Some Interactions of Modern Optimization and Statistics

Xingyuan Fang

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Department of
Operations Research and Financial Engineering
Advisers: Professor Han Liu
Professor Robert J. Vanderbei

May 2016

Abstract

This dissertation attacks several challenging problems using state-of-the-art modern optimization and statistics. We first consider optimal, two stage, adaptive enrichment designs for randomized trials, using sparse linear programming. Adaptive enrichment designs involve preplanned rules for modifying enrollment criteria based on accruing data in a randomized trial. Such designs have been proposed. The goal is to learn which populations benefit from an experimental treatment. Two critical components of adaptive enrichment designs are the decision rule for modifying enrollment, and the multiple testing procedure. We provide the first general framework for simultaneously optimizing both of these components for two stage, adaptive enrichment designs. We minimize expected sample size under constraints on power and the familywise Type I error rate.

Next, we consider high-dimensional spatial graphical model estimation under a total cardinality constraint (i.e., the ℓ_0-constraint). Though this problem is highly nonconvex, we show that its primal-dual gap diminishes linearly with the dimensionality and provide a convex geometry justification of this 'blessing of massive scale' phenomenon. Motivated by this result, we propose an efficient algorithm to solve the dual problem and prove that the solution achieves optimal statistical properties.

Finally, we consider the problem of hypothesis testing and confidence intervals under high dimensional proportional hazards models. Motivated by a geometric projection principle, we propose a unified likelihood ratio inferential framework, including score, Wald and partial likelihood ratio statistics for hypothesis testing. Without assuming model selection consistency, we derive the asymptotic distributions of these test statistics, establish their semiparametric optimality, and conduct power analysis under Pitman alternatives. We also develop procedures to construct pointwise confidence intervals for the baseline hazard function and conditional hazard function.

Acknowledgements

It has been a long way for me to reach this point. Along the way, there are many people deserve my thankfulness. Naturally, let me start with my advisors. No words in any language could properly explain how thankful I am to my advisors, Professor Han Liu and Professor Robert Vanderbei. Without them, I would not have this thesis done. Three years ago, Professor Liu and Professor Vanderbei kindly accepted me as their students. I remember clearly that, when I was at a desperate situation, they told me independently:

> "No matter how others think about you, I believe that you are strong enough to become a successful researcher."

To them, I owe too much. Over the past three years, they have spent countless afternoons and evenings to guide me from all aspects. I have benefited from their insightful suggestions on research, and their unconditional support to my career. Their perpetual encouragement and motivation have illuminated my graduate life. They have tried hard to teach me not only how to conduct interesting research, but also how to behave as a mature independent scholar. From them, I see all the characteristics of successful scholars. I believe that what they have taught me would be my great asset to help me build a more successful career. I feel extremely fortunate to have them as my advisors, and I cannot imagine to have better advisors than them.

Thank you Han! Thank you Bob!

It is my greatest honor to have Professor Jianqing Fan, one of the very top researchers in statistics and one of my heroes since my early undergraduate study, to sit on my dissertation committee. Almost seven years ago, I met Professor Fan for the first time during a conference held in my hometown, and Professor Fan strongly encouraged me to apply for ORFE. Professor Fan has always been kind in discussing research with me and gave me a lot of insightful suggestions, and he has always been supportive in all aspects.

I am sincerely grateful to Professor Mengdi Wang, who agrees to sit on my committee and reads my thesis. Her suggestions greatly improves the quality of this dissertation. I have collaborated with Mengdi on several papers over the past three years. Throughout our collaboration, she has always been nice to share her ideas, and she spent countless hours to discuss with me on technical details. I benefit a lot from her insightful thinking and detailed technical guidance. I am looking forward to more fruitful collaborations with Mengdi.

I would like to thank my great collaborator, Professor Michael Rosenblum from Johns Hopkins. Michael and I have worked on several papers. Michael always share his deep thoughts and ideas with me, and he always encourages me to pursue my ideas and gives his suggestions. I would also like to thank Michael for his financial support for me over the past two years, and hosting me a nice visit to Johns Hopkins. I believe that we will have more fruitful collaborations.

I was fortunate enough to have my research began with my extraordinary undergraduate advisor, Professor Toh Kim-Chuan, who is one of the very top researchers in numerical optimization. From him, I began to learn how to conduct research, and I wrote my first paper with him. His passion and attitude towards research deeply impacted my decision to have an academic career. Professor Toh has always been supportive over the past eight years. He has always been kind to discuss with me about all questions/issues I have from all aspects. I would also like to sincerely thank my other undergraduate advisor, Professor Loh Wei-Liem, who has always been supportive over the past eight years.

I am deeply thankful to my other collaborators including Professor Michael I. Jordan, Professor Xiaoyong Yang, Professor Xiaoming Yuan, Dr. Min-Dian Li, Dr. Wen-Xin Zhou, Ms. Emily Huang and many others. I want to specially thank Dr. Yang Ning, who spent countless hours to help me on many technical details during my early stage. All the collaborations have been enjoyable and fruitful.

I want to thank all the professors I met at Princeton such as Professors Sebastien Bubeck, Patrick Cheridito, Marc Hallin, Samory Kpotufe, Philippe Rigollet, Ramon van Handel. I am especially grateful to Professors William Cook and Alain Kornhauser, who let me be their teaching assistant. I also want to specially thank Professor Amirali Ahmadi for letting me co-organize weekly optimization seminar, and helped me prepare my job talk. All these experience will certainly help my later career. I also own special thanks to all the staffs at ORFE, Michael, Kim, Carol, Tara, Melissa, Connie, Tara, Lisa and Tabitha, who always help me to solve problems perfectly.

I would like to thank all my fellow friends in ORFE who have been supportive over these years: Xu Han, Yuan Liao, Yi Ma, Lei Qi, Xin Tong Sr., Weijie Gu, Tracy Ke, Xin Tong Jr., Wei Dai, Juan Sagredo, Xiaofeng Shi, Jiawei Yao, Weichen Wang, Che-Yu Liu, Xiuneng Zhu, Zhao Chen, Lingzhou Xue, Dan Wang, Yuan Cao, Junwei Lu, Boyang Song, Zhaoran Wang, Tianqi Zhao, Quefeng Li, Yuyan Wang, Peiqi Wang, Ziwei Zhu, Quanquan Gu, Qiang Sun, Huanran Lu, Kean-Meng Tan, Matey Neykov, Ritwik Mitra, Junchi Li, Xialiang Dou, Cong Ma and Zhuoran Yang. All the friends and many others I can not list here make my graduate life enjoyable and memorable.

I deeply appreciate all the teachers I met, from primary school to university. During my studying journey over the past twenty years, I met many great teachers, who helped me along the journey. Due to the space limit, I cannot list all the teachers I met. To name a few of them, NUS: Chu Delin, Tay Yong-Chiang, Victor Tan, Yu Shih-Hsien, Zhu Chen-Bo, UIUC: Slawomir Solecki, UFL: Malay Ghosh, Chengdu No. 7 High School: Fang Tinggang, Fu Yang, Ni Chi, Tang Zhaojun, Wang Zhijian, Xu Chen, Yang Hongji, CEFLS: Ge Renjian, Xu Furong, Yao Bing.

I do want to thank a friend of mine, Lingyi, for her trust, understanding, and support in the past many years. Indeed, our friendship for more than 10 years makes me a better person.

I would like to thank my parents for growing me up through the way they did. I am sure that through the way they treated me, I have become a much better person than I would if they were not there. They have been always supportive in all the aspects through my entire life. It is certain to me that they love me more than anything else in the universe.

I deeply thank my wife, Lin Lin. I want to particularly thank her tolerance, support and love. I feel extremely fortunate to find her as my special one to accompany me throughout my graduate study journey, and the exciting adventures ahead at Penn State.

Finally, I want to thank my grandfather from the past, an old retired professor to whom I owe more than what I realized when he was alive. He was my best friend through my childhood. He illuminated my childhood during the days we spent together. From him, I first learnt stories about Issac Newton, Marie Curie and Galileo Galilei. During his last years, he read One Hundred Thousand Whys for me everyday, and he taught me basic scientific thinking which lightened the fire in my mind to pursue research for the first time. To the memory of my beloved grandfather and my other family members, I dedicate this thesis.

To my family

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Over the past two decades, we have seen significant interactions of optimization and statistics. This dissertation considers three important challenging statistical problems where we solve them by modern optimization techniques. In Chapter 2, we first consider a clinical trial problem. Consider the problem of planning a randomized trial of a new treatment versus control, when the population of interest is partitioned into two subpopulations. The subpopulations could be defined in terms of a biomarker or risk score measured at baseline. Our goal is to test the null hypotheses of no average treatment benefit for each subpopulation and for the combined population. Standard randomized trial designs may have low power to detect a treatment effect if the treatment only benefits one subpopulation. Adaptive enrichment designs have been proposed for this problem, e.g., Follmann (1997), Russek-Cohen and Simon (1997), Jennison and Turnbull (2007), Wang et al. (2007), Wang et al. (2009b), Brannath et al. (2009a), Rosenblum and van der Laan (2011), Jenkins et al. (2011a), Friede et al. (2012), Boessen et al. (2013a), Stallard et al. (2014).

An adaptive enrichment design consists of a decision rule for potentially modifying enrollment at an interim analysis, and a multiple testing procedure. The decision rule is allowed to be an arbitrary, prespecified function from the stage 1 data to a finite set of possible enrollment decisions for stage 2. The multiple testing procedure can be an arbitrary, pre-

specified function from the stage 1 and 2 data to the set of null hypotheses that are rejected. The class of possible designs is therefore quite large. Our goal is to construct new adaptive enrichment designs that minimize expected sample size under constraints on power and Type I error, over this class of possible designs. This is a nonconvex optimization problem that is computationally infeasible to solve directly.

Our approach is to approximate the original optimization problem by a large, sparse linear program. This idea was used in the context of standard (non-adaptive) designs by Rosenblum et al. (2014), where the only feature optimized was the multiple testing procedure. We tackle the substantially more challenging problem of simultaneously optimizing the decision rule and multiple testing procedure in two-stage, adaptive enrichment designs. The difficulty of the latter problem is twofold: it is harder to construct a representation as a sparse, linear program, and the resulting linear program is harder to solve computationally. Another difference between the work here and Rosenblum et al. (2014) is that we consider not only power, but also expected sample size. In practice, both of these are important in trial planning.

We prove that our designs control the familywise Type I error rate in the strong sense defined by Hochberg and Tamhane (1987, pg. 3). Control of the familywise Type I error rate is generally required by regulatory agencies such as the U.S. Food and Drug Administration for confirmatory randomized trials (FDA, 2010).

As in all of the above related work, we require the subpopulations to be defined before the trial starts. Such a definition could be based on prior trial data and scientific understanding of the disease being treated. Designs exist that try to solve the more challenging problem of defining a subpopulation based on accruing data and then testing for a treatment effect in that subpopulation, e.g., Freidlin and Simon (2005); Lai et al. (2014). Optimizing trial designs in this context is an area of future work.

Hampson and Jennison (2015) consider the related problem of optimizing a two-stage adaptive design to determine the optimal treatment among k possible treatments, for a single population. Their general approach of converting the problem to a Bayes decision problem does not work in our setting, since this approach requires that the optimal solution at the global null hypothesis also controls familywise Type I error at all other alternatives.

We focus on designs where the only allowed adaptation is to modify enrollment for stage 2. Other types of adaptive designs involve modifying randomization probabilities (called covariate-adaptive or response-adaptive designs), or modifying the treatment for each individual in response to his/her outcomes over time (called dynamic treatment regimes). In contrast to these types of adaptation, each participant in our designs is randomized with probability 1/2 to treatment or control, and once he/she is randomized there is no change to the treatment received. The only design feature that may be changed is the stage 2 enrollment criteria.

In Chapter 3, we consider the problem of estimating high dimensional spatial graphical models. More specifically, let $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ be a d-dimensional random vector on a spatial field (e.g., a lattice). We aim to find an undirected graph $G = (V, E)$ with vertex set $V = \{1, 2, ..., d\}$ and edge set $E \subset V \times V$ to encode the conditional independence of $\boldsymbol{X}$, i.e., $(j, k) \in E$ if and only if X_j and X_k are conditionally independent given the remaining variables. A spatial graphical model also requires the graph G to be conformed with the spatial proximity. In other words, a necessary condition for the existence of edge $(j, k) \in E$ is that vertices j and k are spatially closed (more details are provided later).

Spatial graphical models find various real-world applications. For example, an important application is to infer the topology of sensor network on a 2D surface. The wireless sensor network is widely used in several applications including agriculture (Langendoen et al., 2006), military (Lee et al., 2009) and environmental science (Howard et al., 2002). See Yick et al. (2008) for a survey. In these applications, it is important to understand how the sensors

interact with one another. In practice, each sensor can only communicate with other sensors that are geographically close. Also, in applications such as agricultural and environmental studies, all sensors' corresponding locations are known. Thus, the spatial proximity information of these sensors are available a priori, and such information is incorporated in spatial graphical model estimation. We provide a numerical example of sensor network estimation in Section 3.5.2.

There are many other applications of spatial graphical model such as estimating short-range brain network. In these networks, the vertices are voxels or ROIs (region of interest) embedded in the 3D space. Given the brain imaging data and the spatial information, we aim to estimate the graphical model under the constraint that each vertex can only connect to the nodes that are physically close as discussed in (Cao and Fei-Fei, 2007). The estimated graph serves a first step for more sophisticated downstream analysis (e.g., long-range brain network construction or functional region partition (Bullmore and Sporns, 2009; Bullmore and Bassett, 2011)).

In Chapter 4, we consider high-dimensional inference under the high-dimensional proportional hazards model. The proportional hazards model (Cox, 1972) is one of the most important tools for analyzing time to event data. It finds wide applications in epidemiology, medicine, economics, and sociology (Kalbfleisch and Prentice, 2011). This model is semi-parametric by treating the baseline hazard function as a nuisance parameter. To infer the finite dimensional parameter of interest, Cox (1972, 1975) proposes the partial likelihood approach which is invariant to the baseline hazard function. In low dimensional settings, Tsiatis (1981) and Andersen and Gill (1982) have established the consistency and asymptotic normality of the maximum partial likelihood estimator.

In high dimensional settings, when the number of covariates d is larger than the sample size n, the partial maximum likelihood estimation is an ill-posed problem. To solve this problem, we resort to the regularized estimators (Tibshirani, 1996, 1997; Fan and Li, 2002;

Antoniadis et al., 2010). Other types of estimation procedures and the theoretical properties are studied by Cai et al. (2005); Zhang and Lu (2007); Wang et al. (2009a); Zhao and Li (2012). In particular, under the ultra-high dimensional regime that $d = o\big\{\exp(s^{-1}n)\big\}$, Bradic et al. (2011); Huang et al. (2013); Kong and Nan (2014) establish the oracle properties and error bounds of penalized maximum partial likelihood estimator, where s denotes the number of nonzero elements in the parametric component of the Cox model.

Though significant progress has been made towards developing the estimation theory, little work exists on the inferential aspects (e.g., testing hypothesis or constructing confidence intervals) of high dimensional proportional hazard models. A notable exception is Bradic et al. (2011), who establish the limiting distribution of the oracle estimator. However, such a result hinges on model selection consistency, which is not always possible in applications. There exist a number of recent works (van de Geer et al., 2014; Belloni et al., 2013; Javanmard and Montanari, 2013; Lockhart et al., 2014; Zhang and Zhang, 2014; Ning and Liu, 2014; Zhong et al., 2015) considering high dimensional inference under the linear, generalized linear and additive hazard models. Compared to these existing inferential results, the analysis of the proportional hazards model is much more challenging. First, to handle the time-dependent covariates, we need to use the counting process formulation, which is a unique challenge in the survival analysis. This formulation "permits a regression analysis of the intensity of a recurrent event allowing for complicated censoring patterns and time-dependent covariate" (Andersen and Gill, 1982). Second, the log-partial likelihood no longer has the sum of i.i.d structure, which is different from the standard regression model. To address this challenge, we need to use the martingale and empirical process theory to control the estimation and normal approximation errors in the high dimensional setting. To the best of our knowledge, uncertainty assessment under the high dimensional proportional hazards model remains an open problem. This work aims to close this gap by developing valid inferential procedures and theory for high dimensional proportional hazards models.

In particular, we test hypotheses and construct confidence regions for a low dimensional component of a d dimensional parameter vector. Compared with Bradic et al. (2011), our method does not require any type of irrepresentable condition or the minimal signal strength condition. Thus, it is more practical in applications.

Chapter 2

Optimal Two Stage Trial Design Using Sparse Linear Programming

2.1 Problem Definition

2.1.1 Null Hypotheses

We assume that the population is partitioned into two subpopulations, defined in terms of variables measured before randomization. Let p_s denote the proportion of the population in subpopulation $s \in \{1, 2\}$, which we assume are known; $p_1 + p_2 = 1$. Each enrolled participant is assigned to treatment ($a = 1$) or control ($a = 0$) with probability 1/2. Below, for clarity, we focus on normally distributed outcomes with known variances for ease of presentation.

For each subpopulation $s \in \{1, 2\}$ and stage $k \in \{1, 2\}$, we assume that exactly half the participants are assigned to each study arm $a \in \{0, 1\}$. This can be approximately achieved in practice by using block randomization stratified by subpopulation. For each participant i from subpopulation $s \in \{1, 2\}$ enrolled in stage $k \in \{1, 2\}$, let $(A_{s,i}^{(k)}, Y_{s,i}^{(k)})$ denote his/her arm assignment $A_{s,i}^{(k)} \in \{0, 1\}$ and outcome $Y_{s,i}^{(k)} \in \mathbb{R}$, respectively. Throughout, the *sub*population indicator s is in the *sub*script, and the stage number k is in the superscript. We assume that

conditioned on $A_{s,i}^{(k)} = a$, the outcome $Y_{s,i}^{(k)} \sim N(\mu_{sa}, \sigma_{sa}^2)$ and is independent of all other participant data. Let $\boldsymbol{\sigma}^2 = (\sigma_{10}^2, \sigma_{11}^2, \sigma_{20}^2, \sigma_{21}^2)$, which we assume is known. Let $X^{(k)}$ denote all the data from stage k, and let $X = X^{(1)} \cup X^{(2)}$ denote the cumulative data at the end of stage 2. Let $\mathcal{X}^{(k)}$ and $\mathcal{X}$ denote the sample spaces corresponding to $X^{(k)}$ and X, respectively.

Denote the population average treatment effect for each subpopulation $s \in \{1, 2\}$ by $\Delta_s = \mu_{s1} - \mu_{s0}$, and for the combined population by $\Delta_C = p_1\Delta_1 + p_2\Delta_2$. Let $\boldsymbol{\Delta} = (\Delta_1, \Delta_2)$. Define H_{01}, H_{02}, H_{0C}, to be the null hypotheses of no average treatment benefit in subpopulation 1, subpopulation 2, and the combined population, respectively, i.e.,

$$H_{01} : \Delta_1 \leq 0; \qquad H_{02} : \Delta_2 \leq 0; \qquad H_{0C} : \Delta_C \leq 0.$$

Let $\mathcal{H} = \{H_{01}, H_{02}, H_{0C}\}$, and let $\mathcal{S}$ denote the power set of $\mathcal{H}$. For any $\boldsymbol{\Delta} \in \mathbb{R}^2$, define $\mathcal{H}_{\text{TRUE}}(\boldsymbol{\Delta})$ to be the set of true null hypotheses at $\boldsymbol{\Delta}$. For each $s \in \{1, 2\}$, this set contains H_{0s} if $\Delta_s \leq 0$; it contains H_{0C} if $p_1\Delta_1 + p_2\Delta_2 \leq 0$.

2.1.2 Two-Stage Adaptive Enrichment Designs

In stage 1, $n_s^{(1)}$ participants are enrolled from each subpopulation s. At the interim analysis following stage 1, a decision rule D determines the number of participants to enroll from each subpopulation in stage 2. This decision is based on the data from stage 1, and there are $K < \infty$ possible decisions denoted by $\mathcal{D} = \{1, \ldots, K\}$. At the end of stage 2, a multiple testing procedure M determines which subset (if any) of the null hypotheses to reject, based on the data from stages 1 and 2. A two stage adaptive enrichment design is defined by the following quantities, which must be specified before the trial starts:

i. The stage 1 sample sizes $n_1^{(1)}, n_2^{(1)}$ for subpopulations 1 and 2, respectively.

ii. The number K of possible stage 2 decisions, and for each decision $d \in \mathcal{D} = \{1, \ldots, K\}$ the stage 2 sample sizes $n_1^{(2),d}, n_2^{(2),d}$ for subpopulations 1 and 2, respectively.

iii. A decision rule D mapping the stage 1 data $X^{(1)}$ to an enrollment decision in $\mathcal{D}$.

iv. A multiple testing procedure M mapping the stage 1 and 2 data X to a set of hypotheses $H \subseteq \mathcal{H}$ to reject.

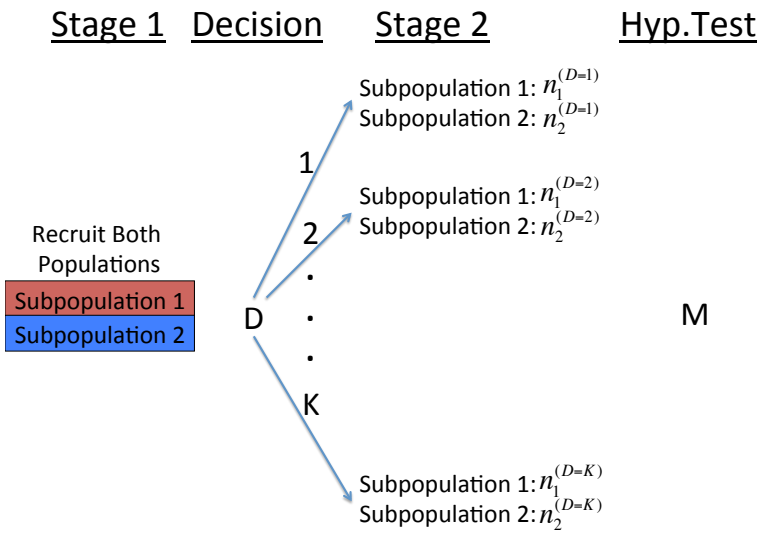
Define an adaptive design template to be the quantities defined in (i)-(ii), i.e., the set of possible decisions and corresponding sample sizes $\mathbf{n} = (\mathcal{D}, n_1^{(1)}, n_2^{(1)}, \{n_1^{(2),d}, n_2^{(2),d}\}_{d\in\mathcal{D}})$. A general adaptive design template is displayed in Figure 2.1a. A specific example of an adaptive design template is given in Figure 2.1b for the case of $p_1 = 1/2$. In this example, for a given $n > 0$, the stage 1 sample sizes satisfy $n_1^{(1)} = n_2^{(1)} = n/4$. There are four choices for stage 2 enrollment: $D = 1$: stop the trial, i.e., $n_1^{(2),1} = n_2^{(2),1} = 0$; $D = 2$: enroll exactly as in stage 1, i.e., $n_1^{(2),2} = n_2^{(2),2} = n/4$; $D = 3$: only enroll from subpopulation 1, i.e., $n_1^{(2),3} = 3n/4, n_2^{(2),3} = 0$; $D = 4$: only enroll from subpopulation 2, i.e., $n_1^{(2),4} = 0, n_2^{(2),4} = 3n/4$.

For a given adaptive design template $\mathbf{n}$, we aim to simultaneously optimize the decision rule D and multiple testing procedure M, in the sense defined in Section 2.1.3. The only constraints on D and M are that they are measurable functions. Let $\mathcal{E}^*$ denote the class of all measurable functions from the sample space $\mathcal{X}^{(1)}$ to $\mathcal{D}$, and let $\mathcal{M}^*$ denote the class of all measurable functions from the sample space and decision $\mathcal{X} \times \mathcal{D}$ to the power set $\mathcal{S}$ of null hypotheses. For given values of $(\mathbf{n}, \boldsymbol{\sigma}, D, M)$, let $P_{\boldsymbol{\Delta}}$ denote the corresponding distribution of X and and let $E_{\boldsymbol{\Delta}}$ denote expectation with respect to this distribution.

2.1.3 General Optimization Problem

The quantity to be minimized, called the objective function, is defined in terms of a loss function L and a distribution Λ on the alternatives $\boldsymbol{\Delta}$. The loss function and distribution are set by the user to determine the quantity of interest to be optimized, e.g., these can be chosen to represent expected sample size and/or power as described below. We allow the loss function L to be any bounded, integrable function of the treatment effect $\boldsymbol{\Delta}$, the enrollment decision D, and the set of hypotheses rejected M. For a given loss function L, the risk at

a.



b.

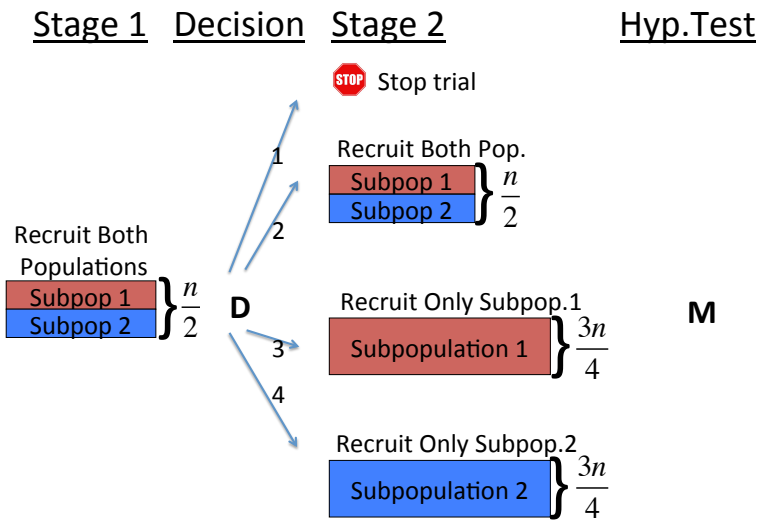


Figure 2.1: (a) Adaptive enrichment design template; (b) Example of adaptive enrichment design.

treatment effect vector $\boldsymbol{\Delta} \in \mathbb{R}^2$ is defined as $R_L(\boldsymbol{\Delta}) = E_{\boldsymbol{\Delta}} L[M\{X, D(X^{(1)})\}, D(X^{(1)}), \boldsymbol{\Delta}]$. The objective function is the Bayes risk $\int R_L(\boldsymbol{\Delta}) d\Lambda(\boldsymbol{\Delta})$. Below, for clarity of notation, we write D for $D(X^{(1)})$ and M for $M\{X, D(X^{(1)})\}$.

The above definition of the objective function allows us to select different quantities of interest (or weighted combinations of these quantities) to optimize, e.g., power, expected sample size, or expected number assigned to an ineffective treatment. For example, the loss function could be set equal to the total sample size $L^{\text{SS}} = n_1^{(1)} + n_2^{(1)} + n_1^{(2),D} + n_2^{(2),D}$; the corresponding risk at $\boldsymbol{\Delta} \in \mathbb{R}^2$ equals the expected sample size of the trial when the treatment effect is $\boldsymbol{\Delta}$. Alternatively, we could encode power to reject different null hypotheses using the following loss functions:

$$\text{For each } s \in \{1, 2\}, L^{(s)} = 1[H_{0s} \notin M; \Delta_s \geq \Delta^{\min}];$$
$$L^{(C)} = 1[H_{0C} \notin M, \Delta_1 \geq \Delta^{\min}, \Delta_2 \geq \Delta^{\min}],$$

where $\Delta^{\min}$ represents the minimum, clinically meaningful treatment effect, which is user-specified. The reason we put the constraint $\Delta_s \geq \Delta^{\min}$ in the loss function $L^{(s)}$ is that we only want to penalize for failing to reject H_{0s} when in truth the treatment effect for subpopulation s is above the clinically meaningful level. For each subpopulation $s \in \{1, 2\}$, if the treatment effect Δ_s equals or exceeds the minimum level $\Delta^{\min}$, then the risk $R_{L^{(s)}}(\boldsymbol{\Delta})$ equals one minus the power to reject H_{0s}. Similarly, if both treatment effects Δ_1, Δ_2 equal or exceed the minimum level, the risk $R_{L^{(C)}}(\boldsymbol{\Delta})$ equals one minus the power to reject H_{0C}. In either case, minimizing risk corresponds to maximizing power.

We aim to minimize the Bayes risk, i.e., the risk integrated with respect to a distribution Λ on the alternatives $\boldsymbol{\Delta}$. For example, we could let Λ denote a weighted sum of the four point masses in the set $Q = \{(0, 0), (\Delta^{\min}, 0), (0, \Delta^{\min}), (\Delta^{\min}, \Delta^{\min})\}$, which correspond to no treatment effect, only subpopulation 1 benefiting at the minimum level, only subpopulation 2 benefiting at the minimum level, and both subpopulations benefiting at the minimum

level, respectively. Let Λ^{pm} denote this distribution with weight 1/4 on each point mass. Let Λ^{mix} denote a mixture of four normal distributions, with one centered at each of the aforementioned point masses, and each having variance σ_Λ^2. Then the Bayes risk corresponding to the pair $(L, \Lambda) = (L^{\text{SS}}, \Lambda^{\text{pm}})$ is the expected sample size under $\boldsymbol{\Delta}$, averaged over the four scenarios $\boldsymbol{\Delta} \in Q$. As another example, the Bayes risk corresponding to the pair $(L, \Lambda) = (a_1 L^{(1)} + a_2 L^{(2)}, \Lambda^{\text{mix}})$ for positive constants a_1, a_2 is the weighted sum of 1 minus the power to reject each H_{0s} when the corresponding treatment effect exceeds the minimum level, integrated over the distribution $\Lambda^{\text{mix}}(\boldsymbol{\Delta})$.

Our optimization problem has two types of constraints. The first are familywise Type I error constraints, and the second are additional constraints involving J triples $(L_j, \Lambda_j, \beta_j)$ of loss function L_j, distribution $\Lambda_j(\boldsymbol{\Delta})$, and threshold $\beta_j \in \mathbb{R}$ defined below.

Constrained Bayes Optimization Problem: For given $\mathbf{n}$, $\alpha > 0$, $\boldsymbol{\sigma}^2$, $\{(L_j, \Lambda_j, \beta_j) : j = 0, \ldots, J\}$, find the adaptive enrichment design $(D, M) \in (\mathcal{E}^* \times \mathcal{M}^*)$ minimizing

$$\int E_{\boldsymbol{\Delta}} \left(L_0[M\{X, D(X^{(1)})\}, D(X^{(1)}), \boldsymbol{\Delta}] \right) d\Lambda_0(\boldsymbol{\Delta}), \tag{2.1.1}$$

under the familywise Type I error constraints: for any $\boldsymbol{\Delta} \in \mathbb{R}^2$,

$$P_{\boldsymbol{\Delta}} \{M \text{ rejects any null hypotheses in } \mathcal{H}_{\text{TRUE}}(\boldsymbol{\Delta})\} \leq \alpha, \tag{2.1.2}$$

and additional constraints: for each $j \in \{1, \ldots, J\}$:

$$\int E_{\boldsymbol{\Delta}} \left(L_j[M\{X, D(X^{(1)})\}, D(X^{(1)}), \boldsymbol{\Delta}] \right) d\Lambda_j(\boldsymbol{\Delta}) \leq \beta_j. \tag{2.1.3}$$

First, consider the case where $J = 0$, i.e., there are no additional constraints (2.1.3). Then the constrained Bayes optimization problem is to minimize the Bayes risk subject to

strong control on the familywise Type I error rate at level α. For example, one can optimize power in the sense described above by setting $(L_0, \Lambda_0) = (a_1 L^{(1)} + a_2 L^{(2)}, \Lambda^{\text{mix}})$.

2.1.4 Example of Optimization Problem

The additional constraints (2.1.3) allow the user to define a broader set of problems, such as optimizing expected sample size subject to power and Type I error constraints. We consider two types of priors in this work. Specifically, as discussed in the following two examples, the first example considers a prior of four point masses, and the second considers a prior of Gaussian mixture.

Example 2.1.1. Consider the problem of minimizing expected sample size averaged over the four point masses in Q, under the Type I error constraints (2.1.2) and the following power constraints for given Type II error $\beta > 0$:

P1. At $(\Delta_1, \Delta_2) = (\Delta^{\min}, 0)$, the power to reject H_{01} is at least $1 - \beta$.

P2. At $(\Delta_1, \Delta_2) = (0, \Delta^{\min})$, the power to reject H_{02} is at least $1 - \beta$.

P3. At $(\Delta_1, \Delta_2) = (\Delta^{\min}, \Delta^{\min})$, the power to reject H_{0C} is at least $1 - \beta$.

This problem can be represented by setting $(L_0, \Lambda_0) = (L^{\text{SS}}, \Lambda^{\text{pm}})$ and $J = 3$ additional constraints of the form $(L_j, \Lambda_j, \beta_j)$ equal to

$$(L^{(1)}, \mathbf{1}(\Delta^{\min}, 0), \beta); \qquad (L^{(2)}, \mathbf{1}(0, \Delta^{\min}), \beta); \qquad (L^{(C)}, \mathbf{1}(\Delta^{\min}, \Delta^{\min}), \beta),$$

where $1(x, y)$ denotes a point mass at $\boldsymbol{\Delta} = (x, y)$. We solve this problem in Section 2.4.

Example 2.1.2. Under the Type I error constraints (2.1.2) and the same power constraints for given Type II error $\beta > 0$ as discussed in the previous example, consider the problem of minimizing the expected sample size averaged over a mixture of four Gaussian components.

Namely, let

$$\Lambda = \sum_{k=1}^{4} \pi_k N(\boldsymbol{\mu}_k, \sigma^2 \cdot \mathbf{I}_2),$$

where $\sum_{k=1}^{4} \pi_k = 1$ and $\boldsymbol{\mu}_1 = (0,0)^T$, $\boldsymbol{\mu}_2 = (0, \Delta^{\min})^T$, $\boldsymbol{\mu}_3 = (\Delta^{\min}, 0)^T$, $\boldsymbol{\mu}_4 = (\Delta^{\min}, \Delta^{\min})^T$, and $\sigma = \Delta^{\min}$.

2.2 Reducing Problem Complexity through Minimal Sufficient Statistics

We show that it suffices to consider decision rules D and multiple testing procedures M that depend only on minimal sufficient statistics. This dramatically reduces the problem complexity from having to search over arbitrarily complex functions of the data X, to the easier (but still very challenging) problem of searching over functions of the 2-dimensional sufficient statistics at each stage. Let $N_s^{(k)}$ denote the number enrolled from subpopulation $s \in \{1,2\}$ during stage $k \in \{1,2\}$. The stage 1 sample sizes are set in advance, while the stage 2 sample sizes are functions of the stage 1 data; specifically, $N_s^{(1)} = n_s^{(1)}$ and $N_s^{(2)} = n_s^{(2),D(X^{(1)})}$ for each $s \in \{1,2\}$.

For each subpopulation $s \in \{1,2\}$ and stage $k \in \{1,2\}$, define the corresponding z-statistic as

$$Z_s^{(k)} = \left\{ \frac{\sum_{i=1}^{N_s^{(k)}} Y_{s,i}^{(k)} A_{s,i}^{(k)}}{\sum_{i=1}^{N_s^{(k)}} A_{s,i}^{(k)}} - \frac{\sum_{i=1}^{N_s^{(k)}} Y_{s,i}^{(k)} (1 - A_{s,i}^{(k)})}{\sum_{i=1}^{N_s^{(k)}} (1 - A_{s,i}^{(k)})} \right\} \left\{ \frac{\sigma_{s1}^2 + \sigma_{s0}^2}{N_s^{(k)}/2} \right\}^{-1/2}, \tag{2.2.1}$$

where the quantity inside curly braces on the right is the variance of the difference between sample means on the left. Let $\mathbf{Z}^{(k)} = (Z_1^{(k)}, Z_2^{(k)})$. Define the final (cumulative) z-statistic

based on all stage 1 and 2 data for subpopulation s by

$$Z_s^{(F)} = \left\{ \frac{\sum_{k=1}^{2} \sum_{i=1}^{N_s^{(k)}} Y_{s,i}^{(k)} A_{s,i}^{(k)}}{\sum_{k=1}^{2} \sum_{i=1}^{N_s^{(k)}} A_{s,i}^{(k)}} - \frac{\sum_{k=1}^{2} \sum_{i=1}^{N_s^{(k)}} Y_{s,i}^{(k)} (1 - A_{s,i}^{(k)})}{\sum_{k=1}^{2} \sum_{i=1}^{N_s^{(k)}} (1 - A_{s,i}^{(k)})} \right\} \left\{ \frac{\sigma_{s1}^2 + \sigma_{s0}^2}{(N_s^{(1)} + N_s^{(2)})/2} \right\}^{-1/2}. \tag{2.2.2}$$

Let $\mathbf{Z}^{(1)} = (Z_1^{(1)}, Z_2^{(1)})^T$ and $\mathbf{Z}^{(F)} = (Z_1^{(F)}, Z_2^{(F)})^T$. The distribution of $\mathbf{Z} = (\mathbf{Z}^{(1)T}, \mathbf{Z}^{(F)T})$ is characterized as follows:

a. $\mathbf{Z}^{(1)}$ is bivariate normal with mean vector $\left(\Delta_1 \left\{ \frac{n_1^{(1)}}{2\left(\sigma_{11}^2 + \sigma_{10}^2\right)} \right\}^{1/2}, \Delta_2 \left\{ \frac{n_2^{(1)}}{2\left(\sigma_{21}^2 + \sigma_{20}^2\right)} \right\}^{1/2} \right)^T$ and covariance matrix I_2, i.e., the 2×2 identity matrix.

b. $\mathbf{Z}^{(2)}$, which uses only stage 2 data, is conditionally independent of $\mathbf{Z}^{(1)}$ given the decision $D(\mathbf{Z}^{(1)}, U_1)$. The conditional distribution of $\mathbf{Z}^{(2)}$ given $D(\mathbf{Z}^{(1)}, U_1) = d$ is bivariate normal with mean vector $\left(\Delta_1 \left\{ \frac{n_1^{(2),d}}{2\left(\sigma_{11}^2 + \sigma_{10}^2\right)} \right\}^{1/2}, \Delta_2 \left\{ \frac{n_2^{(2),d}}{2\left(\sigma_{21}^2 + \sigma_{20}^2\right)} \right\}^{1/2} \right)^T$ and covariance matrix I_2.

c. For each subpopulation $s \in \{1, 2\}$, for $D = D(\mathbf{Z}^{(1)}, U_1)$, we have the following relationship between the final (cumulative) z-statistic and the stagewise z-statistics:

$$Z_s^{(F)} = \left\{ \frac{n_s^{(1)}}{n_s^{(1)} + n_s^{(2),D}} \right\}^{1/2} Z_s^{(1)} + \left\{ \frac{n_s^{(2),D}}{n_s^{(1)} + n_s^{(2),D}} \right\}^{1/2} Z_s^{(2)}. \tag{2.2.3}$$

We show that it suffices to consider decision rules D that depend on the data only through $\mathbf{Z}^{(1)}$, and multiple testing procedures M that depend on the data only through $\mathbf{Z}^{(F)}$ and the decision D. We consider randomized decision rules and multiple testing procedures, i.e., we allow D and M to additionally take as input U_1 and U_2, respectively, which are independent, uniform random variables. For conciseness, we refer to "randomized decision rules" as "decision rules", and refer to "randomized multiple testing procedures" as "multiple testing procedures."

