 10 non-zero values for $\boldsymbol{\phi}$ and $\boldsymbol{\gamma}$, respectively. To do so, we sample 10 out of $p = 200$

Table 3.1: Models for **Example 1** for $K = 2$. $\delta_{l,m}$ denotes Kronecker's delta

	M1	M2	M3	M4
n	200	200	200	200
p	200	200	200	200
k	2	2	2	2
β_1	(3,3,3,3)	(3,3,3,3)	(3,3,3,3)	(3,3,3,3)
β_2	(-1,-1,-1,-1)	(-1,-1,-1,-1)	(-1,-1,-1,-1)	(-1,-1,-1,-1)
γ_2	(3,-2,4)	(3,-2,4)	(3,-2,4)	(3,-2,4)
σ	(0.5,0.3)	(1.5,0.9)	(0.5,0.3)	(1.5,0.9)
$corr_{l,m}$	$\delta_{l,m}$	$\delta_{l,m}$	$\|0.8\|^{l-m}$	$\|0.8\|^{l-m}$
SNR	64.3	7.2	88.5	9.7

numbers to indicate the indices of initial non-zero $\boldsymbol{\phi}$ and $\boldsymbol{\gamma}$. For these 10 non-zero $\boldsymbol{\phi}$ and $\boldsymbol{\gamma}$, we generate their values from $N(0, 12)$. Repeat the procedure 25 times and we get 25 different starting values. For each starting value, we get the estimates of the parameters via optimizing (3.8), and finally we compare the value of (3.8) by plugging these 25 different estimates and choose the one that minimize (3.8) as our final estimates for one realization. For 100 realizations, Table 3.2 shows the probability of the estimated model including all the active covariates and the average RS for both the thresholding and the non-thresholding method. We use a naive threshold $\sqrt{\log p/n}$.

We can see from Table 3.2 that both the thresholding and the non-thresholding methods select the true set of the variables with very high probability for well chosen λ_1 and λ_2. In addition, the thresholding method with a naive threshold $\sqrt{\log p/n}$ provides much sparser models compare to the non-thresholding method, whereas the accuracy does not loss much in the sense that the inclusion probabilities are almost the same for both methods. We also note that the correlated covariates $\boldsymbol{X}$ make the variable selection harder and yield lower inclusion probabilities compare to those models with independent covariates. SNR affects the average RS; the estimated models for $M2$ and $M4$, which have low SNR, tend to contain more active covariates, which leads to larger average RS.

To verify the claim in **Theorem 5** that as n and p go to infinity, with properly chosen λ_1, λ_2 and a threshold value, the probability of the exact recovery of the model goes to one, we examine model M_1 with $n = 500$ and $p = 500$. We also use the thresholding method with a

Table 3.2: Probability of the estimated model containing the true model and the average RS based on 100 realizations for **Example 1** of $K = 2$. The numbers in the parentheses are from the thresholding method with a naive threshold $\sqrt{\log p/n}$.

$K = 2$	(λ_1, λ_2)	(0.1,0.02)	(0.08,0.01)	(0.06,0.03)	(0.12,0.05)	(0.14, 0.04)
M1	$T \subset \tilde{T}$	1(1)	1(1)	1(1)	0.52(0.4)	0.52(0.29)
	ARS	4.8(1.4)	8.84(2.83)	3.48(1.16)	1.22(0.73)	1.41(0.65)
M2	$T \subset \tilde{T}$	1(0.99)	1(1)	1(1)	0.31(0.08)	0.27(0.05)
	ARS	4.8(1.37)	10.8(2.7)	3.2(1.1)	1.22(0.5)	1.29(0.43)
M3	$T \subset \tilde{T}$	0.82(0.62)	0.97(0.96)	0.37(0.11)	0(0)	0(0)
	ARS	4.15(1.47)	7.64(3.07)	3.33(1.04)	1.06(0.86)	1.32(0.89)
M4	$T \subset \tilde{T}$	0.62(0.42)	0.96(0.92)	0.16(0.03)	0(0)	0.02(0)
	ARS	4.45(1.37)	8.71(2.86)	5.42(1.0)	1.12(0.86)	1.27(0.87)

naive threshold $\sqrt{\log p/n}$. Then, with the tuning parameters $(\lambda_1, \lambda_2) = (0.11, 0.028)$, we will *exactly* select the true set of the variables 99 out of 100 times, which demonstrates the claim in **Theorem 5**.

Example 2

It is usually easy to deal with two subgroups, $i.e.$, $K = 2$, since we only need to estimate $\boldsymbol{\gamma}_2$ to identify subgroup membership. When $K > 2$, say $K = 3$, mulitple $\boldsymbol{\gamma}$ come into the model and different combinations of $\boldsymbol{\gamma}$ might give similar subgroup structures. We would like to see if our method still works well for $K = 3$.

We consider 4 logistic normal mixture models with 3 mixture components: M_5, M_6, M_7 and M_8. Similar to **Example 1**, Models M_5 and M_6 have independent baseline covariates $\boldsymbol{X}$, whereas models M_7 and M_8 have correlated baseline covariates $\boldsymbol{X}$. We use the notation $corr_{l,m}$ to denote the correlation between covariates l and m. In all cases, $\boldsymbol{X}$ are simulated from a 200-variate Gaussian distribution, i.e., $p = 200$, however, only 3 of them are included in the true model. We consider small variances for mixture components in models M_5 and M_7, which produce high SNR, whereas models M_6 and M_8 have large variances for the components and hence have low SNR. The ℓ_1 normed penalized estimator is obtained from samples of size $n = 200$. The particular values for parameters $\boldsymbol{\rho}, \boldsymbol{\phi}, \boldsymbol{\gamma}$ are specified in Table 3.3, and SNR is calculated via

Table 3.3: Models for **Example 2** for $K = 3$. $\delta_{l,m}$ denotes Kronecker's delta

	M5	M6	M7	M8
n	200	200	200	200
p	200	200	200	200
k	3	3	3	3
β_1	(3,3,3,3)	(3,3,3,3)	(3,3,3,3)	(3,3,3,3)
β_2	(-1,-1,-1,-1)	(-1,-1,-1,-1)	(-1,-1,-1,-1)	(-1,-1,-1,-1)
β_3	(1,-2,2)	(1,-2,2)	(1,-2,2)	(1,-2,2)
γ_2	(3,-2,4)	(3,-2,4)	(3,-2,4)	(3,-2,4)
γ_3	(-1,1,1)	(-1,1,1)	(-1,1,1)	(-1,1,1)
σ	(0.5,0.3,0.2)	(1.5,0.9,0.6)	(0.5,0.3,0.2)	(1.5,0.9,0.6)
$corr_{l,m}$	$\delta_{l,m}$	$\delta_{l,m}$	$\|0.8\|^{l-m}$	$\|0.8\|^{l-m}$
SNR	65	8.0	112	13.5

Monte Carlo simulations.

We again consider 100 realizations for each model, and use the same procedure discussed in **Example 1** to get the ℓ_1 normed penalized estimates. Table 3.4 shows the probability of the estimated model including all the active covariates and the average RS for both the thresholding and the non-thresholding methods with a naive threshold $\sqrt{\log p/n}$.

The high inclusion probabilities and the small average RSs shown in Table 3.4 proved that our proposed estimator works well even for complicated data structure ($K = 3$). Still in **Example 1** of $K = 2$, we have even higher inclusion probabilities and smaller average RSs. This means the estimations for $K = 2$ are not only sparser but also more accurate than those for $K = 3$. This is reasonable. Roughly speaking, with the same sample size $n = 200$, we have on average 70 observations for each mixture component when $K = 3$, whereas we have on average 100 observations for each mixture component when $K = 2$. Hence, the estimators for $K = 2$ are expected to be more accurate than $K = 3$.

3.4.2 Selection for K

Quite often, we do not know the number of the mixture components K. We need to estimate it from the data. Städler, Bühlmann and van de Geer (2010) proposed to use cross validation, whereas Khalili and Lin (2013) suggested to use BIC to select K in practice. However,

Table 3.4: Probability of the estimated model containing the true model and the average RS based on 100 realizations for **Example 2** of K=3. The numbers in the parentheses are from the thresholding method with a naive threshold $\sqrt{\log p/n}$.

$K=3$	(λ_1,λ_2)	(0.1,0.02)	(0.08,0.01)	(0.06,0.03)	(0.12,0.05)	(0.14, 0.04)
M5	$T\subset\tilde{T}$	1(0.98)	1(1)	1(1)	0.46(0.23)	0.29(0.14)
	ARS	4.66(1.41)	9.91(3.23)	4.21(1.14)	0.96(0.78)	1.08(0.59)
M6	$T\subset\tilde{T}$	0.97(0.8)	0.95(0.93)	0.93(0.65)	0.16(0.03)	0.07(0.04)
	ARS	4.47(1.35)	12(3.5)	6.29(1.12)	0.62(0.31)	0.75(0.25)
M7	$T\subset\tilde{T}$	0.95(0.8)	1(0.99)	0.88(0.64)	0.09(0)	0.13(0.03)
	ARS	4.9(1.66)	9.2(3.7)	4.22(1.18)	1.1(0.87)	1.43(0.90)
M8	$T\subset\tilde{T}$	0.71(0.42)	0.95(0.9)	0.42(0.09)	0.04(0)	0.02(0)
	ARS	4.97(1.52)	10.2(3.5)	5.32(1.06)	1.07(0.86)	1.26(0.88)

these methods do not yet have theoretical justifications. We proposed a new selection criterion $SCMM$ in Chapter 3.3 and showed its consistency. In this chapter, we would like to examine its empirical performance and compare it to that of BIC and EBIC of Chen and Chen (2008) (EBIC is proved to be useful in high dimensional variable selection in the linear regression framework). We show in simulations that BIC tends to choose overfitted models.

Recall the variable selection procedure in Chapter 3.3 that for any $b\in\mathcal{B}$, we first optimize (3.8) with $K=b$ and denote the selected variable set as $\hat{s}(b)$. The estimators of K from SCMM, BIC and EBIC are given as follows.

$$\hat{K}^{SCMM} := \text{argmin}_{b\in\mathcal{B}}\ SCMM(\hat{s}(b)),$$
$$\hat{K}^{BIC} := \text{argmin}_{b\in\mathcal{B}}\ BIC(\hat{s}(b)),$$
$$\hat{K}^{EBIC} := \text{argmin}_{b\in\mathcal{B}}\ EBIC(\hat{s}(b)),$$

where

$$SCMM(\hat{s}(b)) := -l^b(\hat{s}(b),\hat{\boldsymbol{\theta}}) + b|\hat{s}(b)|n^{0.5+\delta_1}\log p/n,$$
$$BIC(\hat{s}(b)) := -l^b(\hat{s}(b),\hat{\boldsymbol{\theta}}) + (2b-1)|\hat{s}(b)|\log n/(2n),$$
$$EBIC(\hat{s}(b)) := BIC(\hat{s}(b)) + (2b-1)|\hat{s}(b)|\log p/(2n).$$

We first consider the 4 models in last chapter, $M_1,\cdots,M_4$, and pretend as if we do not know the right number of the mixture components. We use SCMM, BIC and EBIC selection criterion discussed above to select K from data, with the candidate set $\mathcal{B}=\{1,2,3,4\}$ and the com-

Table 3.5: Frequencies of the estimates $\hat{K}$ for the true K=2 based on SCMM, BIC and EBIC

$\hat{K}(SCMM, BIC, EBIC)$	M1	M2	M3	M4
1	0,0,0	0,0,0	0,0,0	0,0,0
2	98,92,99	99,97,98	100,82,96	100,97,100
3	2,8,1	1,3,2	0,17,4	0,3,0
4	0,0,0	0,0,0	0,1,0	0,0,0

mon tuning parameters $(\lambda_1, \lambda_2) = (0.06, 0.03)$. We use 100 realizations for each model. The frequencies of $\hat{K}$ being 1, 2, 3 and 4 from SCMM, BIC, EBIC are shown in Table 3.5.

We see SCMM and EBIC works well for all 4 models, whereas for model M3, BIC chooses overfitted models 18 out of 100 times.

Note that the linear regression for one single normal component model (K=1) is the degenerate case of the mixture models, whereas it is fundamentally different from the mixture models. The convex log-likelihood function and the explicit formulas of the parameters estimation for linear regression make it attractive to real data applications. Hence, it is important for us to tell if a single normal component model is sufficient for the data or if we need to use the more complicated mixture models structure. The following simulations demonstrate that our proposed selection criterion SCMM can detect $K = 1$ consistently if the data is from the single normal component model, whereas BIC tends to choose $K = 2$.

We consider 4 single normal component models: F_1, F_2, F_3 and F_4. Models F_1 and F_2 have independent baseline covariates $\boldsymbol{X}$, whereas models F_3 and F_4 have correlated baseline covariates $\boldsymbol{X}$. We use the notation $corr_{l,m}$ to denote the correlation between covariates l and m. In all cases, $\boldsymbol{X}$ are simulated from a 200-variate Gaussian distribution, i.e., $p = 200$, however, only 2 of them are included in the true model. We consider small error variances for models F_1 and F_3 and large error variances for models F_2 and F_4. The ℓ_1 normed penalized estimator is obtained from samples of size $n = 200$. The particular values for parameters $\boldsymbol{\rho}, \boldsymbol{\phi}, \boldsymbol{\gamma}$ are specified in Table 3.6.

We again consider the candidate set $\mathcal{B} = \{1, 2, 3, 4\}$ and 100 realizations for each model with the common tuning parameters $(\lambda_1, \lambda_2) = (0.1, 0.02)$. The frequencies of $\hat{K}$ being 1, 2, 3 and 4 from SCMM, BIC and EBIC are shown in Table 3.7. Note that the penalty used in SCMM is larger than those used in BIC and EBIC, hence quite often, if BIC and EBIC choose $K = 1$,

Table 3.6: Models for K=1. $\delta_{l,m}$ denotes Kronecker's delta

	F_1	F_2	F_3	F_4
n	200	200	200	200
p	200	200	200	200
k	1	1	1	1
β	(1,-1,1)	(1,-1,1)	(1,-1,1)	(1,-1,1)
σ	1	2	1	2
$corr_{l,m}$	$\delta_{l,m}$	$\delta_{l,m}$	$\|0.8\|^{l-m}$	$\|0.8\|^{l-m}$
SNR	3	1.5	1.4	1.1

Table 3.7: Frequencies of the estimates $\hat{K}$ for the true K=1 based on SCMM, BIC and EBIC

$\hat{K}(SCMM, BIC, EBIC)$	F_1	F_2	F_3	F_4
1	92,12,27	96,35,93	98,8,40	61,50,60
2	8,87,73	4,62,7	2,90,59	39,48,40
3	0,1,0	0,3,0	0,2,1	0,2,0
4	0,0,0	0,0,0	0,1,0	0,0,0

then SCMM will also do. The simulation results in Table 3.7 show that BIC and EBIC generally loss the power to detect the true $K = 1$, however, our proposed SCMM works very well for F_1, F_2 and F_3. The SNR for F_4 is as low as 1.1, which approaches the lower bound of SNR of 1. Hence, it is reasonable in this case to see less satisfactory performances for all the methods.

3.5 Real Data Example

We apply of our methods to a real data set. The data set concerns telemonitoring of Parkinson's disease (PD), which is a neurological disorder that has affected over one million people in North America. Although the current medications are effective in controlling its symptoms at the early stages of the disease, there is no prescription to cure the disease. Therefore, it is important to diagnose and monitor PD in the early phase. Traditional ways of tracking PD symptoms involves physical examinations. To reduce the cost yet still able to track PD progression, a noninvasive telemonitoring technique, called sustained vowel phonations (SVP) (Little et al, 2009), has been

proposed. The data set at http://archive.ics.uci.edu/ml/datasets/Parkinsons+ Telemonitoring is collected from a clinical trial to assess if SVP can be used as surrogate to study PD severity and progression. There are $n = 5875$ observations of speech signals of 42 patients and a total of 16 characteristics with clinically relevant properties were extracted for each observation (Little et al., 2009). Besides the 16 characteristics, we consider 16 extra covariates which are quadratic forms of the 16 characteristics. The reasons for including those quadratic terms have been discussed in Tsanas et al. (2010).

Using the Unified Parkinson's Disease Rating Scale (UPDRS), which is designed to follow PD progression, as the response variable and $p = 32$ covariates (16 linear and 16 quadratic), Khalili and Lin (2013) fitted the mixture model with constant mixing parameters to the data set. They also assumed constant variability for subgroups. Under their framework, they chose $K = 2$ and selected 6 out of 32 variables as the predictors.

We believe that homogeneity of subgroups is not a reasonable assumption in this study. The purpose of our study is to apply our logistic-normal mixture model to see if we will find something different. To provide a direct contrast with the results of Khalili and Lin (2013), we proceed as if the observations were independent. We note that within-subject correlations need to be examined in any inferential analysis.

We fit the logistic-normal mixtures to the data set with $K = 1, 2, 3, 4$. Based on our proposed selection criterion $SCMM$ with the tuning parameters $(\lambda_1, \lambda_2) = (3\sqrt{\log p/n}, \sqrt{\log p/n}) = (0.073, 0.024)$, we also chose $K = 2$, but selected only 3 variables as the predictors, which are PPE, HNR2 and DFA2, where HNR2 and DFA2 are quadratic terms of HNR and DFA, respectively. Table 3.8 gives the detailed estimation.

Note that $\boldsymbol{\phi}, \rho$ are the reparameterized parameters. We transfer them back and write the fitted model in terms of $\boldsymbol{\beta}$ and σ as follows.

$$
\begin{gathered}
Group_1(G_1): \\
Y(UPDRS) = 29.6 + 1.88PPE - 2.67HNR2 - 3.41DFA2 + \epsilon_1,\ \epsilon_1 \sim N(0, 10.46^2). \\
Group_2(G_2): \\
Y(UPDRS) = 21.32 - 0.95PPE - 3.55HNR2 - 3.4DFA2 + \epsilon_2,\ \epsilon_2 \sim N(0, 2.98^2).
\end{gathered}
$$

Based on the estimated $\boldsymbol{\gamma}_2$, the average mixing probabilities for $Group_1$ and $Group_2$ are 0.94 and 0.06, respectively. The results suggest that the vast majority of the cases come from $Group_1$. We look at the linear regression for a single normal component model with covariates

Table 3.8: Parameter estimates and the corresponding standard errors. Standard errors are obtained via observed Fisher information and are reported inside parentheses purely for reporting purposes. HNR2 and DFA2 are the quadratic terms of HNR and DFA respectively. All covariates are standardized.

Predictors	ϕ_1	ϕ_2	γ_2	ρ_1	ρ_2
Intercept	2.83 (0.01)	7.15 (0.35)	-8.02 (0.85)	9.5×10^{-2} (3×10^{-4})	0.34 (2.4×10^{-3})
PPE	0.18 (0.02)	-0.32 (0.08)	-0.45 (0.2)	-	-
HNR2	-0.26 (0.02)	-1.19 (0.13)	-3.93 (0.45)	-	-
DFA2	-0.33 (0.02)	-1.14 (0.23)	-4.63 (0.43)	-	-

PPE, HNR2 and DFA2. The estimated model from linear regression is

Global linear regression :

$$Y(UPDRS) = 29.02 + 1.19PPE - 1.89HNR2 - 2.34DFA2 + \epsilon,\ \epsilon \sim N(0, 10.29^2).$$

Not surprisingly, the model from the global linear regression is fairly close to that of $Group_1$. In fact, the global linear regression with the covariates PPE, HNR2 and DFA2 has already provided a very high adjusted R-square of 0.8894. However, the QQ-plot for the residuals in Figure 3.1 does not support the normality assumption and statistically, the BIC which works well for small p and large n and SMCC of the global linear regression are worse than those of the two components logistic normal mixture model in Table 3.8. Our analysis indicates that there may exist a small portion of observations that are actually from $Group_2$, which can not be predicted well by the global linear regression.

The FMR model with constant mixing parameters discussed in Khalili and Lin (2013) is unable to provide much information, since the model assigns common mixing probabilities for every observation. A distinct advantage of the logistic normal mixture model is that based on different attributes of different observations, we can calculate individual-specific mixing probabilities and predict their subgroup memberships. Figure 3.2 shows the individual-specific mixing probabilities for $Group_2$. The observations that have large mixing probabilities for $Group_2$ are very likely from $Group_2$.

We select individual observations whose mixing probabilities for $Group_2$ are larger than 0.85 as a subgroup, and we call it S. The size for S is 168. Note that they are not necessarily all from

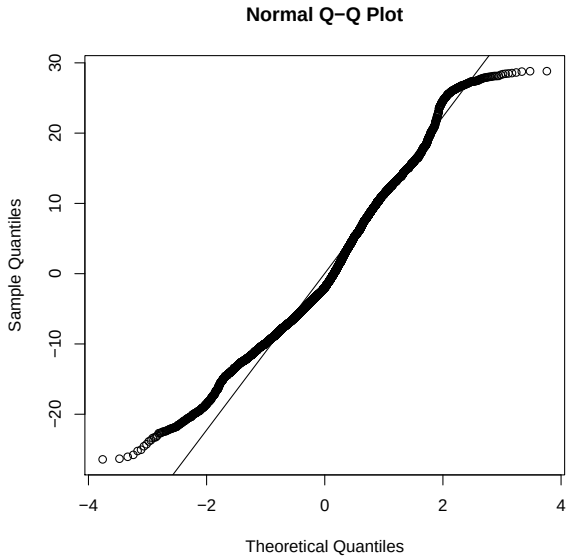

Figure 3.1: QQ-plot for the linear regression residuals

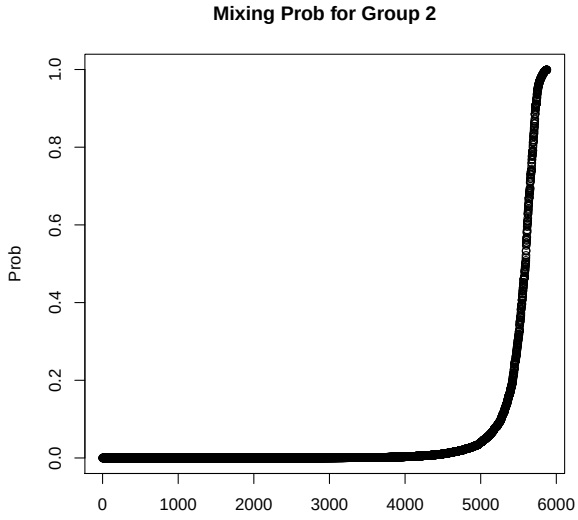

Figure 3.2: Individual-specific mixing probabilities for $Group_2$

$Group_2$, but we believe most of them should be because of the large mixing probabilities, and we expect the majority of the observations from S can not be predicted well by the global linear

regression. Figure 3.3 is the plot for the response Y(UPDRS) against the individual-specific characters X=1.19 PPE -1.89 HNR2 -2.34DFA2 from the global linear regression coefficients. The straight line in black is the fitted global linear regression and the red points are the selected individuals in S. We can see that the points (green) other than red are fairly symmetric with respect to the line, whereas the red points seem to be outliers. Indeed, the global linear regression can not predict these red points well. Figure 3.3 explained the reason intuitively why a single linear regression is not sufficient for the data, and it also demonstrated that the two component logistic normal mixture model is effective in characterizing the subgroup membership based on the mixing probabilities.

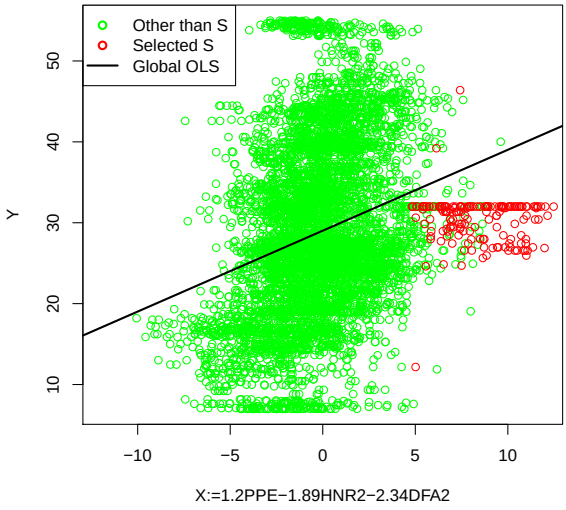

Figure 3.3: The overall data set with the linear regression line in black. The red points are the selected individuals in S based on the mixing probabilities.

Quantitatively, we compare the mean square errors (MSE) of the fitted values to the true values for the observations from S via model G_2 and the global linear regression. The MSE for model G_2 and the global linear regression are 9.6 and 148.5, respectively, whereas the optimal MSE from the least squares of the linear regression based solely on the observations from S is 8.1. Because the MSE for G_2 is close to the optimal MSE and is much smaller than that for the global linear regression, it also demonstrates that most of the observations from S should be centered around the means from $Group_2$.

For the mixture model discussed in Khalili and Lin (2013) (M_{KL}), which assumes both homogeneity for subgroups and constant mixing parameters, we found if we relax the homogeneity condition, the fitted model (M_{KL1}) has better BIC. The estimated mixing parameters are 0.92 and 0.08, which also suggests that most cases are from one single subgroup. However, both the models M_{KL} and M_{KL1} assign common mixing probabilities for every observation and hence fail to capture the underlying subgroup membership for the data set. Consequently, the BIC for M_{KL} and M_{KL1} are worse than the two components logistic normal mixture model in Table 3.8. The BIC for M_{KL} is 2.92, and 2.84 for M_{KL1}, and the BIC for the two components logistic normal mixture model in Table 3.8 is 2.79.

Further inspections of the data show that most observations in S are from one subject, and quite a few others are from another subject. Specifically, 43 out of 168 observations in S are from patient 33 and 110 out of 168 observations in S are from patient 36. The analysis shows that the logistic-normal mixture model is more effective than the traditional FMR model in detecting meaningful subgroups.

When we added 500 noise predictors to the data, each of which is simulated independently from the normal distribution with mean zero and variance one, and choose the tuning parameters $(\lambda_1, \lambda_2) = (\sqrt{\log p/n}, \sqrt{\log p/n}) = (0.033, 0.033)$, we once again selected $K = 2$ and the active covariates PPE, HNR2 and DFA2. This exercise shows that the proposed method in this Chapter can handle a large number of possible predictors.

CHAPTER 4

Discussion

In this dissertation, we considered two issues that are related to the logistic-normal mixture model. In Chapter 2, we proposed a penalized maximum likelihood estimator that can consistently estimate the true parameters when the variances might be unequal for different subgroups. In addition, we proposed a penalized EM test for testing the existence of subgroups. We allow to use data dependent penalty, and we provide a guideline for selecting tuning parameter. The existing literature mainly focuses on homogeneous subgroups or FMR models with no covariates, whereas our setting is geared towards more practical problems where the subgroup membership needs to be characterized by covariates.

In Chapter 3, we proposed a ℓ_1 norm penalized maximum likelihood estimator for the sake of variable selection for the logistic normal mixture models with high dimensional covariates when the number of the mixture components K is given. We proved consistency in terms of Kullback-Leibler information as well as the ℓ_1 norm of the estimated coefficients. We also showed the convergence rate is at the order of $(\log p/n)^{1/2}$, which is the best convergence rate that we could expect. When K is unknown, we proposed a selection criterion $SCMM$ in finding the number of components K and showed its consistency. In addition, we studied the performance of our proposed methods through simulations and a real data example. There are only a few discussions of variable selection in the literature on FMR models with constant mixing parameters in high dimensions; we are the first to study model selection for FMR models with data dependent mixing parameters. Moreover, we provided a selection criterion with theoretical justifications to determine the number of the mixture components for FMR models.

There are a number of possible future directions to further enrich the current state of the research and applications of the FMR models. For example, we required certain compactness of the parameter space when we select variables in high dimensions, but the ℓ_1 penalty might

automatically ensure that. It would be useful if the compactness condition can be relaxed. When selecting the number of the mixture components, the candidate values for K need to have a known upper bound. It would be interesting to know what happens if the upper bound is incorrectly specified. The current theory also requires a larger penalty in $SCMM$ than it might be necessary; it tends to choose a model with a small number of the mixture components. It would be interesting to investigate the local power of $SCMM$, when one of the mixture components vanishes, in the sense that the coefficients in the component shrink to $zero$. Last but not least, our empirical work shows that when λ_1 and λ_2 are appropriately chosen, the results of model selection are quite good, as predicted by the theory. However, it remains a future research problem how we can choose those parameters data adaptively.

APPENDIX A

Proofs for the Main Results in Chapter 2

We first provide a useful lemma which is proved at the end of this section.

Lemma 8. *Suppose* $\{(P_k, Q_k)\}_{k=1}^{\infty}$ *are* $i.i.d.$ *continuous random variables with finite means. Also suppose that the density of* Q_k *and the conditional density of* $P_k|Q_k$ *are bounded by* C*, then uniformly in* σ_n *between* n^{-1} *and* $\exp(-1)$*, there exists a constant* C^* *such that for sufficiently large* n*,*

$$P(\sup_{a,b\in R}\frac{1}{n}\sum_{k=1}^{n}1(|P_k - aQ_k - b| \leq |\sigma_n \log\sigma_n|) > C^*|\sigma_n\log\sigma_n|) \leq Cn^{-2}.$$

A.1 Proof of Theorem 1

We note that **S1** and **S2** play the role of **Lemma 1** in Chen et al. (2008). With these two properties, it then follows from the Borel-Cantelli Lemma that as $n \to \infty$ and almost surely,
1. *for each given* σ *between* n^{-1} *and* $\exp(-1)$*,*

$$\sup_{\boldsymbol{\beta}\in R^{q_1}} n^{-1}\sum_{i=1}^{n}1(|Y_i - \boldsymbol{Z}_i^T\boldsymbol{\beta}| \leq |\sigma\log\sigma|) \leq C|\sigma\log\sigma|,$$

2. *uniformly for* σ *between* 0 *and* n^{-1}*,*

$$\sup_{\boldsymbol{\beta}\in R^{q_1}} n^{-1}\sum_{i=1}^{n}1(|Y_i - \boldsymbol{Z}_i^T\boldsymbol{\beta}| \leq |\sigma\log\sigma|) \leq 4(\log n)^2/n.$$

These almost sure results are stated for a given σ. However, following the arguments in **Lemma 2** of Chen et.al (2008), we have a stronger result as follows.
Except for a zero-probability event not depending on σ, we have for all large enough n:
1. *for σ between n^{-1} and* $\exp(-1)$, $\sup_{\boldsymbol{\beta}\in R^{q_1}} n^{-1}\sum_{i=1}^{n} 1(|Y_i - \boldsymbol{Z}_i^T\boldsymbol{\beta}| \le |\sigma\log\sigma|) \le C|\sigma\log\sigma|$,
2. *for σ between 0 and n^{-1}*, $\sup_{\boldsymbol{\beta}\in R^{q_1}} n^{-1}\sum_{i=1}^{n} 1(|Y_i - \boldsymbol{Z}_i^T\boldsymbol{\beta}| \le |\sigma\log\sigma|) \le 4(\log n)^2/n$.

We partition the parameter space with respect to σ as in Chen et.al. (2008). Let $\Gamma_1 = \{\boldsymbol{\Theta} : \sigma_1 \le \sigma_2 \le \epsilon_0\}$, $\Gamma_2 = \{\boldsymbol{\Theta} : \sigma_1 \le \tau_0, \sigma_2 \ge \epsilon_0\}$, $\Gamma_3 = \Gamma - (\Gamma_1 \bigcup \Gamma_2)$, where ϵ_0, τ_0 and Γ are specified in Chen et.al. (2008). Note that $\boldsymbol{Z}_i^T\boldsymbol{\beta}$ in our setting plays the same role as θ in Chen et.al. (2008), where the model has no covariates. Hence, with the above almost surely results and **Theorem 1** and **Theorem 2** of Chen et.al.(2008), we have as $n \to \infty$ and almost surely, the penalized maximum likelihood estimators of our model will be attained in Γ_3. Note that σ is bounded away from $zero$ in Γ_3, standard techniques of proving the consistency of the maximum likelihood estimators lead to the consistency of our proposed penalized maximum likelihood estimators.

Next, we show **S1** and **S2**. Since the proof of **S2** is essentially the same as that for **S1**, we only provide the details of the proof of **S1**. For convenience, we allow the constants used in the proofs vary line by line.

Recall that $\boldsymbol{Z} = (1, \boldsymbol{U}, \boldsymbol{V})$, where 1 represents the intercept in the model and $\boldsymbol{U}$ consists of only discrete variables with a finite sample space and $\boldsymbol{V}$ consists of only continuous variables. We prove **S1** for the following three cases.
Case 1: If $\boldsymbol{Z}$ only has three dimensions, that is, $\boldsymbol{Z} = (1, U, V)$. Further, we assume $U \sim Ber(1/2)$.
Case 2: If $\boldsymbol{Z} = (1, \boldsymbol{U}, V)$, where $\boldsymbol{U}$ is a random vector taking any finite values and V is one dimensional continuous variable.
Case 3: If $\boldsymbol{Z} = (1, \boldsymbol{U}, \boldsymbol{V})$, where $\boldsymbol{U}$ is a random vector taking finite values and $\boldsymbol{V}$ is a vector of continuous random variables.
From **Case 1** to **Case 3**, we will prove **S1** from the simplest case to the most general situation. Then we complete the proof of **Theorem 1**.
Next we provide the detailed proof under Cases 1-3.

Proof for Case 1: We prove **S1** when $\boldsymbol{Z}$ only has three dimensions, that is, $\boldsymbol{Z} = (1, U, V)$.

Further, we assume $U \sim Ber(1/2)$.

Let $\bar{U}_n = n^{-1}\sum_{i=1}^n U_i$ and let $f_X(x)\ and\ f_{X|Y}(x|y)$ denote the density of X and the conditional density of $X|Y$, respectively. Then, for any given $\sigma_n \in (n^{-1},\ \exp(-1))$, let

$$\begin{gathered}\epsilon_n = \{n^{-1}(8\log n)\}^{1/2},\\ I = P\left(\sup_{\boldsymbol{\beta}\in R^3} W_n(\boldsymbol{\beta}) > C|\sigma_n \log\sigma_n|\,|\,|\bar{U}_n - 1/2| \le \epsilon_n\right),\\ II = P\left(|\bar{U}_n - 1/2| > \epsilon_n\right).\end{gathered}$$

We have,

$$\begin{gathered}P\left(A_n(C)\right) = P\left(\sup_{\boldsymbol{\beta}\in R^3} W_n(\boldsymbol{\beta}) > C|\sigma_n\log\sigma_n|\right)\\ \le P\left(\sup_{\boldsymbol{\beta}\in R^3} W_n(\boldsymbol{\beta}) > C|\sigma_n\log\sigma_n|\,|\,|\bar{U}_n - \tfrac{1}{2}| \le \epsilon_n\right) + P\left(|\bar{U}_n - \tfrac{1}{2}| > \epsilon_n\right)\\ = I + II.\end{gathered} \tag{A.1}$$

We verify the following two claims:

CL1 $II \le Cn^{-2}$;

CL2 $I \le Cn^{-2}$.