Let $\mathcal{E}$ denote the class of all measurable functions D from $\mathbb{R}^2 \times [0,1]$ $\big($representing all possible values of $(\mathbf{Z}^{(1)}, U_1)\big)$ to the set of stage 2 enrollment decisions $\mathcal{D}$. Let $\mathcal{M}$ denote the class of all measurable functions from $\mathbb{R}^2 \times \mathcal{D} \times [0,1]$ (representing all possible values of $\mathbf{Z}^{(F)}, D(\mathbf{Z}^{(1)}, U_1), U_2)$ to $\mathcal{S}$ (indicating the subset of null hypotheses rejected). For conciseness, we let $D = D(\mathbf{Z}^{(1)}, U_1)$ and $M = M\{\mathbf{Z}^{(F)}, D(\mathbf{Z}^{(1)}, U_1), U_2\}$. Define the class of adaptive enrichment designs $\mathcal{A} = \{(D, M) : D \in \mathcal{E}, M \in \mathcal{M}\}$. We prove the following theorem in Section A.4 of the Appendix:

Theorem 2.2.1. If the constrained Bayes optimization problem in Section 2.1.3 is feasible, then there exists an optimal solution $(D, M) \in (\mathcal{E} \times \mathcal{M})$, i.e., for which D depends on the data only through $\mathbf{Z}^{(1)}$, and M depends on the data only through $\mathbf{Z}^{(F)}$ and the decision $D(\mathbf{Z}^{(1)}, U_1)$.

2.3 Transformation of Bayes Optimization Problem into Sparse Linear Program

2.3.1 Discretization of Constrained Bayes Optimization Problem

Even after simplifying the constrained Bayes optimization problem by using only minimal sufficient statistics as in the previous section, the problem is still extremely difficult or impossible to solve directly. This is because the optimization problem is nonconvex, involves infinitely many familywise Type I error constraints (2.1.2), and optimizes over the very large class of decision rules $\mathcal{E}$ and multiple testing procedures $\mathcal{M}$. We propose a novel approach to solve this problem, involving three steps. We first discretize the decision rule and multiple testing procedure, and restrict to a finite subset of familywise Type I error constraints. The resulting problem is still nonconvex, and so is extremely difficult to solve. Step two involves reparametrizing this problem so that it can be represented as a sparse, linear program, a

class of problems that is much easier to solve than nonconvex problems. The third step is to apply advanced optimization methods to solve the sparse, linear program.

The first of the above steps is to discretize the constrained Bayes optimization problem. The decision rule D is discretized by partitioning $\mathbb{R}^2$ into a finite set of rectangles as described below. The intuition for what follows is that we restrict to the subclass of adaptive designs $(D, M) \in (\mathcal{E} \times \mathcal{M})$ such that the following hold: D makes the same decision when the first stage statistics $\mathbf{Z}^{(1)}$ are anywhere within a small rectangle $r \subseteq \mathbb{R}^2$; M rejects the same set of null hypotheses when the first stage statistics $\mathbf{Z}^{(1)}$ are in a rectangle $r \subseteq \mathbb{R}^2$, the enrollment decision is $d \in \mathcal{D}$, and the final statistics $\mathbf{Z}^{(F)}$ are in a rectangle r'. For a fine enough partition of rectangles, we expect the solution to the corresponding discretized optimization problem to be close to that of the original problem.

We consider partitions of $\mathbb{R}^2$ into rectangles. One way to construct such a partition is to start with a box $B = [-b, b] \times [-b, b]$, for a given integer $b > 0$. Let $\tau = (\tau_1, \tau_2)$ be such that b/τ_s is an integer for each $s \in \{1, 2\}$. For each $j, j' \in \mathbb{Z}$, define the rectangle $R_{j,j'} = [j\tau_1, (j+1)\tau_1) \times [j'\tau_2, (j'+1)\tau_2)$. Let $\mathcal{R}_{\mathcal{B}}$ denote the set of such rectangles in the bounded region B, i.e., $\{R_{j,j'} : j, j' \in \mathbb{Z}, R_{j,j'} \subset B\}$. Define the following partition of $\mathbb{R}^2$: $\mathcal{R} = \mathcal{R}_{\mathcal{B}} \cup \{\mathbb{R}^2 \setminus B\}$. Though $\mathbb{R}^2 \setminus B$ is not a rectangle, we still refer to $\mathcal{R}$ as a partition of rectangles, with a slight abuse of notation.

Let $\mathcal{R}_{\text{dec}}$ denote a partition of $\mathbb{R}^2$ into rectangles. We restrict to the subclass of decision rules D with the following property: for any rectangle $r \in \mathcal{R}_{\text{dec}}$ and $u \in [0, 1]$,

$$D(\mathbf{Z}^{(1)}, u) = D(\mathbf{Z}^{(1)'}, u) \text{ whenever } \mathbf{Z}^{(1)} \text{ and } \mathbf{Z}^{(1)'} \text{ are both in } r. \tag{2.3.1}$$

That is, the decision rule only depends on the data through the rectangle that the first stage z-statistics are in.

For each $d \in \mathcal{D}$, let $\mathcal{R}_{\text{mtp},d}$ denote a partition of $\mathbb{R}^2$ into rectangles. Intuitively, we will restrict to multiple testing procedures M that only depend on the data through the

enrollment decision D and the rectangles that the first stage and cumulative statistics are in, respectively. Let $\mathbf{Z} = (\mathbf{Z}^{(1)T}, \mathbf{Z}^{(F)T})^T$ and $\mathbf{Z}' = (\mathbf{Z}^{(1)'}, \mathbf{Z}^{(F)'})^T$, for any $\mathbf{Z}^{(1)}, \mathbf{Z}^{(F)}, \mathbf{Z}^{(1)'}, \mathbf{Z}^{(F)'} \in \mathbb{R}^2$. We restrict to the subclass of multiple testing procedures M such that for any $r \in \mathcal{R}_{\text{dec}}$, $d \in \mathcal{D}$, $r' \in \mathcal{R}_{\text{mtp},d}$, and $u_1, u_2 \in [0,1]$, we have $M(\mathbf{Z}, d, u_2) = M(\mathbf{Z}', d, u_2)$ whenever all of the following hold: $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(1)'}$ are both in r, $D(\mathbf{Z}^{(1)}, u_1) = D(\mathbf{Z}^{(1)'}, u_1)$, and $\mathbf{Z}^C$ and $\mathbf{Z}^{C'}$ are both in r'.

For each $r \in \mathcal{R}_{\text{dec}}$ and $d \in \mathcal{D}$, define x_{rd} to be the probability that decision d is made conditioned on $\mathbf{Z}^{(1)} \in r$, i.e.,

$$x_{rd} = P\left\{D(\mathbf{Z}^{(1)}, U_1) = d | \mathbf{Z}^{(1)} \in r\right\}. \tag{2.3.2}$$

For each $r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{D}, r' \in \mathcal{R}_{\text{mtp},d}, s \in \mathcal{S}$, define $y_{rdr's}$ to be the probability that precisely the subset s is rejected conditioned on $\mathbf{Z}^{(1)} \in r$, $D(\mathbf{Z}^{(1)}, U_1) = d, \mathbf{Z}^{(F)} \in r'$, i.e.,

$$y_{rdr's} = P\left\{M(\mathbf{Z}, d, U_2) = s | \mathbf{Z}^{(1)} \in r, D(\mathbf{Z}^{(1)}, U_1) = d, \mathbf{Z}^{(F)} \in r'\right\}. \tag{2.3.3}$$

The values of all variables x_{rd} and $y_{rdr's}$ are specified by the study designer before the trial, and our goal is to optimize the corresponding Bayes risk under familywise Type I error constraints and the additional constraints (2.1.3).

The probability of rejecting precisely the subset $s \in \mathcal{S}$ at a given vector of population parameters $\boldsymbol{\Delta} = (\Delta_1, \Delta_2) \in \mathbb{R}^2$ is

$$
\begin{aligned}
& P_{\boldsymbol{\Delta}}\left\{M(\mathbf{Z}, D(\mathbf{Z}^{(1)}, U_1), U_2) = s\right\} && (2.3.4)\\
= & \sum_{r,d,r'} P_{\boldsymbol{\Delta}}\left\{\mathbf{Z}^{(1)} \in r, D(\mathbf{Z}^{(1)}, U_1) = d, \mathbf{Z}^{(F)} \in r', M(\mathbf{Z}, d, U_2) = s\right\} \\
= & \sum_{r,d,r'} \left[P_{\boldsymbol{\Delta}}\left\{M(\mathbf{Z}, d, U_2) = s | \mathbf{Z}^{(F)} \in r', \mathbf{Z}^{(1)} \in r, D(\mathbf{Z}^{(1)}, U_1) = d\right\} \times \right. \\
& \left. P_{\boldsymbol{\Delta}}\left\{\mathbf{Z}^{(F)} \in r' | \mathbf{Z}^{(1)} \in r, D(\mathbf{Z}^{(1)}, U_1) = d\right\} P_{\boldsymbol{\Delta}}\left\{D(\mathbf{Z}^{(1)}, U_1) = d | \mathbf{Z}^{(1)} \in r\right\} P_{\boldsymbol{\Delta}}\left\{\mathbf{Z}^{(1)} \in r\right\}\right] \\
= & \sum_{r,d,r'} x_{rd} y_{rdr's} p(\boldsymbol{\Delta}, r, d, r'), && (2.3.5)
\end{aligned}
$$

where (2.3.5) follows from (2.3.2) and (2.3.3), and where we define

$$
p(\boldsymbol{\Delta}, r, d, r') = P_{\boldsymbol{\Delta}}\left\{\mathbf{Z}^{(F)} \in r' | \mathbf{Z}^{(1)} \in r, D(\mathbf{Z}^{(1)}, U_1) = d\right\} P_{\boldsymbol{\Delta}}\left\{\mathbf{Z}^{(1)} \in r\right\}. \qquad (2.3.6)
$$

The value of $p(\boldsymbol{\Delta}, r, d, r')$ does not depend on D, which follows from (2.3.1). This value can be computed to high precision using the multivariate normal distribution function and (a)-(c) from Section 2.1.2.

We can express the objective function (2.1.1) of the constrained Bayes optimization problem in terms of the variables $x_{rd} y_{rdr's}$, since the expectation inside the integral in (2.1.1) satisfies

$$
E_{\boldsymbol{\Delta}}\left\{L(M(\mathbf{Z}, D(\mathbf{Z}^{(1)}, U_1), U_2); \Delta_1, \Delta_2)\right\} = \sum_{s \in \mathcal{S}} \sum_{r,d,r'} x_{rd} y_{rdr's} \left\{L(s; \Delta_1, \Delta_2) p(\boldsymbol{\Delta}, r, d, r')\right\}.
$$

where the second line follows from the equality of (2.3.4) and (2.3.5). The familywise Type I error constraints and additional constraints (2.1.3) can similarly be expressed as a function of $x_{rd} y_{rdr's}$, as shown in Section A.1 of the Appendix.

Let $G \subset \mathbb{R}^2$ denote a discretization of the boundaries of the null spaces of the hypotheses of interest, The discretized version of the Constrained Bayes Optimization Problem above is as follows:

Discretized Problem:

$$\min \sum_{r,d,r',s} x_{rd} y_{rdr's} \int L_0(s;\boldsymbol{\Delta}) p(\boldsymbol{\Delta}, r, d, r') d\Lambda_0(\boldsymbol{\Delta}) \tag{2.3.7}$$

under the following constraints:

$$\text{for each } \boldsymbol{\Delta} \in G, \sum_{r,d,r'} \sum_{s \in \mathcal{S}: s \cap \mathcal{H}_{TRUE}(\boldsymbol{\Delta}) \neq \emptyset} x_{rd} y_{rdr's} p(\boldsymbol{\Delta}, r, d, r') \leq \alpha; \tag{2.3.8}$$

$$\text{for each } j \in \{1, \ldots, J\}, \sum_{r,d,r',s} x_{rd} y_{rdr's} \int L_j(s;\boldsymbol{\Delta}) p(\boldsymbol{\Delta}, r, d, r') d\Lambda_j(\boldsymbol{\Delta}) \leq \beta_j; \tag{2.3.9}$$

$$\text{for each } r \in \mathcal{R}_{\text{dec}}, \sum_{d \in \mathcal{D}} x_{rd} = 1; \tag{2.3.10}$$

$$\text{for each } r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{D}, r' \in \mathcal{R}_{\text{mtp},d}, \sum_{s \in \mathcal{S}} y_{rdr's} = 1; \tag{2.3.11}$$

$$\text{for each } r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{D}, r' \in \mathcal{R}_{\text{mtp},d}, s \in \mathcal{S} : x_{rd} \geq 0, y_{rdr's} \geq 0. \tag{2.3.12}$$

The sum $\sum_{r,d,r',s}$, which appers in (2.3.7) and (2.3.9), is taken over $r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{D}, r' \in \mathcal{R}_{\text{mtp},d}, s \in \mathcal{S}$. The objective function (2.3.7) represents (2.1.1). The constraints (2.3.8) and (2.3.9) represent the familywise Type I error constraints (2.1.2) and additional constraints (2.1.3), respectively. The remaining constraints encode properties of x_{rd} and $y_{rdr's}$ that follow from their definitions (2.3.2)-(2.3.3) as conditional probabilities. Specifically, the constraints (2.3.10) and (2.3.11) follow from the law of total probability; the constraints (2.3.12) encode that each variable must be nonnegative since it represents a probability.

2.3.2 Transformation of (Nonconvex) Discretized Problem into Sparse Linear Program

The discretized problem from Section 2.3 is not linear (and not convex) in the variables $\{x_{rd}, y_{rdr's}\}$. Therefore, this problem is generally computationally intractable to solve, since

only ad hoc methods exist for solving nonconvex optimization problems and even if a local minimum is found there is no general way to determine if it is the global minimum. We transform this problem into a sparse, linear program by defining the new variables:

$$v_{rdr's} = x_{rd} y_{rdr's}, \text{ for all } r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{D}, r' \in \mathcal{R}_{\text{mtp},d}, s \in \mathcal{S}. \tag{2.3.13}$$

The objective function (2.3.7) and familywise Type I error constraints (2.3.8) are linear functions of $v_{rdr's}$. We prove in Section A.2 of the Appendix that the constraints (2.3.10)-(2.3.12) can be equivalently expressed in terms of the linear constraints (2.3.17)-(2.3.19) on $v_{rdr's}$ in the following linear program, where for each $d \in \mathcal{D}$, we let r'_d an arbitrary element in the set $\mathcal{R}_{\text{mtp},d}$ (say, the first element under a fixed ordering of $\mathcal{R}_{\text{mtp},d}$):

Sparse linear program:

$$\min \sum_{r,d,r',s} v_{rdr's} \int L_0(s;\boldsymbol{\Delta}) p(\boldsymbol{\Delta}, r, d, r') d\Lambda_0(\boldsymbol{\Delta}) \tag{2.3.14}$$

under the constraints:

$$\text{for each } \boldsymbol{\Delta} \in G, \sum_{s \in \mathcal{S}: s \cap \mathcal{H}_{TRUE}(\boldsymbol{\Delta}) \neq \emptyset} \sum_{r,d,r'} v_{rdr's} p(\boldsymbol{\Delta}, r, d, r') \leq \alpha; \tag{2.3.15}$$

$$\text{for each } j \in \{1, \ldots, J\}, \sum_{r,d,r',s} v_{rdr's} \int L_j(s;\boldsymbol{\Delta}) p(\boldsymbol{\Delta}, r, d, r') d\Lambda_j(\boldsymbol{\Delta}) \leq \beta_j; \tag{2.3.16}$$

$$\text{for each } r \in \mathcal{R}_{\text{dec}}, \sum_{d} \sum_{s \in \mathcal{S}} v_{rdr'_d s} = 1; \tag{2.3.17}$$

$$\text{for each } r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{D}, \widetilde{r}' \in \mathcal{R}_{\text{mtp},d}, \sum_{s \in \mathcal{S}} v_{rdr'_d s} = \sum_{s \in \mathcal{S}} v_{rd\widetilde{r}'s}; \tag{2.3.18}$$

$$\text{for each } r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{D}, r' \in \mathcal{R}_{\text{mtp},d}, s \in \mathcal{S} : v_{rdr's} \geq 0. \tag{2.3.19}$$

We prove the following theorem in Section A.3 of the Appendix:

Theorem 2.3.1. i. (Equivalence of discretized problem and sparse linear program) The optimum value of the above optimization problem equals the optimum value of the discretized

problem from Section 2.3.1.

ii. (Map from solution of sparse linear program to solution of discretized problem) For any optimal solution $\mathbf{v} = \{v_{rdr's}\}_{r\in\mathcal{R}_{\text{dec}},d\in\mathcal{D},r'\in\mathcal{R}_{\text{mtp},d},s\in\mathcal{S}}$ to the sparse linear program, define the vectors $\mathbf{x}, \mathbf{y}$ by the transformation:

$$x_{rd} = \sum_{s\in\mathcal{S}} v_{rdr's}; \tag{2.3.20}$$

$$y_{rdr's} = \begin{cases} v_{rdr's}/x_{rd}, & \text{if } x_{rd} > 0 \\ 1/|\mathcal{S}|, & \text{otherwise} \end{cases}. \tag{2.3.21}$$

Then $(\mathbf{x}, \mathbf{y})$ is a well-defined, feasible, and optimal solution to the discretized problem from Section 2.3.1.

2.4 Applications

In this section, we apply our proposed method to solve two important and challenging problems in adaptive trials. In particular, we first minimize the expected sample size under power and type I error constraints. Next, we maximize the power under the expected sample size and type I error constraints.

2.4.1 Minimizing Expected Sample Size under Power and Type I Error Constraints

We consider the adaptive design template in Figure 2.1b from Section 2.1.2 and the optimization problem in Section 2.1.4. Let ESS_Q denote the value of the objective function (2.1.1), which equals the expected sample size averaged over the four mass points in Q. The sample sizes $\mathbf{n}$ are a function of n as described in Section 2.1.2, where n is the total sample size if both subpopulations are enrolled during stage 2. Let $p_1 = 1/2$, $\alpha = 0.05$, and let each

σ_{sa}^2 equal a common value $\sigma^2 > 0$. Let Φ denote the cumulative distribution function of the standard normal. For comparison purposes, for $\widetilde{\beta} = 0.05$, we set

$$n = 4\sigma^2\{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\widetilde{\beta})\}^2/(\Delta^{\min})^2, \tag{2.4.1}$$

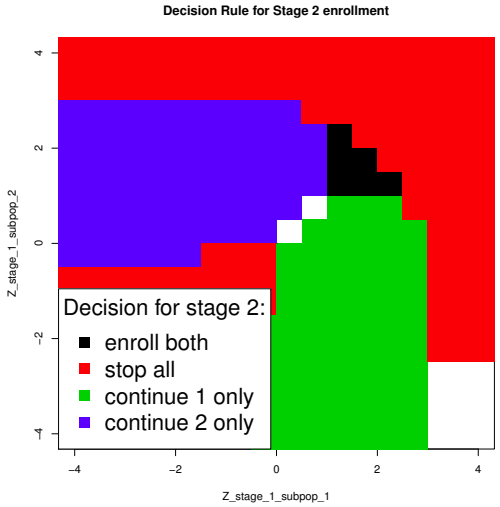
i.e., the smallest n such that in a standard (non-adaptive) design enrolling $n/2$ from each subpopulation, the uniformly most powerful test of H_{0C} at level $\alpha = 0.05$ has power $1-\widetilde{\beta} = 0.95$ at the alternative $\boldsymbol{\Delta} = (\Delta^{\min}, \Delta^{\min})$.

The optimal solution to the above constrained Bayes optimization problem is the same regardless of the choice of $(\sigma^2, \Delta^{\min})$. In brief, the reason is that by (a)-(c) in Section 2.2, both the distribution of $\mathbf{Z}^{(1)}$ and the conditional distribution of $\mathbf{Z}^{(2)}$ given $(D = d, \mathbf{Z}^{(1)})$ depend on $(n, \sigma^2, \boldsymbol{\Delta})$ only through $\Delta_1\{n/(8\sigma^2)\}^{1/2}$ and $\Delta_2\{n/(8\sigma^2)\}^{1/2}$, i.e., the non-centrality parameters for the subpopulations. For each $j \in \{0, 1, 2, 3\}$, the support of the distribution Λ_j is contained in $Q = \{(0,0), (\Delta^{\min}, 0), (0, \Delta^{\min}), (\Delta^{\min}, \Delta^{\min})\}$. Therefore, the probabilities in the objective function (2.1.1) and additional constraints (2.1.3) depend on $(n, \sigma^2, \Delta^{\min})$ only through $\Delta^{\min}\{n/(8\sigma^2)\}^{1/2}$, which equals the constant $2^{-1/2}\{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\widetilde{\beta})\}$ by (2.4.1).

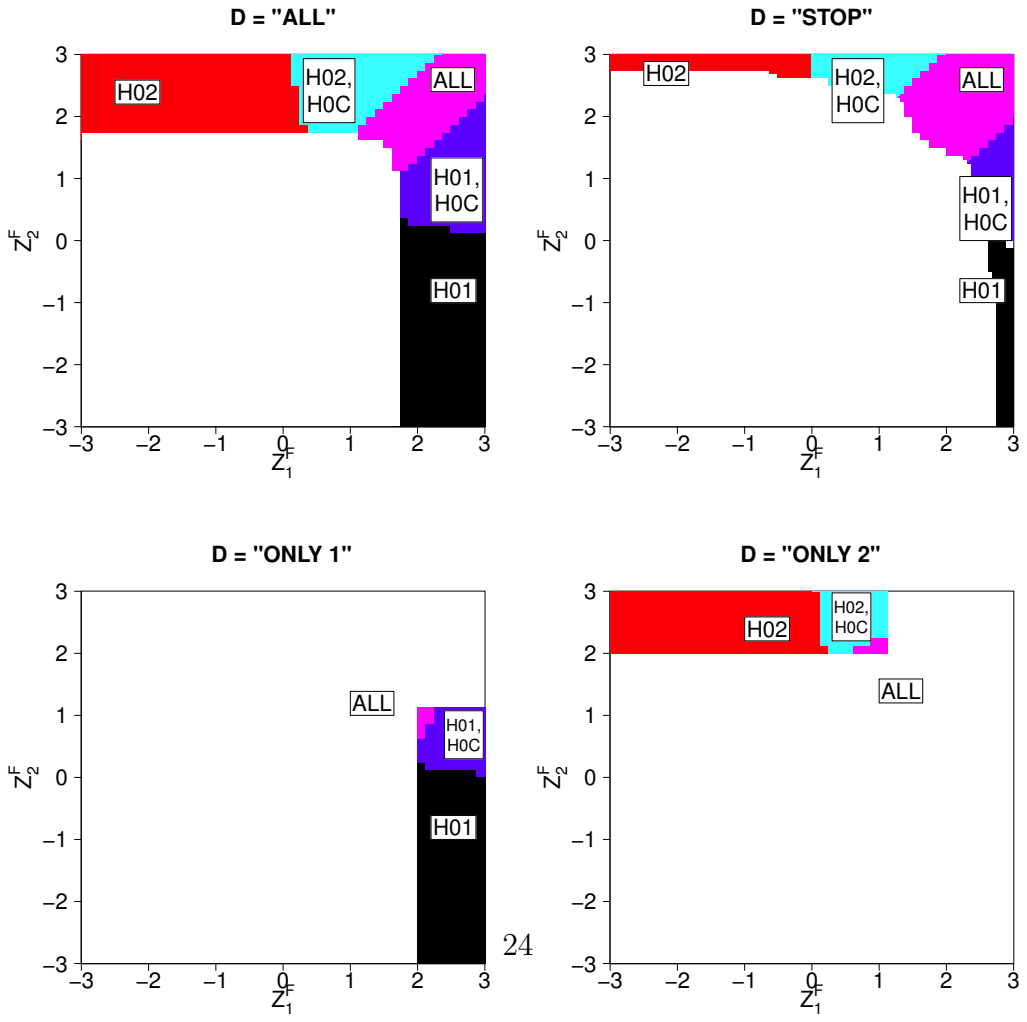
We applied the method from Section 2.3.2 to solve the above problem for each $\beta \in \{0.01, \ldots, 0.99\}$. Our results show the problem is feasible whenever $1 - \beta \leq 0.82$, and is infeasible otherwise. We focus on the case of $1 - \beta = 0.82$, and denote the corresponding decision rule and multiple testing procedure for the optimal solution by D^* and M^*, respectively. These are depicted in Figure 2.2.

Figure 2.2: Optimal Decision rule D^* and Multiple Testing Procedure M^* for Adaptive Enrichment Design Solving Optimization Problem in Section 2.4

Decision Rule D^* for Stage 2 Enrollment (z-statistics correspond to $\mathbf{Z}^{(1)}$):



Rejection Regions of M^* Corresponding to Each Possible Decision:



2.5 Comparison of Optimal Adaptive Enrichment Design Versus Design Based on P-value Combination Approach

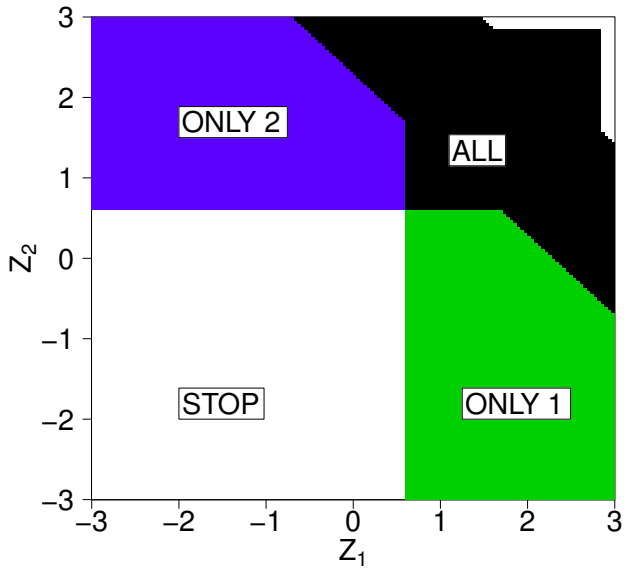
Consider the adaptive design template in Figure 2.1b from Section 2.1.2 and the optimization problem in Section 2.1.4. We apply the p-value combination approach of Bauer (1989), Bauer and Köhne (1994), Lehmacher and Wassmer (1999), with the closed testing principle of Marcus et al. (1976); this approach has been used to construct adaptive enrichment designs by, e.g., Bretz et al. (2006); Schmidli et al. (2006); Jennison and Turnbull (2007); Brannath et al. (2009b); Jenkins et al. (2011b); Boessen et al. (2013b). Multiple testing procedures M based on this approach are flexible in that they strongly control the familywise Type I error rate regardless of what decision rule D is used.

We applied the p-value combination approach to determine the multiple testing procedure M. This approach requires specifying a combination function and local tests for each intersection of null hypotheses. At each stage $k \in \{1, 2\}$, for every subset of null hypotheses $I \subseteq \mathcal{H}$, an adjusted p-value for the intersection null hypothesis $H_I = \bigcap_{H \in I} H$ is computed from the data in stage k, using the Dunnett intersection test (Dunnett, 1955; Jennison and Turnbull, 2007). P-values are then combined across stages by the weighted inverse normal rule with equal weights for each stage. Each elementary null hypothesis $H \in \mathcal{H}$ is rejected if and only if the stage 2 combined p-value is less than 0.05 for every intersection null hypothesis $\bigcap_{H' \in I} H'$ for which $H \in I$. We slightly modified this approach to incorporate early stopping for efficacy after stage 1 as in, e.g., Jennison and Turnbull (2007), using the equivalent of the boundaries of O'Brien and Fleming (1979) for the stage 1 p-values. The resulting multiple testing procedure is denoted by M^{pv}.

The p-value combination approach does not specify a corresponding decision rule D. To construct one, we consider a class of decision rules D that are functions of two thresholds

t_c and t_i, which we approximately optimize in conjunction with M^{pv} as described below. Sample sizes $\mathbf{n}$ are as in the adaptive design template in Figure 2.1b from Section 2.1.2. Define the decision rule $D^{(t_c,t_i)}(\mathbf{Z}^{(1)})$ as follows: If the multiple testing procedure M^{pv} rejects any null hypothesis at the end of stage 1, stop the trial; else, if the combined population statistic $(Z_1^{(1)} + Z_2^{(1)})/\sqrt{2} > t_c$, enroll both subpopulations in stage 2; else, enroll from each subpopulation s for which $Z_s^{(1)} > t_i$. The numerical value of D is then determined as follows: if both subpopulations are enrolled in stage 2, then $D = 2$; else, if only one subpopulation $s \in \{1, 2\}$ is enrolled in stage 2, then $D = 2 + s$; else, the trial stops at the end of stage 1, i.e., $D = 1$. An example of the decision rule $D^{(t_c,t_i)}$ is depicted in Figure 2.5.

Figure 2.3: Decision Rule $D^{(t_c,t_i)}$ for $(t_c, t_i) = (1.6, 0.6)$. (z-statistics correspond to $\mathbf{Z}^{(1)}$). This corresponds to the minimizer of ESS_Q over $\mathcal{A}$ satisfying the power constraints (P1)-(P3) at $1 - \beta = 0.74$. The white area in the upper right corner corresponds to stopping the trial at the end of stage 1.



We next define a set of adaptive enrichment designs $D^{(t_c,t_i)}$ corresponding to pairs (t_c, t_i) in a grid of values; let $\mathcal{A} = \{(D^{(t_c,t_i)}, M^{\text{pv}}) : (t_c, t_i) \in (-3, -2.9, \ldots, 3) \times (-3, -2.9, \ldots, 3)\}$.

Each design in $\mathcal{A}$ strongly controls the familywise Type I error rate at level 0.05, which is a property of the p-value combination approach. For each design in $\mathcal{A}$, we computed ESS_Q and the power to reject each subset of null hypotheses. We used the results to solve the constrained Bayes optimization problem from the previous section restricted to the set of designs $\mathcal{A}$. Specifically, for each value of $1-\beta$ in the top row of Table 2.1, we computed the smallest value of ESS_Q over all designs $\mathcal{A}$ that satisfy (P1)-(P3) at this value of β. Also, for each such value of $1-\beta$, we solved the corresponding problem over the class of designs $\mathcal{E} \times \mathcal{M}$ using the sparse linear programming method from Section 2.3.2; the minimum value of the objective function ESS_Q is given for each class of designs in the bottom rows of Table 2.1 in terms of n. At all values of $1-\beta$ we considered, the minimum value of ESS_Q was substantially lower for the optimal design among $\mathcal{E} \times \mathcal{M}$ computed based on our sparse linear programming approach, compared to the optimal design among $\mathcal{A}$ computed using grid search and p-value combination approach. E.g., at $1-\beta = 0.74$, the value of ESS_Q for the former is 21% smaller than for the latter. In addition, the optimization problem is infeasible for the designs in $\mathcal{A}$ at $1-\beta \geq 0.78$, i.e., it is not possible to simultaneously satisfy the power constraints (P1)-(P3); in contrast, the problem is feasible for the class $\mathcal{E} \times \mathcal{M}$ up to power threshold $1-\beta = 0.82$. This shows that there are substantial gains in expected sample size and power from using the optimal design over the class $\mathcal{E} \times \mathcal{M}$ compared to the optimal design over $\mathcal{A}$. Our sparse linear programming method made it possible to compute the optimal design over $\mathcal{E} \times \mathcal{M}$, which previously was an open problem.

Also, for the comparison purpose, we include standard design using Bergmann and Hommel adjusted p-values (Bergmann and Hommel, 1988) approach in Table 1. Specificaly, in each simulation iteration, we generate $C \cdot n$ samples, where $C \cdot n/2$ samples are from subpopulation 1, and the rest are from subpopulation 2. We compute the adjusted p-values using Bergmann and Hommel's method. We repeat the simulation 100,000 times, and we report

Table 2.1: The minimum value of ESS_Q, among the designs $\mathcal{A}$ (computed using grid search and p-value combination approach) and among the designs $\mathcal{E} \times \mathcal{M}$ (computed using sparse linear programming approach), for various values $1 - \beta$ of the power constraints (P1)-(P3). No value is given for $\mathcal{A}$ when $1 - \beta \geq 0.78$ since the problem is infeasible.

Power Constraint $(1 - \beta)$	58%	62%	66%	70%	74%	78%	82%
Min. ESS_Q over $\mathcal{A}$	$0.86n$	$0.89n$	$0.92n$	$0.97n$	$1.01n$	$\times$	$\times$
Min. $ESS + Q$ over $\mathcal{E} \times \mathcal{M}$	$0.65n$	$0.69n$	$0.73n$	$0.79n$	$0.84n$	$0.92n$	$1.03n$
Min. Standard BH adjusted p values	$1.21n$	$1.29n$	$1.45n$	$1.58n$	$1.75n$	$1.91n$	$2.08n$

the smallest C which achieves the required power empirically. It is seen immediately that the adaptive design significantly outperforms the standard design.

Chapter 3

Blessing of Massive Scale

3.1 Introduction

Under many statistical models, such as Gaussian or Ising models, we estimate spatial graphical models by solving the following problem:

$$\min_{\boldsymbol{\beta}_j \in \mathcal{C}_j} \frac{1}{d} \sum_{j=1}^{d} \mathcal{L}_j(\boldsymbol{\beta}_j), \text{ subject to } \sum_{j=1}^{d} \|\boldsymbol{\beta}_j\|_0 \leq K. \tag{3.1.1}$$

For each vertex j, $\mathcal{L}_j : \mathbb{R}^{d_j} \to \mathbb{R}$ is some convex loss function associated with the statistical model. For example, under the Gaussian model, we let $\mathcal{L}_j(\boldsymbol{\beta}_j) = n^{-1}\|\mathbb{X}_j - \mathbb{X}_{\mathcal{N}_j}\boldsymbol{\beta}_j\|_2^2$, where $\mathbb{X}_j \in \mathbb{R}^n$ is the data corresponding to vertex X_j, and $\mathbb{X}_{\mathcal{N}_j} \in \mathbb{R}^{n \times d_j}$ is the data which corresponds to the potential neighbors of vertex X_j. $\|\boldsymbol{\beta}_j\|_0$ denotes the ℓ_0-(pseudo)norm function, which is defined as the number of nonzero elements of $\boldsymbol{\beta}_j$. The feasible set $\mathcal{C}_j$ is a closed set, and K is some tuning parameter representing the desired sparsity level. Denote the global minimizer of the problem by $\{\widetilde{\boldsymbol{\beta}}_j\}_{j=1}^d$. The corresponding estimated graph is $\widetilde{G} = (V, \widetilde{E})$, where $(j,k) \in \widetilde{E}$ if and only if $\widetilde{\beta}_{jk} \neq 0$ or $\widetilde{\beta}_{kj} \neq 0$. Given the spatial proximity information, we have that each node j can only connect to a set of vertices $\mathcal{N}_j \subset \{1, ..., d\}$,

where $|\mathcal{N}_j| = d_j \ll d$. Then, each set $\mathcal{C}_j \subset \mathbb{R}^{d_j}$ is of small dimensions, and this makes each subproblem $\min_{\boldsymbol{\beta}_j \in \mathcal{C}_j} \mathcal{L}_j(\boldsymbol{\beta}_j)$ small dimensional.