Proof of CL1: By Bernstein's inequality, for sufficient large n,

$$II = 2P\left(\frac{\sum_{i=1}^n U_i}{n} - \frac{1}{2} > \epsilon_n\right) \le \exp\left\{-\frac{\frac{1}{2}n^2\epsilon_n^2}{n + \frac{1}{3}n\epsilon_n}\right\} \le Cn^{-2}.$$

Proof of CL2: Note that

$$\begin{gathered}I = P\left(\sup_{\boldsymbol{\beta}\in R^3} W_n(\boldsymbol{\beta}) > C|\sigma_n\log\sigma_n|\,|\,|\bar{U}_n - \tfrac{1}{2}| \le \epsilon_n\right)\\ = \sum_{u_1,\cdots,u_n} P\left(\sup_{\boldsymbol{\beta}\in R^3} W_n(\boldsymbol{\beta}) > C|\sigma_n\log\sigma_n|\,|\,U_1 = u_1,\cdots,U_n = u_n, |\bar{U}_n - \tfrac{1}{2}| \le \epsilon_n\right)\\ \times f_{(u_1,\cdots,u_n|\bar{U}_n)}(u_1,\cdots,u_n).\end{gathered}$$

For any $u_1,\cdots,u_n$ such that $\bar{U}_n = n^{-1}\sum_{i=1}^n u_i \in [2^{-1}-\epsilon_n, 2^{-1}+\epsilon_n]$, let $\boldsymbol{U} = (U_1,\cdots,U_n)$, $\boldsymbol{u} = (u_1,\cdots,u_n)$, and $\{i_1,\cdots,i_{n\bar{U}_n}\}$ are indices for $\boldsymbol{u} = 1$, and $\{j_1,\cdots,j_{n-n\bar{U}_n}\}$ are indices for $\boldsymbol{u} = 0$. Also let $\boldsymbol{U}_{i_k} = (U_{i_1},\cdots,U_{i_{n\bar{U}_n}})$, $\boldsymbol{U}_{j_k} = (U_{j_1},\cdots,U_{j_{n-n\bar{U}_n}})$ and let the variables

(P_k, Q_k) and (P_k', Q_k') be specified with the following distributions:

$$\{(P_k', Q_k')\}_{k=1}^{n\bar{U}_n} \stackrel{D}{=} \{(Y_{i_k}, V_{i_k})\}_{k=1}^{n\bar{U}_n} | \boldsymbol{U}_{i_k} = \mathbf{1}$$

and

$$\{(P_k, Q_k)\}_{k=1}^{n-n\bar{U}_n} \stackrel{D}{=} \{(Y_{j_k}, V_{j_k})\}_{k=1}^{n-n\bar{U}_n} | \boldsymbol{U}_{j_k} = \mathbf{0}.$$

By the independence of $\{\boldsymbol{Z}_k\}_{k=1}^n = \{(1, U_k, V_k)\}_{k=1}^n$, we have

$$\begin{aligned}
&P\left(\sup_{\boldsymbol{\beta}\in R^3} W_n(\boldsymbol{\beta}) > C|\sigma_n \log \sigma_n| \,|\, U_1 = u_1, \cdots, U_n = u_n, |\bar{U}_n - \tfrac{1}{2}| \le \epsilon_n\right) \\
&= P\left(\sup_{\boldsymbol{\beta}\in R^3} W_n(\boldsymbol{\beta}) > C|\sigma_n \log \sigma_n| \,|\, U_1 = u_1, \cdots, U_n = u_n\right) \\
&= P\big(\sup_{\boldsymbol{\beta}\in R^3} \{\tfrac{1}{n}\sum_{k=1}^{n\bar{U}_n} 1(|Y_{i_k} - \beta_1 - \beta_2 - \beta_3 V_{i_k}| \le |\sigma_n \log \sigma_n|) \\
&+ \tfrac{1}{n}\sum_{k=1}^{n-n\bar{U}_n} 1(|Y_{j_k} - \beta_1 - \beta_3 V_{j_k}| \le |\sigma_n \log \sigma_n|)\} > C|\sigma_n \log \sigma_n| \,|\, \boldsymbol{U} = \boldsymbol{u}\big) \\
&\le P\left(\sup_{\boldsymbol{\beta}\in R^3} \tfrac{1}{n}\sum_{k=1}^{n\bar{U}_n} 1(|Y_{i_k} - \beta_1 - \beta_2 - \beta_3 V_{i_k}| \le |\sigma_n \log \sigma_n|) > (C/2)|\sigma_n \log \sigma_n| \,|\, \boldsymbol{U}_{i_k} = \mathbf{1}\right) \\
&+ P\left(\sup_{\boldsymbol{\beta}\in R^3} \tfrac{1}{n}\sum_{k=1}^{n-n\bar{U}_n} 1(|Y_{j_k} - \beta_1 - \beta_3 V_{j_k}| \le |\sigma_n \log \sigma_n|) > (C/2)|\sigma_n \log \sigma_n| \,|\, \boldsymbol{U}_{j_k} = \mathbf{0}\right) \\
&\le P\left(\sup_{a,b\in R} \tfrac{1}{n\bar{U}_n}\sum_{k=1}^{n\bar{U}_n} 1(|(P_k' - aQ_k' - b| \le |\sigma_n \log \sigma_n|) > (C/2)|\sigma_n \log \sigma_n|\right) \\
&+ P\left(\sup_{a,b\in R} \tfrac{1}{n-n\bar{U}_n}\sum_{k=1}^{n-n\bar{U}_n} 1(|P_k - aQ_k - b| \le |\sigma_n \log \sigma_n|) > (C/2)|\sigma_n \log \sigma_n|\right).
\end{aligned}$$

Since $\{Y_i, \boldsymbol{Z}_i, \boldsymbol{X}_i\}_{i=1}^n$ are $i.i.d.$, $\{(P_k, Q_k)\}_{k=1}^{n-n\bar{U}_n}$ are $i.i.d.$ and so are $\{(P_k', Q_k')\}_{k=1}^{n\bar{U}_n}$. We claim the following two properties under both the null hypothesis and the alternative hypothesis.

CL3 (P_k, Q_k) and (P_k', Q_k') have finite means;

CL4 The densities of $P_k,\ P_k'$ and the conditional densities of $P_k|Q_k,\ P_k'|Q_k'$ are bounded.

Then, by the choice of $\bar{U}_n$, $n\bar{U}_n = O(n/2)$ and $n - n\bar{U}_n = O(n/2)$ almost surely. Taken **CL3** and **CL4** and **Lemma 8** together, we conclude that there exist constant C' such that $I \le C'n^{-2}$ for sufficiently large n. The proof for **CL1** and **CL2** is then complete.

Proof of CL3 and CL4: Recall

$$Y|\{(U, V), \boldsymbol{X}\} \sim \pi(\boldsymbol{X}^T\boldsymbol{\gamma})N(\boldsymbol{Z}^T(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2), \sigma_1^2) + (1 - \pi(\boldsymbol{X}^T\boldsymbol{\gamma}))N(\boldsymbol{Z}^T\boldsymbol{\beta}_1, \sigma_2^2).$$

Note that the null model is just a special case of the above in that $\boldsymbol{\beta}_2 = 0$ and $\sigma_1 = \sigma_2$. By the definitions of (P_k, Q_k) and (P_k', Q_k'), for any $\boldsymbol{\beta}^T = (\boldsymbol{\gamma}^T, \boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \sigma_1, \sigma_2)$, if suffices to show the

following two statements:

S(i) $E\left(|Y|\,|\,U\right) < \infty,\ E\left(|V|\,|\,U\right) < \infty$;

S(ii) the conditional densities of $V|U$ and $Y|U,V$ are bounded.

The statement **S(ii)** is obvious, since $Y|V,U,\boldsymbol{X}$ follows the logistic mixture of normals, its density is uniformly bounded by $\{\sqrt{2\pi}\min\{\sigma_1,\sigma_2\}\}^{-1}$, where $\sigma_1,\ \sigma_2$ are the true parameters in the model. Therefore, the conditional density of $Y|V,U$ is bounded, and by Condition C4, the conditional density of $V|U$ is bounded. For **S(i)**, by Condition C5, $E(|V|\,|\,U) < \infty$, again because $Y|V,U,\boldsymbol{X}$ follows logistic mixture of normals,

$$\begin{aligned} E\left(|Y|\,|\,U\right) &= E\left\{E\left(|Y|\,|\,U,V,\boldsymbol{X}\right)|U\right\} \\ &= E\left\{\left(E\left(\pi(\boldsymbol{X}^T\boldsymbol{\gamma})|Y_1| + (1-\pi(\boldsymbol{X}^T\boldsymbol{\gamma}))|Y_2|\,|\,U,V,\boldsymbol{X}\right)\right)|U\right\} \\ &= E\left\{\left(\pi(\boldsymbol{X}^T\boldsymbol{\gamma})E(|Y_1|\,|\,U,V,\boldsymbol{X}) + (1-\pi(\boldsymbol{X}^T\boldsymbol{\gamma}))E(|Y_2|\,|\,U,V,\boldsymbol{X})\right)|U\right\} \end{aligned}$$

where $Y_1|(U,V,\boldsymbol{X}) \sim N(\boldsymbol{Z}^T(\boldsymbol{\beta}_1+\boldsymbol{\beta}_2),\sigma_1^2)$, $Y_2|(U,V,\boldsymbol{X}) \sim N(\boldsymbol{Z}^T\boldsymbol{\beta}_1,\sigma_2^2)$, and $\boldsymbol{Z} = (1,U,V)$. Note that

$$\begin{aligned} &E(|Y_1|\,|\,U,V,\boldsymbol{X}) \\ &\leq \sigma_1 E\left(|\tfrac{Y_1-\boldsymbol{Z}^T\boldsymbol{\beta}_1-\boldsymbol{Z}^T\boldsymbol{\beta}_2}{\sigma_1}|\,|\,U,V,\boldsymbol{X}\right) + (1+|V|+|U|)||\boldsymbol{\beta}_1+\boldsymbol{\beta}_2||_\infty \\ &= \tfrac{2}{\sqrt{2\pi}}\sigma_1 + (1+|V|+|U|)||\boldsymbol{\beta}_1+\boldsymbol{\beta}_2||_\infty, \end{aligned}$$

where $||\cdot||_\infty$ is the supreme norm and the last equation is due to the fact that $E|Z| = 2(2\pi)^{-1/2}$ if $Z \sim N(0,1)$. Similarly,

$$E(|Y_2|\,|\,U,V,\boldsymbol{X}) \leq \frac{2}{\sqrt{2\pi}}\sigma_2 + (1+|V|+|U|)||\boldsymbol{\beta}_1||_\infty.$$

Therefore,

$$E(|Y|\,|\,U) \leq \frac{2}{\sqrt{2\pi}}\max\{\sigma_1,\sigma_2\} + \max\{||\boldsymbol{\beta}_1||_\infty, ||\boldsymbol{\beta}_1+\boldsymbol{\beta}_2||_\infty\}E(1+|V|+|U|\,|\,U) < \infty,$$

where the last inequality is due to Condition C5.

We have now verified properties **CL3** and **CL4** for any $\boldsymbol{\beta}^T = (\boldsymbol{\gamma}^T,\boldsymbol{\beta}_1^T,\boldsymbol{\beta}_2^T,\sigma_1,\sigma_2)$, under both the null and the alternative hypotheses. By the results from **CL1**, **CL2** and Equation (A.1),

we finished the proof of **Case 1**.

Proof for Case 2: We prove **S1** with $\boldsymbol{Z} = (1, \boldsymbol{U}, V)$, where $\boldsymbol{U}$ is a random vector taking any finite values and V is one dimensional continuous variable.

Let $P(\boldsymbol{U} = \boldsymbol{u}^t) = p_t > 0, t = 1, 2, \cdots, r$, and $\sum_{t=1}^{r} p_t = 1$. Also let $\bar{U}_n^t = n^{-1} \sum_{i=1}^{n} 1(\boldsymbol{U}_i = \boldsymbol{u}^t)$, $t = 1, 2, \cdots, r$. As in the earlier proof, we set $\epsilon_{nt} = \{n^{-1}(8 \log n)\}^{1/2}$, and we bound $P(A_n(C))$ by

$$P\left(A_n(C) | \bar{U}_n^t \in [p_t - \epsilon_{nt}, p_t + \epsilon_{nt}], t = 1, \cdots, r\right) + \sum_{t=1}^{r} P\left(|\bar{U}_n^t - p_t| > \epsilon_{nt}\right) \stackrel{\triangle}{=} I + II.$$

By Bernstein's inequality, we know $II < Cn^{-2}$. For part I, we use arguments conditional on $\boldsymbol{U}_i = \boldsymbol{u}_i$, $i = 1, 2, \cdots, n$, such that the values of $\boldsymbol{u}_i$ satisfy $\bar{U}_n^t \in [p_t - \epsilon_{nt}, p_t + \epsilon_{nt}], t = 1, \cdots, r$. We then group the $\boldsymbol{U}_i = \boldsymbol{u}_i$ which have the same value of $\boldsymbol{u}^t$. Note that the number of the items in each group is of the order of $O(p_t n)$, and by the independence of the vectors of $\mathbf{Z}_i$, we can directly apply **Lemma 8** and get the desired results.

Proof for Case 3: We prove **S1** for general $\boldsymbol{Z} = (1, \boldsymbol{U}, \boldsymbol{V})$, where $\boldsymbol{U}$ is a random vector taking finite values and $\boldsymbol{V}$ is a vector of continuous random variables.

We bound $P(A_n(C))$ by conditioning on the possible values of $\boldsymbol{U}$ as we did previously, then it suffices to show

$$P\left(\sup_{b \in R, \rho \in R^+, ||\alpha|| = 1} \frac{1}{n} \sum_{k=1}^{n} 1(|P_k - \rho \alpha^T \boldsymbol{Q}_k - b| \leq |\sigma_n \log \sigma_n|) > C^* |\sigma_n \log \sigma_n|\right) \leq Cn^{-2},$$

for some C^* and C and sufficiently large n. However, the set of α with $||\alpha|| = 1$ is a compact set, we can prove it by using standard empirical process argument and the same techniques as those used to prove **Lemma 8** in the next subsection.

A.2 Proof of Lemma 8

In this subsection, we prove **Lemma 8** which is needed for the proof of Theorem 1. We allow the constants below to vary line by line. Let

$$
\begin{aligned}
&G_n(a,b,\sigma_n) = n^{-1}\textstyle\sum_{k=1}^{n} 1(|P_k - aQ_k - b| \le |\sigma_n \log \sigma_n|),\\
&L_{n1} = P(\sup_{|a|\le n^2, b\in R} G_n(a,b,\sigma_n) > C^*|\sigma_n \log \sigma_n|),\\
&L_{n2} = P(\sup_{|a|> n^2, b\in R} G_n(a,b,\sigma_n) > C^*|\sigma_n \log \sigma_n|).
\end{aligned}
$$

Note that

$$
\sup_{a,b\in R} G_n(a,b,\sigma_n) = \max\left\{ \sup_{|a|\le n^2, b\in R} G_n(a,b,\sigma_n), \sup_{|a|> n^2, b\in R} G_n(a,b,\sigma_n) \right\}.
$$

Thus we have,

$$
P\left(\sup_{a,b\in R} G_n(a,b,\sigma_n) > C^*|\sigma_n \log \sigma_n| \right) \le L_{n1} + L_{n2}. \tag{A.2}
$$

Step 1: We show $L_{n2} \le Cn^{-2}$.

Note that for any given $\sigma_n \in (n^{-1}, \exp(-1))$, $|\sigma_n \log \sigma_n| \ge n^{-1}\log n$. We have

$$
\begin{aligned}
&G_n(a,b,\sigma_n)\\
&\le \tfrac{1}{n}\textstyle\sum_{k=1}^{n} 1\left(\frac{P_k}{|a|} - \frac{|\sigma_n \log \sigma_n|}{|a|} \le Q_k + \frac{b}{|a|} \le \frac{P_k}{|a|} + \frac{|\sigma_n \log \sigma_n|}{|a|} \right) 1(|P_k| \le (|a|-1)|\sigma_n \log \sigma_n|)\\
&\quad + \tfrac{1}{n}\textstyle\sum_{k=1}^{n} 1(|P_k| > (|a|-1)|\sigma_n \log \sigma_n|).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\sup_{|a|>n^2, b\in R} G_n(a,b,\sigma_n)\\
&\le \sup_{|a|>n^2, b\in R} \left\{ \tfrac{1}{n}\textstyle\sum_{k=1}^{n} 1\left(\frac{P_k}{|a|} - \frac{|\sigma_n \log \sigma_n|}{|a|} \le Q_k + \frac{b}{|a|} \le \frac{P_k}{|a|} + \frac{|\sigma_n \log \sigma_n|}{|a|} \right) \right.\\
&\left. \times 1(|P_k| \le (|a|-1)|\sigma_n \log \sigma_n|) \right\} + \sup_{|a|>n^2} \left\{ \tfrac{1}{n}\textstyle\sum_{k=1}^{n} 1(|P_k| > (|a|-1)|\sigma_n \log \sigma_n|) \right\}\\
&\le \sup_{\theta\in R} \tfrac{1}{n} \left\{ \textstyle\sum_{k=1}^{n} 1(-|\sigma_n \log \sigma_n| \le Q_k - \theta \le |\sigma_n \log \sigma_n|) \right\} + \tfrac{1}{n}\textstyle\sum_{k=1}^{n} 1(|P_k| > n).
\end{aligned}
$$

Let

$$
\begin{aligned}
&L_{n21} = P\left(\sup_{\theta\in R} \{ n^{-1}\textstyle\sum_{k=1}^{n} 1(-|\sigma_n \log \sigma_n| \le Q_k - \theta \le |\sigma_n \log \sigma_n|) \} > (C^*/2)|\sigma_n \log \sigma_n| \right),\\
&\text{and } L_{n22} = P\left(n^{-1}\textstyle\sum_{k=1}^{n} 1(|P_k| > n) > (C^*/2)|\sigma_n \log \sigma_n| \right).
\end{aligned}
$$

Then,

$$L_{n2} = P\left(\sup_{|a|>n^2, b\in R} G_n(a,b,\sigma_n) > C^*|\sigma_n \log \sigma_n|\right) \leq L_{n21} + L_{n22}. \tag{A.3}$$

Step 1-1: We show $L_{n21} \leq Cn^{-2}$.

Observe that in L_{n21},

$$\begin{aligned} &n^{-1}\textstyle\sum_{k=1}^n 1(-|\sigma_n \log \sigma_n| \leq Q_k - \theta \leq |\sigma_n \log \sigma_n|) \\ &= F_n(\theta + |\sigma_n \log \sigma_n|) - F_n(\theta - |\sigma_n \log \sigma_n|), \end{aligned}$$

where F_n is the empirical distribution for Q. Since the density of Q is bounded, a direct application of **Lemma 1** of Chen et al. (2008) yields $L_{n21} \leq Cn^{-2}$.

Step 1-2: We show $L_{n22} \leq Cn^{-2}$.

Note that $E\{1(|P_k| > n)\} \leq n^{-1}E(|P_k|) \leq n^{-1}\log n \leq |\sigma_n \log \sigma_n|$, for sufficiently large n. Then, by Bernstein's inequality, we have

$$\begin{aligned} L_{n22} &\leq P\left(\textstyle\sum_{k=1}^n (1(|P_k| > n) - E(1(|P_k| > n))) > \tilde{C}n|\sigma_n \log \sigma_n|\right) \\ &\leq \exp\left\{-\tfrac{(\tilde{C}n)^2|\sigma_n \log \sigma_n|^2}{2n|\sigma_n \log \sigma_n| + 2\tilde{C}n|\sigma_n \log \sigma_n|}\right\} \leq Cn^{-2}, \end{aligned}$$

where $\tilde{C} = C^*/2 - 1$.

By **Step 1-1**, **Step 1-2** and Equation (B.25), we have

$$L_{n2} \leq L_{n21} + L_{n22} \leq Cn^{-2}, \tag{A.4}$$

which completes the proof of **Step 1**.

Step 2: We show $L_{n1} \leq Cn^{-2}$.