Problem (3.1.1) is highly nonconvex and raises computational challenges. To overcome such challenges, several existing works rely on solving optimization problems derived from convex relaxations, such as the ℓ_1-relaxation. The motivation of different convex relaxations is to avoid solving nonconvex or combinatorial optimization problems while still achieves fast statistical rates. Extensive works study the theoretical guarantees of different relaxations various models, and achieve optimal minimax lower bounds under certain regularity assumptions. See Meinshausen and Bühlmann (2006); Zhao and Yu (2006); Meinshausen and Yu (2009); Zhang (2009); Meinshausen and Bühlmann (2010); Liu and Wang (2012). However, some results prove that there are some unavoidable statistical losses for the estimators derived from these methods. For example, in linear regression, we have that the ℓ_0-constrained estimator $\widehat{\boldsymbol{\beta}}_0$ obtains optimal rate of convergence that $n^{-1}\mathbb{E}\big\{\|\mathbb{X}\widehat{\boldsymbol{\beta}}_0 - \mathbb{X}\boldsymbol{\beta}^*\|_2^2\big\} = \mathcal{O}(n^{-1}s\log d)$, where $s = \|\boldsymbol{\beta}^*\|_0$. In comparison, without restricted eigenvalue-type assumptions, methods based on convex relaxations only achieve a slower rate (Zhang et al., 2014).

Our first contribution is a scalable estimation procedure to solve problem (3.1.1). Although problem (3.1.1) is highly nonconvex, we develop a Splitting-Communicating (SPICA) algorithm which solves the Lagrangian dual of the primal problem (3.1.1). The algorithm utilizes the separable structure of the dual problem and converges to a dual optimal solution geometrically. Since problem (3.1.1) is nonconvex, there exists a positive duality gap between the primal and dual optimal solutions. Suppose that the number of potential neighbors per-vertex is fixed. We prove that the average-per-vertex duality gap diminishes at the rate of $\mathcal{O}(d^{-1})$ as the graph dimension d increases. As a result, if the dimension d is large, the dual optimal solution is close to the primal optimal solution, and achieves optimal statistical properties. This reveals a "blessing of massive scale" phenomenon.

Our second contribution is the characterization of the complexity of problem (3.1.1) in a general setting. We prove that solving problem (3.1.1) is NP-complete and thus difficult in general. For the total cardinality approach, we discover that the problem is polynomial-time solvable by a dynamic programming algorithm, although this algorithm is practically expensive. We further prove that if the constraint of problem (3.1.1) is vector-valued, the problem becomes fundamentally more difficult, in which case even finding a fully polynomial-time approximation scheme to solve the problem is NP-hard. For example, we cannot solve the problem if the constraint in (3.1.1) is changed to $\sum_{j=1}^{d}(\|\boldsymbol{\beta}_j\|_0, \sum_{k=1}^{d_j-1}\|\beta_{j,k+1}-\beta_{jk}\|_0)^T \leq (K_1, K_2)^T$ unless P = NP.

To summarize, our work develops a novel framework to estimate high-dimensional spatial graphical models by directly attacking the nonconvex problem under the total cardinality constraint. The proposed algorithm produces a near-optimal solution to the nonconvex optimization problem, which achieves optimal statistical properties. The characterization of the complexity of problem (3.1.1) provides fundamental understandings and insights of the problem.

Notations. Let $\mathbb{X} \in \mathbb{R}^{n\times d}$ be the data matrix, and $\mathbb{X}_j$ denotes the j-th column of $\mathbb{X}$. Also, $\mathbb{X}_{\mathcal{N}_j}$ denotes the columns of possible neighbors of X_j, where $\mathcal{N}_j$ is the set of possible neighbors of X_j. For a matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, we denote its maximum eigenvalue by $\Lambda_{\max}(\mathbf{A})$, and its minimum eigenvalue by $\Lambda_{\min}(\mathbf{A})$. We denote by $[d] = \{1, \ldots, d\}$. The norms of $\mathbf{v} = (v_1, ..., v_d)^T \in \mathbb{R}^d$ are defined as $\|\mathbf{v}\|_0 = \sum_{j\in[d]} \mathbf{1}\{x_j \neq 0\}$, and $\|\mathbf{v}\|_p = \{\sum_{j\in[d]} v_j^p\}^{1/p}$ for $p \geq 1$.

3.2 SPICA Algorithm for Spatial-Graph Estimation

In this section, we describe an efficient duality-based algorithm to directly attack the nonconvex problem (3.1.1) and prove that it achieves a near-optimal solution, even though problem (3.1.1) belongs to an NP-complete class of problems as shown in Section 3.4. We

illustrate the geometric intuition on why this algorithm generates a near-optimal solution when the dimension d is large. Note that all the analyses of this section can be generalized to general nonconvex constraints. We focus our discussion on the case of ℓ_0-constraint, in which the solution achieves optimal statistical properties. For ease of presentation, in what follows, we assume that all $\boldsymbol{\beta}_j$'s are of identical dimensions d_0, i.e., $\boldsymbol{\beta}_j \in \mathbb{R}^{d_0}$ for all $j \in [d]$, where d_0 is a given constant.

3.2.1 SPICA Algorithm

In this subsection, we propose an algorithm to solve problem (3.1.1) subject to total cardinality constraint. It is practically efficient and can handle problems with large dimension d. We consider the Lagrangian dual of (3.1.1),

$$\max_{\lambda \geq 0} \sum_{j=1}^{d} \mathcal{Q}_j(\lambda) - \lambda K, \text{ where } \mathcal{Q}_j(\lambda) = \inf_{\boldsymbol{\beta}_j \in \mathcal{C}_t} \{\mathcal{L}_j(\boldsymbol{\beta}_j) + \lambda \|\boldsymbol{\beta}_j\|_0\}, \tag{3.2.1}$$

for all $j = 1, ..., d$. The variable λ is the Lagrangian multiplier. According to literatures on duality theory (Bertsekas, 1999), even though the primal problem (3.1.1) is nonconvex, letting $\mathcal{Q}(\lambda) = \sum_{j=1}^{d} \mathcal{Q}_j(\lambda)$, the dual $\widehat{\mathcal{Q}}(\lambda) = \mathcal{Q}(\lambda) - \lambda K$ is a concave function of λ. We aim to obtain the dual optimal solution-multiplier pair $(\{\widehat{\boldsymbol{\beta}}_j\}_{j \in [d]}, \widehat{\lambda})$ defined as

$$\begin{aligned} \widehat{\lambda} &= \operatorname*{argmax}_{\lambda \geq 0} \sum_{j=1}^{d} \mathcal{Q}_j(\lambda) - \lambda K, \text{ where } \mathcal{Q}_j(\lambda) = \mathcal{L}_j(\widehat{\boldsymbol{\beta}}_j) + \lambda \|\widehat{\boldsymbol{\beta}}_j\|_0, \\ \text{and } \ \widehat{\boldsymbol{\beta}}_j &= \operatorname*{argmin}_{\boldsymbol{\beta}_j \in \mathcal{C}_j} \mathcal{L}_j(\boldsymbol{\beta}_j) + \widehat{\lambda} \|\boldsymbol{\beta}_j\|_0, \text{ where } \sum_{j=1}^{d} \|\boldsymbol{\beta}_j\|_0 \leq K. \end{aligned} \tag{3.2.2}$$

We adopt the "golden section search" method to solve the dual problem (3.2.1), which runs iteratively. Let $\xi = (-1 + \sqrt{5})/2$. Given two initial points λ_1 and λ_2, let $\lambda_3 = \lambda_2 + \xi(\lambda_1 - \lambda_2)$ and $\lambda_4 = \lambda_1 + \xi(\lambda_2 - \lambda_1)$. During each iteration, if $\widehat{\mathcal{Q}}(\lambda_3) > \widehat{\mathcal{Q}}(\lambda_4)$, then we move the points $\{\lambda_2, \widehat{\mathcal{Q}}(\lambda_2)\}$ to $\{\lambda_4, \widehat{\mathcal{Q}}(\lambda_4)\}$, and $\{\lambda_4, \widehat{\mathcal{Q}}(\lambda_4)\}$ to $\{\lambda_3, \widehat{\mathcal{Q}}(\lambda_3)\}$, and we

update λ_3 to $\lambda_2 + \xi(\lambda_1 - \lambda_2)$. Otherwise, let $\{\lambda_1, \widehat{\mathcal{Q}}(\lambda_1)\}$ be $\{\lambda_3, \widehat{\mathcal{Q}}(\lambda_3)\}$, and $\{\lambda_3, \widehat{\mathcal{Q}}(\lambda_3)\}$ be $\{\lambda_4, \widehat{\mathcal{Q}}(\lambda_4)\}$, and we update λ_4 to $\lambda_1 + \xi(\lambda_2 - \lambda_1)$. Specifically, at each iteration, we first compute the values of $\{\widehat{\mathcal{Q}}_j(\lambda_i)\}_{i=1}^4$ for all j. This can be conducted efficiently since the dual problem (3.2.1) "splits" the Lagrangian minimization problem into d nonconvex problems with small dimension d_0, and we can compute $\mathcal{Q}_j(\lambda_i)$'s in parallel. We call this a "splitting" step. Next, we centrally update λ_i's according to the golden section search method, which is a "communicating" step. Thus we call it a "splitting-communicating" (SPICA) algorithm, which is summarized in Algorithm 1. This algorithm finds a narrow interval that contains the optimal multiplier of the problem (3.2.1) after some iterations, and the output solution is the midpoint of the interval. It is well known that the golden section search method converges ξ-geometrically to the dual optimal solution $(\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}, \widehat{\lambda})$ (Bertsekas, 1999), i.e., we have

$$\widehat{\lambda}^{(t)} - \widehat{\lambda} \leq \xi^t |\lambda_2^{(0)} - \lambda_1^{(0)}|,$$

where $\lambda_i^{(0)}$'s denote the initial points, $\widehat{\lambda}^{(t)} = |\lambda_1^{(t)} - \lambda_2^{(t)}|/2$, and $(\lambda_1^{(t)}, \lambda_2^{(t)})$ denotes the corresponding point after t iterations.

Algorithm 1 SPICA Algorithm

Input: $\lambda_1, \lambda_2 \in \mathbb{R}_+$, $\epsilon > 0$, $\xi = (-1 + \sqrt{5})/2$
Output: $\widehat{\lambda}$, $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$
$\lambda_3 \leftarrow \lambda_2 + \xi(\lambda_1 - \lambda_2)$, $\lambda_4 \leftarrow \lambda_1 + \xi(\lambda_2 - \lambda_1)$.
while $|\lambda_1 - \lambda_2| > \epsilon$ **do**
  if $\mathcal{Q}(\lambda_3) - \lambda_3 K > \mathcal{Q}(\lambda_4) - \lambda_4 K$ **then**
    $\{\lambda_2, \mathcal{Q}(\lambda_2) - \lambda_2 K\} \leftarrow \{\lambda_4, \mathcal{Q}(\lambda_4) - \lambda_4 K\}$, $\{\lambda_4, \mathcal{Q}(\lambda_4) - \lambda_4 K\} \leftarrow \{\lambda_3, \mathcal{Q}(\lambda_3) - \lambda_3 K\}$.
    $\lambda_3 \leftarrow \lambda_2 + \xi(\lambda_1 - \lambda_2)$.
  else
    $\{\lambda_1, \mathcal{Q}(\lambda_1) - \lambda_1 K\} \leftarrow \{\lambda_3, \mathcal{Q}(\lambda_3) - \lambda_3 K\}$, $\{\lambda_3, \mathcal{Q}(\lambda_3) - \lambda_3 K\} \leftarrow \{\lambda_4, \mathcal{Q}(\lambda_4) - \lambda_4 K\}$.
    $\lambda_4 \leftarrow \lambda_1 + \xi(\lambda_2 - \lambda_1)$.
  end if
end while
$\widehat{\lambda} \leftarrow (\lambda_1 + \lambda_2)/2$, $\widehat{\boldsymbol{\beta}}_j = \operatorname{argmin}_{\boldsymbol{\beta}_j} \mathcal{L}_j(\boldsymbol{\beta}_j) + \widehat{\lambda}\|\boldsymbol{\beta}_j\|_0$ for all $j \in [d]$.

The SPICA algorithm provides significant computational advantages. However, instead of a global optimal solution to problem (3.1.1), it generates an optimal solution to dual problem (3.2.1). Since the total cardinality constraint is nonconvex, there exists some duality gap between the dual and the primal optimal solutions. In the next two subsections, we illustrate a convexification phenomenon that, as d increases, the duality gap diminishes and does not impair any statistical loss in a wide range of problems.

3.2.2 The Convexification Phenomenon

Before rigorously proving that the dual optimal solution obtained by the SPICA algorithm is close to the primal optimal solution of problem (3.1.1), we illustrate some geometric intuition. The intuition traces back to some early convex geometry work, namely, the Shapley-Folkman Lemma (Starr, 1969). Consider the averaged Minkowski sum of d sets $\mathcal{A}_1$,...,$\mathcal{A}_d$ defined as $\{d^{-1}\sum_{j=1}^{d} a_j : a_j \in \mathcal{A}_j \text{ for } j \in [d]\}$. The lemma reveals a geometric fact that the average of many nonconvex sets tends to be convex. In particular, letting $\rho(\mathcal{A})$ be a metric of the nonconvexity of the set $\mathcal{A}$, we have

$$\rho\Big(\frac{\mathcal{A}_1 + \mathcal{A}_2 + \cdots + \mathcal{A}_d}{d}\Big) \to 0, \text{ as } d \to \infty.$$

We provide an example to illustrate this convexification effect. Let the maximum distance between two sets $\mathcal{A}$ and $\mathcal{B}$ be $d(\mathcal{A}, \mathcal{B}) = \sup_{b\in\mathcal{B}} \inf_{a\in\mathcal{A}}\{\|a - b\| : a \in \mathcal{A}, b \in \mathcal{B}\}$. We measure the nonconvexity of a set $\mathcal{A}$ by the maximum distance between $\mathcal{A}$ and its convex hull. Since this distance is 0 if and only if $\mathcal{A}$ is convex, the maximum distance is a reasonable measure of how convex a set is. Considering the discrete set $\mathcal{A} = \{0, 1\}$ and its convex hull $\bar{\mathcal{A}} = [0, 1]$, we have $\rho(\mathcal{A}) = d(\mathcal{A}, \bar{\mathcal{A}}) = 1/2$. The maximum distance between the average of the Minkowski sum of two $\mathcal{A}$'s, which is $\mathcal{A}_2 = \{0, 1/2, 1\}$, and its convex hull is $\rho(\mathcal{A}_2) = d(\mathcal{A}_2, \bar{\mathcal{A}}) = 1/4$. Let the average of d $\mathcal{A}$'s be $\mathcal{A}_d$. We have $\rho(\mathcal{A}_d) = 1/2d$, which converges to 0 as d increases.

We thus conclude that the average of d $\mathcal{A}$'s tend to be more convex as d increases. In Figure 3.1, we provide an geometric illustration of such increase of convexity, and we provide the mathematical description of Shapley-Folkman Lemma in Section B.1 of the Appendix.

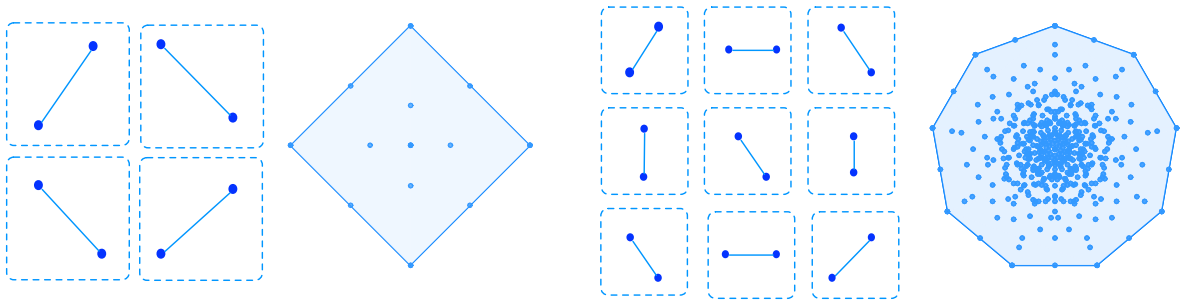


Figure 3.1: Left two: The shaded area of the second figure is the convex hull of the averaged Minkowski sum of four sets illustrated on the first figure. Each of the four sets contains two points, and the line between them represents the convex hull. Right two: The shaded area on the right is the convex hull of the averaged Minkowski sum of nine sets. The maximum distance between the averaged Minkowski sum and its convex hull decreases as the number of sets increases.

Let us return to problem (3.1.1). The duality gap between the primal problem (3.1.1) and its dual can be bounded by the nonconvexity of the set $d^{-1}\sum_{j=1}^{d}\mathcal{A}_j$, where each $\mathcal{A}_j = \left\{\left(\|\boldsymbol{\beta}_j\|_0, \mathcal{L}_j(\boldsymbol{\beta}_j)\right) : \boldsymbol{\beta}_j \in \mathcal{C}_j\right\}$ characterizes the joint nonconvexity of $\left(\|\boldsymbol{\beta}_j\|_0, \mathcal{L}_j(\boldsymbol{\beta}_j)\right)$. By the intuition above, as d increases, the set $d^{-1}\sum_{j=1}^{d}\mathcal{A}_j$ tends to be convex, and we expect a diminishing duality gap. This convexification phenomenon provides a hint that solving the dual problem might be as good as solving the primal problem.

3.2.3 Diminishing Duality Gap

In this subsection, we prove that the average-per-vertex duality gap diminishes at a rate of $\mathcal{O}(1/d)$. This result provides the theoretical justification that the estimator obtained by the SPICA algorithm (Alg. 1) is near-optimal, i.e., it is close to the primal optimal solution

$\{\widetilde{\boldsymbol{\beta}}_j\}_{j\in[d]}$ of problem (3.1.1) defined as

$$\{\widetilde{\boldsymbol{\beta}}_j\}_{j=1}^d = \underset{\boldsymbol{\beta}_j\in\mathcal{C}_j}{\operatorname{argmin}} \frac{1}{d}\sum_{j=1}^d \mathcal{L}_j(\boldsymbol{\beta}_j), \text{ subject to } \sum_{j=1}^d \|\boldsymbol{\beta}_j\|_0 \le K,$$

As we discussed in the previous subsections, we consider the Lagrangian dual problem (3.2.1), and the SPICA algorithm (Alg. 1) finds the dual optimal solution-multiplier pair $\big(\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}, \widehat{\lambda}\big)$ as defined in (3.2.2). Since the problem is nonconvex, strong duality does not hold. In this case, we only have weak duality that

$$\frac{1}{d}\sum_{j=1}^d \mathcal{L}_j(\widehat{\boldsymbol{\beta}}_j) + \widehat{\lambda}\Big(\sum_{j=1}^d \|\widehat{\boldsymbol{\beta}}_j\|_0 - K\Big) \le \frac{1}{d}\sum_{j=1}^d \mathcal{L}_j(\widetilde{\boldsymbol{\beta}}_j), \tag{3.2.3}$$

where both $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$ and $\{\widetilde{\boldsymbol{\beta}}_j\}_{j\in[d]}$ satisfy the total cardinality constraint, but some duality gap might exist in this case. Note that the duality gap is the difference between primal and dual optimal objective values. We provide an example to illustrate that the primal and dual optimal solutions do not necessarily match, which results in a positive duality gap. Let

$$\mathbb{X} = \begin{pmatrix} 1 & 6 & 5 & 5 \\ 8 & 9 & 3 & 2 \\ 7 & 10 & 8 & 8 \end{pmatrix} \text{ and } \mathbf{y} = \begin{pmatrix} 12 \\ 20 \\ 25 \end{pmatrix}.$$

Considering the ℓ_0-constrained problem,

$$\min_{\boldsymbol{\beta}} \|\mathbb{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2, \text{ subject to } \|\boldsymbol{\beta}\|_0 \le 2,$$

the primal solution is $\widetilde{\boldsymbol{\beta}} = (857/497, 0, 292/165, 0)^T$. For the dual solution,

$$\widehat{\boldsymbol{\beta}}(\lambda) = \|\mathbb{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_0.$$

it is not difficult to check that when $\lambda \geq 1169 - 1669/217$, $\widehat{\boldsymbol{\beta}}(\lambda) = \mathbf{0}$, if $\lambda \in [1669/434, 1669/217)$, $\widehat{\boldsymbol{\beta}}(\lambda) = (0, 502/217, 0, 0, 0)^T$, if $\lambda < 1669/434$, $\widehat{\boldsymbol{\beta}}(\lambda) = (1, 1, 1, 0)^T$. This implies that the primal and dual optimal solutions do not match, and there exists a strictly positive duality gap equals 449/894.

The next theorem proves that, as d increases, the average-per-vertex duality gap vanishes at the rate of $\mathcal{O}(1/d)$. This gives a strong evidence that the dual solution obtained by the SPICA algorithm is a fairly good approximation to the primal solution, especially when d is large. Usually, such a large d would cause the "curse of dimensionality" in nonconvex optimization, but our result reveals a "blessing of massive scale" phenomenon.

Theorem 3.2.1. *The solution* $\big(\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}, \widehat{\lambda}\big)$ *obtained by the SPICA algorithm (Alg. 1) is a dual optimal solution-multiplier pair, which solves the dual problem* (3.2.1)*, and satisfies*

$$\frac{1}{d}\sum_{j=1}^{d} \mathcal{L}_j(\widehat{\boldsymbol{\beta}}_j) \leq \frac{1}{d}\sum_{j=1}^{d} \mathcal{L}_j(\widetilde{\boldsymbol{\beta}}_j) + \frac{C_g}{d}, \text{ and } \sum_{j=1}^{d} \|\boldsymbol{\beta}_j\|_0 \leq K, \tag{3.2.4}$$

where $\{\widetilde{\boldsymbol{\beta}}\}_{j\in[d]}$ *is the primal optimal solution, and the constant* C_g *is*

$$C_g = \max_{j\in[d]} \big|\mathcal{L}_j(\mathbf{0}) - \min_{\boldsymbol{\beta}_j \in \mathcal{C}_j} \mathcal{L}_j(\boldsymbol{\beta}_j)\big|. \tag{3.2.5}$$

Proof. First, we prove the existence of the optimal dual solution. This is proved in Lemma B.2.1 in the Appendix. Next, if $\widehat{\lambda} = 0$, we have $\widehat{\boldsymbol{\beta}}_j = \operatorname{argmin}_{\boldsymbol{\beta}_j} \mathcal{L}_j(\boldsymbol{\beta}_j)$ for all j's as defined in (3.2.2). Since $\sum_{j\in[d]} \|\widehat{\boldsymbol{\beta}}_j\|_0 \leq K$ by the feasibility of $\widehat{\boldsymbol{\beta}}_j$'s, we have $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$ is also the primal optimal solution. This implies $\widehat{\boldsymbol{\beta}}_j = \widetilde{\boldsymbol{\beta}}_j$ for all j, and our claim follows as desired.

If $\widehat{\lambda} > 0$, we prove in Lemma B.2.3 in the Appendix that one of the two cases must hold:

1. There exists a dual optimal solution-multiplier pair $(\widehat{\lambda}, \{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]})$, such that $\sum_{j\in[d]} \|\widehat{\boldsymbol{\beta}}_j\|_0 = K$.

2. Case (i) does not hold, and there exist at least two solutions achieve dual optimal objective, denoted as $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$ and $\{\widehat{\boldsymbol{\beta}}'_j\}_{j\in[d]}$, such that $\sum_{j\in[d]}\|\widehat{\boldsymbol{\beta}}_j\|_0 < K$ and $\sum_{j\in[d]}\|\widehat{\boldsymbol{\beta}}'_j\|_0 > K$.

Next, we consider the two cases separately. For case (i), there exists a dual optimal solution $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$ satisfying $\sum_{j\in[d]}\|\widehat{\boldsymbol{\beta}}_j\|_0 = K$. By the weak duality (3.2.3), we have

$$\frac{1}{d}\sum_{j=1}^{d}\mathcal{L}_j(\widehat{\boldsymbol{\beta}}_j) \le \frac{1}{d}\sum_{j=1}^{d}\mathcal{L}_j(\widetilde{\boldsymbol{\beta}}_j) + \widehat{\lambda}\big(\sum_{j=1}^{d}\|\widehat{\boldsymbol{\beta}}_j\|_0 - K\big) = \frac{1}{d}\sum_{j=1}^{d}\mathcal{L}_j(\widetilde{\boldsymbol{\beta}}_j),$$

where the first inequality holds by the definition of dual optimal solution that $\widehat{\boldsymbol{\beta}}_j = \operatorname{argmin}_{\boldsymbol{\beta}_j}\mathcal{L}_j(\boldsymbol{\beta}_j) + \widehat{\lambda}\|\boldsymbol{\beta}_j\|_0$, and the assertion of the theorem follows as desired. We also point out that since $\{\widetilde{\boldsymbol{\beta}}_j\}_{j\in[d]}$ is the primal optimal solution, the above inequality and the feasibility of $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$ guarantee the primal optimality of $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$, i.e., the dual optimal solution $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$ is also a primal optimal solution. This also leads to the certificate of primal optimality result stated in Corollary 3.2.2.

In the remaining proof, we focus our discussion on case (ii). This case is more complicated and requires more careful analysis due to the existence of multiple solutions. Recall that, given the multiplier $\widehat{\lambda}$, a dual solution is obtained by solving d subproblems of the ℓ_0-penalized form:

$$\min_{\boldsymbol{\beta}_j}\mathcal{L}_j(\boldsymbol{\beta}_j) + \widehat{\lambda}\|\boldsymbol{\beta}_j\|_0, \text{ for all } j = 1, ..., d.$$

Since there are multiple solutions which achieve dual optimal objective, as shown in Lemma B.2.3 in the Appendix, we have that there is at least one j, such that the above ℓ_0-penalized optimization problem has multiple solutions, i.e., for some j, there exist $\widehat{\boldsymbol{\beta}}_j^{(1)}$ and $\widehat{\boldsymbol{\beta}}_j^{(2)}$ such that

$$\mathcal{L}_j(\widehat{\boldsymbol{\beta}}_j^{(1)}) + \widehat{\lambda}\|\widehat{\boldsymbol{\beta}}_j^{(1)}\|_0 = \mathcal{L}_j(\widehat{\boldsymbol{\beta}}_j^{(2)}) + \widehat{\lambda}\|\widehat{\boldsymbol{\beta}}_j^{(2)}\|_0. \tag{3.2.6}$$

In addition, any combination of the optimal solutions of the subproblems provides a dual optimal objective without satisfying the feasibility. In what follows, we show that we can select a dual optimal solution from all possible combinations, such that the selected solution achieves the error bound (3.2.4).

Suppose that there exist m solutions achieve dual optimal objective. Let $\{\widehat{\boldsymbol{\beta}}_j^{(1)}\}_{j\in[d]}$, $\{\widehat{\boldsymbol{\beta}}_j^{(2)}\}_{j\in[d]}$,..., $\{\widehat{\boldsymbol{\beta}}_j^{(m)}\}_{j\in[d]}$ be the sequence of solutions ranked by their corresponding primal objective values, i.e.,

$$\sum_{j=1}^{d}\mathcal{L}_j\big(\widehat{\boldsymbol{\beta}}_j^{(1)}\big) \leq \sum_{j=1}^{d}\mathcal{L}_j\big(\widehat{\boldsymbol{\beta}}_j^{(2)}\big) \leq, ..., \leq \sum_{j=1}^{d}\mathcal{L}_j\big(\widehat{\boldsymbol{\beta}}_j^{(m)}\big). \tag{3.2.7}$$

Meanwhile, by the dual optimality, we have,

$$\begin{aligned}\sum_{j=1}^{d}\Big[\mathcal{L}_j\{\widehat{\boldsymbol{\beta}}_j^{(1)}\} + \widehat{\lambda}\|\widehat{\boldsymbol{\beta}}_j^{(1)}\|_0\Big] &= \sum_{j=1}^{d}\Big[\mathcal{L}_j\{\widehat{\boldsymbol{\beta}}_j^{(2)}\} + \widehat{\lambda}\|\widehat{\boldsymbol{\beta}}_j^{(2)}\|_0\Big] \\ &= , ..., = \sum_{j=1}^{d}\Big[\mathcal{L}_j\{\widehat{\boldsymbol{\beta}}_j^{(m)}\} + \widehat{\lambda}\|\widehat{\boldsymbol{\beta}}_j^{(m)}\|_0\Big].\end{aligned}$$

Since $\widehat{\lambda} > 0$ by assumption, we have

$$\sum_{j=1}^{d}\|\widehat{\boldsymbol{\beta}}_j^{(1)}\|_0 \geq \sum_{j=1}^{d}\|\widehat{\boldsymbol{\beta}}_j^{(2)}\|_0 \geq, ..., \geq \sum_{j=1}^{d}\|\widehat{\boldsymbol{\beta}}_j^{(m)}\|_0.$$

Consequently, by the assumption that case (ii) holds, we have

$$\sum_{j=1}^{d}\|\widehat{\boldsymbol{\beta}}_j^{(1)}\|_0 > K > \sum_{j=1}^{d}\|\widehat{\boldsymbol{\beta}}_j^{(m)}\|_0.$$

To prove our claim, a key observation is that, for any $m_1 \in \{1, ..., m-1\}$, $\sum_{j\in[d]}\mathcal{L}_j\{\widehat{\boldsymbol{\beta}}_j^{(m_1+1)}\} - \sum_{j\in[d]}\mathcal{L}_j\{\widehat{\boldsymbol{\beta}}_j^{(m_1)}\} \leq C_g$, where C_g is defined in (3.2.5). This is proved in Lemma B.2.5 in the Appendix.

Thus, by the assumption that case (ii) holds, there exist two consecutive solutions $\{\widehat{\boldsymbol{\beta}}_j^{(m_1)}\}_{j\in[d]}$ and $\{\widehat{\boldsymbol{\beta}}_j^{(m_1+1)}\}_{j\in[d]}$ for some $m_1 \in [m]$, such that

$$\Big|\sum_{j=1}^{d}\mathcal{L}_j\{\widehat{\boldsymbol{\beta}}_j^{(m_1)}\} - \sum_{j=1}^{d}\mathcal{L}_j\{\widehat{\boldsymbol{\beta}}_j^{(m_1+1)}\}\Big| \le C_g,$$
$$\text{and } \sum_{j=1}^{d}\|\widehat{\boldsymbol{\beta}}_j^{(m_1)}\|_0 > K > \sum_{j=1}^{d}\|\widehat{\boldsymbol{\beta}}_j^{(m_1+1)}\|_0.$$

In addition, by the dual optimality of the two solutions, it holds that

$$\begin{aligned}&\sum_{j=1}^{d}\mathcal{L}_j\{\widehat{\boldsymbol{\beta}}_j^{(m_1)}\} + \widehat{\lambda}\{\sum_{j=1}^{d}\|\widehat{\boldsymbol{\beta}}_j^{(m_1)}\|_0 - K\}\\ =&\sum_{j=1}^{d}\mathcal{L}_j\{\widehat{\boldsymbol{\beta}}_j^{(m_1+1)}\} + \widehat{\lambda}\big(\sum_{j=1}^{d}\|\widehat{\boldsymbol{\beta}}_j^{(m_1+1)}\|_0 - K\big).\end{aligned}$$

Consequently, as $\sum_{j\in[d]}\|\widehat{\boldsymbol{\beta}}_j^{(m_1)}\|_0 > K > \sum_{j\in[d]}\|\widehat{\boldsymbol{\beta}}_j^{(m_1+1)}\|_0$, and $\widehat{\lambda} > 0$, we further obtain that

$$\begin{aligned}0 \le -\widehat{\lambda}\big(\sum_{j=1}^{d}\|\widehat{\boldsymbol{\beta}}_j^{(m_1+1)}\|_0 - K\big) &\le \widehat{\lambda}\big(\sum_{j=1}^{d}\|\widehat{\boldsymbol{\beta}}_j^{(m_1)}\|_0 - \sum_{j=1}^{d}\|\widehat{\boldsymbol{\beta}}_j^{(m_1+1)}\|_0\big)\\ &= \sum_{j=1}^{d}\mathcal{L}_j\{\widehat{\boldsymbol{\beta}}_j^{(m_1+1)}\} - \sum_{j=1}^{d}\mathcal{L}_j\{\widehat{\boldsymbol{\beta}}_j^{(m_1)}\} \le C_g.\end{aligned} \tag{3.2.8}$$

We have

$$\sum_{j=1}^{d}\mathcal{L}_j\{\widehat{\boldsymbol{\beta}}_j^{(m_1)}\} \le \sum_{j=1}^{d}\mathcal{L}_j(\widetilde{\boldsymbol{\beta}}_j) - \widehat{\lambda}\big(\sum_{j=1}^{d}\|\widehat{\boldsymbol{\beta}}_j^{(m_1)}\|_0 - K\big) \le \sum_{j=1}^{d}\mathcal{L}_j(\widetilde{\boldsymbol{\beta}}_j) + C_g,$$

where the first inequality holds by the weak duality (3.2.3), and the second inequality holds by (3.2.8).

To conclude, in both cases (i) and (ii), we prove that there exists an dual optimal solution $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$ that achieves the total cardinality constraint, and approximates the primal solution within a constant error bound even if d increases. □

To interpret the constant C_g, each $\left|\mathcal{L}_j(\mathbf{0}) - \min_{\boldsymbol{\beta}_j \in \mathcal{C}_j} \mathcal{L}_j(\boldsymbol{\beta}_j)\right|$ is some "divergence" related to vertex j. It essentially measures the information gain by using neighboring vertices to explain uncertainties of vertex j. The constant C_g is the maximal divergence among all vertices.

This result indicates that when the maximal divergence C_g is bounded, the average-per-vertex duality gap decreases to 0 as d increases. By the proof of Theorem 3.2.1 for case (i), i.e., when $\sum_{j\in[d]} \|\boldsymbol{\beta}_j\|_0 = K$, the next corollary follows immediately. This provides a criterion to determine if the primal optimality holds for $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$.

Corollary 3.2.2. (Certificate for Primal Optimality) Let $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$ be the dual optimal solution for problem (3.1.1) obtained by the SPICA algorithm (Alg. 1). When the equality $\sum_{j=1}^d \|\widehat{\boldsymbol{\beta}}_j\|_0 = K$ holds, it holds that the dual optimal solution also achieves the primal optimality, i.e.,

$$\sum_{j=1}^{d} \mathcal{L}_j\big(\widehat{\boldsymbol{\beta}}_j\big) = \sum_{j=1}^{d} \mathcal{L}_j\big(\widetilde{\boldsymbol{\beta}}_j\big), \text{ if } \sum_{j=1}^{d} \|\widehat{\boldsymbol{\beta}}_j\|_0 = K.$$

In Section 3.5, we find that empirically, the dual solution $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$ satisfies the certificate $\sum_{j=1}^d \|\widehat{\boldsymbol{\beta}}_j\|_0 = K$ with high probability.

3.3 Statistical Properties

In this section, we provide theoretical justifications of the estimators derived from the SPICA algorithm. We provide the statistical guarantee that under weak assumptions, the duality gap does not sacrifice any statistical efficiency when d is large. This matches Theorem 3.2.1. We discuss the rates of convergence for the estimators provided by the SPICA algorithm under Gaussian and Ising graphical models. Most technical proofs are provided in Appendix Section B.3.

3.3.1 Gaussian Graphical Model

We first apply the SPICA algorithm to estimate Gaussian graphical model. Consider a d-dimensional random vector $\boldsymbol{X} = (X_1, X_2, ..., X_d)^T \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. Under the Gaussian assumption, the conditional independence between X_j and X_k holds if and only if $\boldsymbol{\Theta}_{jk} = 0$, where $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$. Extensive literatures study this problem (Meinshausen and Bühlmann, 2006; Yuan and Lin, 2007; Cai et al., 2011; Liu and Wang, 2012). Under the spatial graphical modeling setting, taking a neighborhood pursuit approach, we formulate the graph estimation problem as

$$\min_{\{\boldsymbol{\beta}_j\}_{j\in[d]}} \frac{1}{dn} \sum_{j=1}^{d} \|\mathbb{X}_j - \mathbb{X}_{\mathcal{N}_j}\boldsymbol{\beta}_j\|_2^2, \text{ subject to } \sum_{j=1}^{d} \|\boldsymbol{\beta}_j\|_0 \leq K,$$

where $\mathbb{X} \in \mathbb{R}^{n\times d}$ is the data matrix; $\mathbb{X}_{\mathcal{N}_j} \in \mathbb{R}^{n\times d_j}$ denotes the columns of $\mathbb{X}$ which correspond to the potential neighbors of X_j, and K is a pre-specified total cardinality. Given a solution $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$, we obtain the connected neighbors of each X_j by taking the corresponding nonzero components of $\widehat{\boldsymbol{\beta}}_j$. Consequently, we construct the graph estimator by either "OR" or "AND" rule on combining the neighborhoods for all X_j's. This approach is based on the fact that

$$\begin{aligned} &X_j = \boldsymbol{X}_{\mathcal{N}_j}^T\boldsymbol{\beta}_j^* + \epsilon_j, \text{ where } \boldsymbol{\beta}_j^* = \left(\boldsymbol{\Sigma}_{\mathcal{N}_j,\mathcal{N}_j}\right)^{-1}\boldsymbol{\Sigma}_{\mathcal{N}_j,j} \in \mathbb{R}^{d-1}, \\ &\epsilon_j \sim N\left(0, \sigma_j^2\right), \text{ and } \sigma_j^2 = \boldsymbol{\Sigma}_{jj} - \boldsymbol{\Sigma}_{j,\mathcal{N}_j}\left(\boldsymbol{\Sigma}_{\mathcal{N}_j,\mathcal{N}_j}\right)^{-1}\boldsymbol{\Sigma}_{\mathcal{N}_j,j}, \end{aligned} \tag{3.3.1}$$

and by the block matrix inversion formula, it holds that

$$\boldsymbol{\Theta}_{jj} = \{\mathrm{Var}(\epsilon_j)\}^{-1} = \sigma_j^{-2}, \text{ and } \boldsymbol{\Theta}_{\mathcal{N}_j,j} = -\{\mathrm{Var}(\epsilon_j)\}^{-1}\boldsymbol{\beta}_j^* = -\sigma_j^{-2}\boldsymbol{\beta}_j^*.$$

Thus, $\boldsymbol{\Theta}_{jk} = 0$ if and only if the corresponding component of $\boldsymbol{\beta}_j^*$ is 0.