Let $\delta_n = n^{-1}|\sigma_n \log \sigma_n| \geq n^{-2}(\log n)$. Divide $|a| \leq n^2$ into the union of k_n subsets $\{\Omega_{nj}\}_{j=1}^{k_n}$, such that, the distance between any two points in each subset is no greater than δ_n. It is clear that we can achieve this with $k_n \leq (\log n)^{-1}2n^4 \leq O(n^4)$. Let $U_k(a,b,\sigma_n) =$

$1(|P_k - aQ_k - b| \leq |\sigma_n \log \sigma_n|)$, then

$$\begin{aligned}
&\sup_{|a|\leq n^2, b\in R} G_n(a,b,\sigma_n) \\
&= \max_{1\leq j\leq k_n} \left[\sup_{a\in\Omega_{nj}, b\in R}\{G_n(a,b,\sigma_n)\}\right] \\
&\leq \max_{1\leq j\leq k_n} \left[\sup_{b\in R} G_n(a_j,b,\sigma_n) + \sup_{|a-a_j|\leq\delta_n, b\in R}\{|G_n(a,b,\sigma_n) - G_n(a_j,b,\sigma_n)|\}\right] \\
&\leq \max_{1\leq j\leq k_n} \left[\sup_{b\in R} G_n(a_j,b,\sigma_n) + \sup_{|a-a_j|\leq\delta_n, b\in R}\left\{\tfrac{1}{n}\sum_{k=1}^n |U_k(a,b,\sigma_n) - U_k(a_j,b,\sigma_n)|\right\}\right],
\end{aligned}$$

where a_j is any fixed point in Ω_{nj}. Let

$$\begin{aligned}
L_{n11} &= k_n \sup_{a\in R} P\left(\sup_{b\in R} G_n(a,b,\sigma_n) > (C^*/2)|\sigma_n \log \sigma_n|\right), \\
L_{n12} &= k_n \sup_{a'\in R} P\left(\sup_{|a-a'|\leq\delta_n, b\in R}(1/n)\sum_{k=1}^n |U_k(a,b,\sigma_n) - U_k(a',b,\sigma_n)| > (C^*/2)|\sigma_n \log \sigma_n|\right).
\end{aligned}$$

Then we have

$$L_{n1} \leq L_{n11} + L_{n12}. \tag{A.5}$$

Step 2-1: We show $L_{n11} \leq Cn^{-2}$.

In L_{n11}, for any $a \in R$, let $R_k^a = P_k - aQ_k$. Since P_k, Q_k are continuous, and R_k^a is continuous and its density $f_{R_k^a}(r) = \int f_{R_k^a|Q_k}(r|q_k) f_{Q_k}(q_k) dq_k = \int f_{P_k|Q_k}(r + aq_k|q_k) f_{Q_k}(q_k) dq_k \leq C$. Therefore,

$$\begin{aligned}
G_n(a,b,\sigma_n) &= \tfrac{1}{n}\sum_{k=1}^n 1(|R_k^a - b| \leq |\sigma_n \log \sigma_n|) \\
&= F_n(b + |\sigma_n \log \sigma_n|) - F_n(b - |\sigma_n \log \sigma_n|),
\end{aligned}$$

where F_n is the empirical distribution for R_k^a, $k = 1, \cdots, n$. Since the density of R_k^a is uniformly bounded over a, a direct application of **Lemma 1** of Chen et al. (2008) yields

$$P\left(\sup_{b\in R}(1/n)\sum_{k=1}^n 1(|R_k^a - b| \leq |\sigma_n \log \sigma_n|) > (C^*/2)|\sigma_n \log \sigma_n|\right) < Cn^{-6},$$

for any $a \in R$ and for some fixed constant C^*. By using the order of k_n, we have for some C^*,

$$L_{n11} \leq C^* n^{-2} \tag{A.6}$$

for sufficiently large n.

Step 2-2: We show $L_{n12} \leq Cn^{-2}$.

For any $a^{'} \in R$, let

$$M_{n1}(a,b,a^{'},\sigma_n) = n^{-1}\sum_{k=1}^{n} 1(P_k - a^{'}Q_k - b \geq -|\sigma_n \log \sigma_n|)1(P_k - aQ_k - b \leq -|\sigma_n \log \sigma_n|),$$
$$M_{n2}(a,b,a^{'},\sigma_n) = n^{-1}\sum_{k=1}^{n} 1(P_k - a^{'}Q_k - b \leq |\sigma_n \log \sigma_n|)1(P_k - aQ_k - b \geq |\sigma_n \log \sigma_n|),$$
$$M_{n3}(a,b,a^{'},\sigma_n) = n^{-1}\sum_{k=1}^{n} 1(P_k - a^{'}Q_k - b \leq -|\sigma_n \log \sigma_n|)1(P_k - aQ_k - b \geq -|\sigma_n \log \sigma_n|),$$
$$M_{n4}(a,b,a^{'},\sigma_n) = n^{-1}\sum_{k=1}^{n} 1(P_k - a^{'}Q_k - b \geq |\sigma_n \log \sigma_n|)1(P_k - aQ_k - b \leq |\sigma_n \log \sigma_n|);$$

$$N_{n1}(b,a^{'},\sigma_n) = n^{-1}\sum_{k=1}^{n} 1(-|\sigma_n \log \sigma_n| + \delta_n|Q_k| \geq P_k - a^{'}Q_k - b \geq -|\sigma_n \log \sigma_n|),$$
$$N_{n2}(b,a^{'},\sigma_n) = n^{-1}\sum_{k=1}^{n} 1(-|\sigma_n \log \sigma_n| - \delta_n|Q_k| \leq P_k - a^{'}Q_k - b \leq -|\sigma_n \log \sigma_n|),$$
$$N_{n3}(b,a^{'},\sigma_n) = n^{-1}\sum_{k=1}^{n} 1(|\sigma_n \log \sigma_n| - \delta_n|Q_k| \leq P_k - a^{'}Q_k - b \leq |\sigma_n \log \sigma_n|),$$
$$N_{n4}(b,a^{'},\sigma_n) = n^{-1}\sum_{k=1}^{n} 1(|\sigma_n \log \sigma_n| + \delta_n|Q_k| \geq P_k - a^{'}Q_k - b \geq |\sigma_n \log \sigma_n|).$$

Then,

$$\begin{gathered}\tfrac{1}{n}\sum_{k=1}^{n} |U_k(a,b,\sigma_n) - U_k(a^{'},b,\sigma_n)| \\ = \tfrac{1}{n}\sum_{k=1}^{n} |U_k(a,b,\sigma_n) - U_k(a^{'},b,\sigma_n)|U_k(a^{'},b,\sigma_n) \\ +\tfrac{1}{n}\sum_{k=1}^{n} |U_k(a,b,\sigma_n) - U_k(a^{'},b,\sigma_n)|(1 - U_k(a^{'},b,\sigma_n)) \\ \leq M_{n1}(a,b,a^{'},\sigma_n) + M_{n2}(a,b,a^{'},\sigma_n) + M_{n3}(a,b,a^{'},\sigma_n) + M_{n4}(a,b,a^{'},\sigma_n).\end{gathered}$$

Note that for any a, such that $|a - a^{'}| \leq \delta_n$,

$$P_k - a^{'}Q_k - b = P_k - aQ_k - b - (a^{'} - a)Q_k \in [P_k - aQ_k - b - \delta_n|Q_k|, P_k - aQ_k - b + \delta_n|Q_k|].$$

Thus, for any a, such that $|a - a^{'}| \leq \delta_n$,

$$\begin{gathered}M_{n1}(a,b,a^{'},\sigma_n) + M_{n2}(a,b,a^{'},\sigma_n) + M_{n3}(a,b,a^{'},\sigma_n) + M_{n4}(a,b,a^{'},\sigma_n) \\ \leq N_{n1}(b,a^{'},\sigma_n) + N_{n2}(b,a^{'},\sigma_n) + N_{n3}(b,a^{'},\sigma_n) + N_{n4}(b,a^{'},\sigma_n).\end{gathered}$$

Therefore, for any $a^{'} \in R$,

$$\begin{gathered}\sup_{|a-a'|\leq\delta_n, b\in R} \tfrac{1}{n}\sum_{k=1}^{n} |U_k(a,b,\sigma_n) - U_k(a^{'},b,\sigma_n)| \\ \leq \sup_{b\in R} N_{n1}(b,a^{'},\sigma_n) + \sup_{b\in R} N_{n2}(b,a^{'},\sigma_n) + \sup_{b\in R} N_{n3}(b,a^{'},\sigma_n) + \sup_{b\in R} N_{n4}(b,a^{'},\sigma_n).\end{gathered}$$

Let $L_{n12i} = k_n \sup_{a' \in R} P\left(N_{ni}(b, a', \sigma_n) > (C^*/8)|\sigma_n \log \sigma_n|\right)$, $i = 1, 2, 3, 4$. Then

$$L_{n12} \leq \sum_{i=1}^{4} L_{n12i}. \tag{A.7}$$

By the choice of δ_n,

$$\begin{gathered} N_{n1}(b, c, \sigma_n) \\ \leq \sup_{b \in R} \tfrac{1}{n} \textstyle\sum_{k=1}^{n} 1(-|\sigma_n \log \sigma_n| + \delta_n |Q_k| \geq P_k - a' Q_k - b \geq -|\sigma_n \log \sigma_n|) 1(|Q_k| \leq n) \\ + \tfrac{1}{n} \textstyle\sum_{k=1}^{n} 1(|Q_k| > n) \\ \leq \sup_{b \in R} \tfrac{1}{n} \textstyle\sum_{k=1}^{n} 1(0 \geq P_k - a' Q_k - b \geq -|\sigma_n \log \sigma_n|) \\ + \tfrac{1}{n} \textstyle\sum_{k=1}^{n} 1(|Q_k| > n). \end{gathered}$$

Therefore,

$$\begin{gathered} L_{n121} \leq k_n \sup_{a' \in R} P\left(\sup_{b \in R} \tfrac{1}{n} \textstyle\sum_{k=1}^{n} 1(0 \geq P_k - a' Q_k - b \geq -|\sigma_n \log \sigma_n|) > (C^*/16)|\sigma_n \log \sigma_n|\right) \\ + k_n P\left(\tfrac{1}{n} \textstyle\sum_{k=1}^{n} 1(|Q_k| > n) > (C^*/16)|\sigma_n \log \sigma_n|\right). \end{gathered}$$

Analogous to the proof for (A.4), we have $L_{n121} \leq Cn^{-2}$. Similarly, the results hold for $L_{n12i}, i = 2, 3, 4$. Therefore, by (A.7), $L_{n12} \leq \sum_{i=1}^{4} L_{n12i} \leq Cn^{-2}$.

By **Step 2-1**, **Step 2-2** and Equation (A.5), we have

$$L_{n1} \leq \sum_{i=1}^{2} L_{n1i} \leq Cn^{-2}, \tag{A.8}$$

which completes the proof of **Step 2**.

By **Step1**, **Step 2** and Equation (A.2), we complete the proof of **Lemma 8**.

APPENDIX B

Proofs for the Main Results in Chapter 3

B.1 Proof of Theorem 3

The constants below vary line by line and depend only on fixed numbers, e.g. $K, \tilde{K}, s_1, s_2, c_0$. We are going to show $P\left(\mathcal{E}(T)^c\right) \to 0$. Note that

$$P\left(\mathcal{E}(T)^c\right) = P\left(\sup_{\boldsymbol{\theta}\in\tilde{\boldsymbol{\Theta}}} \frac{|\frac{1}{n}\sum_{i=1}^n\{(l_\theta(Y_i,\boldsymbol{x}_i) - E[l_\theta(Y_i,\boldsymbol{x}_i)]) - (l_{\theta_0}(Y_i,\boldsymbol{x}_i) - E[l_{\theta_0}(Y_i,\boldsymbol{x}_i)])\}|}{(||\boldsymbol{\phi}-\boldsymbol{\phi}_0||_1 + ||\boldsymbol{\gamma}-\boldsymbol{\gamma}_0||_1 + ||\boldsymbol{\eta}-\boldsymbol{\eta}_0||_2) \vee \lambda_0} > T\lambda_0\right),$$

and

$$\begin{aligned} &|l_\theta(Y_i,\boldsymbol{x}_i) - l_{\theta_0}(Y_i,\boldsymbol{x}_i)| \\ &= |S^T_{\boldsymbol{\theta}^*}\{(\boldsymbol{x}_i^T\boldsymbol{\phi}_j - \boldsymbol{x}_i^T\boldsymbol{\phi}_{0,j})_{j=1}^K, (\rho_j - \rho_{0,j})_{j=1}^K, (\pi(\boldsymbol{x}_i^T\boldsymbol{\gamma}_j) - \pi(\boldsymbol{x}_i^T\boldsymbol{\gamma}_{0,j}))_{j=1}^K\}|, \end{aligned} \tag{B.1}$$

where S_{θ^*} is defined in Chapter 3.2.1, and $\boldsymbol{\theta}^*$ is between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$. $\boldsymbol{\phi}_j$, $\boldsymbol{\phi}_{0,j}$ are the j'th component of $\boldsymbol{\phi}$ and $\boldsymbol{\phi}_0$, ρ_j, $\rho_{0,j}$ are the j'th component of $\boldsymbol{\rho}$ and $\boldsymbol{\rho}_0$, and $\boldsymbol{\gamma}_j$, $\boldsymbol{\gamma}_{0,j}$ are the j'th component of $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}_0$.

For any postive number $\bar{M}$, define $\tilde{\boldsymbol{\Theta}}_{\bar{M}} = \left\{\boldsymbol{\theta}; ||\log\boldsymbol{\eta}||_\infty \le \tilde{K}, ||\boldsymbol{\theta}-\boldsymbol{\theta}_0||_1 \le \bar{M}\right\}$. By C5, $||S_{\theta^*}||_\infty \le G_1(Y_i) := C(|Y_i|^2 + |Y_i| + C)$. Then for any given $\bar{M} > 0$ and $\boldsymbol{\theta} \in \tilde{\boldsymbol{\Theta}}_{\bar{M}}$,

$$\begin{aligned} &|l_\theta(Y_i,\boldsymbol{x}_i) - l_{\theta_0}(Y_i,\boldsymbol{x}_i)| \\ &\le CG_1(Y_i)\textstyle\sum_{j=1}^K(|\boldsymbol{x}_i^T(\boldsymbol{\phi}_j - \boldsymbol{\phi}_{0,j})| + |\pi(\boldsymbol{x}_i^T\boldsymbol{\gamma}_j) - \pi(\boldsymbol{x}_i^T\boldsymbol{\gamma}_{0,j})| + |\rho_j - \rho_{0,j}|) \\ &\le CG_1(Y_i)\textstyle\sum_{j=1}^K(|\boldsymbol{x}_i^T(\boldsymbol{\phi}_j - \boldsymbol{\phi}_{0,j})| + \sum_{j=1}^K|\boldsymbol{x}_i^T(\boldsymbol{\gamma}_j - \boldsymbol{\gamma}_{0,j})| + |\rho_j - \rho_{0,j}|) \\ &\le CKG_1(Y_i)(||\boldsymbol{x}_i||_\infty \vee 1)\textstyle\sum_{j=1}^K(||\boldsymbol{\phi}_j - \boldsymbol{\phi}_{0,j}||_1 + |\rho_j - \rho_{j,0}| + ||\boldsymbol{\gamma}_j - \boldsymbol{\gamma}_{0,j}||_1) \\ &= CG_1(Y_i)||\boldsymbol{\theta}-\boldsymbol{\theta}_0||_1 \le CG_1(Y_i)\bar{M}, \end{aligned} \tag{B.2}$$

where we use the fact of the boundedness of $\mathcal{X}$ and $|(\pi(\boldsymbol{x}_i^T\boldsymbol{\gamma}_j) - \pi(\boldsymbol{x}_i^T\boldsymbol{\gamma}_{0,j}))| \leq \sum_{j=1}^K |(\boldsymbol{x}_i^T\boldsymbol{\gamma}_j - \boldsymbol{x}_i^T\boldsymbol{\gamma}_{0,j})|$ for all $j = 1, 2, \cdots, K$.

Let $\rho_\theta^c(Y_i, \boldsymbol{x}_i) = l_\theta(Y_i, \boldsymbol{x}_i) - E[l_\theta(Y_i, \boldsymbol{x}_i)]$, then

$$V_n(\boldsymbol{\theta}) - V_n(\boldsymbol{\theta}_0) = \frac{1}{n}\sum_{i=1}^n (\rho_\theta^c(Y_i, \boldsymbol{x}_i) - \rho_{\theta_0}^c(Y_i, \boldsymbol{x}_i)).$$

Also define

$$Y^\epsilon(\theta, \theta_0) = \frac{1}{n}\sum_{i=1}^n (\rho_\theta^c(Y_i, \boldsymbol{x}_i) - \rho_{\theta_0}^c(Y_i, \boldsymbol{x}_i))\epsilon_i,$$

where $\epsilon_1, \cdots, \epsilon_n$ is a Rademacher sequence independent of $Y_1, \cdots, Y_n$.

Then for any given $\bar{M} > 0$, we are going to establish the bounds for

$$E_n := E\left(\left[\sup_{\boldsymbol{\theta}\in\tilde{\boldsymbol{\Theta}}_{\bar{M}}} |Y^\epsilon(\theta, \theta_0)|\right] \mid \mathbf{Y}\right),$$
$$R_n := \sup_{\boldsymbol{\theta}\in\tilde{\boldsymbol{\Theta}}_{\bar{M}}} \sqrt{\tfrac{1}{n}\textstyle\sum_{i=1}^n |\rho_\theta^c(Y_i, \boldsymbol{x}_i) - \rho_{\theta_0}^c(Y_i, \boldsymbol{x}_i)|^2}.$$

By (B.1),

$$\begin{aligned}
&|\rho_\theta^c(Y_i, \boldsymbol{x}_i) - \rho_{\theta_0}^c(Y_i, \boldsymbol{x}_i)| = |(l_\theta(Y_i, \boldsymbol{x}_i) - l_{\theta_0}(Y_i, \boldsymbol{x}_i)) - E(l_\theta(Y_i, \boldsymbol{x}_i) - l_{\theta_0}(Y_i, \boldsymbol{x}_i))| \\
&\leq (||S_{\theta^*}||_\infty + E(||S_{\theta^*}||_\infty))(\textstyle\sum_{j=1}^K (|\boldsymbol{x}_i^T(\boldsymbol{\phi}_j - \boldsymbol{\phi}_{0,j})| + K|\boldsymbol{x}_i^T(\boldsymbol{\gamma}_j - \boldsymbol{\gamma}_{0,j})| + |\rho_j - \rho_{0,j}|)) \\
&\leq CG_1(Y_i)(\textstyle\sum_{j=1}^K (|\boldsymbol{x}_i^T(\boldsymbol{\phi}_j - \boldsymbol{\phi}_{0,j})| + K|\boldsymbol{x}_i^T(\boldsymbol{\gamma}_j - \boldsymbol{\gamma}_{0,j})| + |\rho_j - \rho_{0,j}|)).
\end{aligned}$$

where the last inequality is because $||S_{\theta^*}||_\infty \leq G_1(Y_i) := C(|Y_i|^2 + |Y_i| + C)$ and we note the fact that $E|G_1(Y_i)| \leq C$ when Y_i follows logistic normal mixtures.