We point out that there are several advantages of the total cardinality approach over the ℓ_1 or other penalized approaches: (i) Imposing the total cardinality constraint directly

handles the estimator's sparsity level. This provides a more intuitive approach than penalized methods, where tuning parameters do not give very interpretable meanings. (ii) Total cardinality constraint approach does not incur any estimation bias. In comparison, penalized approach induces some estimation biases. Although such biases are asymptotically negligible under appropriate scaling, the finite-sample behavior of the penalized approach is indeed outperformed by the total cardinality approach as demonstrated in simulation studies in Section 3.5.

Next, we analyze the statistical properties of the estimator obtained by the SPICA algorithm. As the neighborhood pursuit approach formulates the problem as a regression problem, we first bound the "prediction risk" of the estimator. This leads to the estimator's fast rate of convergence.

In the following discussion, for ease of presentation, we assume that the numbers of potential neighbors of the vertices are the same, i.e., $|\mathcal{N}_1| =, ..., = |\mathcal{N}_d| = d_0$. The next therem guarantees that the average-per-vertex risk of our estimator converges at the minimax optimal rate, and justifies the vanishing gap does not incur statistical loss if the dimension d is large.

Theorem 3.3.1. *Suppose that we have n independent samples of $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma}) \in \mathbb{R}^d$, and the spatial information that each vertex j can only connect to a set of vertices $\mathcal{N}_j \subset \{1, ..., d\}$ and $|\mathcal{N}_j| = d_0$. Let $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$ be the estimator obtained by the SPICA algorithm. Assume $s \leq K$, and $2K \leq dd_0$, where $\sum_{j=1}^d \|\boldsymbol{\beta}_j^*\|_0 = s$, and $\boldsymbol{\beta}_j^*$'s are defined in* (3.3.1)*. We further assume* $\mathrm{diag}(\boldsymbol{\Sigma}) \leq \sigma^2$*. Then, with probability at least $1 - \mathcal{O}(d^{-1})$, we have*

$$\frac{1}{dn}\sum_{j=1}^{d} \|\mathbb{X}_{\mathcal{N}_j}\widehat{\boldsymbol{\beta}}_j - \mathbb{X}_{\mathcal{N}_j}\boldsymbol{\beta}_j^*\|_2^2 \leq C_1 \cdot \frac{K \log d}{dn} + C_2 \cdot \frac{\log d}{d}, \tag{3.3.2}$$

where C_1 and C_2 are two constants, and do not depend on K, d and n.

Proof. The proof is based on the following two lemmas. The first lemma quantifies the risk of the estimator, which involves the duality gap. The second lemma quantifies the duality gap C_g incurred by the SPICA algorithm. The two lemmas are proved in Section B.3.1 of the Appendix.

Lemma 3.3.2. *Suppose we have n independent samples of $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma}) \in \mathbb{R}^d$, and the prior information that each X_j can only connect to a set of nodes $\mathcal{N}_j \subset \{1, ..., d\}$, i.e., $\boldsymbol{\Theta}_{jk} = 0$ if $k \notin \mathcal{N}_j$. Let $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$ be the estimator obtained by the SPICA algorithm. Assume $2K \leq dd_0$, $s \leq K$, $\sum_{j\in[d]} \|\boldsymbol{\beta}_j^*\|_0 = s$, where $\boldsymbol{\beta}_j^*$'s are defined in* (3.3.1)*. We further assume* $\mathrm{diag}(\boldsymbol{\Sigma}) \leq \sigma^2$*. Then, for any $\delta > 0$, with probability at least $1 - \delta$, we have*

$$\begin{aligned} &\frac{1}{n}\sum_{j=1}^{d} \|\mathbb{X}_{\mathcal{N}_j}\widehat{\boldsymbol{\beta}}_j - \mathbb{X}_{\mathcal{N}_j}\boldsymbol{\beta}_j^*\|_2^2 \\ &\leq 64\frac{\sigma^2}{n}\log\Big\{\sum_{j=1}^{2K}\binom{dd_0}{j}\Big\} + \frac{128\sigma^2 K}{n}\log 6 + \frac{64\sigma^2}{n}\log(\sigma^{-1}) + 2C_g, \end{aligned} \tag{3.3.3}$$

where the constant C_g is the duality gap incurred by the SPICA algorithm.

Lemma 3.3.3. *Suppose we have n independent samples of $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma}) \in \mathbb{R}^d$. Let $\mathcal{L}_j(\boldsymbol{\beta}_j) = \|\mathbb{X}_j - \mathbb{X}_{\mathcal{N}_j}\boldsymbol{\beta}_j\|_2^2$ be the least square loss. We have,*

$$\max_j \big\{\mathcal{L}_j(\mathbf{0}) - \min_{\boldsymbol{\beta}_j} \mathcal{L}_j(\boldsymbol{\beta}_j)\big\} \leq n\sigma^2 + C \cdot n\log d,$$

with probability at least $1 - \mathcal{O}(d^{-1})$, where C is a constant.

Combining the above two lemmas, and plugging (B.4.1) in Section B.4 of the Appendix into (3.3.3), our claim follows as desired. □

This theorem proves that if $n < d$ and if the average-per-vertex degree is larger than 1, the estimator $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$ obtains the optimal rate of convergence. Note that this result does not require any restricted-eigenvalue type assumptions on $\mathbb{X}$. In comparison, it is shown in

Zhang et al. (2014) that if we do not impose such assumptions, other estimators based on convex relaxations, such as the Lasso estimator, cannot achieve the optimal rate unless P = NP. In addition, if we impose the sparse eigenvalue condition that the minimum eigenvalue of the sub-covariance matrices $\mathbf{\Sigma}_{\mathcal{N}_j,\mathcal{N}_j}$'s are all bounded below, i.e., there exists a constant $\rho > 0$, such that

$$\Lambda_{\min}\big(\mathbf{\Sigma}_{\mathcal{N}_j,\mathcal{N}_j}\big) > \rho, \text{ for all } j = 1, ..., d.$$

We have that the estimator $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$ obtains the fast rate of convergence that

$$\frac{1}{d}\sum_{j=1}^{d} \|\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*\|_2^2 \leq \underbrace{C_1 \cdot \frac{K \log d}{dn}}_{\text{statistical error}} + \underbrace{C_2 \cdot \frac{\log d}{d}}_{\text{duality gap}},$$

where C_1 and C_2 are two constants, and do not depend on K, d and n. Note that if the certificate of primal optimality (Cor. 3.2.2) holds, the duality gap term disappears.

In graphical model estimation, support recovery is of significant importance. The next corollary provides the support recovery guarantee of the estimator.

Corollary 3.3.4. Assume that all the assumptions in Theorem 3.3.1 and the sparse eigenvalue condition hold and $K = s$, where $\sum_{j\in[d]} \|\boldsymbol{\beta}_j^*\|_0 = s$. Suppose that we have the minimal signal strength that for all j,

$$\|\boldsymbol{\beta}_j^*(\mathcal{S}_j)\|_{\min} > C \cdot \sqrt{\frac{\log d}{n}}, \tag{3.3.4}$$

where $\mathcal{S}_j$ denotes the support of $\boldsymbol{\beta}_j^*$; $\|\mathbf{v}\|_{\min} = \min_j |v_j|$, and C is a constant which does not depend on K, d and n. We have, with probability at least $1 - \mathcal{O}(d^{-1})$,

$$\big|\mathrm{supp}\big(\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}\big) \cap \mathrm{supp}\big(\{\boldsymbol{\beta}_j^*\}_{j\in[d]}\big)\big| \geq s - d_0,$$

where $\mathrm{supp}(\mathbf{v})$ denotes the support of the vector $\mathbf{v}$.

Proof. By the proof of Theorem 3.2.1, the estimator obtained by the SPICA algorithm is an optimal solution under the constraint $\sum_{j\in[d]} \|\boldsymbol{\beta}_j\|_0 = s - s'$ for some $s' \in \{0, ..., d_0\}$. Thus, the corollary follows by analyzing the property of such minimizer. See Section B.3.3 of the Appendix for the proof. □

This corollary proves that the SPICA algorithm almost exactly recovers the support of the graph with high probability. As d and s increase, if d_0 is fixed, the ratio between the number of correctly estimated support over the number of true support converges to 1 with high probability. Also, similar to the estimation results, if the certificate of primal optimality (Cor. 3.2.2) holds, the estimator exactly recovers the true support with high probability.

3.3.2 Ising Graphical Model

In this subsection, we consider the spatial Ising graphical model. Ising graphical model studies the conditional independences among random variables $X_j \in \{\pm 1\}$ for $j \in [d]$. Under Ising graphical model, the joint distribution of $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ is

$$\mathbb{P}(X_1 = x_1, ..., X_d = x_d) = \frac{1}{Z(\boldsymbol{\beta})} \exp\Big(\sum_{j\neq k} \frac{\beta_{jk} x_j x_k}{4}\Big),$$

where $Z(\boldsymbol{\beta})$ is some unknown partition function; each β_{jk} describes the interaction between vertex j and vertex k, and $\beta_{jk} = \beta_{kj}$.

Since the function $Z(\boldsymbol{\beta})$ is not given, directly estimating β_{jk}'s is not tractable. For the i-th observation $\mathbf{x}_i = (x_{i1}, ..., x_{id})^T \in \{\pm 1\}^d$, let $\theta_{ij} = \mathbb{P}(X_j = x_{ij} | \boldsymbol{X}_{\setminus j} = \mathbf{x}_{i,\setminus j})$ be the conditional distribution of the j-th vertex given others. Adopting the composite likelihood idea, we have

$$\theta_{ij} = \frac{\exp\big(\sum_{k:k\neq j} \beta_{jk} x_{ij} x_{ik}\big)}{\exp\big(\sum_{k:k\neq j} \beta_{jk} x_{ij} x_{ik}\big) + 1}.$$

We have that the negative conditional log-likelihood of the j-th vertex is

$$\mathcal{L}_j(\boldsymbol{\beta}_j) = -\frac{1}{n}\sum_{i=1}^{n}\log(\theta_{ij}),$$

Incorporating the spatial information, we have the prior information that $\boldsymbol{\beta}_{jk} = 0$ if $(j,k) \notin \mathcal{N}_j$ for each j, where $|\mathcal{N}_j| = d_0$. Adopting the total cardinality approach, we estimate $\boldsymbol{\beta}_j$'s by solving the following problem

$$\min_{\boldsymbol{\beta}_j} \frac{1}{d}\sum_{j=1}^{d}\mathcal{L}_j(\boldsymbol{\beta}_j), \text{ subject to } \sum_{j=1}^{d}\|\boldsymbol{\beta}_j\|_0 \leq K.$$

Next, we analyze the statistical properties of the estimators $\{\widehat{\boldsymbol{\beta}}_j\}_{j=1}^{d}$ obtained by the SPICA algorithm. We impose the following mild assumptions:

Assumption 3.3.5. Under Ising model with parameters $\{\boldsymbol{\beta}_j^*\}_{j\in[d]}$, assume:

- (A.1) $\|\boldsymbol{\beta}_j^*\|_\infty \leq R$ for some $R \in (0,\infty)$.
- (A.2) The population Hessian matrix with respect to any subset $\mathcal{K} \subset \{1,...,dd_0\}$, satisfies the local sparse eigenvalue condition that $\Lambda_{\min}\{\mathbb{E}[\nabla^2_{\mathcal{K}\mathcal{K}}\mathcal{L}(\boldsymbol{\beta}^*)]\} > 2\rho$, where $|\mathcal{K}| = 2K$, $\mathcal{L}(\boldsymbol{\beta}^*) = \sum_{j=1}^{d}\mathcal{L}_j(\boldsymbol{\beta}_j^*)$, $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^{*T},...,\boldsymbol{\beta}_d^{*T})^T$, and $\rho > 0$ is a constant.

Note that assumption (A.1) is used in most literatures. For assumption (B.2), we only assume such a sparse eigenvalue condition at the point $\boldsymbol{\beta}^*$. This is essential for the identifiability of $\boldsymbol{\beta}^*$. Existing work (Ravikumar et al., 2010; Xue et al., 2012) imposes additional assumptions such as incoherence condition on the population Hessian matrix. Thus, our assumption is weaker than existing work. The next theorem provides the fast rate of convergence of the estimator obtained by the SPICA algorithm.

Theorem 3.3.6. *Suppose that Assumption 3.3.5 holds, and assume that the n independent samples are generated from a Ising model with parameters* $\boldsymbol{\beta}_j^* \in \mathbb{R}^{d_0}$ *for all* $j \in [d]$ *and*

$\sum_{j\in[d]} \|\boldsymbol{\beta}_j^*\|_0 = s \leq K$. *We have, with probability at least* $1 - \mathcal{O}(d^{-1})$,

$$\frac{1}{d}\sum_{j=1}^{d} \|\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*\|_2^2 \leq \underbrace{C_1 \cdot \frac{K \log d}{dn}}_{\textit{statistical error}} + \underbrace{C_2 \cdot \frac{1}{d}}_{\textit{duality gap}},$$

where C_1 *and* C_2 *are two constants, and do not depend on* K, d *and* n.

Proof. See Section B.3.4 of the Appendix for the proof. □

3.4 The Complexity of Spatial-Graphical Model Problem

In this section, we study the complexity of problem (3.1.1) in the general case that we consider problem

$$\min_{\boldsymbol{\beta}_j \in \mathcal{C}_j} \frac{1}{d}\sum_{j=1}^{d} \mathcal{L}_j(\boldsymbol{\beta}_j), \text{ subject to } \sum_{j=1}^{d} \mathcal{R}_j(\boldsymbol{\beta}_j) \leq K, \tag{3.4.1}$$

where $\mathcal{R}_j(\cdot)$ is some nonconvex function. We prove that problem (3.4.1) is NP-complete by relating it to a classical discrete NP-complete problem - the knapsack problem. In the special case of $\mathcal{R}_j(\boldsymbol{\beta}_j) = \|\boldsymbol{\beta}_j\|_0$, we show that the problem admits a polynomial-time algorithm. In the general setting where the nonconvex constraints $\mathcal{R}_j$'s are vector-valued, we prove that the problem does not admit a fully polynomial-time approximation scheme unless P = NP, and the problem is fundamentally more difficult. For example, we cannot solve problems under the constraint $\sum_{j\in[d]} \mathcal{R}_j(\boldsymbol{\beta}_j) \leq (K_1, K_2)^T$, where $\mathcal{R}_j(\boldsymbol{\beta}_j) = (\|\boldsymbol{\beta}_j\|_0, \sum_{k=1}^{d_0-1} \|\beta_{j,k+1} - \beta_{jk}\|_0)^T$.

3.4.1 Knapsack Problem and Complexity

The knapsack problem plays an important role in combinatorics. It is motivated from applications in resource allocation, where the goal is to maximize the total utility under

capacity constraints. Its simplest form is the following 0-1 knapsack problem:

$$\max_{x_j} \sum_{j=1}^{d} c_j x_j, \text{ subject to } \sum_{j=1}^{d} b_j x_j \leq b_0, \ x_j \in \{0,1\}, \quad \text{for } j = 1,..,d, \tag{3.4.2}$$

where c_j's, b_j's and b_0 are positive integers. The input to the 0-1 knapsack problem includes: the constant c_j which is the value of the j-th item; the constant b_j which is the cost of the j-th item, and the constant b_0 which is the total budget. Let $\mathbf{c} = (c_1, ..., c_d)^T$ and $\mathbf{b} = (b_1, ..., b_d)^T \in \mathbb{R}^d$. We refer to problem (3.4.2) as the 0-1 knapsack problem with input $(\mathbf{c}, \mathbf{b}, b_0)$. This problem is known to be NP-complete (Williamson and Shmoys, 2011).

An important variant of the 0-1 knapsack problem is the *multiple-row knapsack problem* (also known as the multiple-dimensional knapsack problem):

$$\begin{aligned} &\max_{x_j} \sum_{j=1}^{d} c_j x_j, \\ &\text{subject to } \sum_{j=1}^{d} b_j^{(\ell)} x_j \leq b_0^{(\ell)}, \quad \text{for all } \ell = 1, ..., L, \ x_j \in \{0,1\}. \end{aligned} \tag{3.4.3}$$

In comparison with the 0-1 knapsack problem, this problem has multiple-row constraints. The multiple-row knapsack problem is fundamentally more difficult than the 0-1 knapsack problem. It is NP-hard to solve the problem to an arbitrary precision. More specifically, it is shown that finding a fully polynomial time approximation scheme for the multiple-row knapsack problem is NP-hard (Magazine and Chern, 1984), which is defined below.

Definition 3.4.1. An approximation scheme for a maximization problem (P) is an algorithm that takes two inputs: One is the problem instance P, and the other is a desired numerical accuracy $\epsilon > 0$. Denote by $f^* > 0$ the optimal value of P. The algorithm produces a solution for P with objective value $f(P)$ such that $\{f^* - f(P)\}/f^* \leq \epsilon$. If the running time for the algorithm is bounded by a polynomial function of $1/\epsilon$ and the problem size, it is a fully polynomial time approximation scheme.

To facilitate our discussion, we briefly review some definitions in computational complexity theory in Section B.5 of the Appendix. We refer to Williamson and Shmoys (2011) for more detailed discussion about the knapsack problem and computational complexity theory.

3.4.2 NP-Completeness of Problem (3.4.1)

In this subsection, we prove that problem (3.4.1) is NP-complete. To prove that problem (3.4.1) is NP-complete, we shall construct a two-way polynomial time reduction between problem (3.4.1) and 0-1 knapsack problem (3.4.2). We show that given one instance of the 0-1 knapsack problem or problem (3.4.1), we can construct another instance of the other problem within a polynomial-time, and by solving the other instance we can recover the solution to the original instance. As we discussed in the introduction, the form of loss functions $\mathcal{L}_j$'s depends on the specific statistical model. Without loss of generality, we assume all $\mathcal{L}_j$'s are of a same form (least square or logistic loss for example), and each $\mathcal{L}_j$ only depends on some input data $(\mathbb{X}_j, \mathbb{Y}_j)$. Thus, each $\mathcal{L}_j(\boldsymbol{\beta}_j)$ can be represented as $\mathcal{L}(\boldsymbol{\beta}_j; \mathbb{X}_j, \mathbb{Y}_j)$. We consider finding an ϵ-optimal solution to the problem

$$\min_{\boldsymbol{\beta}_j} \sum_{j=1}^{d} \mathcal{L}(\boldsymbol{\beta}_j; \mathbb{X}_j, \mathbb{Y}_j), \text{ subject to } \sum_{j=1}^{d} \mathcal{R}_j(\boldsymbol{\beta}_j) \leq b_0, \tag{3.4.4}$$

with input $\big(\{\mathbb{X}_j, \mathbb{Y}_j\}_{j\in[d]}, b_0, \epsilon\big)$, where we say a solution is ϵ-optimal if its corresponding objective value is within ϵ of the optimal value, and the solution is feasible. Problem (3.4.4) can be continuous since both objective and constraint functions in (3.4.4) can be continuous, and 0-1 knapsack problem is discrete. To connect the two problems, we need to "discretize" problem (3.4.1). We first consider the loss functions. We impose the following assumption.

- (B.1) Given positive constants c_j's for all $j \in [d]$, we can find $\mathbb{X}_j$, $\mathbb{Y}_j$ and a constant c_0 within a polynomial time such that

$$\mathcal{L}(0; \mathbb{X}_j, \mathbb{Y}_j) = c_0 \quad \text{and} \quad \min_{\beta_j \in \mathbb{R}} \mathcal{L}(\beta_j; \mathbb{X}_j, \mathbb{Y}_j) = -c_j + c_0.$$

This assumption is satisfied for most statistical models. For example, if $\mathcal{L}(\beta_j; \mathbb{X}_j, \mathbb{Y}_j) = \|\mathbb{Y}_j - \mathbb{X}_j\beta_j\|_2^2/n$, it is easy to verify that letting $\mathbb{Y}_j = (\sqrt{c_0}, \sqrt{c_0})^T$ and $\mathbb{X}_j = \left(\sqrt{c_j + c_j'}, \sqrt{c_j - c_j'}\right)^T$, where $c_j' = \sqrt{2c_jc_0 - c_0^2}$, satisfy this assumption.

Next, we look at constraint functions $\mathcal{R}_j$'s. Given a problem instance of (3.4.4), we need to efficiently construct a knapsack problem of which the constraint is similar to the problem instance (3.4.4). Since the knapsack problem is discrete, and problem (3.4.4) is possibly continuous, we assume that we can efficiently "discretize" the constraint functions $\mathcal{R}_j$'s, where we impose the following assumption.

- (A.2) For any j, given any $\delta > 0$ and any set $[-r, r]^{d_0}$ for some $r > 0$, we can find a finite discretization $\mathcal{B}$ of the set that for any point $\boldsymbol{\beta} \in [-r, r]^{d_0}$, there exists a point $\mathbf{p} \in \mathcal{B}$ such that $\|\mathbf{p} - \boldsymbol{\beta}\|_2 \leq \delta$ and $\mathcal{R}_j(\mathbf{p}) \leq \mathcal{R}_j(\boldsymbol{\beta})$ in a polynomial time.

This assumption holds for most common $\mathcal{R}_j$'s in statistical applications. For example, suppose $\mathcal{R}_j$'s are SCAD functions. We have that the discretization $\{0, \pm\sqrt{\delta/d_0}, \pm 2\sqrt{\delta/d_0}, ..., \pm p^*\sqrt{\delta/d_0}\}^{d_0}$ satisfies the assumption, where $p^* = \text{argmax}_{p \in \mathbb{N}}\{p\sqrt{\delta/d_0} \leq r\}$, and $\mathbb{N}$ is the set of natural numbers.

Next, we provide the main theorem of this section. We show that given one instance of 0-1 knapsack problem (3.4.2), we can construct an instance of problem (3.4.4) within a polynomial-time, and by solving the instance of problem (3.4.4) we can recover the solution to the original instance of 0-1 knapsack problem, and vice versa. This proves that problem (3.4.4) is NP-complete since 0-1 knapsack problem (3.4.2) is known to be NP-complete.

Theorem 3.4.2. *Under assumptions (B.1)-(B.2), the nonconvex constrained optimization problem* (3.4.4) *is NP-complete.*

Proof. See Section B.6 of the Appendix for the detailed proof. □

3.4.3 Polynomial-Time Algorithm in the Case of ℓ_0-Constrained Problem

Though problem (3.4.4) is NP-complete, we show that the special case of problem (3.4.4) under a total cardinality constraint, i.e., the case where $\mathcal{R}_j(\boldsymbol{\beta}_j) = \|\boldsymbol{\beta}_j\|_0$, admits a polynomial-time algorithm. Specifically, given an instance of problem (3.4.4), we map it to an instance of multiple-choice knapsack problem, and by solving the instance of multiple-choice knapsack problem efficiently, we recover the solution to the instance of problem (3.4.4).

Let us first introduce multiple-choice knapsack problem. Denote by $\mathbf{C} = (\mathbf{c}_1, ..., \mathbf{c}_d)^T \in \mathbb{R}^{d \times d_0}$ and $\mathbf{B} = (\mathbf{b}_1, ..., \mathbf{b}_d)^T \in \mathbb{R}^{d \times d_0}$, where $\mathbf{c}_j = (c_{j1}, ..., c_{jd_0})^T$ and $\mathbf{b}_j = (b_{j1}, ..., b_{jd_0})^T$. Consider the multiple-choice knapsack problem with input $(\mathbf{C}, \mathbf{B}, b_0)$, where all b_{jk}'s and b_0 are positive integers:

$$\begin{aligned} &\max_{x_{jk}} \sum_{j=1}^{d} \sum_{k=1}^{d_0} c_{jk} x_{jk} \\ &\text{subject to } \sum_{j=1}^{d} \sum_{k=1}^{d_0} b_{jk} x_{jk} \leq b_0, \sum_{k=1}^{d_0} x_{jk} \leq 1,\ x_{jk} \in \{0, 1\}, \end{aligned} \tag{3.4.5}$$

for all $j \in [d]$ and all $k \in [d_0]$. Given an instance of problem (3.4.4) under the ℓ_0-constraint, we map the instance to an instance of multiple-choice knapsack problem (3.4.5). For each j, we solve the subproblems

$$\widehat{\boldsymbol{\beta}}_j(k) = \underset{\boldsymbol{\beta}_j \in \mathcal{C}_j}{\operatorname{argmin}} \mathcal{L}(\boldsymbol{\beta}_j; \mathbb{X}_j, \mathbb{Y}_j), \quad \text{subject to } \|\boldsymbol{\beta}_j\|_0 \leq k, \text{ for } k = 0, 1, ..., d_0.$$

Since we assume that d_0 is a constant, the cost of computing all $\widehat{\boldsymbol{\beta}}_j(k)$'s increases linearly as d increases. Let $b_{jk} = k$, $b_0 = K$ and $c_{jk} = -\mathcal{L}\big(\widehat{\boldsymbol{\beta}}_j(k); \mathbb{X}_j, \mathbb{Y}_j\big) + c_0$, where $c_0 > \max_{j,k} \mathcal{L}\big(\widehat{\boldsymbol{\beta}}_j(k); \mathbb{X}_j, \mathbb{Y}_j\big)$ for $j \in [d]$ and $k \in [d_0]$. We obtain a multiple-choice knapsack problem of the form (3.4.5). Denote by $\{x^*_{jk}\}$ an optimal solution to the multiple-choice knapsack problem. We have that $\{x^*_{jk}\}$ recovers an optimal solution to the ℓ_0-constrained problem by setting $\widehat{\boldsymbol{\beta}}_j = \widehat{\boldsymbol{\beta}}_j(k)$ if $x^*_{jk} = 1$.

Next, we present a dynamic programming approach to solve the multiple-choice knapsack problem, which is a variant of Pisinger (1995). We formulate a dynamic program with the state variable (d', k'), where $1 \le d' \le d$ and $0 \le k' \le K$. The dimension of the state space is $d(K+1)$. We define the value function of a state (d', k') to be the optimal value for the multiple-choice knapsack problem considering only multiple-choice sets 1 to d' with constraint k'. In another words, let

$$\begin{aligned} V(d', k') &= \max_{x_{jk}} \sum_{j=1}^{d'} \sum_{k=0}^{d_0} c_{jk} x_{jk}, \\ \text{subject to } &\sum_{j=1}^{d'} \sum_{k=0}^{d_0} k x_{jk} \le k', \sum_{k=0}^{d_0} x_{jk} \le 1,\ x_{jk} \in \{0,1\}. \end{aligned}$$

Thus, $V(d, K)$ is the optimal value for the original multiple-choice knapsack problem. To facilitate our discussion, fixing c_{jk}'s, we denote the knapsack problem with first d' multiple choice set and constraint variable k' by $(MK_{d',k'})$, i.e., we let

$$\begin{aligned} (MK_{d',k'}): \quad &\max_{x_{jk}} \sum_{j=1}^{d'} \sum_{k=0}^{d_0} c_{jk} x_{jk}, \\ \text{subject to } &\sum_{j=1}^{d'} \sum_{k=0}^{d_0} k x_{jk} \le k', \sum_{k=0}^{d_0} x_{jk} \le 1,\ x_{jk} \in \{0,1\}. \end{aligned}$$

Denote by $\{x^*_{jk}\}$ the optimal solution to the problem $(MK_{d,K})$. Let $k' = \sum_{j=1}^{d'} \sum_{k=0}^{d_0} k x^*_{jk}$. To efficiently solve the problem, a key observation is that the partial solution $\{x^*_{jk}\}$ for

$j = 1, ..., d'$ and $k = 0, ..., d_0$ is the optimal solution solution for the problem $(MK_{d',k'})$. This can be proved by contradiction that if the assertion does not hold, we can replace the partial solution with the optimal solution to $(MK_{d',k'})$, and we keep the rest of the original optimal solution to $(MK_{d,K})$ the same. The sum of the corresponding objectives of the two partial solutions is greater than the original optimal objective. This leads to a contradiction.

Based on this observation, we find the optimal value $V(d, K)$ by a recursive algorithm based on following recursive equations:

$$V(1, k') = \max\Big\{V(1, k' - 1), \max\{c_{1k} : k \leq k'\}\Big\},$$

and for $d' > 1$

$$V(d', k') = \max_k \Big\{V(d', k' - 1), \max\{V(d' - 1, k' - k) + c_{d'k} : k \leq k'\Big\}. \tag{3.4.6}$$

The dynamic programming algorithm for solving the problem $(MK_{d,K})$ is summarized in Algorithm 2. The total number of states is $d(K+1)$ for the problem, and the computational

Algorithm 2 Dynamic Programming Algorithm for Problem (3.4.5)

Input: $c_{jk} \in \mathbb{R}_+$, K
Output: x^*_{jk}
$V(d', -1) \leftarrow 0$, $V(0, k') \leftarrow 0$, $\mathcal{S}(d', k') = 0$ for all d' and k'. $d' \leftarrow 0$.
Let $\mathcal{S}(1, k') \leftarrow \text{argmax}_k\{c_{1k} : k \leq k'\}$.
while $d' < d$ **do**
    Let $d' \leftarrow d' + 1$. Solve (3.4.6) for $1 \leq k' \leq K$.
    for $k' = 0 : K$ **do**
        if $\max_k \big\{V(d' - 1, k' - k) + c_{d'k} : k \leq k', V(d' - 1, k' - k) > 0\big\} > V(d', k' - 1)$ **then**
            Let $\mathcal{S}(d', k') \leftarrow \text{argmax}_k \big\{V(d' - 1, k' - k) + c_{d'k} : k \leq k', V(d' - 1, k' - k) > 0\big\}$.
        end if
    end for
end while
$k' \leftarrow K$.
while $d' > 0$ **do**
    Let $x^*_{d',\mathcal{S}(d',k')} = 1$, $k' \leftarrow k' - \mathcal{S}(d', k')$, $d' \leftarrow d' - 1$.
end while

complexity for computing each $V(d', k')$ is $\mathcal{O}(d_0 + 1)$. Thus, the complexity of computing $V(d, K)$ is of the order $\mathcal{O}(dd_0K)$. In our problem, the number K is upper-bounded by dd_0, so the computational complexity is of the order $\mathcal{O}(d^2d_0^2)$. Note that this does not include the computation for the coefficients c_{jk}'s. To compute all c_{jk}'s, for each sub-problem $\mathcal{L}_j$, we need to enumerate all 2^{d_0} possible combinations of the support of $\boldsymbol{\beta}_j \in \mathbb{R}^{d_0}$. Thus, applying dynamic programming techniques, the total computational complexity for solving the ℓ_0-constraint problem is of order $\mathcal{O}(2^{d_0}d + d^2d_0^2)$, which is still a polynomial order of the dimension d.

In summary, Algorithm 2 is a dynamic programming algorithm that runs in a polynomial-time. However, it can be very expensive in practice as it requires enumerating and solving all subproblems. In comparison, SPICA algorithm (Alg. 1) avoids solving all subproblems and is more practical. We compare the computational complexities of the dynamic programming approach and the SPICA algorithm in Section B.7 of the Appendix.

Meanwhile, we point out that the dynamic programming approach becomes significantly more expensive when $\mathcal{R}_j$'s are some continuous functions instead of the ℓ_0-norm. When $\mathcal{R}_j$ is continuous, our reduction to the multiple-choice knapsack problem requires a fine discretization of $\mathcal{R}_j$. This may result in a large number of choices in the constructed knapsack problem, making the dynamic programming approach inefficient. In comparison, when $\mathcal{R}_j$ is the ℓ_0-constraint, the values of $\mathcal{R}_j$ are naturally discrete. The resulting knapsack problem has at most d_0 choices, which is a relatively small number. In general, the dynamic programming approach to problem (3.4.4) is practically slow, even though it is a polynomial-time algorithm.

3.4.4 A "Harder" Result in the Case of Vector-Valued Constraint

In this subsection, we consider the case where the functions $\mathcal{R}_j(\boldsymbol{\beta}_j)$'s are vector-valued. Specifically, we consider the problem:

$$\min_{\beta_j} \sum_{j=1}^{d} \mathcal{L}(\boldsymbol{\beta}_j; \mathbb{X}_j, \mathbb{Y}_j), \text{ subject to } \sum_{j=1}^{d} \mathcal{R}_j^{(\ell)}(\boldsymbol{\beta}_j) \leq b_0^{(\ell)}, \tag{3.4.7}$$

for $\ell = 1, ..., L$, where $L \geq 1$. This problem contains (3.4.4) as a special case. In practice, one of the row-constraints can be the total sparsity constraint, and the other can be a fused-type constraint. Intuitively speaking, finding an ϵ-optimal solution to the problem (3.4.7) should not be more difficult than problem (3.4.4). However, the next theorem proves that the multiple-row constraints case (3.4.7) is fundamentally more difficult.

Theorem 3.4.3. *Under assumptions (B.1)-(B.2), if $L > 1$, finding a fully polynomial-time approximation scheme for problem* (3.4.7) *is NP-hard.*

Proof. The proof is based on constructing a two-way polynomial-time reduction between the multiple-row knapsack problem (3.4.3) and the problem (3.4.7). The argument is analogous to the proof of Theorem 3.4.2, and we omit it to avoid repetition. Then, as shown in Magazine and Chern (1984), there does not exist a fully polynomial-time approximation scheme to solve the two-row multiple-choice knapsack problem unless we assume P = NP. □

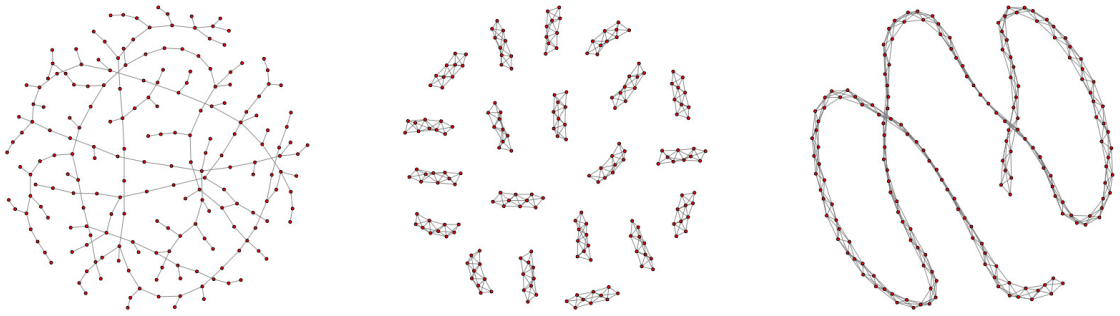
This theorem establishes one of the strongest forms of complexity, and shows the problem (3.4.7) is fundamentally hard to solve. In comparison, when there exists only one total cardinality constraint, we can solve the problem within a polynomial-time by dynamic programming.

3.5 Numerical Results

In this section, we conduct extensive numerical experiments to test the SPICA algorithm in comparison with ℓ_1-penalized method. We compare the parameter estimation and graph recovery performances of these two methods using both synthetic and real datasets. For ease of presentation, we provide the numerical performances under the Gaussian graphical model.

3.5.1 Synthetic Data

We first use synthetic data. We consider three different sets of parameters: (i) $n = 100$, $d = 1,000$; (ii) $n = 100$, $d = 2,000$; (iii) $n = 100$, $d = 5,000$, and we let the number of potential neighbors $d_0 = 10$. We further consider three different models for generating undirected graphs and precision matrices. Figure 3.2 illustrates sample graphs under these models. We repeat each setting for 100 times and report the averaged performance.



(a) Scale-Free (b) Block (c) Band

Figure 3.2: Examples of the three graph patterns we consider in the simulation study.

$*$ **Scale-free graph**. We generate the graph by the preferential attachment mechanism. We begin with a graph with a chain of 2 vertices. At iteration j, we add a new vertex to the graph. The new vertex j connects to one of the previous d_0 vertices, with a probability which is proportional to the number of degrees of the existing vertex. Mathematically,

let p_i be the probability that the new vertex j will connect to the existing vertex i is, $p_i = k_i / \sum_{i'=\min\{1,j-d_0\}}^{j-1} k_{i'}$, where k_i is the current degree of the vertex i. Thus, the resulting graph has $d-1$ edges. Given the graph, we generate the corresponding adjacency matrix $\mathbf{A}$ by setting the diagonal elements to be 0, and we set the nonzero off-diagonal elements to be $\rho = 0.3$, 0.5 or 0.7. Then, we construct the precision matrix $\mathbf{\Theta}$ as

$$\mathbf{\Theta} = \mathbf{D}\big[\mathbf{A} + \big\{\big|\Lambda_{\min}(\mathbf{A})\big| + 0.2\big\} \cdot \mathbf{I}_d\big]\mathbf{D}, \tag{3.5.1}$$

where $\Lambda_{\min}(\mathbf{A})$ denotes the smallest eigenvalue of $\mathbf{A}$; $\mathbf{I}_d$ denotes the identity matrix, and $\mathbf{D}$ is a diagonal matrix with $D_{jj} = 1$ for $j = 1, ..., d/2$ and $D_{jj} = 3$ for $j = d/2+1, ..., d$. Finally, we generate the multivariate Gaussian samples: $\mathbf{x}_1, ..., \mathbf{x}_n \sim N_d(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = \mathbf{\Theta}^{-1}$.

$*$ **Block graph**. We construct the precision matrix $\mathbf{A}$ as a block diagonal matrix. Each block is of the size 8. We set the nonzero off-diagonal entries to be ρ and diagonal entries to be 1. This matrix is positive definite. The graph has $3.5d$ edges, and we let the precision matrix be $\mathbf{\Theta} = \mathbf{DAD}$.

$*$ **Band graph**. Given d vertices indexed by $j = 1, ..., d$, we generate edges between the vertices whose corresponding coordinates are at distance less than or equal to 3. The resulting graph has $3d-6$ edges. Given the graph, we construct the precision matrix same as (3.5.1).

We first consider the graph recovery performances of SPICA algorithm and the ℓ_1-penalized method. In particular, we evaluate the graph recovery performance by looking at the false positive and false negative rates. In particular, let $\widehat{G}^K = (V, \widehat{E}^K)$ be an estimated graph under the total cardinality constraint with tuning parameter K. The number of false positive discoveries using tuning parameter K is $\mathrm{FP}(K) = |\widehat{E}^K \backslash E|$, where $A \backslash B = \{a : a \in A \text{ and } a \notin B\}$, and the number of false negative discoveries with K is $\mathrm{FP}(K) = |E \backslash \widehat{E}^K|$. Consequently, we define the corresponding false positive rate (FPR) and

the false negative rate (FNR) as

$$\mathrm{FPR}(K) = \frac{\mathrm{FP}(K)}{\binom{d}{2} - |E|} \text{ and } \mathrm{FNR}(K) = \frac{\mathrm{FN}(K)}{|E|}.$$

We plot the receiver operating characteristic (ROC) curves using $\{\mathrm{FNR}(K), 1-\mathrm{FPR}(K)\}$ for the SPICA algorithm. Also, we plot the averaged ROC curves for the ℓ_1-penalized method for comparisons.

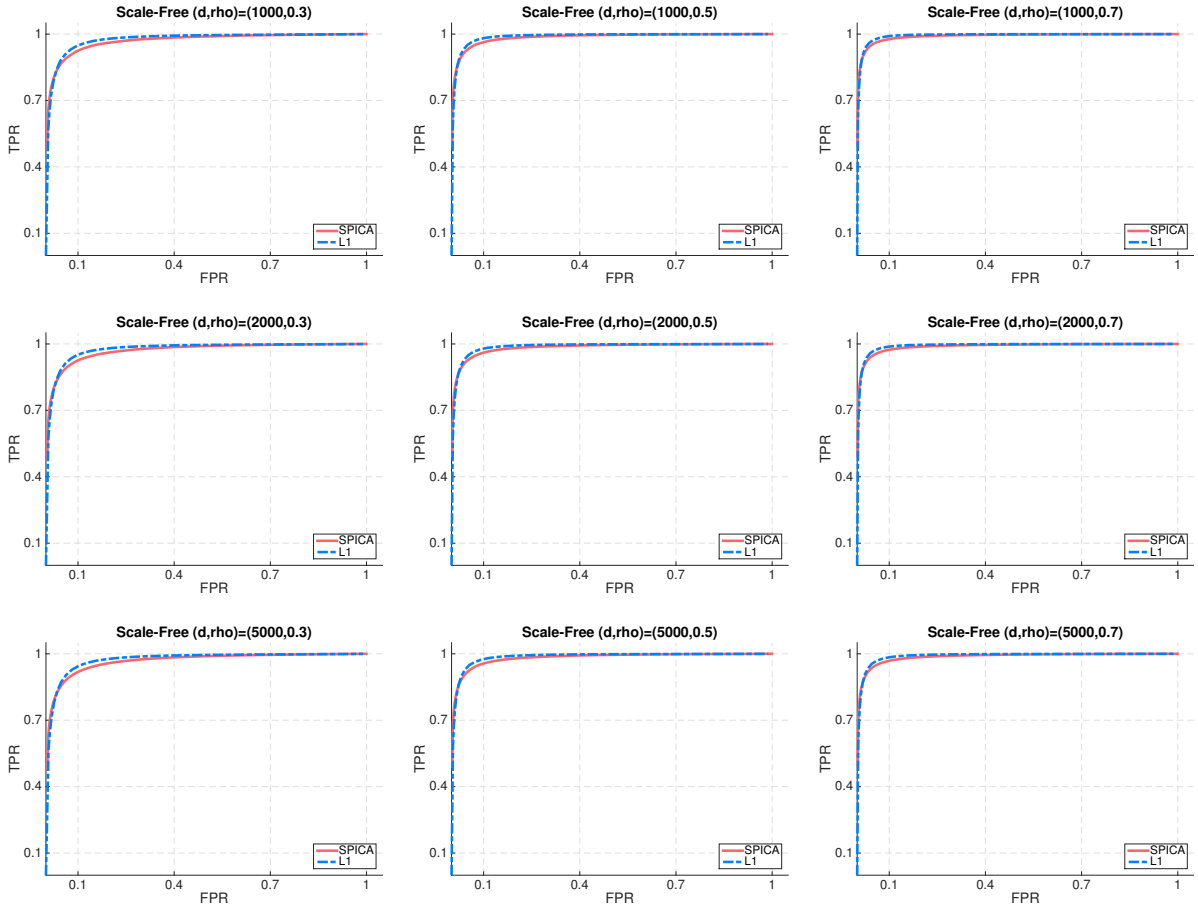


Figure 3.3: ROC curves for Scale-Free Model under different settings.

By Figures 3.3, 3.4 and 3.5, we see that the SPICA algorithm performs better than ℓ_1-penalized method under the block and band models, and the two methods perform similarly under the scale-free model. Thus, we conclude that the SPICA algorithm works better than

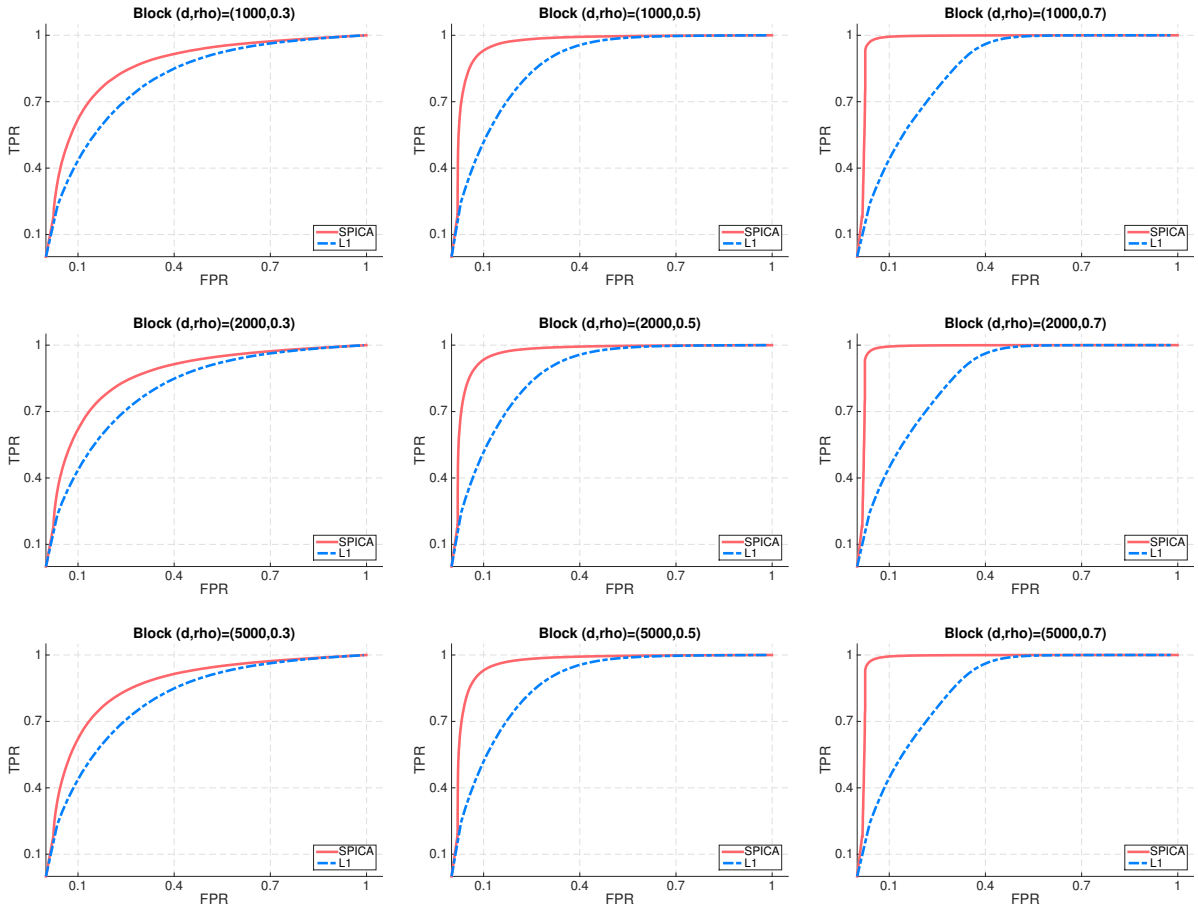


Figure 3.4: ROC curves for Block Model under different settings.

the ℓ_1-penalized method when the number of edges of the graph is larger. Also, for block and band models, we observe that the margin of the SPICA algorithm over the ℓ_1-penalized method increases when ρ increases. This phenomenon has an intuitive explanation that the penalization term $\lambda \sum_{j\in[d]} \|\boldsymbol{\beta}_j\|_1$ increases with the signal strength of $\boldsymbol{\beta}_j^*$'s, which induces more estimation bias, and results a worse performance in graph recovery.

We then compare the SPICA algorithm with the ℓ_1-penalized method from the perspective of parameter estimation. We select the tuning parameters by stability selection (Meinshausen and Bühlmann, 2010), and we report the error $\sum_{j\in[d]} \|\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*\|_2^2$ under all settings mentioned above. We observe that the SPICA algorithm performs better than the ℓ_1-penalized method as the degree or the signal strength increases. In addition, we observe

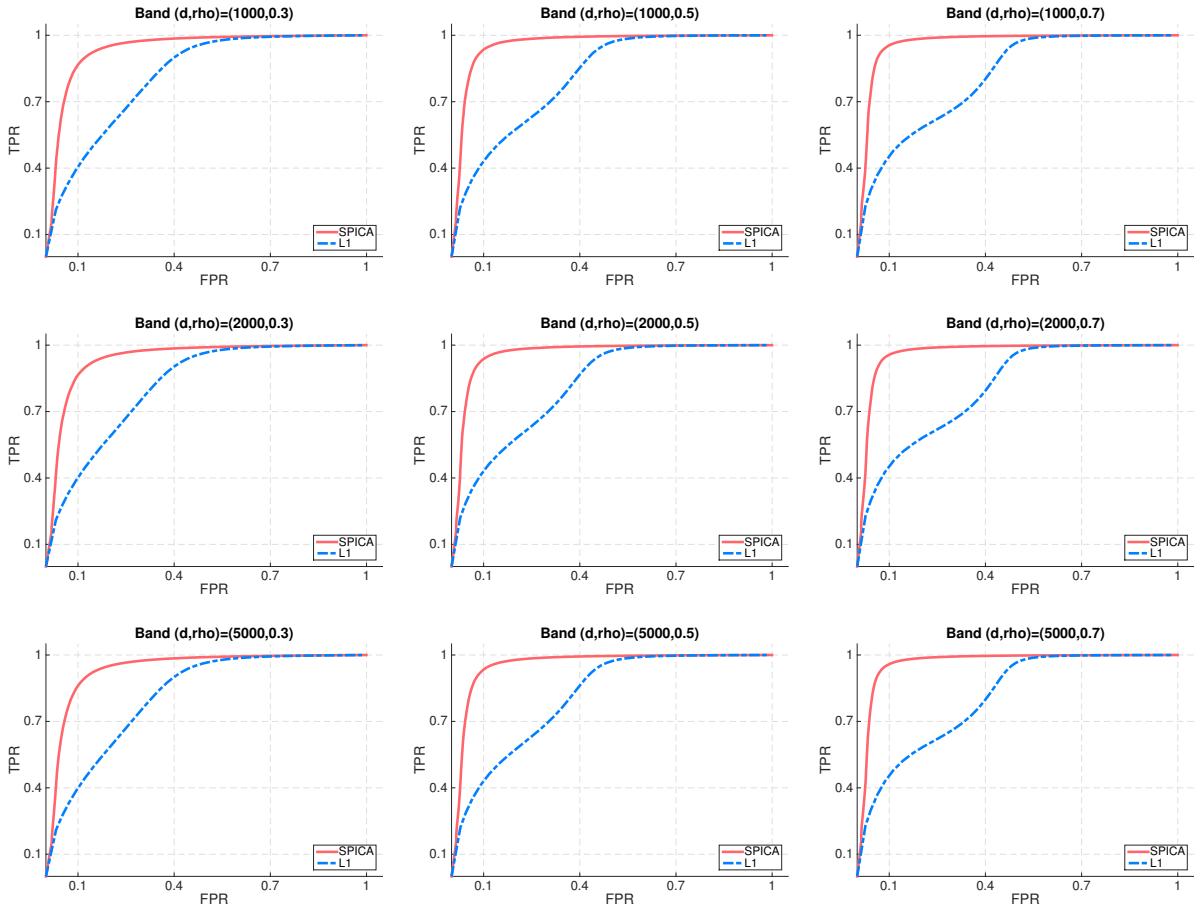


Figure 3.5: ROC curves for Band Model under different settings.

some interesting phenomenon. In the scale-free and block models, the errors decrease as ρ increases for both methods. This is intuitive that as the increase of signal strength helps graph recovery, and consequently it also helps parameters estimation. In the band model, same as the scale-free and block models, the errors decrease as ρ increases for the SPICA algorithm. However, the errors increase as ρ increases for the ℓ_1-penalized method. This again confirms the intuition that as the ℓ_1-norm $\sum_{j\in[d]} \|\boldsymbol{\beta}_j^*\|_1$ increases, the penalization terms induces more biases. In comparison, the total cardinality constraint approach does not induce any biases.

To summarize, the advantage of the SPICA algorithm over the ℓ_1-penalized method is well illustrated from the perspectives of both graph support recovery and parameters estimation.

We also point out that the certificate of primal optimality, i.e., $\sum_{j\in[d]} \|\widehat{\boldsymbol{\beta}}_j\|_0 = K$, holds in more than 98% cases, which means that the SPICA algorithm generates the optimal solution to the problem with total cardinality constraint in these cases. This further shows the reliability of the SPICA algorithm.

3.5.2 Sensor Network Data

We also use wireless sensor network data to conduct tests. Our goal is to estimate how the sensors are connected. In practical applications, depending on the sensor type, the communication network of sensors might be known or not. In our data, the communication network is given. The reason we choose this type of data is that our primary goal is to evaluate the two different methods, and without such information, it is difficult to tell which method works better. In the implementation of different methods, we do not use the information of how the sensors are connected, and we only use such information to evaluate the results at a later stage.

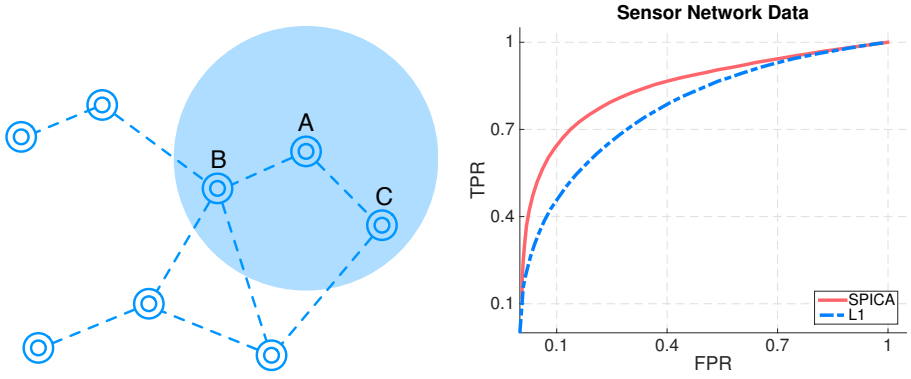


Figure 3.6: Left: In the sensor network, each sensor can only connect to another sensor if they are physically close on the plane. Each dashed line represents a possible connection. For example, sensor A can only possibly connect to B or C, but not others. Right: ROC curve for sensor network data.

As discussed in the introduction, in a sensor network, each sensor can only connect to another if they are sufficiently close as illustrated in Figure 3.6. Thus, estimating the

Table 3.1: Quantitative comparisons of the SPICA and ℓ_1-penalized method on different models. We report the averaged Frobenius norm $\sum_{j\in[d]} \|\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*\|_2^2$ with sample variance in the parenthes after repeating the simulation 100 times.

			$\rho = 0.3$		$\rho = 0.5$		$\rho = 0.7$	
Model	n	d	SPICA	ℓ_1	SPICA	ℓ_1	SPICA	ℓ_1
Scale-Free	100	1,000	61.329	61.670	54.871	54.656	51.3893	51.7357
			(1.886)	(1.881)	(2.045)	(2.150)	(2.092)	(2.195)
		2,000	122.402	122.635	110.085	110.168	104.119	105.225
			(3.170)	(3.291)	(3.084)	(3.276)	(2.448)	(2.756)
		5,000	310.646	312.157	280.662	280.588	266.137	263.420
			(4.266)	(3.843)	(5.722)	(5.353)	(9.237)	(11.315)
Block	100	1,000	102.757	101.108	86.691	93.019	74.326	80.351
			(2.319)	(1.953)	(2.684)	(2.507)	(1.930)	(1.994)
		2,000	205.958	204.684	172.149	185.602	147.552	159.882
			(4.228)	(3.923)	(2.964)	(3.952)	(3.531)	(4.019)
		5,000	519.001	519.991	434.613	467.328	368.587	399.727
			(5.806)	(5.541)	(5.650)	(5.913)	(6.207)	(6.108)
Band	100	1,000	89.666	91.822	83.516	97.139	80.624	103.440
			(2.598)	(1.788)	(2.798)	(2.407)	(2.405)	(3.390)
		2,000	180.204	183.039	164.144	189.883	160.774	206.789
			(3.320)	(3.073)	(4.948)	(4.876)	(3.981)	(4.964)
		5,000	452.647	460.142	417.141	483.318	398.361	513.636
			(7.767)	(7.230)	(4.835)	(7.448)	(5.677)	(5.911)

network of sensors fits into the spatial graphical model framework. In our data, we have $d = 3,592$ sensors, and each sensor can only connect with another if they are within 3 meters. On average, each sensor has 24 potential neighbors. We have in total $n = 98$ samples. Each sample contains a signal strength of each sensor. Taking a Gaussian graphical model approach, we test the SPICA algorithm and ℓ_1-penalized method. We plot the ROC curves of the SPICA algorithm and the ℓ_1-penalized method in Figure 3.6 . It is clear that the SPICA algorithm performs significantly better than the ℓ_1-penalized method. This shows that the SPICA algorithm is capable of estimating spatial graphical models in practice.

Chapter 4

High Dimensional Inference for the Cox Model

In this chapter, we develop a unified inferential framework by extending the classical score, Wald and partial likelihood ratio tests to high dimensional proportional hazards models. Our first step is to construct a decorrelated score function by applying a high dimensional projection approach. Towards the goal of performing the likelihood based inference, we further obtain the least favorable direction and propose a new type of least favorable partial likelihood function which is used to construct the likelihood ratio test.

Theoretically, we establish the asymptotic distributions of score, Wald and partial likelihood ratio statistics under both the null and Pitman alternatives. Empirically, we find that the partial likelihood ratio test is more powerful than the Wald and score tests, which shows the advantage of our likelihood ratio inference in finite samples. We also construct point-wise confidence intervals for the baseline hazard function and the conditional hazard function, and establish their asymptotic properties.

4.1 Background

We start with an introduction of notation. Let $\mathbf{a} = (a_1, ..., a_d)^T \in \mathbb{R}^d$ be a d dimensional vector and $\mathbf{A} = [a_{jk}] \in \mathbb{R}^{d\times d}$ be a d by d matrix. Let $\text{supp}(\mathbf{a}) = \{j : a_j \neq 0\}$. For $0 < q < \infty$, we define ℓ_0, ℓ_q and ℓ_∞ vector norms as $\|\mathbf{a}\|_0 = \text{card}\{\text{supp}(\mathbf{a})\}$, $\|\mathbf{a}\|_q = (\sum_{j=1}^d \|\mathbf{a}_j\|^q)^{1/q}$ and $\|\mathbf{a}\|_\infty = \max_{1\leq j\leq d} |a_j|$. We define the matrix ℓ_∞-norm as the elementwise sup-norm that $\|\mathbf{A}\|_\infty = \max_{1\leq j,k\leq d} |a_{jk}|$, and let $\|\mathbf{A}\|_0 = \sum_{1\leq j,k\leq d} \mathbf{1}\{a_{jk} \neq 0\}$ and $\|\mathbf{A}\|_1 = \sum_{1\leq j,k\leq d} |a_{jk}|$. Let $\mathbf{I}_d$ be the identity matrix in $\mathbb{R}^{d\times d}$. For a sequence of random variables $\{X_n\}_{n=1}^\infty$ and a random variable Y, we denote X_n weakly converges to Y by $X_n \xrightarrow{d} Y$. We denote $[n] = \{1, \ldots, n\}$.

4.1.1 Cox's Proportional Hazards Model

We briefly review the Cox's proportional hazards model. Let Q be the time to event; R be the censoring time, and $\boldsymbol{X}(t) = (X_1(t), ..., X_d(t))^T$ be the d dimensional time-dependent covariates at time t. We consider the non-informative censoring setting that Q and R are conditionally independent given $\boldsymbol{X}(t)$. Let $W = \min\{Q, R\}$ and $\Delta = \mathbf{1}\{Q \leq R\}$ denote the observed survival time and censoring indicator, where $\mathbf{1}\{\cdot\}$ denotes the indicator function. Let τ be the end of study time. We observe n independent copies of $\{(\boldsymbol{X}(t), W, \Delta) : 0 \leq t \leq \tau\}$,

$$\Big\{(\boldsymbol{X}_i(t), W_i, \Delta_i) : 0 \leq t \leq \tau\Big\}_{i\in[n]}.$$

We denote $\lambda\{t|\boldsymbol{X}(t)\}$ as the conditional hazard rate function at time t given the covariates $\boldsymbol{X}(t)$. Under the proportional hazards model, we assume that

$$\lambda\{t|\boldsymbol{X}(t)\} = \lambda_0(t) \exp\{\boldsymbol{X}^T(t)\boldsymbol{\beta}^*\},$$

where $\lambda_0(t)$ is an unknown baseline hazard rate function, and $\boldsymbol{\beta}^* \in \mathbb{R}^d$ is an unknown parameter.

4.1.2 Penalized Estimation

Following Andersen and Gill (1982), we introduce some counting process notation. For each i, let $N_i(t) := \mathbf{1}\{W_i \leq t, \Delta_i = 1\}$ be the counting process, and $Y_i(t) := \mathbf{1}\{W_i \geq t\}$ be the at risk process for subject i. Assume that the process $Y_i(t)$ is left continuous with its right-hand limits satisfying $\mathbb{P}(Y_i(t) = 1, 0 \leq t \leq \tau) > C_\tau$ for some positive constant C_τ. The negative log-partial likelihood is

$$\mathcal{L}(\boldsymbol{\beta}) = -\frac{1}{n}\Big(\sum_{i=1}^{n} \int_0^\tau \boldsymbol{X}_i^T(u)\boldsymbol{\beta} dN_i(u) - \int_0^\tau \log\big[\sum_{i=1}^{n} Y_i(u) \exp\{\boldsymbol{X}_i^T(u)\boldsymbol{\beta}\}\big] d\bar{N}(u)\Big),$$

where $\bar{N}(t) = \sum_{i=1}^n N_i(t)$.

When the dimension d is fixed and smaller than the sample size n, $\boldsymbol{\beta}^*$ can be estimated by the maximum partial likelihood estimator (Andersen and Gill, 1982). However, in high dimensional settings with $n < d$, the maximum partial likelihood estimator is not well defined. To solve this problem, Tibshirani (1997) and Fan and Li (2002) impose the sparsity assumption and propose the following penalized estimator

$$\widehat{\boldsymbol{\beta}} := \underset{\boldsymbol{\beta} \in \mathbb{R}^d}{\operatorname{argmin}} \big\{ \mathcal{L}(\boldsymbol{\beta}) + \mathcal{P}_\lambda(\boldsymbol{\beta}) \big\}, \tag{4.1.1}$$

where $\mathcal{P}_\lambda(\cdot)$ is a sparsity-inducing penalty function, and λ is a tuning parameter. Bradic et al. (2011) and Huang et al. (2013) establish the rates of convergence and oracle properties of the penalized maximum partial likelihood estimators $\widehat{\boldsymbol{\beta}}$ using SCAD and lasso penalties. For notational simplicities, we focus on the lasso penalized estimator (Tibshirani, 1997) in this work and indicate that similar properties hold for the SCAD penalized estimator. Existing works generally impose the following assumptions.

Assumption 4.1.1. The difference of the covariates is uniformly bounded:

$$\sup_{0\le t\le \tau}\max_{i,i'\le n}\max_{1\le j\le d}|X_{ij}(t)-X_{i'j}(t)|\le C_X,$$

for some constant $C_X>0$.

Assumption 4.1.2. For any set $\mathcal{S}\subset\{1,...,d\}$ where $|\mathcal{S}|\asymp s$ and any vector $\mathbf{v}$ belonging to the cone $\mathcal{C}(\xi,\mathcal{S})=\{\mathbf{v}\in\mathbb{R}^d:\|\mathbf{v}_{\mathcal{S}^C}\|_1\le\xi\|\mathbf{v}_{\mathcal{S}}\|_1\}$, it holds that

$$\kappa\big(\xi,\mathcal{S};\nabla^2\mathcal{L}(\boldsymbol{\beta}^*)\big)=\inf_{\mathbf{0}\neq\mathbf{v}\in\mathcal{C}(\xi,\mathcal{S})}\frac{s^{1/2}\big\{\mathbf{v}^T\nabla^2\mathcal{L}(\boldsymbol{\beta}^*)\mathbf{v}\big\}^{1/2}}{\|\mathbf{v}_{\mathcal{S}}\|_1}\ge\lambda_{\min}>0.$$

Assumption 4.1.1 is the bounded covariate condition, which is imposed by both Bradic et al. (2011) and Huang et al. (2013), and holds in most real applications. Assumption 4.1.2 is known as the compatibility factor condition which is also used by Huang et al. (2013). This assumption essentially bounds the minimal eigenvalue of the Hessian matrix $\nabla^2\mathcal{L}(\boldsymbol{\beta}^*)$ from below for those directions within the cone $\mathcal{C}(\xi,\mathcal{S})$. In particular, the validity of this assumption has been verified in Theorem 4.1 of Huang et al. (2013). Under these assumptions, Huang et al. (2013) derive the rate of convergence of the Lasso estimator $\widehat{\boldsymbol{\beta}}$ under the ℓ_1-norm. More specifically, they prove that under Assumptions 4.1.1 and 4.1.2, if $\|\boldsymbol{\beta}^*\|_0=s$ and $\lambda\asymp\sqrt{n^{-1}\log d}$, it holds that

$$\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1=\mathcal{O}_{\mathbb{P}}(s\lambda),\tag{4.1.2}$$

which establishes the estimation consistency in the high dimensional regime.

For the theoretical development, we introduce some additional notations. For a vector u, we denote $\mathbf{u}^{\otimes 0} = 1$, $\mathbf{u}^{\otimes 1} = \mathbf{u}$ and $\mathbf{u}^{\otimes 2} = \mathbf{u}\mathbf{u}^T$. Denote

$$\begin{aligned} S^{(r)}(t,\boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i^{\otimes r}(t) Y_i(t) \exp\{\boldsymbol{X}_i^T(t)\boldsymbol{\beta}\} \text{ for } r = 0,1,2, \quad \bar{\boldsymbol{Z}}(t,\boldsymbol{\beta}) = \frac{S^{(1)}(t,\boldsymbol{\beta})}{S^{(0)}(t,\boldsymbol{\beta})}, \\ \mathbf{V}_n(t,\boldsymbol{\beta}) &= \sum_{i=1}^{n} \frac{Y_i(t)\exp\{\boldsymbol{X}_i(t)^T\boldsymbol{\beta}\}}{nS^{(0)}(t,\boldsymbol{\beta})}\{\boldsymbol{X}_i(t) - \bar{\boldsymbol{Z}}(t,\boldsymbol{\beta})\}^{\otimes 2} = \frac{S^{(2)}(t,\boldsymbol{\beta})}{S^{(0)}(t,\boldsymbol{\beta})} - \bar{\boldsymbol{Z}}(t,\boldsymbol{\beta})^{\otimes 2}. \end{aligned} \tag{4.1.3}$$

The gradient of $\mathcal{L}(\boldsymbol{\beta})$ is

$$\nabla\mathcal{L}(\boldsymbol{\beta}) = \frac{\partial\mathcal{L}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \big\{\boldsymbol{X}_i(u) - \bar{\boldsymbol{Z}}(u,\boldsymbol{\beta})\big\} dN_i(u), \tag{4.1.4}$$

and the Hessian matrix of $\mathcal{L}(\boldsymbol{\beta})$ is

$$\nabla^2\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{n}\int_0^\tau \mathbf{V}_n(u,\boldsymbol{\beta}) d\bar{N}(u) = \frac{1}{n}\int_0^\tau \Big\{\frac{S^{(2)}(u,\boldsymbol{\beta})}{S^{(0)}(u,\boldsymbol{\beta})} - \bar{\boldsymbol{Z}}(u,\boldsymbol{\beta})^{\otimes 2}\Big\} d\bar{N}(u). \tag{4.1.5}$$

We denote the population versions of above defined quantities by

$$\mathbf{s}^{(r)}(t,\boldsymbol{\beta}) = \mathbb{E}\big[Y(t)\boldsymbol{X}(t)^{\otimes r}\exp\{\boldsymbol{X}(t)^T\boldsymbol{\beta}\}\big] \text{ for } r = 0,1,2;\ \mathbf{e}(t,\boldsymbol{\beta}) = \mathbf{s}^{(1)}(t,\boldsymbol{\beta})/\mathbf{s}^{(0)}(t,\boldsymbol{\beta}), \tag{4.1.6}$$

and

$$\mathbf{H}(\boldsymbol{\beta}) = \mathbb{E}\Big[\int_0^\tau \Big\{\frac{\mathbf{s}^{(2)}(t,\boldsymbol{\beta})}{\mathbf{s}^{(0)}(t,\boldsymbol{\beta})} - \mathbf{e}(t,\boldsymbol{\beta})^{\otimes 2}\Big\} dN(t)\Big], \text{ and } \mathbf{H}^* = \mathbf{H}(\boldsymbol{\beta}^*), \tag{4.1.7}$$

where $\mathbf{H}^*$ is the Fisher information matrix based on the partial likelihood.

4.2 Hypothesis Test and Confidence Interval

While the estimation consistency has been established in high dimensions, it remains challenging to develop inferential procedures (e.g., valid confidence intervals and hypotheses testing) for the high dimensional proportional hazards model. In this section, we propose

three novel hypothesis testing procedures. The proposed tests can be viewed as high dimensional counterparts of the conventional score, Wald, and partial likelihood ratio tests. Hereafter, for notational simplicity, we partition the vector $\boldsymbol{\beta}$ as $\boldsymbol{\beta} = (\alpha, \boldsymbol{\theta}^T)^T$, where $\alpha = \beta_1 \in \mathbb{R}$ is the parameter of interest; $\boldsymbol{\theta} = (\beta_2, ..., \beta_d)^T \in \mathbb{R}^{d-1}$ is the vector of nuisance parameters, and we denote $\mathcal{L}(\boldsymbol{\beta})$ by $\mathcal{L}(\alpha, \boldsymbol{\theta})$. Let $\nabla^2_{\alpha\alpha}\mathcal{L}(\boldsymbol{\beta})$, $\nabla^2_{\alpha\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\beta})$ and $\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\beta})$ be the corresponding partitions of $\nabla^2\mathcal{L}(\boldsymbol{\beta})$. Let $\mathbf{H}^*_{\alpha\alpha}$, $\mathbf{H}^*_{\alpha\boldsymbol{\theta}}$ and $\mathbf{H}^*_{\boldsymbol{\theta}\boldsymbol{\theta}}$ be the corresponding partitions of $\mathbf{H}^*$, where $\mathbf{H}^*$ is defined in (4.1.7). For instances, $\mathbf{H}^*_{\boldsymbol{\theta}\alpha} = \mathbf{H}^*_{2:d,1} \in \mathbb{R}^{d-1}$ and $\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\beta}) = \nabla^2_{2:d,2:d}\mathcal{L}(\boldsymbol{\beta}) \in \mathbb{R}^{(d-1)\times(d-1)}$. In this section, without loss of generality, we test the hypothesis H_0: $\alpha^* = 0$ versus H_1: $\alpha^* \neq 0$ for some univariate parameter of interest α.

4.2.1 Decorrelated Score Test

In the classical low dimensional setting, we can exploit the profile partial score function

$$S(\alpha) = \nabla_\alpha \mathcal{L}(\alpha, \boldsymbol{\theta})|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}(\alpha)}$$

to conduct test, where $\widehat{\boldsymbol{\theta}}(\alpha) = \text{argmin}_\theta \, \mathcal{L}(\alpha, \boldsymbol{\theta})$ is the maximum partial likelihood estimator for $\boldsymbol{\theta}$ with a fixed α. Under the null hypothesis that $\alpha^* = 0$, when d is fixed while n goes to infinity, it holds that $\sqrt{n}S(0) \xrightarrow{d} N(0, \mathbf{H}^*_{\alpha\alpha})$. If $n(\mathbf{H}^*_{\alpha\alpha})^{-1}S^2(0)$ is larger than the $(1-\eta)$th quantile of a chi-squared distribution with one degree of freedom, we reject the null hypothesis. Classical asymptotic theory shows that this procedure controls type I error with significance level η.

However, in high dimensions, the profile partial score function $S(\alpha)$ with $\widehat{\boldsymbol{\theta}}(\alpha)$ replaced by a penalized estimator, say the corresponding components of $\widehat{\boldsymbol{\beta}}$ in (4.1.1), does not yield a tractable limiting distribution due to the existence of a large number of nuisance parameters. To address this problem, we construct a new score function for α that is asymptotically

normal even in high dimensions. The key component is a high dimensional decorrelation method, aiming to handle the impact of the high dimensional nuisance vector.