Write $\boldsymbol{x}_i = (x_{ir})_{r=1}^p$, $\boldsymbol{\phi}_j = (\phi_{rj})_{r=1}^p$, $\boldsymbol{\phi}_{0,j} = (\phi_{0,rj})_{r=1}^p$, and $\boldsymbol{\gamma}_j = (\gamma_{rj})_{r=1}^p, \boldsymbol{\gamma}_{0,j} = (\gamma_{0,rj})_{r=1}^p$ for $j = 1, 2, \cdots, K$. Then,

$$\begin{aligned}
&|\rho_\theta^c(Y_i, \boldsymbol{x}_i) - \rho_{\theta_0}^c(Y_i, \boldsymbol{x}_i)| \\
&\leq \textstyle\sum_{j=1}^K |\sum_{r=1}^p (\phi_{rj} - \phi_{0,rj}) C x_{ir} G_1(Y_i)| + \sum_{j=1}^K |(\rho_j - \rho_{0,j}) C G_1(Y_i)| \\
&\quad + \textstyle\sum_{j=1}^K |\sum_{r=1}^p (\gamma_{rj} - \gamma_{0,rj}) C K x_{ir} G_1(Y_i)|.
\end{aligned}$$

Let $\Psi_{j,r}(Y_i,i) = CKx_{ir}G_1(Y_i)$ and $\Psi_j(Y_i,i) = CG_1(Y_i)$, then from the above, we have

$$\begin{gathered}|\rho_\theta^c(Y_i,\boldsymbol{x}_i) - \rho_{\theta_0}^c(Y_i,\boldsymbol{x}_i)| \\ \leq \sum_{j=1}^K |\sum_{r=1}^p (\phi_{rj}-\phi_{0,rj})\Psi_{j,r}(Y_i,i)| + \sum_{j=1}^K |\sum_{r=1}^p (\gamma_{rj}-\gamma_{0,rj})\Psi_{j,r}(Y_i,i)| \\ + \sum_{j=1}^K |(\rho_j-\rho_{0,j})\Psi_j(Y_i,i)|.\end{gathered} \tag{B.3}$$

Let $K_n := \max_{j,r}\{||\Psi_{j,r}||_n, ||\Psi_j||_n\} = \max_{j,r}\left\{(n^{-1}\sum_{i=1}^n \Psi_{j,r}^2(Y_i,i))^{1/2}, (n^{-1}\sum_{i=1}^n \Psi_j^2(Y_i,i))^{1/2}\right\}$, by **Theorem 4.1** of van de Geer (2013), for any $\bar{M} > 0$,

$$E_n := E\left(\left[\sup_{\boldsymbol{\theta}\in\tilde{\boldsymbol{\Theta}}_{\bar{M}}} |Y^\epsilon(\theta,\theta_0)|\right] \mid \mathbf{Y}\right) \leq C\bar{M}\sqrt{\frac{\log p}{n}}K_n = C\bar{M}\lambda_0 K_n. \tag{B.4}$$

Next, for any $\bar{M} > 0$, we consider R_n.

Let $a_{jr} = |\phi_{rj}-\phi_{0,rj}|$, $b_{j,r} = |\gamma_{rj}-\gamma_{0,rj}|$, and $c_j = |\rho_j - \rho_{j,0}|$. By (B.3), for any $\boldsymbol{\theta} \in \tilde{\boldsymbol{\Theta}}_{\bar{M}}$,

$$\begin{gathered}\sum_{i=1}^n |\rho_\theta^c(Y_i,\boldsymbol{x}_i) - \rho_{\theta_0}^c(Y_i,\boldsymbol{x}_i)|^2 \\ \leq \sum_{i=1}^n \left(\sum_{j=1}^K\sum_{r=1}^p a_{j,r}|\Psi_{j,r}(Y_i,i)| + \sum_{j=1}^K\sum_{r=1}^p b_{j,r}|\Psi_{j,r}(Y_i,i)| + \sum_{j=1}^K c_j|\Psi_j(Y_i,i)|\right)^2.\end{gathered}$$

For notational convenience, we use uniform sequence $\{u_l\}_{l=1}^{2Kp+K}$ to re-label $a_{j,r}, b_{j,r}, c_j$, and re-label $|\Psi_{j,r}|, |\Psi_j|$ by $\{h_l\}_{l=1}^{2Kp+K}$. Then from the above, we have

$$\begin{gathered}\sum_{i=1}^n |\rho_\theta^c(Y_i,\boldsymbol{x}_i) - \rho_{\theta_0}^c(Y_i,\boldsymbol{x}_i)|^2 \\ \leq \sum_{i=1}^n \left(\sum_{l=1}^{2Kp+K} u_l h_l(Y_i,i)\right)^2 = \sum_{i=1}^n \left(\sum_{l=1}^{2Kp+K} u_l^2 h_l^2(Y_i,i) + 2\sum_{l<l'} u_l u_{l'} h_l(Y_i,i)h_{l'}(Y_i,i)\right) \\ = \sum_{l=1}^{2Kp+K} u_l^2 \sum_{i=1}^n h_l^2(Y_i,i) + \sum_{l<l'} u_l u_{l'} \sum_{i=1}^n 2h_l(Y_i,i)h_{l'}(Y_i,i) \\ \leq \sum_{l=1}^{2Kp+K} u_l^2 \sum_{i=1}^n h_l^2(Y_i,i) + \sum_{l<l'} u_l u_{l'} \sum_{i=1}^n (h_l^2(Y_i,i) + h_{l'}^2(Y_i,i))\end{gathered}$$

Note that $\max_l ||h_l||_n = \max_l (n^{-1}\sum_{i=1}^n h_l^2(Y_i,i))^{1/2} = K_n$, then

$$\begin{gathered}\sum_{l=1}^{2Kp+K} u_l^2 \sum_{i=1}^n h_l^2(Y_i,i) + \sum_{l<l'} u_l u_{l'} \sum_{i=1}^n (h_l^2(Y_i,i) + h_{l'}^2(Y_i,i)) \\ \leq n\sum_{l=1}^{2Kp+K} u_l^2 K_n^2 + 2n\sum_{l<l'} u_l u_{l'} K_n^2 = nK_n^2(\sum_{l=1}^{2Kp+K} u_l)^2 \\ \leq nK_n^2\bar{M}^2,\end{gathered}$$

where the last line is because

$$\sum_{l=1}^{2Kp+K} u_l = \sum_{j,r}(|\phi_{rj} - \phi_{0,rj}| + |\gamma_{\theta,rj} - \gamma_{0,rj}|) + \sum_j |\rho_j - \rho_{j,0}| \leq \bar{M}.$$

Henceforth,

$$R_n = \sup_{\boldsymbol{\theta} \in \tilde{\boldsymbol{\Theta}}_{\bar{M}}} \sqrt{\frac{1}{n}\sum_{i=1}^n |\rho_\theta^c(Y_i, \boldsymbol{x}_i) - \rho_{\theta_0}^c(Y_i, \boldsymbol{x}_i)|^2} \leq \bar{M} K_n. \tag{B.5}$$

Note that the space for $\mathcal{X}$ is bounded, we have $K_n \leq C||G_1||_n = C\left\{(n^{-1}\sum_{i=1}^n G_1^2(Y_i)\right\}^{1/2}$. By (B.4) and (B.5), for any given $\bar{M}$,

$$E_n \leq C\bar{M}\lambda_0||G_1||_n \text{ and } R_n \leq C\bar{M}||G_1||_n. \tag{B.6}$$

With (B.6), for any given $\bar{M} > 0$,we are ready find an upper bound for

$$P\left(\sup_{\boldsymbol{\theta} \in \tilde{\boldsymbol{\Theta}}_{\bar{M}}} |\frac{1}{n}\sum_{i=1}^n (\rho_\theta^c(Y_i, \boldsymbol{x}_i) - \rho_{\theta_0}^c(Y_i, \boldsymbol{x}_i))| > C\lambda_0\bar{M}\right).$$

By Chebyshev's inequality, for any $\boldsymbol{\theta} \in \tilde{\boldsymbol{\Theta}}_{\bar{M}}$ with $\lambda_0 = \{\log p/n\}^{1/2}$,

$$\begin{aligned} &P\left(\tfrac{1}{n}\textstyle\sum_{i=1}^n (\rho_\theta^c(Y_i, \boldsymbol{x}_i) - \rho_{\theta_0}^c(Y_i, \boldsymbol{x}_i)) > \tfrac{C}{2}\lambda_0\bar{M}\right) \\ &\leq (C^2n^2\lambda_0^2\bar{M}^2)^{-1}[\textstyle\sum_{i=1}^n Var\left(l_\theta(Y_i, \boldsymbol{x}_i) - l_{\theta_0}(Y_i, \boldsymbol{x}_i)\right)] \\ &\leq (C^2 n \log p\bar{M}^2)^{-1}\textstyle\sum_{i=1}^n E[(l_\theta(Y_i, \boldsymbol{x}_i) - l_{\theta_0}(Y_i, \boldsymbol{x}_i))^2]. \end{aligned}$$

By (B.2),

$$E[(l_\theta(Y_i, \boldsymbol{x}_i) - l_{\theta_0}(Y_i, \boldsymbol{x}_i))^2] \leq C^2\bar{M}^2 E[G_1^2(Y_i)] \leq C^*\bar{M}^2,$$

where we note that $E[G_1^2(Y_i)] \leq C$ for $i = 1, 2, \cdots, n$.

Therefore, there exists a constant C, such that, for any $\boldsymbol{\theta} \in \tilde{\boldsymbol{\Theta}}_{\bar{M}}$,

$$P\left(\frac{1}{n}\sum_{i=1}^n (\rho_\theta^c(Y_i, \boldsymbol{x}_i) - \rho_{\theta_0}^c(Y_i, \boldsymbol{x}_i)) > \frac{C}{2}\lambda_0\bar{M}\right) \leq \frac{nC^*\bar{M}^2}{C^2 n\log p\bar{M}^2} = \frac{C^*}{C^2\log p} \leq \frac{1}{2}. \tag{B.7}$$

By symmetrization of Pollard (1984),

$$P\left(\sup_{\boldsymbol{\theta}\in\tilde{\boldsymbol{\Theta}}_{\bar{M}}}|\frac{1}{n}\sum_{i=1}^{n}(\rho_{\theta}^{c}(Y_i,\boldsymbol{x}_i)-\rho_{\theta_0}^{c}(Y_i,\boldsymbol{x}_i))|>C\lambda_0\bar{M}\right)\leq 4P\left(\sup_{\boldsymbol{\theta}\in\tilde{\boldsymbol{\Theta}}_{\bar{M}}}|Y^{\epsilon}(\boldsymbol{\theta},\boldsymbol{\theta}_0)|>\frac{C}{4}\lambda_0\bar{M}\right). \tag{B.8}$$

By **Theorem 3** of Massart (2000), we have

$$P\left(\sup_{\boldsymbol{\theta}\in\tilde{\boldsymbol{\Theta}}_{\bar{M}}}|Y^{\epsilon}(\boldsymbol{\theta},\boldsymbol{\theta}_0)|>C\left(E_n+\lambda_0R_n\right)\right)\leq\frac{1}{p}. \tag{B.9}$$

Then,

$$\begin{aligned}&P\left(\sup_{\boldsymbol{\theta}\in\tilde{\boldsymbol{\Theta}}_{\bar{M}}}|Y^{\epsilon}(\boldsymbol{\theta},\boldsymbol{\theta}_0)|>\tfrac{C}{4}\lambda_0\bar{M}\right)\\ \leq\ &P\left(\{\sup_{\boldsymbol{\theta}\in\tilde{\boldsymbol{\Theta}}_{\bar{M}}}|Y^{\epsilon}(\boldsymbol{\theta},\boldsymbol{\theta}_0)|>\tfrac{C}{4}\lambda_0\bar{M}\}\bigcap\{||G_1||_n\leq C^*\}\right)+P\left(||G_1||_n>C^*\right)\end{aligned} \tag{B.10}$$

Note that by (B.6), $||G_1||_n\leq C^*$ implies

$$C(E_n+\lambda_0R_n)\leq C(C\lambda_0\bar{M}||G_1||_n+C\lambda_0\bar{M}||G_1||_n)\leq\tilde{C}\lambda_0\bar{M}. \tag{B.11}$$

Choose C in (B.10) large enough so that $C/4>\tilde{C}$, then from (B.11)

$$\left\{\boldsymbol{y}:\{\sup_{\boldsymbol{\theta}\in\tilde{\boldsymbol{\Theta}}_{\bar{M}}}|Y^{\epsilon}(\boldsymbol{\theta},\boldsymbol{\theta}_0)|>\frac{C}{4}\lambda_0\bar{M}\}\bigcap\{||G_1||_n\leq C^*\}\right\}\subset\left\{\boldsymbol{y}:\sup_{\boldsymbol{\theta}\in\tilde{\boldsymbol{\Theta}}_{\bar{M}}}|Y^{\epsilon}(\boldsymbol{\theta},\boldsymbol{\theta}_0)|>C\left(E_n+\lambda_0R_n\right)\right\}.$$

Therefore by (B.9),

$$\begin{aligned}&P\left(\{\sup_{\boldsymbol{\theta}\in\tilde{\boldsymbol{\Theta}}_{\bar{M}}}|Y^{\epsilon}(\boldsymbol{\theta},\boldsymbol{\theta}_0)|>\tfrac{C}{4}\lambda_0\bar{M}\}\bigcap\{||G_1||_n\leq C^*\}\right)\\ &\leq P\left(\sup_{\boldsymbol{\theta}\in\tilde{\boldsymbol{\Theta}}_{\bar{M}}}|Y^{\epsilon}(\boldsymbol{\theta},\boldsymbol{\theta}_0)|>C\left(E_n+\lambda_0R_n\right)\right)\leq p^{-1}.\end{aligned} \tag{B.12}$$

Also note that,

$$P(||G_1||_n>C^*)=P\left(\frac{1}{n}\sum_{i=1}^{n}(G_1^2(Y_i)-E[G_1^2(Y_i)])>C\right)\leq\frac{\sum_{i=1}^{n}E[G_1^4(Y_i)]}{n^2C^2}\leq\frac{C}{n}, \tag{B.13}$$

where we use the fact that $E[G_1^2(Y_i)]$ and $E[G_1^4(Y_i)]$ are uniformly bounded over $i=1,\cdots,n$.

Then, from (B.12),(B.13) and together with (B.10), we have

$$P\left(\sup_{\boldsymbol{\theta}\in\tilde{\boldsymbol{\Theta}}_{\bar{M}}}|Y^{c}(\boldsymbol{\theta},\boldsymbol{\theta}_0)|>\frac{C}{4}\lambda_0\bar{M}\right)\leq C\left(\frac{1}{p}+\frac{1}{n}\right).$$

Therefore from (B.8), there exists C such that for any given $\bar{M}>0$,

$$P\left(\sup_{\boldsymbol{\theta}\in\tilde{\boldsymbol{\Theta}}_{\bar{M}}}|\frac{1}{n}\sum_{i=1}^{n}(\rho_{\theta}^{c}(Y_i,\boldsymbol{x}_i)-\rho_{\theta_0}^{c}(Y_i,\boldsymbol{x}_i))|>C\lambda_0\bar{M}\right)\leq C\left(\frac{1}{p}+\frac{1}{n}\right). \tag{B.14}$$

Finally, to bound $P(\mathcal{E}(T)^c)$, we invoke the peeling device and choose the constant T in $\mathcal{E}(T)$ at least as large as C$\exp(1)$, where C is from (B.14).

Then, for any given $M>0$, devide $\tilde{\boldsymbol{\Theta}}_M$ to $\tilde{\boldsymbol{\Theta}}_F\bigcup\left\{\tilde{\boldsymbol{\Theta}}_{M_j}\right\}_{j=1,2,\cdots}$, where

$$\left\{\tilde{\boldsymbol{\Theta}}_{M_j}\right\}=\left\{\boldsymbol{\theta};||\log\boldsymbol{\eta}||_\infty\leq\tilde{K},e^{-j}M\leq||\boldsymbol{\theta}-\boldsymbol{\theta}_0||_1\leq e^{1-j}M\right\},$$

and $\left\{\tilde{\boldsymbol{\Theta}}_F\right\}=\left\{\boldsymbol{\theta};||\log\boldsymbol{\eta}||_\infty\leq\tilde{K},||\boldsymbol{\theta}-\boldsymbol{\theta}_0||_1\leq\lambda_0\right\}$. It can be seen that the number of these sets is at most $C\log n$. Then, by (B.14),

$$\begin{aligned}
&P(\mathcal{E}(T)^c)\\
\leq&\sum_j P\left(\sup_{\boldsymbol{\theta}\in\tilde{\boldsymbol{\Theta}}_{M_j}}\frac{|\frac{1}{n}\sum_{i=1}^{n}\{(l_\theta(Y_i,\boldsymbol{x}_i)-E[l_\theta(Y_i,\boldsymbol{x}_i)])-(l_{\theta_0}(Y_i,\boldsymbol{x}_i)-E[l_{\theta_0}(Y_i,\boldsymbol{x}_i)])\}|}{||\boldsymbol{\theta}-\boldsymbol{\theta}_0||_1}>C\exp(1)\lambda_0\right)\\
&+P\left(\sup_{\boldsymbol{\theta}\in\tilde{\boldsymbol{\Theta}}_F}\frac{|\frac{1}{n}\sum_{i=1}^{n}\{(l_\theta(Y_i,\boldsymbol{x}_i)-E[l_\theta(Y_i,\boldsymbol{x}_i)])-(l_{\theta_0}(Y_i,\boldsymbol{x}_i)-E[l_{\theta_0}(Y_i,\boldsymbol{x}_i)])\}|}{\lambda_0}>C\lambda_0\right)\\
\leq&\sum_j P\left(\sup_{\boldsymbol{\theta}\in\tilde{\boldsymbol{\Theta}}_{M_j}}|\frac{1}{n}\sum_{i=1}^{n}(\rho_\theta^c(Y_i,\boldsymbol{x}_i)-\rho_{\theta_0}^c(Y_i,\boldsymbol{x}_i))|>C\lambda_0e^{1-j}M\right)\\
&+P\left(\sup_{\boldsymbol{\theta}\in\tilde{\boldsymbol{\Theta}}_F}|\frac{1}{n}\sum_{i=1}^{n}(\rho_\theta^c(Y_i,\boldsymbol{x}_i)-\rho_{\theta_0}^c(Y_i,\boldsymbol{x}_i))|>C\lambda_0^2\right)\\
\leq&C\log n\left(\frac{1}{p}+\frac{1}{n}\right)\to 0,
\end{aligned}$$

and we complete the proof of **Theorem 1**.

Before proving **Theorem 4**, we list a useful lemma.