More specifically, we propose a decorrelated score test for H_0: $\alpha^* = 0$. We first estimate $\boldsymbol{\theta}^*$ by $\widehat{\boldsymbol{\theta}}$ using the ℓ_1 penalized estimator $\widehat{\boldsymbol{\beta}}$ in (4.1.1). Next, we calculate a linear combination of the partial score function $\nabla_\theta \mathcal{L}(0, \widehat{\boldsymbol{\theta}})$ to best approximate $\nabla_\alpha \mathcal{L}(0, \widehat{\boldsymbol{\theta}})$. The population version of the vector of coefficients in the best linear combination can be calculated as

$$\begin{aligned} \mathbf{w}^* &= \underset{\mathbf{w}}{\operatorname{argmin}}\, \mathbb{E}\big\{\nabla_\alpha \mathcal{L}(0, \boldsymbol{\theta}^*) - \mathbf{w}^T \nabla_{\boldsymbol{\theta}} \mathcal{L}(0, \boldsymbol{\theta}^*)\big\}^2 \\ &= \mathbb{E}\{\nabla_{\boldsymbol{\theta}} \mathcal{L}(0, \boldsymbol{\theta}^*) \nabla_{\boldsymbol{\theta}} \mathcal{L}(0, \boldsymbol{\theta}^*)^T\}^{-1} \mathbb{E}\big\{\nabla_{\boldsymbol{\theta}} \mathcal{L}(0, \boldsymbol{\theta}^*) \nabla_\alpha \mathcal{L}(0, \boldsymbol{\theta}^*)\big\} = \mathbf{H}^{*-1}_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathbf{H}^*_{\boldsymbol{\theta}\alpha}, \end{aligned} \tag{4.2.1}$$

where the last equality is by the second Bartlett identity (Tsiatis, 1981). Note that $-\mathbf{w}^*/H_{\alpha|\boldsymbol{\theta}}$, where $H_{\alpha|\boldsymbol{\theta}} = \mathbf{H}^*_{\alpha\alpha} - \mathbf{H}^*_{\alpha\boldsymbol{\theta}} \mathbf{H}^{*-1}_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathbf{H}^*_{\boldsymbol{\theta}\alpha}$, equals the corresponding column of $\mathbf{H}^{*-1}$ by the block matrix inversion formula. In fact, $\mathbf{w}^{*T} \nabla_{\boldsymbol{\theta}} \mathcal{L}(0, \boldsymbol{\theta}^*)$ can be interpreted as the projection of $\nabla_\alpha \mathcal{L}(0, \boldsymbol{\theta}^*)$ onto the linear span of the partial score function $\nabla_\theta \mathcal{L}(0, \boldsymbol{\theta}^*)$. In high dimensions, one cannot directly estimate $\mathbf{w}^*$ by the corresponding sample version since the problem is ill-posed. Motivated by the definition of $\mathbf{w}^*$ in (4.2.1), we estimate it by the Dantzig type estimator,

$$\widehat{\mathbf{w}} = \underset{\mathbf{w} \in \mathbb{R}^{d-1}}{\operatorname{argmin}} \|\mathbf{w}\|_1, \text{ subject to } \|\nabla^2_{\alpha\boldsymbol{\theta}} \mathcal{L}(\widehat{\boldsymbol{\beta}}) - \mathbf{w}^T \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(\widehat{\boldsymbol{\beta}})\|_\infty \le \lambda', \tag{4.2.2}$$

where λ' is a tuning parameter. Since $\mathbf{w}^*$ is of high dimension $d-1$, we impose the sparsity condition on $\mathbf{w}^*$. Given $\widehat{\boldsymbol{\theta}}$ and $\widehat{\mathbf{w}}$, we propose a decorrelated score function for α as

$$\widehat{U}(\alpha, \widehat{\boldsymbol{\theta}}) = \nabla_\alpha \mathcal{L}(\alpha, \widehat{\boldsymbol{\theta}}) - \widehat{\mathbf{w}}^T \nabla_{\boldsymbol{\theta}} \mathcal{L}(\alpha, \widehat{\boldsymbol{\theta}}). \tag{4.2.3}$$

Note that the decorrelated score function in equation (4.2.3) can be re-written as

$$\widehat{U}(\alpha, \widehat{\boldsymbol{\theta}}) = -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \Big\{[X_{i1}(u) - \widehat{\mathbf{w}}^T \boldsymbol{X}_{i2:d}(u)] - [\bar{Z}_1(u, \alpha, \widehat{\boldsymbol{\theta}}) - \widehat{\mathbf{w}}^T \bar{\boldsymbol{Z}}_{2:d}(u, \alpha, \widehat{\boldsymbol{\theta}})]\Big\} dN_i(u).$$

Recall that the standard score function is given by

$$\nabla\mathcal{L}(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \big\{\boldsymbol{X}_i(u) - \bar{\boldsymbol{Z}}(u, \boldsymbol{\beta})\big\} dN_i(u).$$

By the definition of $\bar{\boldsymbol{Z}}(u, \boldsymbol{\beta})$, it is easy to see that $\widehat{U}(\alpha, \widehat{\boldsymbol{\theta}})$ has the same structure as $\nabla\mathcal{L}(\boldsymbol{\beta})$ with $\boldsymbol{X}_i(u)$ replaced by $X_{i1}(u) - \widehat{\mathbf{w}}^T \boldsymbol{X}_{i2:d}(u)$ and the risk set average $\bar{\boldsymbol{Z}}(u, \boldsymbol{\beta})$ of $\boldsymbol{X}_i(u)$ replaced by the risk set average of $X_{i1}(u) - \widehat{\mathbf{w}}^T \boldsymbol{X}_{i2:d}(u)$. Hence, the proposed method essentially constructs a new univariate covariate $\widetilde{X}_i(u) := X_{i1}(u) - \widehat{\mathbf{w}}^T \boldsymbol{X}_{i2:d}(u)$ such that the decorrelated score function $\widehat{U}(\alpha, \widehat{\boldsymbol{\theta}})$ can be interpreted as the integrated difference between the new covariate $\widetilde{X}_i(u)$ and its risk set average. So, the key is to construct a "good" covariate. The rationale behind the construction of the new covariate $\widetilde{X}_i(u)$ is to reduce the (weighted) correlation between the original covariate $X_{i1}(u)$ and the remaining ones $\boldsymbol{X}_{i2:d}(u)$, where the weight is introduced to account for the nonlinearity of the Cox model. If the (weighted) correlation is weak enough, one can perform the marginal analysis to infer the regression coefficient of $\widetilde{X}_i(u)$. Thus, the standard score function with the new covariate $\widetilde{X}_i(u)$ can be used for inference. That also explains why we call our method as the decorrelation based method.

Geometrically, the decorrelated score function is approximately orthogonal to any component of the nuisance score function $\nabla_\theta\mathcal{L}(0, \boldsymbol{\theta}^*)$. This orthogonality property, which does not hold for the original score function $\nabla_\alpha\mathcal{L}(\alpha, \widehat{\boldsymbol{\theta}})$, reduces the variability caused by the nuisance parameters. A geometric illustration of the decorrelation-based methods is provided in Figure 4.1, which also incorporates the illustration of the decorrelated Wald and partial likelihood ratio tests to be introduced in the following subsections. Technically, the uncer-

tainty of estimating $\boldsymbol{\theta}$ in the partial score function $\nabla_\alpha \mathcal{L}(\alpha, \widehat{\boldsymbol{\theta}})$ can be reduced by subtracting the decorrelation term $\widehat{\mathbf{w}}^T \nabla_{\boldsymbol{\theta}} \mathcal{L}(\alpha, \widehat{\boldsymbol{\theta}})$. As will be shown in the next section, this is a key step to establish the result that the decorrelated score function $\widehat{U}(0, \widehat{\boldsymbol{\theta}})$ weakly converges to $N(0, H_{\alpha|\boldsymbol{\theta}})$ under the null, where $H_{\alpha|\boldsymbol{\theta}} = \mathbf{H}^*_{\alpha\alpha} - \mathbf{H}^*_{\alpha\boldsymbol{\theta}} \mathbf{H}^{*-1}_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathbf{H}^*_{\boldsymbol{\theta}\alpha}$. This further explains why the decorrelated score function $\widehat{U}(\alpha, \widehat{\boldsymbol{\theta}})$ rather than the original score function $\nabla_\alpha \mathcal{L}(\alpha, \widehat{\boldsymbol{\theta}})$ should be used as the inferential function in high dimensions. On the other hand, in the low dimensional setting, it can be shown that the decorrelated score function $\widehat{U}(\alpha, \widehat{\boldsymbol{\theta}})$ is asymptotically equivalent to the profile partial score function $S(\alpha)$.

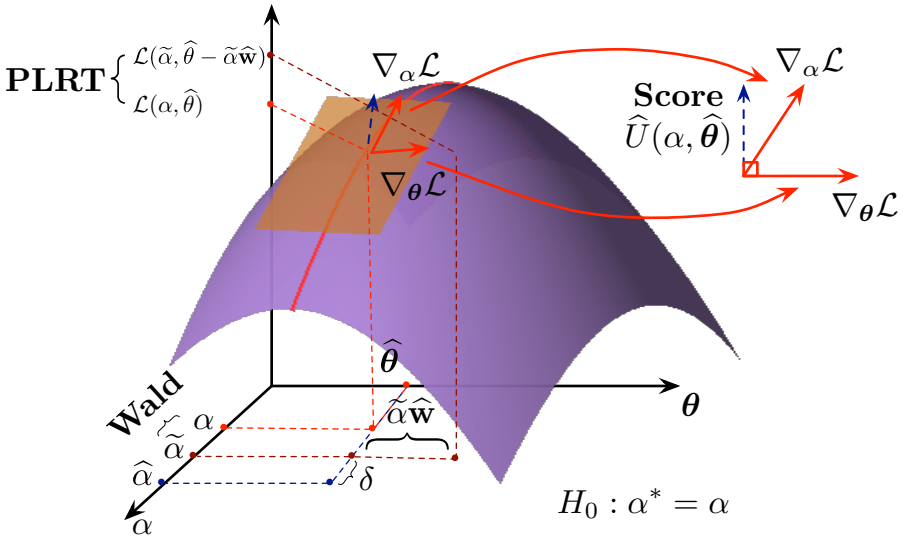


Figure 4.1: Geometric illustration of the decorrelated score, Wald and partial likelihood ratio tests. The purple surface corresponds to the log-partial likelihood function. The orange plane is the tangent plane of the surface at point $(\alpha, \widehat{\boldsymbol{\theta}})$. The two red arrows in the orange plane represent $\nabla_\alpha \mathcal{L}$ and $\nabla_{\boldsymbol{\theta}} \mathcal{L}$. The correlated score function in blue is the projection of $\nabla_\alpha \mathcal{L}$ onto the space orthogonal to $\nabla_{\boldsymbol{\theta}} \mathcal{L}$. Given Lasso estimator $\widehat{\alpha}$, the decorrelated Wald estimator is $\widetilde{\alpha} = \widehat{\alpha} - \delta$, where $\delta = \left\{\partial \widehat{U}(\widehat{\alpha}, \widehat{\boldsymbol{\theta}}) / \partial \alpha\right\}^{-1} \widehat{U}(\widehat{\alpha}, \widehat{\boldsymbol{\theta}})$. The decorrelated partial likelihood ratio test compares the log-partial likelihood function values at $(\alpha, \widehat{\boldsymbol{\theta}})$ and $(\widetilde{\alpha}, \widehat{\boldsymbol{\theta}} - \widetilde{\alpha}\widehat{\mathbf{w}})$.

To test if $\alpha^* = 0$, we need to standardize $\widehat{U}(0, \widehat{\boldsymbol{\theta}})$ in order to construct the test statistic. We estimate $H_{\alpha|\boldsymbol{\theta}}$ by

$$\widehat{H}_{\alpha|\boldsymbol{\theta}} = \nabla^2_{\alpha\alpha} \mathcal{L}(\widehat{\alpha}, \widehat{\boldsymbol{\theta}}) - \widehat{\mathbf{w}}^T \nabla^2_{\boldsymbol{\theta}\alpha} \mathcal{L}(\widehat{\alpha}, \widehat{\boldsymbol{\theta}}). \tag{4.2.4}$$

Hence, we define the decorrelated score test statistic as

$$\widehat{S}_n = n\widehat{H}^{-1}_{\alpha|\boldsymbol{\theta}}\widehat{U}^2(0,\widehat{\boldsymbol{\theta}}), \text{where } \widehat{U}(0,\widehat{\boldsymbol{\theta}}) \text{ and } \widehat{H}_{\alpha|\boldsymbol{\theta}} \text{ are defined in (4.2.3) and (4.2.4).} \quad (4.2.5)$$

In the next section, we show that under the null, $\widehat{S}_n$ converges weakly to a chi-squared distribution with one degree of freedom. Given a significance level $\eta \in (0,1)$, the score test $\psi_S(\eta)$ is

$$\psi_S(\eta) = \begin{cases} 0 & \text{if } \widehat{S}_n \leq \chi^2_1(1-\eta) \\ 1 & \text{otherwise} \end{cases}, \quad (4.2.6)$$

where $\chi^2_1(1-\eta)$ denotes the $(1-\eta)$-th quantile of a chi-squared random variable with one degree of freedom, and the null hypothesis $\alpha^* = 0$ is rejected if and only if $\psi_S(\eta) = 1$.

Remark 4.2.1. To estimate $\mathbf{w}^*$, we may also adopt other regularized estimators. For instance, we can adopt the following Lasso estimator

$$\widehat{\mathbf{w}} = \underset{\mathbf{w}\in\mathbb{R}^{d-1}}{\operatorname{argmin}} \frac{1}{2n}\sum_{i=1}^{n}(\mathbf{w}^T\nabla_{\boldsymbol{\theta}}\mathcal{L}_i(\widehat{\boldsymbol{\beta}}) - \nabla_\alpha\mathcal{L}_i(\widehat{\boldsymbol{\beta}}))^2 + \lambda'\|\mathbf{w}\|_1,$$

where $\mathcal{L}_i$ denotes the negative log-partial likelihood of the i-th sample. Similarly, other nonconvex penalized methods such as SCAD or MCP can be applied.

4.2.2 Confidence Intervals and Decorrelated Wald Test

The proposed score test does not directly provide a confidence interval for α^*. In low dimensions, by looking at the limiting distribution of the maximum partial likelihood estimator, we can get a confidence interval for α^* (Andersen and Gill, 1982), which is equivalent to the classical Wald test. In this subsection, we extend the classical Wald test under the proportional hazards model to high dimensional settings to construct confidence intervals for α^*.

The key idea of performing Wald test is to derive a regular estimator for α^*. Our procedure is based on the deccorelated score function $\widehat{U}(\alpha, \widehat{\boldsymbol{\theta}})$ in (4.2.3). Since $\widehat{U}(\alpha, \widehat{\boldsymbol{\theta}})$ serves as an approximately unbiased estimating equation for α, the root of the equation $\widehat{U}(\alpha, \widehat{\boldsymbol{\theta}}) = 0$ with respect to α defines an estimator for α^*. However, searching for the root may be computationally intensive, especially when α is multi-dimensional. To reduce the computational cost, we exploit a closed-form estimator $\widetilde{\alpha}$ obtained by linearizing $\widehat{U}(\alpha, \widehat{\boldsymbol{\theta}}) = 0$ at the initial estimator $\widehat{\alpha}$. More specifically, let $\widehat{\boldsymbol{\beta}} = (\widehat{\alpha}, \widehat{\boldsymbol{\theta}}^T)^T$ be the ℓ_1 penalized estimator in (4.1.1), we adopt the following one-step estimator,

$$\widetilde{\alpha} = \widehat{\alpha} - \Big\{ \frac{\partial \widehat{U}(\widehat{\alpha}, \widehat{\boldsymbol{\theta}})}{\partial \alpha} \Big\}^{-1} \widehat{U}(\widehat{\alpha}, \widehat{\boldsymbol{\theta}}), \text{ where } \widehat{U}(\widehat{\alpha}, \widehat{\boldsymbol{\theta}}) = \nabla_\alpha \mathcal{L}(\widehat{\alpha}, \widehat{\boldsymbol{\theta}}) - \widehat{\mathbf{w}}^T \nabla_{\boldsymbol{\theta}} \mathcal{L}(\widehat{\alpha}, \widehat{\boldsymbol{\theta}}). \tag{4.2.7}$$

In the next section, we prove that $\sqrt{n}(\widetilde{\alpha} - \alpha^*)$ converges weakly to $N(0, H^{-1}_{\alpha|\boldsymbol{\theta}})$. Hence, let $Z_{1-\eta/2}$ be the $(1-\eta/2)$-th quantile of $N(0,1)$. We have

$$\Big[\widetilde{\alpha} - n^{-1/2} Z_{1-\eta/2} \widehat{H}^{-1/2}_{\alpha|\boldsymbol{\theta}}, \widetilde{\alpha} + n^{-1/2} Z_{1-\eta/2} \widehat{H}^{-1/2}_{\alpha|\boldsymbol{\theta}} \Big]$$

is a $100(1-\eta)\%$ confidence interval for α^*. From the perspective of hypothesis testing, the decorrelated Wald test statistic for H_0: $\alpha^* = 0$ versus H_1: $\alpha^* \neq 0$ is

$$\widehat{W}_n = n \widehat{H}_{\alpha|\boldsymbol{\theta}} \widetilde{\alpha}^2, \text{ where } \widetilde{\alpha} \text{ and } \widehat{H}_{\alpha|\boldsymbol{\theta}} \text{ are defined in (4.2.7) and (4.2.4), respectively.} \tag{4.2.8}$$

Consequently, the decorrelated Wald test at significance level η is

$$\psi_W(\eta) = \begin{cases} 0 & \text{if } \widehat{W}_n \leq \chi^2_1(1-\eta), \\ 1 & \text{otherwise,} \end{cases} \tag{4.2.9}$$

and the null hypothesis $\alpha^* = 0$ is rejected if and only if $\psi_W(\eta) = 1$.

4.2.3 Decorrelated Partial Likelihood Ratio Test

Under low dimensional settings, the likelihood ratio inference enjoys great success in the statistical literature. Under the proportional hazards model, the partial likelihood ratio test statistic is PLRT $= 2n\big[\mathcal{L}\{0, \widehat{\boldsymbol{\theta}}_P(0)\} - \mathcal{L}(\widehat{\alpha}_P, \widehat{\boldsymbol{\theta}}_P)\big]$, where $\widehat{\boldsymbol{\theta}}_P(0) = \text{argmin}_{\boldsymbol{\theta}}\, \mathcal{L}(0, \boldsymbol{\theta})$ and $(\widehat{\alpha}_P, \widehat{\boldsymbol{\theta}}_P) = \text{argmin}_{\alpha, \boldsymbol{\theta}}\, \mathcal{L}(\alpha, \boldsymbol{\theta})$ are the maximum partial likelihood estimators under the null and alternative. Hence, PLRT evaluates the validity of the null hypothesis by comparing the partial likelihood under H_0 with that under H_1. Similar to the partial score test, the partial likelihood ratio test also fails in the high dimensional setting due to the presence of a large number of nuisance parameters. In this section, we propose a new type of the partial likelihood ratio test which is valid in high dimensions.

To handle the impact of high dimensional nuisance parameters, we define the (negative) decorrelated partial likelihood for α as $\mathcal{L}_{\text{decor}}(\alpha) = \mathcal{L}(\alpha, \widehat{\boldsymbol{\theta}} - \alpha\widehat{\mathbf{w}})$. The aim of this decorrelated partial likelihood is to approximate the directional likelihood of the model along with the direction $(1, -\mathbf{w}^*)$, which also corresponds to the least favorable direction for estimating α. In the low dimensional setting, the decorrelated partial likelihood $\mathcal{L}_{\text{decor}}(\alpha)$ is asymptotically equivalent to the profile partial likelihood $\mathcal{L}\{\alpha, \widehat{\boldsymbol{\theta}}(\alpha)\}$. Hence, we view $\mathcal{L}_{\text{decor}}(\alpha)$ as an extension of the classical profile partial likelihood to high dimensions. Given $\mathcal{L}_{\text{decor}}(\alpha)$, the decorrelated partial likelihood ratio test statistic is defined as

$$\widehat{L}_n = 2n\big\{\mathcal{L}_{\text{decor}}(0) - \mathcal{L}_{\text{decor}}(\widetilde{\alpha})\big\}, \quad \text{where} \;\; \mathcal{L}_{\text{decor}}(\alpha) = \mathcal{L}(\alpha, \widehat{\boldsymbol{\theta}} - \alpha\widehat{\mathbf{w}}), \tag{4.2.10}$$

and $\widetilde{\alpha}$ is given in (4.2.7). As discussed in the previous subsection, $\widetilde{\alpha}$ is a one-step approximation of the global minimizer of $\mathcal{L}_{\text{decor}}(\alpha)$. Hence, the log-likelihood ratio $\widehat{L}_n$ evaluates the validity of the null hypothesis by comparing the decorrelated partial likelihood under H_0 with that under H_1.

In the next section, we show that $\widehat{L}_n$ converges weakly to a chi-squared distribution with one degree of freedom. Therefore, a decorrelated partial likelihood ratio test with significance level η is

$$\psi_L(\eta) = \begin{cases} 0 & \text{if } \widehat{L}_n \leq \chi_1^2(1-\eta) \\ 1 & \text{otherwise} \end{cases}, \tag{4.2.11}$$

and $\psi_L(\eta) = 1$ indicates a rejection of the null hypothesis.

4.3 Asymptotic Properties

In this section, we derive the limiting distributions of the decorrelated test statistics under the null and alternative hypotheses.

4.3.1 Limiting Distributions under the Null

In our analysis, we make the following regularity assumptions.

Assumption 4.3.1. The true hazard is uniformly bounded, i.e., $\sup_{t\in[0,\tau]} \max_{i\in[n]} |\boldsymbol{X}_i^T(t)\boldsymbol{\beta}^*| = \mathcal{O}(1)$.

Assumption 4.3.2. It holds that $\|\mathbf{w}^*\|_0 = s' \asymp s$, and $\sup_{t\in[0,\tau]} \max_{i\in[n]} |\boldsymbol{X}_{i,2:d}^T(t)\mathbf{w}^*| = \mathcal{O}(1)$.

Assumption 4.3.3. The Fisher information matrix is bounded, $\|\mathbf{H}^*\|_\infty = \mathcal{O}(1)$, and its minimum eigenvalue is also bounded from below, $\Lambda_{\min}(\mathbf{H}^*) \geq C_h > 0$, which implies that $H_{\alpha|\boldsymbol{\theta}} = \mathbf{H}^*_{\alpha\alpha} - \mathbf{H}^*_{\alpha\boldsymbol{\theta}}\mathbf{H}^{*-1}_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathbf{H}^*_{\boldsymbol{\theta}\alpha} \geq C_h$.

To connect these assumptions with existing literature, Assumptions 4.3.1 and 4.3.2 extend Assumption (iv) of Theorem 3.3 in van de Geer et al. (2014) to the proportional hazards model. In particular, the sparsity of $\mathbf{w}^*$ is assumed in order to show the consistency of $\widehat{\mathbf{w}}$ to $\mathbf{w}^*$. By the block matrix inversion formula, the sparsity assumption of $\mathbf{w}^*$ is equivalent to the corresponding row/column of $\mathbf{H}^{*-1}$ being sparse. This assumption is weaker than van de

Geer et al. (2014); see Remark 4.3.13 for details. Assumption 4.3.3 is related to the Fisher information matrix, which is essential even in low dimensional settings.

The following result characterizes the asymptotic normality of the decorrelated score function $\widehat{U}(0,\widehat{\theta})$ in (4.2.3) under the null.

Theorem 4.3.4. *Under Assumptions 4.1.1, 4.1.2, 4.3.1, 4.3.2 and 4.3.3, let* $\lambda \asymp \sqrt{n^{-1}\log d}$, $\lambda' \asymp s\sqrt{n^{-1}\log d}$ *and* $n^{-1/2}s^3 \log d = o(1)$. *Under the null hypothesis that* $\alpha^* = 0$, *the decorrelated score function* $\widehat{U}(0,\widehat{\boldsymbol{\theta}})$ *defined in* (4.2.3) *satisfies*

$$\sqrt{n}\widehat{U}(0,\widehat{\boldsymbol{\theta}}) \xrightarrow{d} Z, \text{ where } Z \sim N(0, H_{\alpha|\boldsymbol{\theta}}), \tag{4.3.1}$$

and $H_{\alpha|\boldsymbol{\theta}} = \mathbf{H}^*_{\alpha\alpha} - \mathbf{H}^*_{\alpha\boldsymbol{\theta}}\mathbf{H}^{*-1}_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathbf{H}^*_{\boldsymbol{\theta}\alpha}$.

As we have discussed before, the limiting variance of the decorrelated score function can be estimated by $\widehat{H}_{\alpha|\boldsymbol{\theta}} = \nabla^2_{\alpha\alpha}\mathcal{L}(\widehat{\alpha},\widehat{\boldsymbol{\theta}}) - \widehat{\mathbf{w}}^T\nabla^2_{\boldsymbol{\theta}\alpha}\mathcal{L}(\widehat{\alpha},\widehat{\boldsymbol{\theta}})$. The next lemma shows the consistency of $\widehat{H}_{\alpha|\boldsymbol{\theta}}$.

Lemma 4.3.5. *Suppose that Assumptions 4.1.1, 4.1.2, 4.3.1, 4.3.2 and 4.3.3 hold. If* $\lambda \asymp \sqrt{n^{-1}\log d}$ *and* $\lambda' \asymp s\sqrt{n^{-1}\log d}$, *we have*

$$|H_{\alpha|\boldsymbol{\theta}} - \widehat{H}_{\alpha|\boldsymbol{\theta}}| = \mathcal{O}_{\mathbb{P}}\Big(s^2\sqrt{\frac{\log d}{n}}\Big),$$

where $\widehat{H}_{\alpha|\boldsymbol{\theta}}$ *is defined in* (4.2.4).

By Theorem 4.3.4 and Lemma 4.3.5, the next corollary shows that under the null hypothesis, type I error of the decorrelated score test $\psi_S(\eta)$ in (4.2.6) converges asymptotically to the significance level η. Let the associated p-value of the decorrelated score test be $P_S = 2\big\{1 - \Phi(\widehat{S}_n)\big\}$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal random variable and $\widehat{S}_n$ is the score test statistic defined in (4.2.5). The distribution of P_S converges to a uniform distribution asymptotically.

Corollary 4.3.6. Suppose Assumptions 4.1.1, 4.1.2, 4.3.1, 4.3.2 and 4.3.3 hold, $\lambda \asymp \sqrt{n^{-1}\log d}$, $\lambda' \asymp s\sqrt{n^{-1}\log d}$, and $n^{-1/2}s^3 \log d = o(1)$. The decorrelated score test and the its corresponding p-value satisfy

$$\lim_{n\to\infty} \mathbb{P}\big(\psi_S(\eta) = 1 | \alpha^* = 0\big) = \eta, \text{ and } P_S \xrightarrow{d} \text{Unif}[0,1], \text{ when } \alpha^* = 0,$$

where Unif$[0,1]$ denotes a random variable uniformly distributed in $[0,1]$.

We then analyze the decorrelated Wald test under the null. We derive the limiting distribution of the one-step estimator $\widetilde{\alpha}$ defined in (4.2.7) in the next theorem.

Theorem 4.3.7. *Suppose Assumptions 4.1.1, 4.1.2, 4.3.1, 4.3.2 and 4.3.3 hold, and* $\lambda \asymp \sqrt{n^{-1}\log d}$, $\lambda' \asymp s\sqrt{n^{-1}\log d}$, $n^{-1/2}s^3\log d = o(1)$. *When the null hypothesis* $\alpha^* = 0$ *holds, the decorrelated estimator* $\widetilde{\alpha}$ *satisfies*

$$\sqrt{n}\widetilde{\alpha} \xrightarrow{d} Z, \text{ where } Z \sim N(0, H^{-1}_{\alpha|\boldsymbol{\theta}}). \tag{4.3.2}$$

Utilizing the asymptotic normality of $\widetilde{\alpha}$, we can establish the limiting type I error of $\psi_W(\eta)$ in (4.2.9), in the next corollary. Note that, it is straightforward to generalize the result to $\sqrt{n}(\widetilde{\alpha} - \alpha^*) \xrightarrow{d} Z$, where $Z \sim N(0, H^{-1}_{\alpha|\boldsymbol{\theta}})$ for any α^*. This gives us a confidence interval of α^*.

Corollary 4.3.8. Under Assumptions 4.1.1, 4.1.2, 4.3.1, 4.3.2 and 4.3.3, suppose $\lambda \asymp \sqrt{n^{-1}\log d}$, $\lambda' \asymp s\sqrt{n^{-1}\log d}a$ and $n^{-1/2}s^3\log d = o(1)$. The type I error of the decorrelated Wald test $\psi_W(\eta)$ and its corresponding p-value $P_W = 2\big\{1 - \Phi(\widehat{W}_n)\big\}$ satisfy

$$\lim_{n\to} \mathbb{P}(\psi_W(\eta) = 1 | \alpha^* = 0) = \eta, \text{ and } P_W \xrightarrow{d} \text{Unif}[0,1] \text{ when } \alpha^* = 0.$$

In addition, an asymptotic $(1-\eta)\times 100\%$ confidence interval of α^* is

$$\left(\widetilde{\alpha}-\frac{\Phi^{-1}(1-\eta/2)}{\sqrt{n\widehat{H}_{\alpha|\boldsymbol{\theta}}}},\widetilde{\alpha}+\frac{\Phi^{-1}(1-\eta/2)}{\sqrt{n\widehat{H}_{\alpha|\boldsymbol{\theta}}}}\right).$$

Finally, we present our main result on the limiting distribution of the decorrelated partial likelihood ratio test statistic $\widehat{L}_n$ introduced in (4.2.10).

Theorem 4.3.9. *Suppose Assumptions 4.1.1, 4.1.2, 4.3.1, 4.3.2 and 4.3.3 hold,* $\lambda \asymp \sqrt{n^{-1}\log d}$, $\lambda' \asymp s\sqrt{n^{-1}\log d}$ *and* $n^{-1/2}s^3\log d = o(1)$. *If the null hypothesis* $\alpha^* = 0$ *holds, the decorrelated likelihood ratio test statistic* $\widehat{L}_n$ *in* (4.2.10) *satisfies*

$$\widehat{L}_n \xrightarrow{d} Z_\chi, \text{ where } Z_\chi \sim \chi^2_1. \tag{4.3.3}$$

This theorem justifies the decorrelated partial likelihood ratio test $\psi_L(\eta)$ in (4.2.11). Also, let the p-value associated with the decorrelated partial likelihood ratio test be $P_L = 1 - F(\widehat{L}_n)$, where $F(\cdot)$ is the cumulative distribution function of χ^2_1. Similar to Corollaries 4.3.6 and 4.3.8, we characterize the type I error of the test $\psi_L(\eta)$ in (4.2.11) and its corresponding p-value below.

Corollary 4.3.10. Suppose Assumptions 4.1.1, 4.1.2, 4.3.1, 4.3.2 and 4.3.3 hold, $\lambda \asymp \sqrt{n^{-1}\log d}$, $\lambda' \asymp s\sqrt{n^{-1}\log d}$ and $n^{-1/2}s^3\log d = o(1)$. The type I error of the decorrelated partial likelihood ratio test $\psi_L(\eta)$ with significance level η and its associated p-value P_L satisfy

$$\lim_{n\to\infty}\mathbb{P}\big(\psi_L(\eta)=1|\alpha^*=0\big)=\eta, \text{ and } P_L \xrightarrow{d} \text{Unif}[0,1] \text{ when } \alpha^*=0.$$

By Corollaries 4.3.6, 4.3.8 and 4.3.10, we see that the decorrelated score, Wald and partial likelihood ratio tests are asymptotically equivalent as summarized in the next corollary.

Corollary 4.3.11. Suppose Assumptions 4.1.1, 4.1.2, 4.3.1, 4.3.2 and 4.3.3 hold, $\lambda \asymp \sqrt{n^{-1}\log d}$, $\lambda' \asymp s\sqrt{n^{-1}\log d}$ and $n^{-1/2}s^3 \log d = o(1)$. If the null hypothesis $\alpha^* = 0$ holds, the test statistics $\widehat{S}_n$ in (4.2.5), $\widehat{W}_n$ in (4.2.8), and $\widehat{L}_n$ in (4.2.10) are asymptotically equivalent, i.e.,

$$\widehat{S}_n = \widehat{W}_n + o_{\mathbb{P}}(1) = \widehat{L}_n + o_{\mathbb{P}}(1).$$

To summarize this subsection, Corollaries 4.3.6, 4.3.8 and 4.3.10 characterize the asymptotic distributions of the proposed decorrelated test statistics under the null hypothesis. It is known that $H_{\alpha|\theta}$ is the semiparametric information lower bound for inferring α. Theorem 4.3.7 shows that $\widetilde{\alpha}$ achieves the semiparametric information bound, which indicates the semiparametric efficiency of $\widetilde{\alpha}$. Using the asymptotic equivalence in Corollary 4.3.11, all of our test statistics are semiparametrically efficient (van der Vaart, 2000). Although the three tests are asymptotically equivalent, our numerical results suggest that the partial likelihood ratio test outperforms the remain tests empirically.

Remark 4.3.12. All the theoretical results in this section are still valid if we replace the Lasso estimator $\widehat{\boldsymbol{\beta}}$ with nonconvex (SCAD or MCP) estimators.

Remark 4.3.13. Existing works mainly consider high dimensional inferences for linear and generalized models; see Lockhart et al. (2014); van de Geer et al. (2014), and Zhang and Zhang (2014). More specifically, Lockhart et al. (2014) consider conditional inference given the event that a set of covariates is selected, while we consider unconditional inference. We defer the detailed comparison of the conditional inference and the unconditional inference to the discussion section. Recently, Zhang and Zhang (2014) propose a novel LDP method for inference in high dimensional linear models. However, the method in Zhang and Zhang (2014) strongly relies on the linear structure of the model. For instance, their method is motivated by the decomposition of a closed form expression of the univariate least square estimator. It is unclear whether the method can be easily extended to the proportional hazards model, because the maximum partial likelihood estimator does not have a closed form expression

and the decomposition seems difficult to apply due to the model's nonlinearity. The method in van de Geer et al. (2014) is based on inverting KKT condition of the Lasso estimator in generalized linear models. Compared to van de Geer et al. (2014) which only focuses on Lasso estimator, our approach can be also applied to nonconvex estimators such as SCAD and MCP. In addition, our inference on α allows the submatrix of $\mathbf{H}^{*-1}$ corresponding to the nuisance parameter to be non-sparse, which is weaker than the sparsity of the whole matrix $\mathbf{H}^{*-1}$ in van de Geer et al. (2014). Thus, the proposed decorrelation method is more general than van de Geer et al. (2014) and requires weaker technical assumptions.

4.3.2 Limiting Distributions under the Alternative

Statistical power under the alternative hypothesis is one of the most important criteria to compare different tests. Due to the root n convergence of the estimator $\widetilde{\alpha}$ in (4.2.7) as illustrated in Theorem 4.3.7, it is of interest to examine the corresponding tests under the alternative where α^* shrinks to the null in a suitable rate.

This subsection investigates the power of the decorrelated score, Wald and partial likelihood ratio tests under a sequence of local alternatives, named as Pitman alternatives. In particular, denote by H_a^n: $\alpha^* = n^{-1/2}c$ the alternative hypothesis, where c is a nonzero constant. Under H_a^n, as n goes to infinity, α^* approaches to 0 as specified in the null hypothesis. We first derive the asymptotic distribution of the decorrelated score function $\widehat{U}(0, \widehat{\theta})$ in (4.2.3) under Pitman alternatives.

Theorem 4.3.14. *Suppose that Assumptions 4.1.1, 4.1.2, 4.3.1, 4.3.2 and 4.3.3 hold,* $\lambda \asymp \sqrt{n^{-1}\log d}$, $\lambda' \asymp s\sqrt{n^{-1}\log d}$ *and* $n^{-1/2}s^3 \log d = o(1)$. *Under the alternative* H_a^n: $\alpha^* = n^{-1/2}c$, *the decorrelated score function* $\widehat{U}(0, \widehat{\boldsymbol{\theta}})$ *in* (4.2.3) *satisfies*

$$\sqrt{n}\widehat{U}(0, \widehat{\boldsymbol{\theta}}) \xrightarrow{d} Z', \text{ where } Z' \sim N(-cH_{\alpha|\boldsymbol{\theta}}, H_{\alpha|\boldsymbol{\theta}}).$$

By Theorem 4.3.14, we have the power of the decorrelated score test under the alternative, H_a^n: $\alpha^* = n^{-1/2}c$, is defined as

$$\mathbb{P}\big(\psi_S(\eta) = 1 \big| \alpha^* = n^{-1/2}c\big) = \mathbb{P}\big(\widehat{S}_n > \chi_1^2(1-\eta) \big| \alpha^* = n^{-1/2}c\big),$$

where $\widehat{S}_n$ is defined in (4.2.5), and $\chi_1^2(1-\eta)$ is the $(1-\eta)$-th quantile of a chi-squared distribution with one degree of freedom. Denote by $NC_{\chi_1}(\xi)$ the noncentral chi-squared distribution with one degree of freedom and noncentrality parameter ξ. By Theorem 4.3.14, it follows that $\widehat{S}_n \xrightarrow{d} NC_{\chi_1}(c^2 H_{\alpha|\boldsymbol{\theta}})$. The following corollary of Theorem 4.3.14 provides the power of the decorrelated score test $\psi_S(\eta)$ in (4.2.6) at a significance level η.

Corollary 4.3.15. Suppose Assumptions 4.1.1, 4.1.2, 4.3.1, 4.3.2 and 4.3.3 hold, $\lambda \asymp \sqrt{n^{-1}\log d}$, $\lambda' \asymp s\sqrt{n^{-1}\log d}$ and $n^{-1/2}s^3 \log d = o(1)$. Under Pitman alternative hypothesis H_a^n: $\alpha^* = n^{-1/2}c$, the power of the decorrelated score test $\psi_S(\eta)$ at a significance level η is

$$\lim_{n\to\infty} \mathbb{P}\big(\psi_S(\eta) = 1 | \alpha^* = n^{-1/2}c\big) = \mathbb{P}\big(NC_{\chi_1}(c^2 H_{\alpha|\boldsymbol{\theta}}) > \chi_1^2(1-\eta)\big).$$

This corollary implies the intuitive fact that the decorrelated score test is asymptotically more powerful when the null and alternative hypotheses become further separated (i.e., $|c|$ increases).

The next theorem provides the limiting distribution of the decorrelated Wald statistic $\widetilde{\alpha}$ in (4.2.7) and partial likelihood ratio test statistic $\widehat{L}_n$ in (4.2.11) under Pitman alternative H_a^n: $\alpha^* = n^{-1/2}c$. We also obtain the asymptotic power of these two tests.