Lemma 9. *For all $\boldsymbol{\theta}=(\boldsymbol{\gamma},\boldsymbol{\phi},\boldsymbol{\eta})$, $\tilde{\boldsymbol{\theta}}=(\tilde{\boldsymbol{\gamma}},\tilde{\boldsymbol{\phi}},\tilde{\boldsymbol{\eta}})\in\tilde{\boldsymbol{\Theta}}$, and $\boldsymbol{x}\in\mathcal{X}$, there exists a constant C, such that*

$$\sum_{k=1}^{K}|\boldsymbol{x}^T(\tilde{\boldsymbol{\gamma}}_k-\boldsymbol{\gamma}_k)|^2\leq C\sum_{k=1}^{K}\left\{\pi(\boldsymbol{x}^T\tilde{\boldsymbol{\gamma}}_k)-\pi(\boldsymbol{x}^T\boldsymbol{\gamma}_k)\right\}^2,$$

B.2 Proof of Theorem 4

It can be seen that on $\mathcal{E}(T)$, we have

$$\begin{aligned}&\bar{\varepsilon}(\hat{\theta}|\theta_0) + \lambda_1||\hat{\boldsymbol{\phi}}||_1 + \lambda_2||\hat{\boldsymbol{\gamma}}||_1 \\ \leq T\lambda_0 &\left[(||\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0||_1 + ||\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0||_1 + ||\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0||_2) \vee \lambda_0\right] + \lambda_1||\boldsymbol{\phi}_0||_1 + \lambda_2||\boldsymbol{\gamma}_0||_1.\end{aligned}$$

Following the arguments used for **Theorem 3** of Städler, Bühlmann and van de Geer (2010), we discuss 4 different cases.

Case 1: If $||\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0||_1 + ||\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0||_1 + ||\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0||_2 \leq \lambda_0$, then we have,

$$\bar{\varepsilon}(\hat{\theta}|\theta_0) \leq T\lambda_0^2 + \lambda_1||\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0||_1 + \lambda_2||\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0||_1 \leq (\lambda_1 \vee \lambda_2 + T\lambda_0)\lambda_0.$$

Note that $||\hat{\boldsymbol{\phi}}_{S_1^c}||_1 \leq ||\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0||_1$ and $||\hat{\boldsymbol{\gamma}}_{S_2^c}||_1 \leq ||\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0||_1$, we have,

$$\begin{aligned}&\bar{\varepsilon}(\hat{\theta}|\theta_0) + 2(\lambda_1 - T\lambda_0)||\hat{\boldsymbol{\phi}}_{S_1^c}||_1 + 2(\lambda_2 - T\lambda_0)||\hat{\boldsymbol{\gamma}}_{S_2^c}||_1 \\ &\leq (\lambda_1 \vee \lambda_2 + T\lambda_0)\lambda_0 + 2(\lambda_1 \vee \lambda_2 - T\lambda_0)(||\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0||_1 + ||\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0||_1) \\ &\leq (\lambda \vee \lambda_2 + T\lambda_0)\lambda_0 + 2(\lambda_1 \vee \lambda_2 - T\lambda_0)\lambda_0 \\ &\leq 3(\lambda \vee \lambda_2 + T\lambda_0)\lambda_0,\end{aligned}$$

and $||\hat{\boldsymbol{\phi}}_{S_1} - (\boldsymbol{\phi}_0)_{S_1}||_1 + ||\hat{\boldsymbol{\gamma}}_{S_2} - (\boldsymbol{\gamma}_0)_{S_2}||_1 \leq ||\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0||_1 + ||\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0||_1 \leq \lambda_0$.

Case 2: When $||\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0||_1 + ||\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0||_1 + ||\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0||_2 > \lambda_0$, on $\mathcal{E}(T)$ we have,

$$\begin{aligned}&\bar{\varepsilon}(\hat{\theta}|\theta_0) + \lambda_1||\hat{\boldsymbol{\phi}}||_1 + \lambda_2||\hat{\boldsymbol{\gamma}}||_1 \\ \leq T\lambda_0&(||\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0||_1 + ||\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0||_1 + ||\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0||_2) + \lambda_1||\boldsymbol{\phi}_0||_1 + \lambda_2||\boldsymbol{\gamma}_0||_1.\end{aligned}$$

Note that

$$\begin{aligned}||\hat{\boldsymbol{\phi}}||_1 &= ||\hat{\boldsymbol{\phi}}_{S_1}||_1 + ||\hat{\boldsymbol{\phi}}_{S_1^c}||_1, \\ ||\hat{\boldsymbol{\gamma}}||_1 &= ||\hat{\boldsymbol{\gamma}}_{S_2}||_1 + ||\hat{\boldsymbol{\gamma}}_{S_2^c}||_1, \\ ||\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0||_1 &= ||(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)_{S_1}||_1 + ||\hat{\boldsymbol{\phi}}_{S_1^c}||_1, \\ ||\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0||_1 &= ||(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)_{S_2}||_1 + ||\hat{\boldsymbol{\gamma}}_{S_2^c}||_1.\end{aligned}$$

On $\mathcal{E}(T)$, we have

$$\begin{gathered}\bar{\varepsilon}(\hat{\theta}|\theta_0) + (\lambda_1 - T\lambda_0)||\hat{\boldsymbol{\phi}}_{S_1^c}||_1 + (\lambda_2 - T\lambda_0)||\hat{\boldsymbol{\gamma}}_{S_2^c}||_1 \\ \leq T\lambda_0(||(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)_{S_1}||_1 + ||(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)_{S_2}||_1 + ||\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0||_2) + \lambda_1||(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)_{S_1}||_1 + \lambda_2||(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)_{S_2}||_1 \\ \leq (\lambda_1 + T\lambda_0)||(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)_{S_1}||_1 + (\lambda_2 + T\lambda_0)||(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)_{S_2}||_1 + T\lambda_0||\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0||_2.\end{gathered}$$

We study 3 sub-cases of **Case 2.**

Case 2.1: If $(\lambda_1 + T\lambda_0)||(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)_{S_1}||_1 + (\lambda_2 + T\lambda_0)||(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)_{S_2}||_1 \leq T\lambda_0||\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0||_2$, then on $\mathcal{E}(T)$,

$$\bar{\varepsilon}(\hat{\theta}|\theta_0) + (\lambda_1 - T\lambda_0)||\hat{\boldsymbol{\phi}}_{S_1^c}||_1 + (\lambda_2 - T\lambda_0)||\hat{\boldsymbol{\gamma}}_{S_2^c}||_1 \leq 2T\lambda_0||\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0||_2.$$

By (B.21), $||\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0||_2^2 \leq c_0^2\varepsilon(\hat{\theta}(\boldsymbol{x}_i)|\theta_0(\boldsymbol{x}_i))$, for $i = 1, \cdots, n$. Then, $||\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0||_2^2 \leq c_0^2\bar{\varepsilon}(\hat{\theta}|\theta_0)$. Therefore, by Cauchy-Schwarz inequality,

$$\begin{gathered}\bar{\varepsilon}(\hat{\theta}|\theta_0) + (\lambda_1 - T\lambda_0)||\hat{\boldsymbol{\phi}}_{S_1^c}||_1 + (\lambda_2 - T\lambda_0)||\hat{\boldsymbol{\gamma}}_{S_2^c}||_1 \\ \leq 2T\lambda_0 c_0\sqrt{\bar{\varepsilon}(\hat{\theta}|\theta_0)} \\ \leq 2T^2\lambda_0^2 c_0^2 + \tfrac{1}{2}\bar{\varepsilon}(\hat{\theta}|\theta_0).\end{gathered}$$

Then, we can conclude that

$$\bar{\varepsilon}(\hat{\theta}|\theta_0) + 2(\lambda_1 - T\lambda_0)||\hat{\boldsymbol{\phi}}_{S_1^c}||_1 + 2(\lambda_2 - T\lambda_0)||\hat{\boldsymbol{\gamma}}_{S_2^c}||_1 \leq 4T^2\lambda_0^2 c_0^2. \tag{B.15}$$

Also note $(\lambda_1 \wedge \lambda_2 + T\lambda_0)(||(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)_{S_1}||_1 + ||(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)_{S_2}||_1) \leq T\lambda_0||\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0||_2$, by (B.21) and (B.15), we have

$$||(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)_{S_1}||_1 + ||(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)_{S_2}||_1 \leq ||\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0||_2 \leq c_0\sqrt{\bar{\varepsilon}(\hat{\theta}|\theta_0)} \leq 2Tc_0^2\lambda_0.$$

Case 2.2: If $(\lambda_1 + T\lambda_0)||(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)_{S_1}||_1 + (\lambda_2 + T\lambda_0)||(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)_{S_2}||_1 > T\lambda_0||\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0||_2$, and $(\lambda_1 + T\lambda_0)||(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)_{S_1}||_1 \geq (\lambda_2 + T\lambda_0)||(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)_{S_2}||_1$, then on $\mathcal{E}(T)$,

$$\bar{\varepsilon}(\hat{\theta}|\theta_0) + (\lambda_1 - T\lambda_0)||\hat{\boldsymbol{\phi}}_{S_1^c}||_1 + (\lambda_2 - T\lambda_0)||\hat{\boldsymbol{\gamma}}_{S_2^c}||_1 \leq 4(\lambda_1 + T\lambda_0)||(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)_{S_1}||_1.$$

By the choice of λ_1, we have

$$||(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)_{S_1^c}||_1 = ||\hat{\boldsymbol{\phi}}_{S_1^c}||_1 \leq \frac{4(\lambda_1 + T\lambda_0)}{\lambda_1 - T\lambda_0}||(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)_{S_1}||_1 \leq 6||(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)_{S_1}||_1.$$

By C4,

$$\begin{aligned} &||(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)_{S_1}||_2^2 \\ &\leq \kappa^2 \textstyle\sum_{k=1}^K (\hat{\boldsymbol{\phi}}_k - \boldsymbol{\phi}_{0k})^T \Sigma_n (\hat{\boldsymbol{\phi}}_k - \boldsymbol{\phi}_{0k}) \\ &= \kappa^2 \tfrac{1}{n} \textstyle\sum_{i=1}^n \sum_{k=1}^K |\boldsymbol{x}_i^T (\hat{\boldsymbol{\phi}}_k - \boldsymbol{\phi}_{0k})|^2. \end{aligned}$$

By (B.21),

$$\sum_{k=1}^K |\boldsymbol{x}_i^T (\hat{\boldsymbol{\phi}}_k - \boldsymbol{\phi}_{0k})|^2 \leq c_0^2 \varepsilon(\hat{\theta}(\boldsymbol{x}_i)|\theta_0(\boldsymbol{x}_i)), \text{ for } i = 1, \cdots, n.$$

Therefore, $||(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)_{S_1}||_2^2 \leq \kappa^2 c_0^2 \bar{\varepsilon}(\hat{\theta}|\theta_0)$. By Cauchy-Schwarz inequality,

$$\begin{aligned} &\bar{\varepsilon}(\hat{\theta}|\theta_0) + (\lambda_1 - T\lambda_0)||\hat{\boldsymbol{\phi}}_{S_1^c}||_1 + (\lambda_2 - T\lambda_0)||\hat{\boldsymbol{\gamma}}_{S_2^c}||_1 \\ &\leq 4(\lambda_1 + T\lambda_0)\sqrt{s_1}||(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)_{S_1}||_2 \\ &\leq 4(\lambda_1 + T\lambda_0)\kappa c_0 \sqrt{s_1}\sqrt{\bar{\varepsilon}(\hat{\theta}|\theta_0)} \\ &\leq 8(\lambda_1 + T\lambda_0)^2 s_1 \kappa^2 c_0^2 + \tfrac{1}{2}\bar{\varepsilon}(\hat{\theta}|\theta_0). \end{aligned}$$

We then can conclude that

$$\bar{\varepsilon}(\hat{\theta}|\theta_0) + 2(\lambda_1 - T\lambda_0)||\hat{\boldsymbol{\phi}}_{S_1^c}||_1 + 2(\lambda_2 - T\lambda_0)||\hat{\boldsymbol{\gamma}}_{S_2^c}||_1 \leq 16(\lambda_1 + T\lambda_0)^2 s_1 \kappa^2 c_0^2. \tag{B.16}$$

Also note

$$\begin{aligned} &(\lambda_1 \wedge \lambda_2 + T\lambda_0)(||(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)_{S_1}||_1 + ||(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)_{S_2}||_1) \\ &\leq 2(\lambda_1 + T\lambda_0)||(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)_{S_1}||_1 \leq 2(\lambda_1 + T\lambda_0)\kappa c_0 \sqrt{\bar{\varepsilon}(\hat{\theta}|\theta_0)}. \end{aligned}$$

By (B.16), we have

$$(\lambda_1 \wedge \lambda_2 + T\lambda_0)(||(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)_{S_1}||_1 + ||(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)_{S_2}||_1) \leq C(\lambda_1 + T\lambda_0)^2.$$

Case 2.3: If $(\lambda_1 + T\lambda_0)||(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)_{S_1}||_1 + (\lambda_2 + T\lambda_0)||(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)_{S_2}||_1 > T\lambda_0||\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0||_2$, and $(\lambda_1 + T\lambda_0)||(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)_{S_1}||_1 < (\lambda_2 + T\lambda_0)||(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)_{S_2}||_1$, then on $\mathcal{E}(T)$,

$$\bar{\varepsilon}(\hat{\theta}|\theta_0) + (\lambda_1 - T\lambda_0)||\hat{\boldsymbol{\phi}}_{S_1^c}||_1 + (\lambda_2 - T\lambda_0)||\hat{\boldsymbol{\gamma}}_{S_2^c}||_1 \leq 4(\lambda_2 + T\lambda_0)||(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)_{S_2}||_1.$$

Similar to the arguments used in **Case 2.2**,

$$\begin{aligned}
&\bar{\varepsilon}(\hat{\theta}|\theta_0) + (\lambda_1 - T\lambda_0)||\hat{\boldsymbol{\phi}}_{S_1^c}||_1 + (\lambda_2 - T\lambda_0)||\hat{\boldsymbol{\gamma}}_{S_2^c}||_1 \\
&\leq 4(\lambda_2 + T\lambda_0)||(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)_{S_2}||_1 \\
&\leq 4(\lambda_2 + T\lambda_0)\sqrt{s_2}||(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)_{S_2}||_2 \\
&\leq 4(\lambda_2 + T\lambda_0)\kappa\sqrt{s_2}\sqrt{\tfrac{1}{n}\textstyle\sum_{i=1}^n \sum_{k=1}^K |\boldsymbol{x}_i^T(\hat{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}_{0k})|^2}.
\end{aligned}$$

By **Lemma 9** and the argument in **Case2.2**, we continue the above inequality to have,

$$\begin{aligned}
&\bar{\varepsilon}(\hat{\theta}|\theta_0) + (\lambda_1 - T\lambda_0)||\hat{\boldsymbol{\phi}}_{S_1^c}||_1 + (\lambda_2 - T\lambda_0)||\hat{\boldsymbol{\gamma}}_{S_2^c}||_1 \\
&\leq 4(\lambda_2 + T\lambda_0)\kappa\sqrt{s_2}\sqrt{\tfrac{1}{n}\textstyle\sum_{i=1}^n \sum_{k=1}^K |\boldsymbol{x}_i^T(\hat{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}_{0k})|^2} \\
&\leq 4(\lambda_2 + T\lambda_0)\kappa\sqrt{s_2 C}\sqrt{\tfrac{1}{n}\textstyle\sum_{i=1}^n \sum_{k=1}^K \{\pi(\boldsymbol{x}_i^T\hat{\boldsymbol{\gamma}}_k) - \pi(\boldsymbol{x}_i^T\boldsymbol{\gamma}_{0k})\}^2} \\
&\leq 4(\lambda_2 + T\lambda_0)\kappa c_0\sqrt{s_2 C}\sqrt{\bar{\varepsilon}(\hat{\theta}|\theta_0)} \\
&\leq 8(\lambda_2 + T\lambda_0)^2\kappa^2 c_0^2 C s_2 + \tfrac{1}{2}\bar{\varepsilon}(\hat{\theta}|\theta_0).
\end{aligned}$$

Therefore,

$$\bar{\varepsilon}(\hat{\theta}|\theta_0) + 2(\lambda_1 - T\lambda_0)||\hat{\boldsymbol{\phi}}_{S_1^c}||_1 + 2(\lambda_2 - T\lambda_0)||\hat{\boldsymbol{\gamma}}_{S_2^c}||_1 \leq 16(\lambda_2 + T\lambda_0)^2\kappa^2 c_0^2 C s_2.$$

Similar analog in **Case2.2** yields

$$(\lambda_1 \wedge \lambda_2 + T\lambda_0)(||(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)_{S_1}||_1 + ||(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)_{S_2}||_1) \leq C(\lambda_2 + T\lambda_0)^2.$$

Combining the above 4 cases, we finish the proof of **Theorem 4**.

B.3 Proof of Lemma 9

Let $a_k = \boldsymbol{x}^T\tilde{\boldsymbol{\gamma}}_k$, $b_k = \boldsymbol{x}^T\boldsymbol{\gamma}_k$. Note that $\tilde{\boldsymbol{\gamma}}_1 = \boldsymbol{\gamma}_1 = \mathbf{0}_{p\times 1}^T$, we have $a_1 = b_1 \equiv 0$. Since $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{\eta})$, $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\phi}}, \tilde{\boldsymbol{\eta}}) \in \tilde{\boldsymbol{\Theta}}$, and $\boldsymbol{x} \in \mathcal{X}$, and both $\tilde{\boldsymbol{\Theta}}$ and $\mathcal{X}$ are bounded, we have $|a_k| \leq Q$, $|b_k| \leq Q$, for $k = 2, 3, \cdots, K$, for some finite number Q. What we need to show is then

$$\sum_{k=2}^K (a_k - b_k)^2 \leq C\sum_{k=1}^K \left(\frac{e^{a_k}}{\sum_{k=1}^K e^{a_k}} - \frac{e^{b_k}}{\sum_{k=1}^K e^{b_k}}\right)^2,$$

for some C. We consider $C = 4K^3 \exp(8Q)\{\exp(2Q) \vee 1\}$.