Theorem 4.3.16. *Assume that Assumptions 4.1.1, 4.1.2, 4.3.1, 4.3.2 and 4.3.3 hold,* $\lambda \asymp \sqrt{n^{-1}\log d}$, $\lambda' \asymp s\sqrt{n^{-1}\log d}$ *and* $n^{-1/2}s^3\log d = o(1)$. *Suppose that the alternative* H_a^n: $\alpha^* = n^{-1/2}c$ *holds. We have:*

1. *The one-step estimator* $\widetilde{\alpha}$ *in* (4.2.7) *satisfies*

$$\sqrt{n}\widetilde{\alpha} \xrightarrow{d} Z', \text{ where } Z' \sim N(c, H^{-1}_{\alpha|\boldsymbol{\theta}}).$$

The decorrelated Wald test $\psi_W(\eta)$ *in* (4.2.9) *with a significance level* η *has asymptotic power*

$$\lim_{n\to\infty} \mathbb{P}\big(\psi_W(\eta) = 1 \big| \alpha^* = n^{-1/2}c\big) = \mathbb{P}\big(NC_{\chi_1}(c^2 H_{\alpha|\boldsymbol{\theta}}) > \chi^2_1(1-\eta)\big).$$

2. *The decorrelated partial likelihood ratio statistic* $\widehat{L}_n$ *satisfies*

$$\widehat{L}_n \xrightarrow{d} Z'_\chi, \text{ where } Z'_\chi \sim NC_{\chi_1}(c^2 H_{\alpha|\boldsymbol{\theta}}).$$

The power of the decorrelated partial likelihood ratio test at a significance level η *satisfies*

$$\lim_{n\to\infty} \mathbb{P}\big(\psi_L(\eta) = 1 \big| \alpha^* = n^{-1/2}c\big) = \mathbb{P}\big(NC_{\chi_1}(c^2 H_{\alpha|\boldsymbol{\theta}}) > \chi^2_1(1-\eta)\big).$$

In summary, Corollary 4.3.15 and Theorem 4.3.16 imply that the decorrelated score, Wald and partial likelihood ratio tests have the same local asymptotic power. This observation coincides with the conventional asymptotic equivalence among these tests.

4.4 Inference on the Baseline Hazard Function

The baseline hazard function

$$\Lambda_0(t) = \int_0^t \lambda_0(u)du$$

is treated as a nuisance function in the log-partial likelihood method. In practice, inferences on the baseline hazard function can be also of interest. To the best of our knowledge, such problems remain unexplored in high dimensional settings. In this section, we aim to construct confidence intervals for the baseline hazard function and the survival function. In addition, we extend the procedure to conduct inference on the conditional hazard function in Section C.5 of the Appendix.

We consider the following Breslow-type estimator for the baseline hazard function. Given an ℓ_1-penalized estimator $\widehat{\boldsymbol{\beta}}$ derived from (4.1.1), the direct plug-in estimator for the baseline hazard function at time t is

$$\widehat{\Lambda}_0(t, \widehat{\boldsymbol{\beta}}) = \int_0^t \frac{\sum_{i=1}^n dN_i(u)}{\sum_{i=1}^n Y_i(u) \exp\{\boldsymbol{X}_i^T(u)\widehat{\boldsymbol{\beta}}\}}. \tag{4.4.1}$$

Since the plug-in estimator $\widehat{\boldsymbol{\beta}}$ does not have a tractable distribution, inference based on the estimator $\widehat{\Lambda}_0(t, \widehat{\boldsymbol{\beta}})$ is difficult. To handle this problem, we adopt a bias correction procedure to reduce the uncertainty caused by plugging $\widehat{\boldsymbol{\beta}}$ into $\widehat{\Lambda}_0(t, \boldsymbol{\beta})$. Specifically, we estimate $\Lambda_0(t)$ by the sample version of $\widehat{\Lambda}_0(t, \widehat{\boldsymbol{\beta}}) - \left\{\nabla\Lambda_0(t, \boldsymbol{\beta}^*)\right\}^T \mathbf{H}^{*-1}\nabla\mathcal{L}(\widehat{\boldsymbol{\beta}})$, where

$$\Lambda_0(t, \boldsymbol{\beta}) = \mathbb{E}\int_0^t \frac{dN_i(u)}{S^{(0)}(u, \boldsymbol{\beta})},$$

and the gradient $\nabla\Lambda_0(t, \boldsymbol{\beta}^*)$ is taken with respect to the corresponding $\boldsymbol{\beta}$ component, and $\mathbf{H}^*$ is the Fisher information matrix defined in (4.1.7). Similar to Section 4.2.1, we directly estimate $\mathbf{H}^{*-1}\nabla\widehat{\Lambda}_0(t, \widehat{\boldsymbol{\beta}})$ by the following Dantzig type selector

$$\widehat{\mathbf{u}}(t) = \operatorname{argmin} \|\mathbf{u}(t)\|_1, \text{ subject to } \|\nabla\widehat{\Lambda}_0(t, \widehat{\boldsymbol{\beta}}) - \nabla^2\mathcal{L}(\widehat{\boldsymbol{\beta}})\mathbf{u}(t)\|_\infty \leq \delta, \tag{4.4.2}$$

where δ is a tuning parameter. It can be shown that the estimator $\widehat{\mathbf{u}}(t)$ converges to $\mathbf{u}^*(t) = \mathbf{H}^{*-1}\nabla\Lambda_0(t, \boldsymbol{\beta}^*)$ under the following regularity assumption.

Assumption 4.4.1. It holds that $\|\mathbf{u}^*(t)\|_0 = s' \asymp s$ for all $0 \le t \le \tau$.

This Assumption plays the same role as Assumption 4.3.2 in the previous section. Hence, the decorrelated baseline hazard function estimator at time t is

$$\widetilde{\Lambda}_0(t, \widehat{\boldsymbol{\beta}}) = \widehat{\Lambda}_0(t, \widehat{\boldsymbol{\beta}}) - \widehat{\mathbf{u}}(t)^T \nabla \mathcal{L}(\widehat{\boldsymbol{\beta}}), \text{ where } \widehat{\mathbf{u}}(t) \text{ is defined in (4.4.2).} \tag{4.4.3}$$

Based on the estimator (4.4.3), the survival function $S_0(t) = \exp\{-\Lambda_0(t)\}$ is estimated by $\widetilde{S}(t, \widehat{\boldsymbol{\beta}}) = \exp\{-\widetilde{\Lambda}_0(t, \widehat{\boldsymbol{\beta}})\}$. The main theorem of this section characterizes the asymptotic normality of $\widetilde{\Lambda}_0(t, \widehat{\boldsymbol{\beta}})$ and $\widetilde{S}(t, \widehat{\boldsymbol{\beta}})$ as follows.

Theorem 4.4.2. *Suppose Assumptions 4.1.1, 4.1.2, 4.3.1, 4.3.3 and 4.4.1 hold,* $\lambda \asymp \sqrt{n^{-1} \log d}$, $\delta \asymp s' \sqrt{n^{-1} \log d}$ *and* $n^{-1/2} s^3 \log d = o(1)$. *We have, for any* $t \in [0, \tau]$, *the decorrelated baseline hazard function estimator* $\widetilde{\Lambda}_0(t, \widehat{\boldsymbol{\beta}})$ *in* (4.4.3) *satisfies*

$$\sqrt{n}\big\{\Lambda_0(t) - \widetilde{\Lambda}_0(t, \widehat{\boldsymbol{\beta}})\big\} \xrightarrow{d} Z, \text{ where } Z \sim N\big(0, \sigma_1^2(t) + \sigma_2^2(t)\big),$$

and

$$\sigma_1^2(t) = \int_0^t \frac{\lambda_0(u) du}{\mathbb{E}\big[\exp\{\boldsymbol{X}^T(u)\boldsymbol{\beta}^*\} Y(u)\big]} \text{ and } \sigma_2^2(t) = \nabla \Lambda_0(t, \boldsymbol{\beta}^*)^T \mathbf{H}^{*-1} \nabla \Lambda_0(t, \boldsymbol{\beta}^*). \tag{4.4.4}$$

The estimated survival function $\widetilde{S}(t, \widehat{\boldsymbol{\beta}})$ *satisfies*

$$\sqrt{n}\big\{\widetilde{S}(t, \widehat{\boldsymbol{\beta}}) - S_0(t)\big\} \xrightarrow{d} Z', \text{ where } Z' \sim N\bigg(0, \frac{\sigma_1^2(t) + \sigma_2^2(t)}{\exp(2\Lambda_0(t))}\bigg).$$

Given Theorem 4.4.2, we further need to estimate the limiting variances $\sigma_1^2(t)$ and $\sigma_2^2(t)$. To this end, we use

$$\widehat{\sigma}_1^2(t) = \int_0^t \frac{d\widehat{\Lambda}_0(u, \widehat{\boldsymbol{\beta}})}{n^{-1} \sum_{i'=1}^n \exp\{\boldsymbol{X}_{i'}^T(u)\widehat{\boldsymbol{\beta}}\} Y_{i'}(u)} \text{ and } \widehat{\sigma}_2^2(t) = \big\{\nabla \widehat{\Lambda}_0(t, \widehat{\boldsymbol{\beta}})\big\}^T \widehat{\mathbf{u}}(t),$$

where $\widehat{\Lambda}_0(t, \widehat{\boldsymbol{\beta}})$ is defined in (4.4.1).

We conclude this section by the following corollary which provides confidence intervals for $\Lambda_0(t)$ and $S_0(t)$.

Corollary 4.4.3. Suppose Assumptions 4.1.1, 4.1.2, 4.3.2, 4.3.3 and 4.4.1 hold, $\lambda \asymp \sqrt{n^{-1}\log d}$, $\delta \asymp s\sqrt{n^{-1}\log d}$ and $n^{-1/2}s^3 \log d = o(1)$. For any $t > 0$ and $0 < \eta < 1$,

$$\lim_{n\to\infty} \mathbb{P}\bigg(\big|\Lambda_0(t) - \widetilde{\Lambda}_0(t, \widehat{\boldsymbol{\beta}})\big| \leq \frac{\Phi^{-1}(1-\eta/2)\big\{\widehat{\sigma}_1^2(t) + \widehat{\sigma}_2^2(t)\big\}^{1/2}}{\sqrt{n}} \bigg) = 1 - \eta,$$

and

$$\lim_{n\to\infty} \mathbb{P}\bigg(\big|S_0(t) - \widetilde{S}_0(t, \widehat{\boldsymbol{\beta}})\big| \leq \frac{\Phi^{-1}(1-\eta/2)\big\{\widehat{\sigma}_1^2(t) + \widehat{\sigma}_2^2(t)\big\}^{1/2} \exp\{-\widetilde{\Lambda}_0(t, \widehat{\boldsymbol{\beta}})\}}{\sqrt{n}} \bigg) = 1 - \eta.$$

4.5 Numerical Results

This section reports numerical results of our proposed methods using both simulated and real data.

4.5.1 Simulated Data

We first investigate empirical performances of the decorrelated score, Wald and partial likelihood ratio tests on the parametric component $\boldsymbol{\beta}^*$ as proposed in Section 4.2. In our simulation, we let $\beta_1^* = 0$, and randomly select s out of the next $d-1$ components to be non-zero, where rest entries are set to be 0. To estimate $\boldsymbol{\beta}^*$ and $\mathbf{w}^*$, we choose the tuning parameters λ by 10-fold cross-validation and set $\lambda' = \frac{1}{2}\sqrt{n^{-1}\log d}$. We find that our simulation results are insensitive to the choice of λ'. We conduct decorrelated score, Wald and partial likelihood ratio tests for β_1 which is set to be 0 under null hypothesis H_0: $\beta_1^* = 0$ versus alternative H_a: $\beta_1^* \neq 0$, where we set the significance level to be $\eta = 0.05$. In each setting, we simulate $n = 150$ independent samples from a multivariate Gaussian distribution

$N_d(\mathbf{0}, \mathbf{\Sigma})$ for $d = 100, 200,$ or 500, where $\mathbf{\Sigma}$ is a Toeplitz matrix with $\Sigma_{jk} = \rho^{|j-k|}$ and $\rho = 0.25$, 0.4, 0.6 or 0.75. The cardinality of the active set s is either 2 or 3, and the regression coefficients in the active set are either all 1's (Dirac) or drawn randomly from the uniform distribution Unif$[0, 2]$. We set the baseline hazard rate function to be identity. Thus, the i-th survival time follows an exponential distribution with mean $\exp(\boldsymbol{X}_i^T \boldsymbol{\beta}^*)$. The i-th censoring time is independently generated from an exponential distribution with mean $U \times \exp(\boldsymbol{X}_i^T \boldsymbol{\beta}^*)$, where $U \sim$ Unif$[1, 3]$. As discussed in Fan and Li (2002), this censoring scheme results in about 30% censored samples.

The above simulation is repeated 1,000 times. The empirical type I errors of the decorrelated score, Wald and partial likelihood ratio tests are summarized in Tables 4.1 and 4.2. We see that the empirical type I errors of all three tests are close to the desired 5% significance level, which supports our theoretical results. This observation holds for the whole range of ρ, s and d specified in the data generating procedures. In addition, as expected, the empirical type I errors further deviate from the significance level as d increases for all three tests, illustrating the effects of dimensionality d on finite sample performance.

Table 4.1: Average Type I error of the decorrelated tests with $\eta = 5\%$ where $(n, s) = (150, 2)$.

		$\rho = 0.25$		$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.75$	
Method	d	Dirac	Unif$[0, 2]$	Dirac	Unif$[0, 2]$	Dirac	Unif$[0, 2]$	Dirac	Unif$[0, 2]$
Score	100	5.1%	5.2%	5.1%	4.9%	5.2%	5.1%	4.9%	5.0%
	200	5.2%	4.8%	5.3%	4.8%	5.3%	5.6%	4.7%	4.6%
	500	6.1%	6.4%	5.5%	4.6%	4.2%	4.4%	3.9%	3.7%
Wald	100	5.2%	5.3%	5.0%	5.0%	5.2%	4.9%	5.0%	5.1%
	200	5.4%	4.7%	5.2%	4.8%	4.6%	4.7%	4.3%	4.6%
	500	6.3%	6.1%	5.9%	5.5%	5.8%	4.2%	4.5%	3.9%
PLRT	100	4.9%	4.8%	5.2%	5.2%	5.0%	5.2%	4.8%	4.7%
	200	5.7%	5.5%	5.2%	5.5%	4.8%	5.6%	4.6%	4.5%
	500	6.2%	6.2%	5.9%	5.3%	4.5%	4.2%	3.8%	3.6%

We also investigate the empirical power of the proposed tests. Instead of setting $\beta_1 = 0$, we generate the data with $\beta_1 = 0.05, 0.1, 0.15, ..., 0.55$, following the same simulation scheme

Table 4.2: Average type I error of the decorrelated tests with $\eta = 5\%$ where $(n, s) = (150, 3)$.

		$\rho = 0.25$		$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.75$	
	d	Dirac	Unif$[0, 2]$	Dirac	Unif$[0, 2]$	Dirac	Unif$[0, 2]$	Dirac	Unif$[0, 2]$
Score	100	5.2%	5.2%	4.8%	5.3%	5.3%	4.9%	5.3%	4.8%
	200	5.2%	4.6%	4.7%	5.3%	5.4%	5.8%	4.5%	4.8%
	500	6.3%	6.5%	5.8%	4.4%	5.2%	4.6%	3.6%	3.4%
Wald	100	5.1%	4.9%	5.3%	4.7%	5.2%	4.9%	5.0%	5.1%
	200	4.8%	4.6%	4.9%	5.1%	5.2%	5.7%	4.2%	4.4%
	500	6.5%	6.8%	6.2%	5.9%	5.1%	4.5%	3.9%	4.2%
PLRT	100	5.3%	5.2%	5.0%	5.3%	5.4%	5.2%	4.9%	4.8%
	200	5.5%	5.3%	5.4%	4.6%	5.2%	5.7%	5.4%	4.3%
	500	6.5%	6.3%	5.7%	5.5%	4.8%	4.1%	3.7%	3.2%

introduced above. We plot the rejection rates of the three decorrelated tests for testing $H_0 : \beta_1 = 0$ with significance level 0.05 and $\rho = 0.25$ in Figure 4.2. We see that when $d = 100$, the three tests share similar power. However, for larger d (e.g., $d = 500$), the decorrelated partial likelihood ratio test is the most powerful test. In addition, the Wald test is less effective for problems with higher dimensionality. Based on our simulation results, we recommend the decorrelated partial likelihood ratio test for inference in high dimensional problems.

4.5.2 Analyzing a Gene Expression Dataset

We apply the proposed testing procedures to analyze a genomic dataset, which is collected from a diffuse large B-cell lymphoma study analyzed by Alizadeh et al. (2000). The dataset can be downloaded from `http://llmpp.nih.gov/lymphoma/data.shtml`. One of the goals in this study is to investigate how different genes in B-cell malignancies are associated with the survival time. The expression values for over 13,412 genes in B-cell malignancies are measured by microarray experiments. The dataset contains 40 patients with diffuse large B-cell lymphoma who are recruited and followed until death or the end of the study. A small

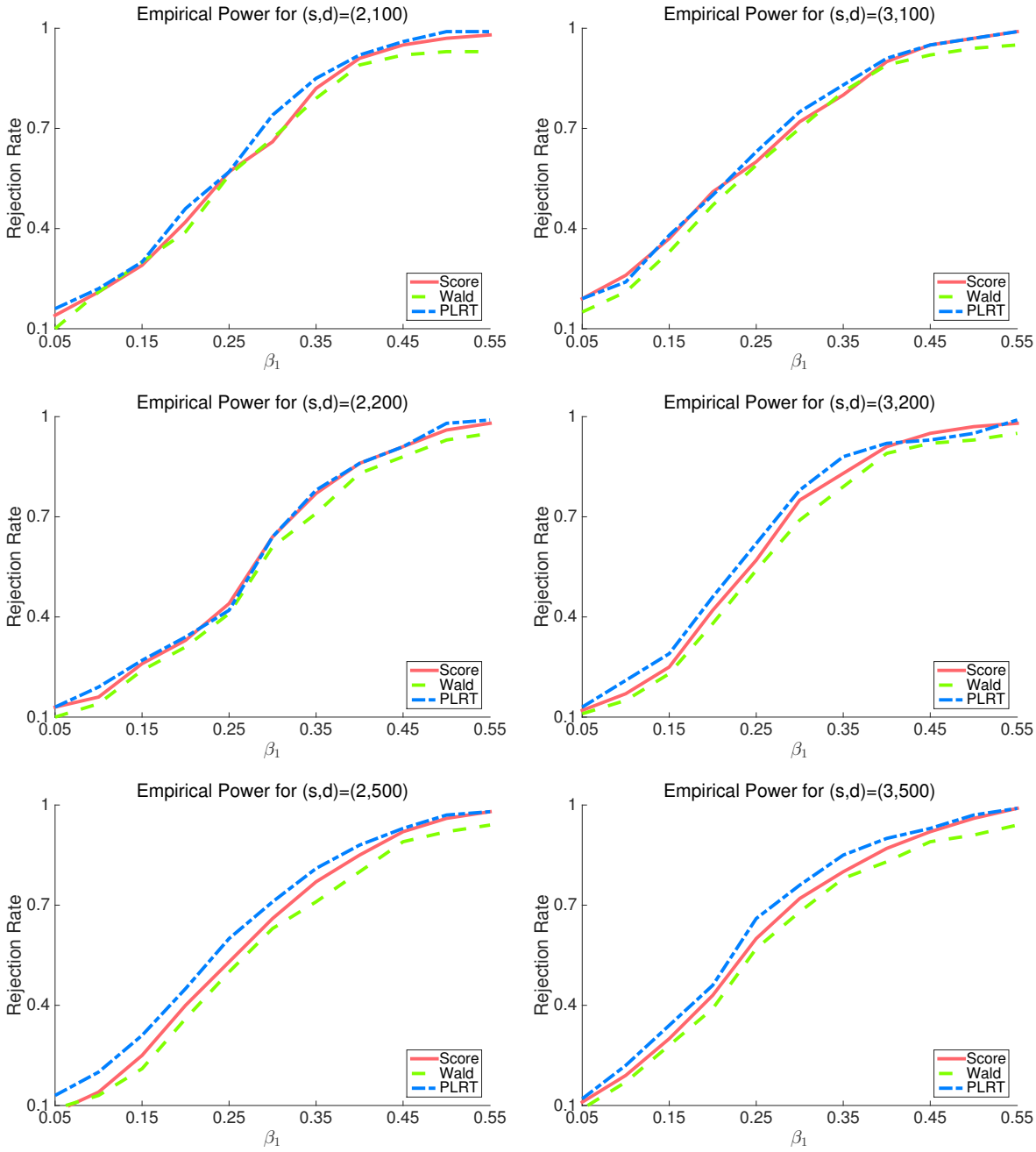


Figure 4.2: Empirical rejection rates of the decorrelated score, Wald and partial likelihood ratio tests on simulated data with different active set sizes and dimensionality.

proportion ($\approx 5\%$) of the gene expression values are not well measured and are treated as missing values by Alizadeh et al. (2000). For simplicity, we impute the missing values of

each gene by the median of the observed values of the same gene. The average survival time is 43.9 months and the censoring rate is 55%.

We apply the proposed score, Wald and partial likelihood ratio tests to the data. The same strategy for choosing the tuning parameters as that in the simulation studies is adopted. We repeatedly apply the testing procedures for all parameters. To control the family-wise error rate due to the multiple testing, the p-values are adjusted by the Bonferroni's method. To be more conservative, we only report the genes with adjusted p-values less than 0.05 by all of the three methods in Table 4.3. Many of the genes which are significant in the hypothesis tests are biologically related to lymphoma. For instance, the relation between lymphoma and genes FLT3 (Meierhoff et al., 1995), CDC10 (Di Gaetano et al., 2003), CHN2 (Nishiu et al., 2002), Emv11 (Hiai et al., 2003), CD137 (Alizadeh et al., 2011) and TAP2 (Nielsen et al., 2015) have been experimentally confirmed. This provides evidence that our methods can be used to discover findings in scientific applications involving high dimensional covariates. We further plot the estimated baseline hazard function and its 95% confidence interval in Figure 4.3 for illustration.

Table 4.3: Genes with the adjusted p-values less than 0.05 using score, Wald and partial likelihood ratio tests for the large B-cell lymphoma gene expression dataset.

Gene	Score	Wald	PLRT
SP1	6.20×10^{-6}	3.84×10^{-5}	9.61×10^{-6}
PTMAP1	1.07×10^{-4}	8.12×10^{-5}	4.50×10^{-4}
Emv11	5.16×10^{-4}	1.55×10^{-3}	6.53×10^{-4}
CDC10	1.13×10^{-3}	3.24×10^{-4}	1.76×10^{-4}
NR2E3	1.30×10^{-3}	2.15×10^{-2}	4.45×10^{-3}
FLT3	1.56×10^{-3}	1.49×10^{-4}	3.68×10^{-4}
GPD2	4.34×10^{-3}	1.63×10^{-3}	3.32×10^{-4}
TAP2	6.07×10^{-3}	2.19×10^{-2}	7.61×10^{-3}
CHN2	7.19×10^{-3}	9.74×10^{-3}	3.58×10^{-3}
CD137	1.18×10^{-2}	1.75×10^{-3}	1.67×10^{-4}
CYTB	2.21×10^{-2}	9.34×10^{-3}	1.15×10^{-2}

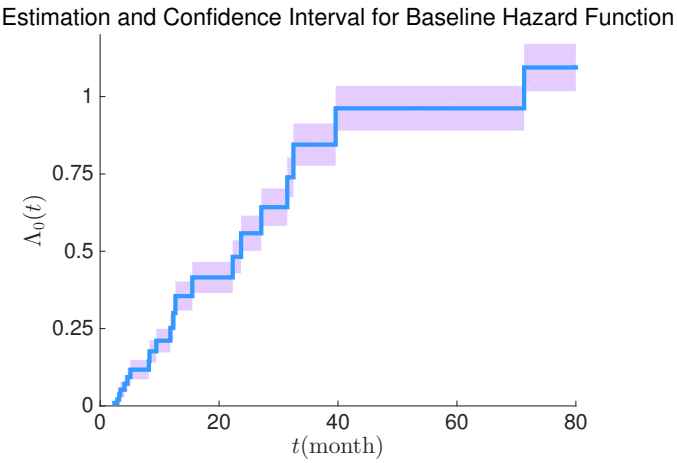


Figure 4.3: Estimation and 95% confidence interval of the baseline hazard function.

4.6 Discussion

The proposed methods involve two tuning parameters λ and λ'. The presence of multiple tuning parameters in the inferential procedures is encountered in many recent works even under high dimensional linear models (van de Geer et al., 2014; Zhang and Zhang, 2014). Theoretically, we establish the asymptotic normality of the test statistics when $\lambda \asymp (n^{-1}\log d)^{1/2}$ and $\lambda' \asymp s(n^{-1}\log d)^{1/2}$. Empirically, our numerical results suggest that cross-validation seems to be a practical procedure for the choice of λ.

The post-selection conditional inference (Lockhart et al., 2014) and the proposed unconditional inference address different inferential problems. To be specific, let us consider the linear regression model $Y_i = \boldsymbol{\beta}^T \boldsymbol{X}_i + \epsilon_i$ $(i = 1, ..., n)$. The post-selection conditional inference aims to construct a 95% confidence interval of $\boldsymbol{\beta}^M$, where

$$\boldsymbol{\beta}^M = \arg\min_{\mathbf{b}^M} \mathbb{E}(Y_i - \boldsymbol{X}_{iM}^T \mathbf{b}^M)^2,$$

where $M \subset \{1, ..., d\}$ denotes the index of selected variables and $\boldsymbol{X}_{iM}$ denotes the components of $\boldsymbol{X}_i$ in the set M. However, it is important to note that in general $\boldsymbol{\beta}^M \neq \boldsymbol{\beta}^*_M$, where $\boldsymbol{\beta}^*_M$ is the components of the true value $\boldsymbol{\beta}^*$ in set M. This makes the interpretation of $\boldsymbol{\beta}^M$

difficult, since this target value highly depends on the selected variables $\boldsymbol{X}_{iM}$. In contrast, our unconditional inference aims to construct confidence intervals for the unknown true value β_j^* for $1 \leq j \leq d$.

Whether conditional or unconditional inference is more appropriate depends on the context. For instance, in our real data applications, the goal is to investigate how different genes in B-cell malignancies are associated with the survival time. That means we are interested in constructing confidence intervals (or testing hypotheses) for the unknown true value β_j^* for all $1 \leq j \leq d$. However, the conditional inference on $\boldsymbol{\beta}^M$ does not directly address this scientific question. Thus, it seems that the proposed unconditional inference is more appropriate in our real data application.

In practice, the proposed method has the following two added values, compared to the standard variable selection method. First, given the p-values produced by our testing-based procedures, we can address an important practical problem that how do practitioners prioritize genes for follow-up experiments. The ranking of p-values provides an efficient approach for prioritizing all genes based on their statistical significance. In contrast, the standard variable selection method (e.g., Lasso estimator) is not informative for prioritizing genes. In fact, the ranking based on the Lasso estimator could be misleading, because some genes may have large coefficients as well as large standard deviations. Second, as seen in our real data analysis, the p-values can be adjusted by the Bonferroni method to control the family-wise error rate at any given level (e.g., 0.05). This is a commonly-used procedure in the analysis of genomic data, because it provides the explicit confidence level (e.g., 0.05) on quantifying the probability of the false discoveries. However, such a measure of uncertainty is not provided by the standard variable selection method.

Following this work, Ning and Liu (2014) further extended the decorrelated score test to the general log-likelihood with i.i.d samples. Due to the presence of censored data, our technical development is quite different from Ning and Liu (2014). In addition, we propose

a novel partial likelihood ratio test, which is empirically more powerful. One future research direction is to extend this framework to other types of survival models such as accelerated failure time model and semiparametric transformation model.

Chapter 5

Conclusion

In this dissertation, we study several challenging statistical problems using optimization tools. We first study optimal two stage adaptive enrichment design problems. Given a new treatment, our goal is to learn if different subpopulations benefit from the treatment. The adaptive enrichment design requires us to have decision rules for modifying enrollment and the multiple testing procedure. With a novel sparse linear programing formulation of the problem, we make computing the optimal design possible, which was an important open problem in the past.

Next, we consider a total cardinality constraint approach to estimate high-dimensional spatial graphical models. As we discussed, spatial Graphical model can be adopted in various applications. We propose a practical SPICA algorithm for spatial graphical model estimation for the total cardinality constraint approach. We solve the problem by considering the Lagrangian dual problem. Though the problem is nonconvex, we prove that the average-per-vertex duality gap decreases as the dimension d increases, and we achieve optimal statistical properties if the dimension d is sufficiently large. We conduct thorough numerical experiments to backup our theory. We further provide several new fundamental results to better understand the total cardinality constraint approach for the spatial graphical model. We prove that problem (3.4.1) is NP-complete. For the case $\mathcal{R}_j(\boldsymbol{\beta}_j) = \|\boldsymbol{\beta}_j\|_0$, we show that the

problem is polynomial-time solvable. For future work, we will continue to develop efficient algorithms to attack different cardinality constrained problems without sacrificing statistical efficiencies. We also plan to apply the proposed method to conduct real-world applications, such as brain functional region partition.

We then study the problem of high-dimensional inference for the Cox model. Under the high-dimensional setting, the inference for Cox model is challenging. Our work develops a unified decorrelation framework to conduct score, Wald and partial likelihood ratio tests for the low-dimensional parametric component of the Cox model. We further extends the method to infer the nonparametric component. We provide thorough numerical investigation to support our proposed method.

Appendices

Appendix A

Appendix to Chapter 2

A.1 Representation of Familywise Type I error constraints

Consider any vector $\boldsymbol{\delta} \in \mathbb{R}^2$, and let $\mathcal{H}_{\text{TRUE}}(\boldsymbol{\delta})$ denote the subset of null hypotheses $\mathcal{H}$ that are true under $\boldsymbol{\delta}$. The familywise Type I error constraint corresponding to $\boldsymbol{\delta}$ has the form:

$$\sum_{s\in\mathcal{S}:s\cap\mathcal{H}_{TRUE}(\boldsymbol{\delta})\neq\emptyset} P_{\boldsymbol{\delta}}\left\{A \text{ rejects } s\right\} = \sum_{s\in\mathcal{S}:s\cap\mathcal{H}_{TRUE}(\boldsymbol{\delta})\neq\emptyset} \sum_{r,d,r'} x_{rd} y_{rdr's} p(\boldsymbol{\delta}, r, d, r'). \tag{A.1.1}$$

A.2 Representation of (2.3.10)-(2.3.12) by linear constraints

For any $r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{D}, r' \in \mathcal{R}_{\text{mtp},d}$, we have by (2.3.13) and (2.3.11) that

$$\sum_{s\in\mathcal{S}} v_{rdr's} = \sum_{s\in\mathcal{S}} x_{rd} y_{rdr's} = x_{rd} \sum_{s\in\mathcal{S}} y_{rdr's} = x_{rd}. \tag{A.2.1}$$

For each $r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{D}$, and any two rectangles $r', \widetilde{r}' \in \mathcal{R}_{\text{mtp},d}$, we have

$$\sum_{s \in \mathcal{S}} v_{rdr's} = \sum_{s \in \mathcal{S}} v_{rd\widetilde{r}'s}, \tag{A.2.2}$$

since the left and right sides of the above display both equal x_{rd} by (A.2.1).

Then for each $r \in \mathcal{R}_{\text{dec}}$, we have

$$\sum_{d \in \mathcal{D}} \sum_{s \in \mathcal{S}} v_{rdr'_d s} = \sum_{d} x_{rd} = 1, \tag{A.2.3}$$

where the first equality follows from (A.2.1) and the second follows from (2.3.10).
The set of constraints (2.3.18) is equivalent to:

$$\text{for each } r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{D}, \text{ and each pair } r', \widetilde{r}' \in \mathcal{R}_{\text{mtp},d}, \sum_{s \in \mathcal{S}} v_{rdr's} = \sum_{s \in \mathcal{S}} v_{rd\widetilde{r}'s}. \tag{A.2.4}$$

The reason we use (2.3.18) instead of the above is that the former has fewer constraints, which makes the corresponding linear program smaller.

A.3 Proof of Theorem 2.3.1

Proof: Consider any feasible solution $\mathbf{z}$ to the sparse LP. Define the corresponding solution $\mathbf{x}, \mathbf{y}$ to the discretized problem from Section 2.3 through the mapping (2.3.20), (2.3.1). This mapping is well defined, since by (A.2.4) we have $\sum_{s \in \mathcal{S}} v_{rdr's}$ does not depend on r'. The equations (2.3.20) and (2.3.1) imply (2.3.13) holds. The solution $\mathbf{x}, \mathbf{y}$ is feasible for the discretized problem from Section 2.3 since (2.3.15) and (2.3.13) imply (2.3.8); (2.3.17) implies (2.3.10); and (2.3.11) follows from (2.3.1). The value of the objective function (2.3.7) of the discretized problem from Section 2.3 equals the value of the objective function (2.3.14) of the above linear program, which follows by (2.3.13). This shows $\mathbf{x}, \mathbf{y}$ is a feasible solution to

the discretized problem from Section 2.3 with with the same value (of the objective function) as the corresponding solution to the sparse LP evaluated at $\mathbf{z}$. Therefore, the value of the optimal solution to the sparse LP is a lower bound on the value of the optimal solution to the discretized problem from Section 2.3.

Next, consider any feasible solution $\mathbf{x}, \mathbf{y}$ to the discretized problem from Section 2.3, and define the corresponding solution $\mathbf{z} = \{v_{rdr's}\}_{r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{D}, r' \in \mathcal{R}_{\text{mtp},d}, s \in \mathcal{S}}$ to the sparse LP using the transformation (2.3.13). We next show that $\mathbf{z}$ is a feasible solution to the sparse LP with the same value (of the objective function) as the corresponding discretized problem from Section 2.3 evaluated at $\mathbf{x}, \mathbf{y}$. By (2.3.13), the constraints (2.3.15) are equivalent to the constraints (2.3.8); the constraints (2.3.19) are equivalent to (2.3.12); and, the value of the objective function (2.3.14) equals (2.3.7). As argued in the second paragraph of this section, (2.3.10), (2.3.11), and (2.3.13) together imply (A.2.1)-(A.2.2); these latter equations imply (2.3.17) and (A.2.4). We have shown $\mathbf{z}$ is a feasible solution to the sparse LP with the same value (of the objective function) as the corresponding discretized problem from Section 2.3 evaluated at $\mathbf{x}, \mathbf{y}$. Therefore, the value of the optimal solution to the discretized problem from Section 2.3 is a lower bound on the value of the optimal solution to the sparse LP.

The results of the above two paragraphs prove the claim (i) in the theorem. Claim (ii) then follows from the result in the first paragraph.

A.4 Multiple Testing Procedure Depend Only on Sufficient Statistics

At the end of stage 2, minimal sufficient statistics for our hypotheses are the stage 2 cumulative z-statistics $Z^{(C)}$ and the enrollment decision made at the end of stage 1. This follows from Lemma A.6.1. It is inefficient to let our hypothesis testing procedure depend on statistics beyond these. We introduce linear constraints for the discretized problem that

force the multiple testing procedure to only use these sufficient statistics, and therefore to ignore the first stage data, conditional on the cumulative z-statistics and the enrollment decision.

$$\text{For each } r_1, r_2 \in \mathcal{R}_{\text{dec}}, d \in \mathcal{D}, s \in \mathcal{S}, r' \in \mathcal{R}_{\text{mtp},d}, v_{r_1 dr's} \leq v_{r_2 dr's} + \sum_{\widetilde{d} \in \mathcal{D}\setminus\{d\}} \sum_{s' \in \mathcal{S}} v_{r_2 \widetilde{d} r'_{\widetilde{d}} s'}. \tag{A.4.1}$$

The double summation on the right side represents $\sum_{\widetilde{d} \in \mathcal{D}\setminus\{d\}} x_{rd}$, i.e., the probability that if the first stage statistic $\mathbf{Z}^{(1)} \in r$ that a decision other than d will be made. To explain the intuition behind this constraint, consider the case where $v_{r_1 dr's} = 1$ and $x_{rd} = 1$; then the constraint implies for all rectangles $r_2 \in \mathcal{R}_{\text{dec}}$ we have $v_{r_2 dr's} = 1$. Roughly speaking, the multiple testing procedure rejects the same set of null hypotheses if the decision d is made and $\mathbf{Z}^{(F)} \in r'$, regardless of the first stage rectangle r.

We express (A.4.1) with a smaller set of constraint as follows. For each $d \in \mathcal{D}, r' \in \mathcal{R}_{\text{mtp},d}, s \in \mathcal{S}$, define the new variable $w_{dr's}$. For each $r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{D}, r' \in \mathcal{R}_{\text{mtp},d}, s \in \mathcal{S}$, set the constraint

$$w_{dr's} \quad \leq \quad v_{rdr's} + \sum_{\widetilde{d} \in \mathcal{D}\setminus\{d\}} \sum_{s' \in \mathcal{S}} v_{r \widetilde{d} r'^*_{\widetilde{d}} s'}. \tag{A.4.2}$$

For each $d \in \mathcal{D}, r' \in \mathcal{R}_{\text{mtp},d}$, set the constraint

$$\sum_{s \in \mathcal{S}} w_{dr's} = 1.$$

Suppose we impose the integer constraints that $w_{dr's}, v_{rdr',s} \in \{0, 1\}$ for all r, r', d, s. It is not difficult to see that (A.4.1) and (A.4.2) are indeed equivalent. Without the integer constraint, (A.4.2) is a relaxation of (A.4.1), which is much more conducive for efficient computation as it significantly reduces the number of constraints. In our numerical studies, we observe

that most solutions ($> 95\%$) solutions are integers. This justifies the rationality of replacing (A.4.1) by (A.4.2). The interpretation is that $w_{dr's}$ represents the probability that the design rejects the subset s of null hypotheses if the decision is d and the cumulative z-statistic $\mathbf{z}^{(F)}$ is in r'; this does not depend on which rectangle r the first stage statistics $\mathbf{z}^{(1)}$ are in, thereby simplifying the procedure and making computation of the familywise Type I error faster as described below.