Let $R = \sum_{k=1}^{K} \exp(a_k)$ and $S = \sum_{k=1}^{K} \exp(b_k)$. Since $|a_k| \leq Q$, $|b_k| \leq Q$, we have $R \leq K \exp(Q)$ and $S \leq K \exp(Q)$. Therefore,

$$\begin{aligned}
&\sum_{k=1}^{K} \left(\frac{e^{a_k}}{\sum_{k=1}^{K} e^{a_k}} - \frac{e^{b_k}}{\sum_{k=1}^{K} e^{b_k}} \right)^2 \\
&= \sum_{k=1}^{K} \left(\frac{e^{a_k} S - e^{b_k} R}{SR} \right)^2 \\
&\geq \frac{1}{K^2 e^{2Q}} \sum_{k=1}^{K} (e^{a_k} S - e^{b_k} R)^2 \\
&= \frac{1}{K^2 e^{2Q}} \left[(S-R)^2 + \sum_{k=2}^{K} (e^{a_k} S - e^{b_k} R)^2 \right].
\end{aligned}$$

Let $t = \max_{2 \leq k \leq K} |a_k - b_k|$, and we just need to consider the case of $t > 0$. Let $j = \min_{2 \leq k \leq K} \{k : |a_k - b_k| = t\}$, then,

Case I: If $|S - R| > \{2 \exp(4Q)\}^{-1} t$, by the choice of C, we have

$$C \sum_{k=1}^{K} \left(\frac{e^{a_k}}{\sum_{k=1}^{K} e^{a_k}} - \frac{e^{b_k}}{\sum_{k=1}^{K} e^{b_k}} \right)^2 \geq \frac{C}{K^2 e^{2Q}} (S-R)^2 \geq \frac{C t^2}{4 K^2 e^{10Q}} \geq K t^2 \geq \sum_{k=2}^{K} (a_k - b_k)^2.$$

Case II: If $|S - R| \leq \{2 \exp(4Q)\}^{-1} t$, it follows

$$\sum_{k=1}^{K} \left(\frac{e^{a_k}}{\sum_{k=1}^{K} e^{a_k}} - \frac{e^{b_k}}{\sum_{k=1}^{K} e^{b_k}} \right)^2 \geq \frac{C}{K^2 e^{2Q}} (e^{a_j} S - e^{b_j} R)^2.$$

Note that $S \geq 1$, it holds

$$\begin{aligned}
&|e^{a_j}S - e^{b_j}R| \\
&= |(e^{a_j} - e^{b_j})S + e^{b_j}(S - R)| \\
&\geq |e^{a_j} - e^{b_j}| - |e^{b_j}(S - R)| \\
&\geq |e^{a_j} - e^{b_j}| - e^{Q}\frac{t}{2e^{4Q}} \\
&= |e^{b_j}||e^{a_j - b_j} - 1| - \frac{t}{2e^{3Q}}.
\end{aligned}$$

Since $|b_j| \leq Q$, we have $\exp(b_j) \geq \exp(-Q)$. Also note that $|\exp(x) - 1| \geq \exp(-2Q)|x|$ for all $|x| \leq 2Q$ and $|a_j - b_j| \leq 2Q$. It follows

$$|e^{a_j}S - e^{b_j}R| \geq |e^{b_j}||e^{a_j - b_j} - 1| - \frac{t}{2e^{3Q}} \geq \frac{|a_j - b_j|}{e^{Q}} - \frac{t}{2e^{3Q}} = \frac{t}{2e^{3Q}}.$$

Furthermore, by the choice of C, we have

$$C\sum_{k=1}^{K}\left(\frac{e^{a_k}}{\sum_{k=1}^{K}e^{a_k}} - \frac{e^{b_k}}{\sum_{k=1}^{K}e^{b_k}}\right)^2 \geq \frac{C}{K^2e^{2Q}}(e^{a_j}S - e^{b_j}T)^2 \geq \frac{Ct^2}{4K^2e^{8Q}} \geq Kt^2 \geq \sum_{k=2}^{K}(a_k - b_k)^2.$$

Based on **Case I** and **Case II**, the proof of **Lemma 9** is completed.

B.4 Proof of Theorem 6

There are only two classes of SCMM that we need to consider; one is for $b < K_0$ and the other one is for $b > K_0$. The former one is for under-fitted models and the latter one is for overfitted models. Hence, without loss of generality, we assume $K_0 = 3$ and $\mathcal{B} = \{2, 3, 4\}$. Then, it is sufficient to show

$$\begin{aligned}
&P\left(SCMM(S_0(3)) < \inf_{|M| \leq s} SCMM(M(2))\right) \to 1 \\
&P\left(SCMM(S_0(3)) < \inf_{|M| \leq s} SCMM(M(4))\right) \to 1
\end{aligned} \tag{B.17}$$

We first show the second statement in equation (B.17). Note that

$$\begin{aligned} &P\left(SCMM(S_0(3)) > \inf_{|M|\le s} SCMM(M(4))\right) \\ =\ &P\left(SCMM(S_0(3)) > \min\{\inf_{|M|\le s, S_0\subset M} SCMM(M(4)),\ \inf_{|M|\le s, S_0\not\subset M} SCMM(M(4))\}\right) \\ \le\ &P\left(SCMM(S_0(3)) > \inf_{|M|\le s, S_0\subset M} SCMM(M(4))\right) \\ &+P\left(SCMM(S_0(3)) > \inf_{|M|\le s, S_0\not\subset M} SCMM(M(4))\right) \\ :=\ &U_1 + U_2 \end{aligned} \tag{B.18}$$

By definition,

$$\begin{aligned} SCMM(S_0(3)) &= -l^3(S_0, \hat{\boldsymbol{\theta}}_{S_0}) + 3|S_0|n^{0.5+\delta_1}\log p/n, \\ SCMM(M(4)) &= -l^4(M, \hat{\boldsymbol{\theta}}_M) + 4|M|n^{0.5+\delta_1}\log p/n. \end{aligned}$$

Note that if $S_0 \subset M$, $|S_0| \le |M|$. Hence,

$$U_1 \le P\left(\sup_{|M|\le s, S_0\subset M} l^4(M, \hat{\boldsymbol{\theta}}_M) - l^3(S_0, \hat{\boldsymbol{\theta}}_{S_0}) > n^{0.5+\delta_1}\log p/n\right). \tag{B.19}$$

Let $g_i^M(x, y|\boldsymbol{\theta})$ as the log of logistic-normal density function for $K = i$ with covariates set M given $\boldsymbol{\theta}$. Let $f_i^M(\boldsymbol{\theta}) = E_0[g_i^M(x, y|\boldsymbol{\theta})]$ and $W_i^M = \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} f_i^M(\boldsymbol{\theta})$, where E_0 is the expectation taken under the truth. Note that the dimension for $\boldsymbol{\theta}$ is different for different i and M, however, because M includes the true set of covariates, we know $W_3^M = W_4^M :\equiv W$ for all M. Another important fact is that W_3^M can only be obtained at the true value $\boldsymbol{\theta}_0$ but W_4^M may be attained at multiple $\boldsymbol{\theta}'s$.

Note that

$$\begin{aligned} &P\left(\sup_{|M|\le s, S_0\subset M} l^4(M, \hat{\boldsymbol{\theta}}_M) - l^3(S_0, \hat{\boldsymbol{\theta}}_{S_0}) > n^{0.5+\delta_1}\log p/n\right) \\ \le\ &Cp^s \sup_{S_0\subset M, |M|\le s} P\left(l^4(M, \hat{\boldsymbol{\theta}}_M) - l^3(S_0, \hat{\boldsymbol{\theta}}_{S_0}) > n^{0.5+\delta_1}\log p/n\right). \end{aligned} \tag{B.20}$$

For any M such that $S_0 \subset M, |M| \le s$, define $R_4^M = \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} ||l^4(M, \boldsymbol{\theta}) - f_4^M(\boldsymbol{\theta})||$ and $R_3^M = \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} ||l^3(M, \boldsymbol{\theta}) - f_3^M(\boldsymbol{\theta})||$. We show

$$l^4(M, \hat{\boldsymbol{\theta}}_M) - l^3(S_0, \hat{\boldsymbol{\theta}}_{S_0}) \le R_4^M + R_3^{S_0}. \tag{B.21}$$

Note that

$$l^4(M,\hat{\boldsymbol{\theta}}_M) - l^3(S_0,\hat{\boldsymbol{\theta}}_{S_0}) \leq \left| l^4(M,\hat{\boldsymbol{\theta}}_M) - W \right| + \left| l^3(S_0,\hat{\boldsymbol{\theta}}_{S_0}) - W \right|.$$

By definition,

$$\begin{aligned} l^4(M,\hat{\boldsymbol{\theta}}_M) - W = l^4(M,\hat{\boldsymbol{\theta}}_M) - l^4(M,\boldsymbol{\theta}_0) + l^4(M,\boldsymbol{\theta}_0) - W \geq -R_4^M, \\ l^4(M,\hat{\boldsymbol{\theta}}_M) - W = l^4(M,\hat{\boldsymbol{\theta}}_M) - f_4^M(\hat{\boldsymbol{\theta}}_4^M) + f_4^M(\hat{\boldsymbol{\theta}}_4^M) - W \leq R_4^M. \end{aligned}$$

Therefore, $\left| l^4(M,\hat{\boldsymbol{\theta}}_M) - W \right| \leq R_4^M$ and similarly we have $\left| l^3(S_0,\hat{\boldsymbol{\theta}}_{S_0}) - W \right| \leq R_3^{S_0}$. Then (B.21) is proved. Together with (B.20) and (B.19), to prove $U_1 \to 0$, it is sufficient to prove for any M such that $M_0 \subset M, |M| \leq s$, we have

$$\begin{cases} p^s P\left(R_4^M > \frac{1}{2} n^{0.5+\delta_1} \log p/n\right) \to 0 \\ p^s P\left(R_3^M > \frac{1}{2} n^{0.5+\delta_1} \log p/n\right) \to 0 \end{cases}. \tag{B.22}$$

Let $Z_i^M(\boldsymbol{\theta}) = g_4^M(x_i, y_i|\boldsymbol{\theta}) - f_4^M(\boldsymbol{\theta})$, then Z_i^M is mean 0 and $l^4(M,\boldsymbol{\theta}) - f_4^M(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n Z_i^M(\boldsymbol{\theta})$. By equation (B.2), we know $|g_4^M(x_i, y_i|\boldsymbol{\theta}) - g_4^M(x_i, y_i|\boldsymbol{\theta}')| \leq C G_1(Y_i)||\boldsymbol{\theta} - \boldsymbol{\theta}'||_1$, where $G_1(Y_i) = C(|Y_i|^2 + |Y_i| + C)$. It is easy to verify that $E_0(G_1(Y_i)) \leq C$, and hence,

$$|f_4^M(\boldsymbol{\theta}) - f_4^M(\boldsymbol{\theta}')| \leq E_0(|g_4^M(x_i, y_i|\boldsymbol{\theta}) - g_4^M(x_i, y_i|\boldsymbol{\theta}')|) \leq C E_0(G_1(Y_i)||\boldsymbol{\theta} - \boldsymbol{\theta}'||_1) \leq C||\boldsymbol{\theta} - \boldsymbol{\theta}'||_1.$$

Therefore

$$|Z_i^M(\boldsymbol{\theta}) - Z_i^M(\boldsymbol{\theta}')| \leq |f_4^M(\boldsymbol{\theta}) - f_4^M(\boldsymbol{\theta}')| + |g_4^M(x_i, y_i|\boldsymbol{\theta}) - g_4^M(x_i, y_i|\boldsymbol{\theta}')| \leq C G_1(Y_i)||\boldsymbol{\theta} - \boldsymbol{\theta}'||_1. \tag{B.23}$$

Divide the parameter space $\boldsymbol{\Theta}$ into d_n pieces, namely, $\{\Omega_j\}_{j=1}^{d_n}$, where the maximum distance for each piece is no more than n^{-1}. Since $\boldsymbol{\Theta}$ is compact with at most $8s$ dimension, we know

$d_n \leq Cn^{8s}$. Then,

$$\begin{aligned}
P\left(R_4^M > \tfrac{1}{2}n^{0.5+\delta_1}\log p/n\right) &= P\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |\tfrac{1}{n}\textstyle\sum_{i=1}^n Z_i^M(\boldsymbol{\theta})| > \tfrac{1}{2}n^{0.5+\delta_1}\log p/n\right) \\
&= P\left(\max_{j\leq d_n}\sup_{\boldsymbol{\theta}\in\Omega_j} |\tfrac{1}{n}\textstyle\sum_{i=1}^n Z_i^M(\boldsymbol{\theta})| > \tfrac{1}{2}n^{0.5+\delta_1}\log p/n\right) \\
&\leq Cn^{8s}\max_{j\leq d_n} P\left(\sup_{\boldsymbol{\theta}\in\Omega_j} |\tfrac{1}{n}\textstyle\sum_{i=1}^n Z_i^M(\boldsymbol{\theta})| > \tfrac{1}{2}n^{0.5+\delta_1}\log p/n\right) \\
&\leq Cn^{8s}\max_{j\leq d_n} P\big(|\tfrac{1}{n}\textstyle\sum_{i=1}^n Z_i^M(\boldsymbol{\theta}_{\Omega_j})| + \\
&\sup_{\boldsymbol{\theta}\in\Omega_j} |\tfrac{1}{n}\textstyle\sum_{i=1}^n Z_i^M(\boldsymbol{\theta}) - \tfrac{1}{n}\textstyle\sum_{i=1}^n Z_i^M(\boldsymbol{\theta}_{\Omega_j})| > \tfrac{1}{2}n^{0.5+\delta_1}\log p/n\big) \\
&\leq Cn^{8s}\max_{j\leq d_n} P\left(|\tfrac{1}{n}\textstyle\sum_{i=1}^n Z_i^M(\boldsymbol{\theta}_{\Omega_j})| > \tfrac{1}{4}n^{0.5+\delta_1}\log p/n\right) + \\
&Cn^{8s}\max_{j\leq d_n} P\left(\sup_{\boldsymbol{\theta}\in\Omega_j} |\tfrac{1}{n}\textstyle\sum_{i=1}^n Z_i^M(\boldsymbol{\theta}) - \tfrac{1}{n}\textstyle\sum_{i=1}^n Z_i^M(\boldsymbol{\theta}_{\Omega_j})| > \tfrac{1}{4}n^{0.5+\delta_1}\log p/n\right) \\
&:= L_1 + L_2,
\end{aligned} \tag{B.24}$$

where $\boldsymbol{\theta}_{\Omega_j}$ is a point in Ω_j.

By (B.23),

$$\begin{aligned}
L_2 &\leq Cn^{8s}\max_{j\leq d_n} P\left(\tfrac{1}{n}\textstyle\sum_{i=1}^n G_1(Y_i)||\boldsymbol{\theta}_{\Omega_j} - \boldsymbol{\theta}||_1 > Cn^{0.5+\delta_1}\log p/n\right) \\
&\leq Cn^{8s} P\left(\tfrac{1}{n}\textstyle\sum_{i=1}^n G_1(Y_i) > Cn^{0.5+\delta_1}\log p\right) \\
&\leq Cn^{8s} P\left(\tfrac{1}{n}\textstyle\sum_{i=1}^n G_1(Y_i)1(|Y_i| \leq \sqrt{20s\log p}) > Cn^{0.5+\delta_1}\log p\right) \\
&\quad + Cn^{8s} P\left(\textstyle\bigcup_{i=1}^n \left\{|Y_i| > \sqrt{20s\log p}\right\}\right) \\
&:= L_{21} + L_{22}.
\end{aligned}$$

Note that if Y follows standard normal, $P(|Y| > t) \leq C\exp\{-t^2/2\}$ for large t. The same results can be readily extended to the mixture normal case by noting that the component means and variances are uniformly bounded over $x \in \mathcal{X}$. Therefore,

$$L_{22} \leq Cn^{8s+1} P\left(|Y| > \sqrt{20s\log p}\right) \leq Cn^{8s+1}\exp\{-10s\log p\} \leq Cp^{1-2s}.$$

Since $G_1(Y_i)1(|Y_i| \leq \sqrt{20s\log p})$ are $i.i.d.$ and bounded by $20s\log p$, and also note that $E(G_1(Y_i)) \leq C = o(n^{0.5+\delta_1}\log p)$ and $E(G_1^2(Y_i)) \leq C = o(n^{0.5+\delta_1}\log p)$. By Bernstein inequality,

$$L_{21} \leq Cn^{8s}\exp\left\{-C\frac{n^{3+2\delta_1}(\log p)^2}{Cn(\log p)^2 + Cn^{1.5+\delta_1}(\log p)^2}\right\} \leq Cn^{8s}\exp\{-n\}.$$

Hence,

$$L_2 \le L_{21} + L_{22} \le C(p^{1-2s} + n^{8s}\exp\{-n\}). \tag{B.25}$$

Next, for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, we calculate $P\left(|\frac{1}{n}\sum_{i=1}^n Z_i^M(\boldsymbol{\theta})| > \frac{1}{4}n^{0.5+\delta_1}\log p/n\right)$. Note that

$$Z_i^M(\boldsymbol{\theta}) = g_4^M(x_i, y_i|\boldsymbol{\theta}) - f_4^M(\boldsymbol{\theta}) = \log\left(\sum_{k=1}^4 \frac{\pi(\boldsymbol{x}^T\boldsymbol{\gamma}_k)\rho_k}{\sqrt{2\pi}}\exp\left\{-\frac{(\rho_k y - \boldsymbol{x}^T\boldsymbol{\phi}_k)^2}{2}\right\}\right) - f_4^M(\boldsymbol{\theta}),$$

and since $\boldsymbol{\Theta}$ is compact, $f_4^M(\boldsymbol{\theta})$ is uniformly bounded over $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Then for large p,

$$\begin{aligned}
|Z_i^M(\boldsymbol{\theta})| > C\log p &\Rightarrow \left|\log\left(\textstyle\sum_{k=1}^4 \frac{\pi(\boldsymbol{x}^T\boldsymbol{\gamma}_k)\rho_k}{\sqrt{2\pi}}\exp\left\{-\frac{(\rho_k y - \boldsymbol{x}^T\boldsymbol{\phi}_k)^2}{2}\right\}\right)\right| > C\log p \\
&\Rightarrow \textstyle\sum_{k=1}^4 \frac{\pi(\boldsymbol{x}^T\boldsymbol{\gamma}_k)\rho_k}{\sqrt{2\pi}}\exp\left\{-\frac{(\rho_k y - \boldsymbol{x}^T\boldsymbol{\phi}_k)^2}{2}\right\} < Cp^{-1} \\
&\Rightarrow \exp\left\{-\tfrac{(\rho_1 y - \boldsymbol{x}^T\boldsymbol{\phi}_1)^2}{2}\right\} < Cp^{-1} \Rightarrow |y| > C\sqrt{\log p},
\end{aligned}$$

where the last line is again by noting that $\boldsymbol{\Theta}$ is compact and $\mathcal{X}$ is bounded. For any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, choose C large enough, such that $\left\{|Z_i^M(\boldsymbol{\theta})| > C\log p\right\} \subset \left\{|y| > \sqrt{20s\log p}\right\}$, then

$$\begin{aligned}
&P\left(|\tfrac{1}{n}\textstyle\sum_{i=1}^n Z_i^M(\boldsymbol{\theta})| > \frac{1}{4}n^{0.5+\delta_1}\log p/n\right) \\
&\le P\left(|\tfrac{1}{n}\textstyle\sum_{i=1}^n Z_i^M(\boldsymbol{\theta})1(|Z_i^M(\boldsymbol{\theta})| \le C\log p)| > \frac{1}{4}n^{0.5+\delta_1}\log p/n\right) + P\left(\textstyle\bigcup_{i=1}^n\{|Z_i^M(\boldsymbol{\theta})| > C\log p\}\right) \\
&\le P\left(|\tfrac{1}{n}\textstyle\sum_{i=1}^n Z_i^M(\boldsymbol{\theta})1(|Z_i^M(\boldsymbol{\theta})| \le C\log p)| > \frac{1}{4}n^{0.5+\delta_1}\log p/n\right) + P\left(\textstyle\bigcup_{i=1}^n\{|Y_i| > \sqrt{20s\log p}\}\right) \\
&=: L_{11} + L_{12}.
\end{aligned} \tag{B.26}$$

From the bound of L_{22}, we know $L_{12} \le Cp^{1-10s}$.

Note that $Z_i^M(\boldsymbol{\theta})1(|Z_i^M(\boldsymbol{\theta})| \le C\log p)$ are $i.i.d.$ and bounded by $C\log p$. In order to use Bernstein inequality, we need to show $E(Z_i^M(\boldsymbol{\theta})1(|Z_i^M(\boldsymbol{\theta})| \le C\log p)) = o(n^{0.5+\delta_1}\log p/n)$. Note that $E(Z_i^M(\boldsymbol{\theta})) = 0$, then it is equivalent to show

$$E(Z_i^M(\boldsymbol{\theta})1(|Z_i^M(\boldsymbol{\theta})| > C\log p)) = o(n^{0.5+\delta_1}\log p/n). \tag{B.27}$$

Note that

$$
\begin{aligned}
&|E(Z_i^M(\boldsymbol{\theta})1(|Z_i^M(\boldsymbol{\theta})| > C\log p))| \\
\leq &\left|E\left(\log\left(\textstyle\sum_{k=1}^4 \frac{\pi(\boldsymbol{X}_i^T\boldsymbol{\gamma}_k)\rho_k}{\sqrt{2\pi}}\exp\left\{-\frac{(\rho_k Y_i - \boldsymbol{X}_i^T\boldsymbol{\phi}_k)^2}{2}\right\}\right)1(|Z_i^M(\boldsymbol{\theta})| > C\log p)\right)\right| \\
&+|f_4^M(\boldsymbol{\theta})E(1(|Z_i^M(\boldsymbol{\theta})| > C\log p))| \\
\leq &\left|E\left(\log\left(\textstyle\sum_{k=1}^4 \frac{\pi(\boldsymbol{X}_i^T\boldsymbol{\gamma}_k)\rho_k}{\sqrt{2\pi}}\exp\left\{-\frac{(\rho_k Y_i - \boldsymbol{X}_i^T\boldsymbol{\phi}_k)^2}{2}\right\}\right)1(|Y_i| > \sqrt{20s\log p})\right)\right| \\
&+|f_4^M(\theta)|E(1(|Y_i| > \sqrt{20s\log p})) \\
= &\left|E\left(\log\left(\textstyle\sum_{k=1}^4 \frac{\pi(\boldsymbol{X}_i^T\boldsymbol{\gamma}_k)\rho_k}{\sqrt{2\pi}}\exp\left\{-\frac{(\rho_k Y_i - \boldsymbol{X}_i^T\boldsymbol{\phi}_k)^2}{2}\right\}\right)1(|Y_i| > \sqrt{20s\log p})\right)\right| \\
&+o(p^{-10s}),
\end{aligned}
\tag{B.28}
$$

where the inequality holds because $\{|Z_i^M(\boldsymbol{\theta})| > C\log p\} \subset \{|Y_i| > \sqrt{20s\log p}\}$ and the log likelihood function is negative when $|Y_i|$ is large.

Note again the compactness of $\boldsymbol{\Theta}$, when $|Y_i|$ is large we have

$$
\textstyle\sum_{k=1}^4 \frac{\pi(\boldsymbol{X}_i^T\boldsymbol{\gamma}_k)\rho_k}{\sqrt{2\pi}}\exp\left\{-\frac{(\rho_k Y_i - \boldsymbol{X}_i^T\boldsymbol{\phi}_k)^2}{2}\right\} \geq \textstyle\sum_{k=1}^4 \frac{\pi(\boldsymbol{X}_i^T\boldsymbol{\gamma}_k)\rho_k}{\sqrt{2\pi}}\exp\left\{-CY_i^2\right\} \geq C\exp\left\{-CY_i^2\right\}.
$$

Then,

$$
\begin{aligned}
&\left|\left(\log\left(\textstyle\sum_{k=1}^4 \frac{\pi(\boldsymbol{X}_i^T\boldsymbol{\gamma}_k)\rho_k}{\sqrt{2\pi}}\exp\left\{-\frac{(\rho_k Y_i - \boldsymbol{X}_i^T\boldsymbol{\phi}_k)^2}{2}\right\}\right)\right|1(|Y_i| > \sqrt{20s\log p})\right) \\
&\quad\leq \left|\log\left(C\exp\left\{-CY_i^2\right\}\right)\right|1(|Y_i| > \sqrt{20s\log p}) \\
&\quad\leq (|\log C| + CY_i^2)1(|Y_i| > \sqrt{20s\log p}).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\left|E\left(\log\left(\textstyle\sum_{k=1}^4 \frac{\pi(\boldsymbol{X}_i^T\boldsymbol{\gamma}_k)\rho_k}{\sqrt{2\pi}}\exp\left\{-\frac{(\rho_k Y_i - \boldsymbol{X}_i^T\boldsymbol{\phi}_k)^2}{2}\right\}\right)1(|Y_i| > \sqrt{20s\log p})\right)\right| \\
&\quad\leq E((|\log C| + CY_i^2)1(|Y_i| > \sqrt{20s\log p})) \\
&\quad= CE(Y_i^2 1(|Y_i| > \sqrt{20s\log p})) + o(p^{-10s}).
\end{aligned}
\tag{B.29}
$$

When Y is standard normal, direct integration yields $E(Y^2 1(Y > C\sqrt{\log p})) \leq C\sqrt{\log p}/p$. It can be seen that the same result can be obtained when Y follows mixture of normals. Hence by (B.28) and (B.29),

$$
|E(Z_i^M(\boldsymbol{\theta})1(|Z_i^M(\boldsymbol{\theta})| > C\log p))| \leq O(\sqrt{\log p}/p) + o(p^{-5s}) = o(n^{0.5+\delta_1}\log p/n).
$$

Therefore (B.27) is proved and thus we have $E(Z_i^M(\boldsymbol{\theta})1(|Z_i^M(\boldsymbol{\theta})| \leq C\log p)) = o(n^{0.5+\delta_1}\log p/n)$.

We then use Bernstein inequality to the sequence of $\{Z_i^M(\boldsymbol{\theta})1(|Z_i^M(\boldsymbol{\theta})| \le C\log p))\}_{i=1}^n$ and we have

$$L_{11} \le \exp\left\{-C\frac{n^{1+2\delta_1}(\log p)^2}{n(\log p)^2+n^{0.5+\delta_1}(\log p)^2}\right\} \le \exp\{-n^{2\delta_1}\}. \tag{B.30}$$

By (B.26),

$$P\left(|\tfrac{1}{n}\textstyle\sum_{i=1}^n Z_i^M(\boldsymbol{\theta})| > \tfrac{1}{4}n^{0.5+\delta_1}\log p/n\right) \le \exp\{-n^{2\delta_1}\} + p^{1-10s}.$$

Hence

$$\begin{aligned} L_1 &= Cn^{8s}\max_{j\le d_n} P\left(|\tfrac{1}{n}\textstyle\sum_{i=1}^n Z_i^M(\boldsymbol{\theta}_{\Omega_j})| > \tfrac{1}{4}n^{0.5+\delta_1}\log p/n\right) \\ &\le Cn^{8s}(\exp\{-n^{2\delta_1}\} + p^{1-10s}). \end{aligned} \tag{B.31}$$

Together (B.31), (B.25) with (B.24), we have,

$$\begin{aligned} p^s P\left(R_4^M > \tfrac{1}{2}n^{0.5+\delta_1}\log p/n\right) &\le p^s(L_1+L_2) \\ &\le Cp^s(n^{8s}(\exp\{-n^{2\delta_1}\} + p^{1-10s}) + p^{1-2s} + n^{8s}\exp\{-n\}) \\ &\le Cp^s(\exp\{-n^{2\delta_1} + 8s\log n\} + 2p^{1-2s} + \exp\{-n + 8s\log n\}) \\ &= C(\exp\{-n^{2\delta_1} + 8s\log n + s\log p\} + 2p^{1-s} + \exp\{-n + 8s\log n + s\log p\}) \\ &\to 0, \end{aligned} \tag{B.32}$$

where the last line is because $p \le n^C$ for some C.

By (B.32), we have proved the first statement in (B.22). Similar arguments yield the second statement in (B.22). Hence $U_1 \to 0$.

Next we are going to show $U_2 \to 0$. Note that

$$\begin{aligned} &SCMM(S_0(3)) - \inf_{|M|\le s, S_0\not\subset M} SCMM(M(4)) \\ &\le -l^3(S_0, \hat{\boldsymbol{\theta}}_{S_0}) + 3|S_0|n^{0.5+\delta_1}\log p/n + \sup_{|M|\le s, S_0\not\subset M} l^4(M, \hat{\boldsymbol{\theta}}_M) \\ &= -W + \sup_{|M|\le s, S_0\not\subset M} l^4(M, \hat{\boldsymbol{\theta}}_M) + 3|S_0|n^{0.5+\delta_1}\log p/n + o_p(1) \\ &\le -W + \sup_{|M|\le s, S_0\not\subset M}\{l^4(M, \hat{\boldsymbol{\theta}}_M) - f_4^M(\hat{\boldsymbol{\theta}}_M)\} \\ &+ \sup_{|M|\le s, S_0\not\subset M} f_4^M(\hat{\boldsymbol{\theta}}_M) + 3|S_0|n^{0.5+\delta_1}\log p/n + o_p(1) \\ &\le -W + \sup_{|M|\le s, S_0\not\subset M} R_4^M \\ &+ \sup_{|M|\le s, S_0\not\subset M} W_4^M + 3|S_0|n^{0.5+\delta_1}\log p/n + o_p(1). \end{aligned} \tag{B.33}$$

By condition (3.15) in **Theorem 6**, there exists a positive constant C, such that

$$W - \sup_{|M|\leq s, S_0 \not\subset M} W_4^M > C.$$

Hence by (B.33),

$$\begin{aligned} U_2 &= P(SCMM(S_0(3)) - \inf_{|M|\leq s, S_0\not\subset M} SCMM(M(4)) > 0) \\ &\leq P(\sup_{|M|\leq s, S_0\not\subset M} R_4^M > C - 3|S_0|n^{0.5+\delta_1}\log p/n - o_p(1)) \to 0, \end{aligned} \tag{B.34}$$

where the last line is by noting (B.22) and the fact that $C - 3|S_0|n^{0.5+\delta_1}\log p/n - o_p(1) > n^{0.5+\delta_1}\log p/n$ eventually.

By (B.18), we proved the second statement in equation (B.17). Similar arguments in proving $U_2 \to 0$ yields the first statement in equation (B.17). Hence, we proved **Theorem 6**.

BIBLIOGRAPHY

[1] Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of Econometrics*, 16(1), 3-14.

[2] Altstein, L. L., Li, G., and Elashoff, R. M. (2011). A method to estimate treatment efficacy among latent subgroups of a randomized clinical trial. *Statistics in medicine*, 30(7), 709-717.

[3] Berger, J. O., Wang, X., and Shen, L. (2014). A Bayesian approach to subgroup identification. *Journal of biopharmaceutical statistics*, 24(1), 110-129.

[4] Bickel, P. J., Ritov, Y. A., and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 1705-1732

[5] Cai, T., Tian, L., Wong, P. H., and Wei, L. J. (2011). Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics*, 12(2), 270-282.

[6] Chen, H., Chen, J., and Kalbfleisch, J. D. (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(1), 19-29.

[7] Chen, J., and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3), 759-771.

[8] Chen, J., and Li, P. (2009). Hypothesis test for normal mixture models: The EM approach. *The Annals of Statistics*, 2523-2542.

[9] Chen, J., Tan, X., and Zhang, R. (2008). Inference for normal mixtures in mean and variance. *Statistica Sinica*, 443-465.

[10] Corduneanu, A., and Bishop, C. M. (2001, January). Variational Bayesian model selection for mixture distributions. *In Artificial intelligence and Statistics* (Vol. 2001, pp. 27-34). Waltham, MA: Morgan Kaufmann.

[11] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B*

(methodological), 1-38.

[12] Fan, A., Lu, W., and Song, R. (2015). Change-Plane Analysis for Subgroup Detection and Sample Size Calculation.

[13] Fan, J., and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456), 1348-1360.

[14] Foster, J. C., Taylor, J. M., and Ruberg, S. J. (2011). Subgroup Identication from Randomized Clinical Trial Data. *Statistics in Medicine* 30,2867-2880.

[15] Friedman, J., Hastie, T., Hfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2), 302-332.

[16] Goeffinet, B., Loisel, P., and Laurent, B. (1992). Testing in Normal Mixture Models when the Proportions are Known. *Biometrika* 79,842-846.

[17] Greenshtein, E., and Ritov, Y. A. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6), 971-988.

[18] Huang, T., Peng, H., and Zhang, K. (2013). Model Selection for Gaussian Mixture Models. *arXiv preprint arXiv*:1301.3558.

[19] Imai, K. and Ratkovic, M. (2013). Estimating Treatment Eect Heterogeneity in Randomized Program Evaluation. *The Annals of Applied Statistics* 7,443-470.

[20] Jiang, W., and Tanner, M. A. (1999). Hierarchical mixtures-of-experts for exponential family regression models: approximation and maximum likelihood estimation. *Annals of Statistics*, 987-1011.

[21] Jordan, M. I., and Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural computation*, 6(2), 181-214.

[22] Kang, C., Janes, H., and Huang, Y. (2014). Combining Biomarkers to Optimize Patient Treatment Recommendations. *Biometrics* 70,695-707.

[23] Khalili, A., and Chen, J. (2012). Variable selection in finite mixture of regression models. *Journal of the american Statistical association*.

[24] Khalili, A., and Lin, S. (2013). Regularization in finite mixture of regression models with diverging number of parameters. *Biometrics*, 69(2), 436-446.

[25] Lehmann, E. L., Casella, G., and Casella, G. (1991). *Theory of point estimation*. Wadsworth and Brooks/Cole Advanced Books and Software.

[26] Lipkovich, I. and Dmitrienkoa, A. (2014). Strategies for Identifying Predictive Biomarkers and Subgroups with Enhanced Treatment Eect in Clinical Trials Using SIDES. *Journal of*

Biopharmaceutical Statistics 24,130-153.

[27] Lipkovich, I., Dmitrienko, A., Denne, J., and Enas, G. (2011). Subgroup identication based on dierential eect searcha recursive partitioning method for establishing response-totreatmentinpatientsubpopulations. *Statistics in Medicine* 30,2601-2621.

[28] Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J., and Ramig, L. O. (2009). Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *Biomedical Engineering, IEEE Transactions on*, 56(4), 1015-1022.

[29] Lounici, K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of statistics*, 2, 90-102.

[30] Massart, P. (2000). About the constants in Talagrand's concentration inequalities for empirical processes. *Annals of Probability*, 863-884.

[31] Massart, P. (2000). Some applications of concentration inequalities to statistics. *In Annales de la Facult des sciences de Toulouse: Mathmatiques* (Vol. 9, No. 2, pp. 245-303).

[32] McLachlan, G., and Peel, D. (2004). *Finite mixture models*. John Wiley and Sons.

[33] Muthn, B., and Asparouhov, T. (2009). Multilevel regression mixture analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(3), 639-657.

[34] Muthn, B., and Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, 55(2), 463-469.

[35] Pollard, D. (2012). *Convergence of stochastic processes*. Springer Science and Business Media.

[36] Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2), 461-464.

[37] Shen, J. and He, X. (2015). Inference for Subgroup Analysis With a Structured Logistic Normal Mixture Model. *Journal of the American Statistical Association* 110, 303- 312.

[38] Simon, R. (2002). Bayesian Subset Analysis: Application to Studying Treatment-byGender Interactions. *Statistics in Medicine* 21,2909-2916.

[39] Sleight,P.(2000). Debate: Subgroupanalysesinclinicaltrials: funtolookat-butdont believe them! *Current Controlled Trials in Cardiovascular Medicine* 1,25-27.

[40] Song, Y. and Chi, G. Y. (2007). A Method for Testing a Pre-specied Subgroup in Clinical Trials. *Statistics in Medicine* 26,3535-3549.

[41] Städler, N., Bühlmann, P., and van de Geer, S. (2010). ℓ_1-penalization for mixture regression models. *Test*, 19(2), 209-256.

[42] Su, X., Tsai, C. L., Wang, H., Nickerson, D. M., and Li, B. (2009). Subgroup Analysis via Recursive Partitioning. *Journal of Machine Learning Research* 10,141-158.

[43] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

[44] Tsanas, A., Little, M. A., McSharry, P. E., and Ramig, L. O. (2010). Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. *Biomedical Engineering, IEEE Transactions on*, 57(4), 884-893.

[45] van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 614-645.

[46] van de Geer, S. (2013). Generic chaining and the ℓ_1-penalty. *Journal of Statistical Planning and Inference*, 6(143), 1026-1028.

[47] Van Horn, M. L., Jaki, T., Masyn, K., Howe, G., Feaster, D. J., Lamont, A. E., George, M. R. W., and Kim, M. (2015). Evaluating dierential eects using regression interactions and regression mixture models. *Educational and Psychological Measurement* 75,677-714.

[48] Wong, C. S., and Li, W. K. (2001). On a logistic mixture autoregressive model. *Biometrika*, 88(3), 833-846.

[49] Zhang, T. (2009). Some sharp performance bounds for least squares regression with ℓ_1 regularization. *The Annals of Statistics*, 37(5A), 2109-2144.

[50] Zhao, L., Tian, L., Cai, T., Claggett, B., and Wei, L. J. (2013). Eectively Selecting a Target Population for a Future Comparative Study. *Journal of the American Statistical Association* 108,527-539.

[51] Zhao, P., and Yu, B. (2006). On model selection consistency of Lasso. *The Journal of Machine Learning Research*, 7, 2541-2563.