A.5 Monotonicity Constraints

This set of constraints is conjectured to be satisfied at the optimum. This can be tested by solving the optimization problem without these constraints, and then with them, and checking if the value of the optimal solutions are equal. The advantage of imposing these constraints is that it will make the solution more interpretable, and will "fill in" rejection regions that contribute such a small amount to the objective function and constraints that the solver ignores them, but that in practice are important.

For each $r' \in \mathcal{R}_{\text{mtp},d},$, let r'_R, r'_A denote the rectangle immediately to its right, and the rectangle immediately above it, respectively (if one exists).

The following set of constraints encodes that if the null hypothesis H_{01} is rejected in a rectangle $r' \in \mathcal{R}_{\text{mtp},d}$, then it is also rejected in all rectangles to the right of r' (corresponding to an even stronger signal against H_{01} than r'):

$$\text{For each } r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{D}, r' \in \mathcal{R}_{\text{mtp},d}, \sum_{s \in \mathcal{S}: H_{01} \in s} v_{rdr's} \leq \sum_{s \in \mathcal{S}: H_{01} \in s} v_{rdr'_R s}. \tag{A.5.1}$$

The following set of constraints encodes that if the null hypothesis H_{0C} is rejected in a rectangle $r' \in \mathcal{R}_{\text{mtp},d}$, then it is also rejected in all rectangles to the right and/or above r'

(corresponding to an even stronger signal against H_{0C} than r'):

$$\text{For each } r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{D}, r' \in \mathcal{R}_{\text{mtp},d}:$$
$$\sum_{s\in\mathcal{S}:H_{0C}\in s} v_{rdr's} \le \sum_{s\in\mathcal{S}:H_{0C}\in s} v_{rdr'_R s} \text{ and } \sum_{s\in\mathcal{S}:H_{0C}\in s} v_{rdr's} \le \sum_{s\in\mathcal{S}:H_{0C}\in s} v_{rdr'_A s}. \quad \text{(A.5.2)}$$

A.6 Proof of Theorem 2.2.1

Lemma A.6.1. At the interim analysis, the sample means $\{\widehat{\mu}_{sa}^{(1)}\}$, where $s \in \{1,2\}$ and $a \in \{0,1\}$, is a minimal sufficient statistic. At the end of the trial, the set of $D(\mathbf{Z}^{(1)})$ and the cumulative sample means $\widehat{\mu}_{sa}^{(F)}$'s is a minimal sufficient statistic.

Proof. Let n_{sa}^k denotes the corresponding sample size of of subpopulation s with arm assignment a at stage k. At the interim stage, the likelihood function is

$$\begin{aligned}
\mathcal{L}(Y) &= \exp\Big\{-\sum_{i=1}^{N^{(1)}} \frac{\big(Y_{s,i}^{(1)} - \mu_{s,A_{s,i}^{(1)}}\big)^2}{2\sigma^2_{s,A_{s,i}^{(1)}}}\Big\} \prod_{i=1}^{N^{(1)}} \big\{2\pi\sigma^2_{s,A_{s,i}^{(1)}}\big\}^{-1/2} \\
&= \prod_{s=1}^{2}\prod_{a=0}^{1} \exp\Big\{-\sum_{i:A_{s,i}^{(1)}=a} \frac{\big(Y_{s,i}^{(1)} - \mu_{s,A_{s,i}^{(1)}}\big)^2}{2\sigma^2_{sa}}\Big\} \prod_{i=1}^{N^{(1)}} \big\{2\pi\sigma^2_{s,A_{s,i}^{(1)}}\big\}^{-1/2} \\
&= \prod_{s=1}^{2}\prod_{a=0}^{1} \Big[\exp\Big\{-\frac{n_{sa}^{(1)}\big(\mu_{sa} - \widehat{\mu}_{sa}^{(1)}\big)\big)^2}{2\sigma^2_{sa}}\Big\} \exp\Big\{-\frac{\sum_{i:A_{s,i}^{(1)}=a}\big(Y_{s,i}^{(1)} - \widehat{\mu}_{sa}^{(1)}\big)^2}{2\sigma^2_{sa}}\Big\}\Big] \prod_{i=1}^{N^{(1)}} \big\{2\pi\sigma^2_{s,A_{s,i}^{(1)}}\big\}^{-1/2}.
\end{aligned}$$

By Fisher-Neyman factorization theorem, it is immediately seen that $\big\{\widehat{\mu}_{sa}^{(1)} : s = 1,2; a = 0,1\big\}$ is a sufficient statistic.

To prove the minimal sufficiency, let Y' be an independent set of outcomes. We look at the likelihood ratio that

$$\begin{aligned}\frac{\mathcal{L}(Y)}{\mathcal{L}(Y')} &= \prod_{s=1}^{2}\prod_{a=0}^{1}\Big[\exp\Big\{-\frac{n_{sa}^{(1)}\big\{(\mu_{sa}-\widehat{\mu}_{sa}^{(1)})^2-(\mu_{sa}-\widetilde{\mu}_{sa}^{(1)})^2\big\}}{2\sigma_{sa}^2}\Big\}\cdot \\ &\qquad \exp\Big\{-\frac{\sum_{i:A_{s,i}^{(1)}=a}\big(Y_{s,i}^{(1)}-\widehat{\mu}_{sa}^{(1)}\big)^2-\big(Y'^{(1)}_{s,i}-\widetilde{\mu}_{sa}^{(1)}\big)^2}{2\sigma_{sa}^2}\Big\}\Big] \\ &= \prod_{s=1}^{2}\prod_{a=0}^{1}\Big[\exp\Big\{-\frac{n_{sa}^{(1)}\big\{2\mu_{sa}(\widetilde{\mu}_{sa}^{(1)}-\widehat{\mu}_{sa}^{(1)})+\big(\{\widehat{\mu}_{sa}^{(1)}\}^2-\{\widetilde{\mu}_{sa}^{(1)}\}^2\big)\big\}}{2\sigma_{sa}^2}\Big\}\cdot \\ &\qquad \exp\Big\{-\frac{\sum_{i:A_{s,i}^{(1)}=a}\big(Y_{s,i}^{(1)}-\widehat{\mu}_{sa}^{(1)}\big)^2-\big(Y'^{(1)}_{s,i}-\widetilde{\mu}_{sa}^{(1)}\big)^2}{2\sigma_{sa}^2}\Big\}\Big],\end{aligned}$$

where $\widetilde{\mu}_{sa}^{(1)}$ denotes the corresponding sample mean generated from Y'. It is immediately seen that the likelihood ratio is independent of μ_{sa}'s if and only if $\widehat{\mu}_{sa}^{(1)}=\widetilde{\mu}_{sa}^{(1)}$ for all $(s,a)\in\{(s,a):s=1,2;a=0,1\}$, which shows the minimal sufficiency of the statistic $\big\{\widehat{\mu}_{sa}^{(1)}:s=1,2;a=0,1\big\}$ by Lehmann-Scheffé Theorem (Lehmann and Scheffé, 1950).

At the end of the trial, the minimal sufficiency of the statistic $\big(D(\mathbf{Z}^{(1)}),\{\widehat{\mu}_{sa}^{(F)}\}\big)$ follows by the same argument. The only difference is that the sample sizes $n_{sa}^{(F)}$'s at the end of the trial depend on $D(\boldsymbol{Z}^{(1)})$. In particular, at the end of stage 2, the likelihood function is

$$\begin{aligned}\mathcal{L}(Y) &= \exp\Big\{-\sum_{k=1}^{2}\sum_{i=1}^{N^{(k)}}\frac{\big(Y_{s,i}^{(k)}-\mu_{s,A_{s,i}^{(k)}}\big)^2}{2\sigma^2_{s,A_{s,i}^{(k)}}}\Big\}\prod_{k=1}^{2}\prod_{i=1}^{N^{(k)}}\big\{2\pi\sigma^2_{s,A_{s,i}^{(k)}}\big\}^{-1/2} \\ &= \prod_{s=1}^{2}\prod_{a=0}^{1}\exp\Big\{-\sum_{k=1}^{2}\sum_{i:A_{s,i}^{(k)}=a}\frac{\big(Y_{s,i}^{(k)}-\mu_{s,A_{s,i}^{(k)}}\big)^2}{2\sigma^2_{sa}}\Big\}\prod_{k=1}^{2}\prod_{i=1}^{N^{(k)}}\big\{2\pi\sigma^2_{s,A_{s,i}^{(k)}}\big\}^{-1/2} \\ &= \prod_{s=1}^{2}\prod_{a=0}^{1}\Big[\exp\Big\{-\frac{\big(n_{sa}^{(1)}+n_{sa}^{(2)}\big)\big(\mu_{sa}-\widehat{\mu}_{sa}\big)\big)^2}{2\sigma_{sa}^2}\Big\}\exp\Big\{-\frac{\sum_{(i,k):A_{s,i}^{(k)}=a}\big(Y_{s,i}^{(k)}-\widehat{\mu}_{sa}\big)^2}{2\sigma_{sa}^2}\Big\}\Big] \\ &\quad\cdot\prod_{k=1}^{2}\prod_{i=1}^{N^{(k)}}\big\{2\pi\sigma^2_{s,A_{s,i}^{(k)}}\big\}^{-1/2}.\end{aligned}$$

By Fisher-Neyman factorization theorem, we have that $\big\{\big(\widehat{\mu}_{sa}, n_{sa}^{(2)}\big) : s = 1, 2; a = 0, 1\big\}$ is a sufficient statistic.

To prove the minimal sufficiency, let Y' be an independent set of outcomes. The likelihood ratio is

$$\frac{\mathcal{L}(Y)}{\mathcal{L}(Y')} = \prod_{s=1}^{2}\prod_{a=0}^{1}\Big[\exp\Big\{-\frac{(n_{sa}^{(1)}+n_{sa}^{(2)})(\mu_{sa}-\widehat{\mu}_{sa})^2-(n'^{(1)}_{sa}+n'^{(2)}_{sa})(\mu_{sa}-\widetilde{\mu}_{sa})^2}{2\sigma_{sa}^2}\Big\}\cdot \exp\Big\{-\frac{\sum_{(i,k):A_{s,i}=a}\big(Y_{s,i}^{(k)}-\widehat{\mu}_{sa}^{(k)}\big)^2-\big(Y'^{(k)}_{s,i}-\widetilde{\mu}_{sa}^{(k)}\big)^2}{2\sigma_{sa}^2}\Big\}\Big].$$

The likelihood ratio is independent of μ_{sa}'s if and only if $\widehat{\mu}_{sa} = \widetilde{\mu}_{sa}$ for all $(s, a) \in \{(s, a) : s = 1, 2; a = 0, 1\}$ and $n_{sa}^{(2)} = n'^{(2)}_{sa}$, which shows the minimal sufficiency of the statistic $\big\{\big(\widehat{\mu}_{sa}, D(\mathbf{Z}^{(1)}, U_1)\big) : s = 1, 2; a = 0, 1\big\}$ by Lehmann-Scheffé Theorem (Lehmann and Scheffé, 1950). □

Next, we prove a theorem shows that to get an optimal design (D, M) which minimizes the constrained Bayes optimization, M only needs to depend on $(\mathbf{Z}^{(F)}, D(\mathbf{Z}^{(1)}, U_1), U_2)$.

Theorem A.6.1. If the constrained Bayes optimization problem in Section 2.1.3 is feasible, then there exists an optimal solution $(D, M) \in (\mathcal{E} \times \mathcal{M})$, i.e., for which D depends on the data only through $\mathbf{Z}^{(1)}$, and M depends only the data through $\mathbf{Z}^{(F)}$ and the decision $D(\mathbf{Z}^{(1)}, U_1)$.

Proof. Since we assume independences among the distributions. The test of interests are equivalent to $H'_{0s} : \Delta_s \leq 0$ for random variables $N(\delta_s, 1)$ for $s \in \{1, 2\}$ and $N\big(\Delta_C, p_1^2(\sigma_{11}^2 + \sigma_{10}^2) + p_2^2(\sigma_{21}^2 + \sigma_{20}^2)\big)$. Note that we assume $n_{s1}^{(k)} = n_{s0}^{(k)} = N_s^{(k)}/2$ for all s and k. Without loss of generality, by reindexing the sample, we let the samples generated from arm $a = 1$ by $Y_{s,i}^{(k)}$, and let samples generated from arm $a = 0$ by $X_{s,i}^{(k)}$. We consider the samples $\{Y_{s,i}^{(k)} - X_{s,i}^{(k)}\}$ for $s \in \{1, 2\}$, $k \in \{1, 2\}$ and $i = 1, ..., N_s^{(k)}/2$. Each sample follows the distribution $N(\Delta_s, \sigma_{s1}^2 + \sigma_{s0}^2)$, and all the samples are independent. By a proof analogous to

that in Lemma A.6.1, we have the z-statistics $\mathbf{Z}^{(1)}$ is a minimal sufficient statistic for (δ_1, δ_2) at the interim stage. Similarly, At the end of the trial, the set of $D(\mathbf{Z}^{(1)}, U_1)$ and $\mathbf{Z}^{(F)}$ is a minimal sufficient statistic for (δ_1, δ_2). Note that these z-statistics do not depend on how we reindex the samples.

Suppose there exists an optimal feasible rule (D', M'), where D depends on $X^{(1)}$, and M' depends on X, and the rule satisfies the constraint $\mathbb{P}_{\boldsymbol{\delta}}\{M \text{ rejects any null hypotheis}\} \leq \alpha$ for given $\boldsymbol{\delta} = (\delta_1, \delta_2)$ and α. Let the exogenous information be U_1 and U_2. By Theorem 1 of Berger (1985), since $\{D(\mathbf{Z}^{(1)}, U_1), \mathbf{Z}^{(F)}\}$ is a sufficient statistic for the parameters of interest, there exists a randomized rule $(D, M) \in (\mathcal{E} \times \mathcal{M})$, which is L-equivalent to (D', M') as defined in Definition A.6.1 and also satisfies the constraint. We have that there exists an optimal randomized decision rule (D, M), where D only depends on $\mathbf{Z}^{(1)}$ and U_1 and M only depends on $D(\mathbf{Z}^{(1)}, U_1)$, $\mathbf{Z}^{(F)}$ and U_2. Our claim holds as desired. □

Definition A.6.1. A decision rule M is L-equivalent to M' if $E_\delta\big\{L(M, \delta)\big\} = E_\delta\big\{L(M', \delta)\big\}$, where L is a loss function, and δ is the parameter of interests.

Appendix B

Appendix to Chapter 3

B.1 Shapley-Folkman Lemma

In this section, we provide the mathematical details of Shapley-Folkman Lemma. This lemma studies the geometry of Minkowski sum of sets in vector space. The Minkowski sum of sets is defined as follows.

Definition B.1.1. The Minkowski sum of sets of vectors is formed by summing one vector of each set, i.e., the Minkowski sum of sets $\mathcal{A}_i$'s, for $i = 1, ..., I$ is

$$\mathcal{A}_1 \oplus \mathcal{A}_2, ..., \oplus \mathcal{A}_I = \Big\{ \sum_{i=1}^{I} a_i : a_i \in \mathcal{A}_i \text{ for all } i \Big\}.$$

Shapley-Folkman Lemma is formally stated as follows.

Theorem B.1.2 (Shapley-Folkman Lemma)**.** *Let $\{\mathcal{A}_i\}_{i=1}^{I}$ be a collection of sets of $\mathbb{R}^k$. Let a belongs to the convex hull of the Minkowski sum $\oplus_{i=1}^{I} \mathcal{A}_i$, i.e., $a \in Conv\big(\oplus_{i=1}^{I} \mathcal{A}_i \big)$. If $k < I$, we have, a belongs to the sum of convex hulls of k sets and the Minkowski sum of the rest sets, i.e.,*

$$a \in \sum_{1 \le i \le k} Conv(\mathcal{A}_i^a) + \mathcal{A}_{k+1}^a \oplus \mathcal{A}_{k+2}^a, ..., \oplus \mathcal{A}_I^a,$$

where $\{\mathcal{A}_i^a\}_{i=1}^I$ *is some re-indexing of* $\{\mathcal{A}_i\}_{i=1}^I$ *depends on the point* a*.*

This Lemma is used to prove the following theorem which bounds the distance between a Minkowski sum and its convex hull.

Theorem B.1.3. *Let* $\{\mathcal{A}_i\}_{i=1}^I$ *be a collection of sets of* $\mathbb{R}^k$*. If* $I \geq k$*, for any point in the convex hull of a Minkowski sum, Conv*$(\sum_{i=1}^I \mathcal{A}_i)$*, its distance to the Minkowski sum* $\oplus_{i=1}^I \mathcal{A}_i$ *is upper bounded by the sum of the squares of the squares of the* k *largest circumradii of the sets* $\mathcal{A}_i$*'s, where the circumradii of a set is defined as the radii of the smallest sphere in* $\mathbb{R}^k$ *enclosing the set.*

We point out the the above theorem is independent of the number I as long as $I > k$ holds. In the constraint of our problem (3.1.1), the dimension of each set is small, and the number of sets d is very large. The above theorem provides the intuition behind the decrease of the distance between the feasible set and its convex hull. This further results the decrease of duality gap, and the distance is upper bounded in Section 3.2.3.

B.2 Lemmas for Proving Theorem 3.2.1

In this section, we provide lemmas used in the proof of Theorem 3.2.1.

Lemma B.2.1. *Assume that there exists a feasible primal optimal solution to the problem* (3.1.1)*. Then, the dual optimal solution* $\left(\widehat{\lambda}, \{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}\right)$ *also exists, which is defined as in* (3.2.2)*.*

Proof. We first prove the existence of the dual optimal Lagrangian multiplier $\widehat{\lambda}$. By the definition of $\sum_{j\in[d]} \mathcal{Q}_j(\lambda) - \lambda K$ in (3.2.1), we have that $\sum_{j\in[d]} \mathcal{Q}_j(\lambda) - \lambda K$ is the infimum of a collection of sum of linear functions. This implies that $\sum_{j\in[d]} \mathcal{Q}_j(\lambda) - \lambda K$ is a concave function. Also, we observe that if $\lambda \to \infty$, then $\sum_{j\in[d]} \mathcal{Q}_j(\lambda) - \lambda K \to -\infty$ as shown in Lemma B.2.2. Thus, $\sum_{j\in[d]} \mathcal{Q}_j(\lambda) - \lambda K$ has compact level sets due to its concavity, i.e., for

any $\alpha \in \mathbb{R}$, the set $\{\lambda : \sum_{j\in[d]} \mathcal{Q}_j(\lambda) - \lambda K \geq \alpha\}$ is compact. By Bolzano-Weiestrass Theorem, there exits at least one optimal Lagrangian multiplier $\widehat{\lambda}$ that maximizes $\sum_{j\in[d]} \mathcal{Q}_j(\lambda) - \lambda K$ over $\lambda \geq 0$.

Next, we prove the existence of $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$. Given the optimal Lagrangian multiplier $\widehat{\lambda}$, for each j, the dual solution $\widehat{\boldsymbol{\beta}}_j$ is

$$\widehat{\boldsymbol{\beta}}_j = \operatorname*{argmin}_{\boldsymbol{\beta}_j} \mathcal{L}_j(\boldsymbol{\beta}_j) + \widehat{\lambda}\|\boldsymbol{\beta}_j\|_0, \text{ or equivalently, } \widehat{\boldsymbol{\beta}}_j = \operatorname*{argmin}_{\widehat{\boldsymbol{\beta}}_j(k)} \mathcal{L}_j\big\{\boldsymbol{\beta}_j(k)\big\} + \widehat{\lambda}k,$$

where $\widehat{\boldsymbol{\beta}}_j(k) = \operatorname{argmin}_{\|\boldsymbol{\beta}_j\|_0=k} \mathcal{L}_j(\boldsymbol{\beta}_j)$. Since the primal solution exists, all $\mathcal{L}_j$'s are bounded below. Thus each subproblem j can be solved by branch-and-bound for each $k \in \{1, ..., d_0\}$, and the solution $\widehat{\boldsymbol{\beta}}_j$ is obtained by minimizing the objective above over k. Therefore there exists at least one dual optimal solution $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$. □

Lemma B.2.2. *Consider the Lagrangian dual $\sum_{j\in[d]} \mathcal{Q}_j(\lambda) - \lambda K$ of the problem (3.1.1). If a primal optimal solution $\{\widetilde{\boldsymbol{\beta}}_j\}_{j=1}^d$ exists for the primal problem, we have $\sum_{j\in[d]} \mathcal{Q}_j(\lambda) - \lambda K \to -\infty$ as $\lambda \to \infty$.*

Proof. We consider each $\mathcal{Q}_j(z)$ separately. For each j, let

$$\widehat{\boldsymbol{\beta}}_j(k) = \operatorname*{argmin}_{\boldsymbol{\beta}_j} \mathcal{L}_j(\boldsymbol{\beta}_j), \text{ subject to } \|\boldsymbol{\beta}_j\|_0 = k.$$

Let $\widehat{\boldsymbol{\beta}}_j(k) = \operatorname{argmin}_{\|\boldsymbol{\beta}_j\|_0=k} \mathcal{L}_j(\boldsymbol{\beta}_j)$ for $k = 0, 1, ..., d_0$. Since the primal solution exits, it is not difficult to check that, if $\lambda > \mathcal{L}_j\big\{\widehat{\boldsymbol{\beta}}_j(k)\big\} - \mathcal{L}_j\big\{\widehat{\boldsymbol{\beta}}_j(k+1)\big\}$ for all $k = 0, ..., d_0 - 1$, we have $\mathbf{0} = \operatorname{argmin}_{\boldsymbol{\beta}_j} \mathcal{L}_j(\boldsymbol{\beta}_j) + \lambda\|\boldsymbol{\beta}_j\|_0$, which implies $\mathcal{Q}_j(\lambda) = \mathcal{L}_j(\mathbf{0})$. Thus, when λ is large, $\sum_{j\in[d]} \mathcal{Q}_j(\lambda) - \lambda K = \sum_{j\in[d]} \mathcal{L}_j(\mathbf{0}) - \lambda K$, which decreases linearly. Since $K > 0$, our claim follows as desired. □

Lemma B.2.3. *For the dual problem (3.2.2), suppose that the optimal Lagrangian multiplier $\widehat{\lambda} > 0$. One of the following two cases holds:*

1. *There exists a dual optimal solution-multiplier pair* $(\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}, \widehat{\lambda})$, *where* $\sum_{j\in[d]} \|\widehat{\boldsymbol{\beta}}_j\|_0 = K$.

2. *There exists at least two solutions achieve optimal dual objective, denoted as* $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$ *and* $\{\widehat{\boldsymbol{\beta}}'_j\}_{j\in[d]}$, *such that* $\sum_{j=1}^d \|\widehat{\boldsymbol{\beta}}_j\|_0 < K$ *and* $\sum_{j=1}^d \|\widehat{\boldsymbol{\beta}}'_j\|_0 > K$.

Proof. We prove the lemma by contradiction. Assume to the contrary that either $\sum_{j\in[d]} \|\widehat{\boldsymbol{\beta}}_j\|_0 > K$ for all solutions $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$ achieve dual optimal objective (for the case $\sum_{j\in[d]} \|\widehat{\boldsymbol{\beta}}_j\|_0 < K$, the claim follows by similar arguments.). We have

$$\begin{aligned}\sum_{j=1}^d \mathcal{Q}_j(\widehat{\lambda}) - \widehat{\lambda}K &= \min_{\boldsymbol{\beta}_j} \Big\{ \sum_{j=1}^d \mathcal{L}_j(\boldsymbol{\beta}_j) + \widehat{\lambda}\big(\sum_{j=1}^d \|\boldsymbol{\beta}_j\|_0 - K\big)\Big\} \\ &= \min_{k_j\in\{0,1,\ldots,d_0\}} \Big\{ \sum_{j=1}^d \mathcal{L}_j\big\{\widehat{\boldsymbol{\beta}}_j(k_j)\big\} + \widehat{\lambda}(\sum_{j=1}^d k_j - K)\Big\},\end{aligned}$$

where $\widehat{\boldsymbol{\beta}}_j(k_j) = \operatorname{argmin}_{\|\boldsymbol{\beta}_j\|_0=k_j} \mathcal{L}_j(\boldsymbol{\beta}_j)$.

By the assumption, it holds that $\sum_{j\in[d]} k_j$ is strictly greater than K. As shown in Lemma B.2.4, we have that for small $\epsilon > 0$ (or $\epsilon < 0$ if $\sum_{j\in[d]} k_j < K$), we obtain

$$\begin{aligned}\sum_{j=1}^d \mathcal{Q}_j(\widehat{\lambda}+\epsilon) - (\widehat{\lambda}+\epsilon)K &= \min_{k_j=1,\ldots,d} \Big[\sum_{j=1}^d \mathcal{L}_j\big\{\widehat{\boldsymbol{\beta}}_j(k_j)\big\} + (\widehat{\lambda}+\epsilon)\big(\sum_{j=1}^d k_j - K\big)\Big] \\ &> \min_{k_j=1,\ldots,d} \Big[\sum_{j=1}^d \mathcal{L}_j\big\{\widehat{\boldsymbol{\beta}}_j(k_j)\big\} + \widehat{\lambda}\big(\sum_{j=1}^d k_j - K\big)\Big] \\ &= \sum_{j=1}^d \mathcal{Q}_j(\widehat{\lambda}) - \widehat{\lambda}K.\end{aligned}$$

Meanwhile, since $\widehat{\lambda}$ is the optimal Lagrangian multiplier by assumption, it maximizes the function $\sum_{j\in[d]} \mathcal{Q}_j(\lambda) - K$. The above result yields a contradiction. □

Lemma B.2.4. *Suppose that the dual optimal Lagrangian multiplier* $\widehat{\lambda} > 0$. *Let* $k_j = \|\widehat{\boldsymbol{\beta}}_j\|_0$ *for some optimal dual solution* $\widehat{\boldsymbol{\beta}}_j$*'s, and let* $\widehat{\boldsymbol{\beta}}_j(k_j) = \operatorname{argmin}_{\|\boldsymbol{\beta}_j\|_0=k_j} \mathcal{L}_j(\boldsymbol{\beta}_j)$. *It holds that if*

$0 < \epsilon < \min_{j\in[d]} \{\mathcal{L}_j(\widehat{\boldsymbol{\beta}}_j(k_j-1)) - \mathcal{L}_j(\widehat{\boldsymbol{\beta}}_j(k_j))\}$,

$$\sum_{j=1}^{d} \mathcal{Q}_j(\widehat{\lambda}+\epsilon) - (\widehat{\lambda}+\epsilon)K = \min_{k_j=1,\ldots,d} \Big\{ \sum_{j=1}^{d} \mathcal{L}_j\{\boldsymbol{\beta}_j(k_j)\} + (\widehat{\lambda}+\epsilon)\big(\sum_{j=1}^{d} k_j - K\big)\Big\}.$$

Proof. By the definition of $\mathcal{Q}_j(\cdot)$ in (3.2.1), when we perturb the optimal Lagrangian multiplier $\widehat{\lambda}$ to $\widehat{\lambda}+\epsilon$, the corresponding dual solutions become

$$\widehat{\boldsymbol{\beta}}'_j = \operatorname*{argmin}_{\boldsymbol{\beta}_j} \mathcal{L}_j(\boldsymbol{\beta}_j) + (\widehat{\lambda}+\epsilon)\|\boldsymbol{\beta}_j\|_0, \text{ for all } j = 1, ..., d.$$

We have, if

$$\mathcal{L}_j\{\widehat{\boldsymbol{\beta}}_j(k_j)\} + (\widehat{\lambda}+\epsilon)k_j < \mathcal{L}_j\{\widehat{\boldsymbol{\beta}}_j(k_j-1)\} + (\widehat{\lambda}+\epsilon)(k_j-1) \text{ for all } j,$$

letting $\widehat{\boldsymbol{\beta}}'_j = \widehat{\boldsymbol{\beta}}_j$ for all j provides dual optimal solutions. Meanwhile, by our assumption of ϵ, it is not difficult to check that the above inequality holds for all j. Thus, our claim holds as desired. □

Lemma B.2.5. *Let* $\{\widehat{\boldsymbol{\beta}}_j^{(1)}\}_{j\in[d]},\ldots,\{\widehat{\boldsymbol{\beta}}_j^{(m)}\}_{j\in[d]}$ *be the sequence of dual solutions ranked by their corresponding primal objective values as defined in* (3.2.7)*. We have that for any* $m_1 \in \{1, ..., m-1\}$*, it holds that the difference of primal objective values of two consecutive solutions is bounded by* C_g *as defined in* (3.2.5)*, i.e.,*

$$\sum_{j=1}^{d} \mathcal{L}_j(\widehat{\boldsymbol{\beta}}_j^{(m_1+1)}) - \sum_{j=1}^{d} \mathcal{L}_j(\widehat{\boldsymbol{\beta}}_j^{(m_1)}) \le C_g.$$

Proof. By the dual optimality of the solutions and the separable structure of the dual objective, given the optimal Lagrangian multiplier $\widehat{\lambda} > 0$, we have

$$\widehat{\boldsymbol{\beta}}_j^{(m_1)} \in \operatorname*{argmin}_{\widehat{\boldsymbol{\beta}}_j} \mathcal{L}_j(\boldsymbol{\beta}_j) + \widehat{\lambda}\|\widehat{\boldsymbol{\beta}}_j\|_0, \text{ for all } j = 1, ..., d \text{ and } m_1 = 1, ..., m.$$

In addition, due to the ℓ_0-penalization term, we have

$$\widehat{\boldsymbol{\beta}}_j^{(m_1)} \in \{\widehat{\boldsymbol{\beta}}_j(k)\}_{k=0}^{d_0}, \text{ for all } j = 1, ..., d, \text{ and } m_1 = 1, ..., m,$$

where $\widehat{\boldsymbol{\beta}}_j(k) = \text{argmin}_{\|\boldsymbol{\beta}_j\|_0=k} \mathcal{L}_j(\boldsymbol{\beta}_j)$.

Thus, for each optimal solution $\{\boldsymbol{\beta}_j^{(m_1)}\}_{j\in[d]}$, it is a combination of the optimal solutions of the ℓ_0-penalized subproblems. By the dual decomposition structure as seen in (3.2.2), we have that for any solution achieves dual optimal objective $\{\widehat{\boldsymbol{\beta}}_j^{(m_1)}\}_{j\in[d]}$, there exists another solution achieves dual optimal objective $\{\widehat{\boldsymbol{\beta}}_j^{(m_2)}\}_{j\in[d]}$, such that these two solutions differ by at most one $\widehat{\boldsymbol{\beta}}_{j'}$ for some j', i.e., there exists an $m_2 \in \{1, ..., m-1\}$ and a $j' \in \{1, ..., d\}$ such that

$$\widehat{\boldsymbol{\beta}}_{j'}^{(m_1)} \neq \widehat{\boldsymbol{\beta}}_{j'}^{(m_2)}, \text{ and } \widehat{\boldsymbol{\beta}}_j^{(m_1)} = \widehat{\boldsymbol{\beta}}_j^{(m_2)} \text{ for all } j \in \{1, ..., j'-1, j'+1, ..., d\}. \tag{B.2.1}$$

Consequently, we have

$$\begin{aligned}
\Big|\sum_{j=1}^{d} \mathcal{L}_j\big(\widehat{\boldsymbol{\beta}}_j^{(m_1)}\big) - \sum_{j=1}^{d} \mathcal{L}_j\big(\widehat{\boldsymbol{\beta}}_j^{(m_2)}\big)\Big| &= \Big|\mathcal{L}_{j'}\big(\widehat{\boldsymbol{\beta}}_{j'}^{(m_1)}\big) - \mathcal{L}_j\big(\widehat{\boldsymbol{\beta}}_{j'}^{(m_1)}\big)\Big| \\
&\leq \big|\mathcal{L}_{j'}\big(\mathbf{0}\big) - \min_{\boldsymbol{\beta}_{j'}\in\mathcal{C}_{j'}} \mathcal{L}_{j'}\big(\boldsymbol{\beta}_{j'}\big)\big| \\
&\leq C_g,
\end{aligned}$$

where the first inequality holds by $\mathcal{L}_j\big(\mathbf{0}\big) \geq \mathcal{L}_j\big(\widehat{\boldsymbol{\beta}}_j(1)\big) \geq, ..., \geq \mathcal{L}_j\big(\widehat{\boldsymbol{\beta}}_j(d_0)\big) = \min_{\boldsymbol{\beta}_j\in\mathcal{C}_j} \mathcal{L}_j\big(\boldsymbol{\beta}_j\big)$ for any j, and the second inequality holds by the definition of C_g in (3.2.5).

Since the sequence $\big\{\widehat{\boldsymbol{\beta}}_j^{(1)}\big\}_{j\in[d]}$, $\big\{\widehat{\boldsymbol{\beta}}_j^{(2)}\big\}_{j\in[d]}$,..., $\big\{\widehat{\boldsymbol{\beta}}_j^{(m)}\big\}_{j\in[d]}$ is ordered by their corresponding objective values, It follows immediately from the above inequality that the difference between any two consecutive objective values is bounded by C_g also, i.e.,

$$\Big|\sum_{j=1}^{d} \mathcal{L}_j\big(\widehat{\boldsymbol{\beta}}_j^{(m_1+1)}\big) - \sum_{j=1}^{d} \mathcal{L}_j\big(\widehat{\boldsymbol{\beta}}_j^{(m_1)}\big)\Big| \leq C_g, \text{ for any } m_1 = 1, ..., m-1,$$

and our claim holds as desired. □

B.3 Proofs in Section 3.3

B.3.1 Proof of Lemma 3.3.2

Proof. By Theorem 3.2.1, we have

$$\frac{1}{n}\sum_{j=1}^{d}\|\mathbb{X}_{\mathcal{N}_j}\widehat{\boldsymbol{\beta}}_j-\mathbb{X}_j\|_2^2\leq\frac{1}{n}\sum_{j=1}^{d}\|\mathbb{X}_{\mathcal{N}_j}\boldsymbol{\beta}_j^*-\mathbb{X}_j\|_2^2+C_g.$$

Let

$$\begin{aligned}\widehat{\mathbb{X}}&=\begin{pmatrix}\mathbb{X}_{\mathcal{N}_1}&\mathbf{0}&\cdots&\mathbf{0}\\\mathbf{0}&\mathbb{X}_{\mathcal{N}_2}&\cdots&\mathbf{0}\\\vdots&\vdots&\ddots&\vdots\\\mathbf{0}&\mathbf{0}&\cdots&\mathbb{X}_{\mathcal{N}_d}\end{pmatrix}\in\mathbb{R}^{nd\times d_0d},\\\mathbf{y}&=\begin{pmatrix}\mathbb{X}_1\\\mathbb{X}_2\\\vdots\\\mathbb{X}_d\end{pmatrix}\in\mathbb{R}^{nd},\text{ and }\widehat{\boldsymbol{\beta}}=\begin{pmatrix}\widehat{\boldsymbol{\beta}}_1\\\widehat{\boldsymbol{\beta}}_2\\\vdots\\\widehat{\boldsymbol{\beta}}_d\end{pmatrix},\boldsymbol{\beta}^*=\begin{pmatrix}\boldsymbol{\beta}_1^*\\\boldsymbol{\beta}_2^*\\\vdots\\\boldsymbol{\beta}_d^*\end{pmatrix}\in\mathbb{R}^{d_0d}.\end{aligned}\tag{B.3.1}$$

Let $\mathbf{y}-\widehat{\mathbb{X}}\boldsymbol{\beta}^*=(\boldsymbol{\epsilon}_1^T,...,\boldsymbol{\epsilon}_d^T)^T=\boldsymbol{\epsilon}^T\in\mathbb{R}^{nd}$. It holds that each component of $\boldsymbol{\epsilon}$ follows a Gaussian distribution with marginal variance at most σ^2. We have

$$\|\mathbf{y}-\widehat{\mathbb{X}}\widehat{\boldsymbol{\beta}}\|_2^2\leq\|\mathbf{y}-\widehat{\mathbb{X}}\boldsymbol{\beta}^*\|_2^2+nC_g=\|\boldsymbol{\epsilon}\|_2^2+nC_g.$$

Meanwhile, it holds that

$$\|\mathbf{y}-\widehat{\mathbb{X}}\widehat{\boldsymbol{\beta}}\|_2^2=\|\widehat{\mathbb{X}}\boldsymbol{\beta}^*+\boldsymbol{\epsilon}-\widehat{\mathbb{X}}\widehat{\boldsymbol{\beta}}\|_2^2=\|\widehat{\mathbb{X}}\widehat{\boldsymbol{\beta}}-\widehat{\mathbb{X}}\boldsymbol{\beta}^*\|_2^2-2\boldsymbol{\epsilon}^T\widehat{\mathbb{X}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)+\|\boldsymbol{\epsilon}\|_2^2.$$

Combining the above two relations, we have

$$\|\widehat{\mathbb{X}}\widehat{\boldsymbol{\beta}} - \widehat{\mathbb{X}}\boldsymbol{\beta}^*\|_2^2 \leq 2\boldsymbol{\epsilon}^T\widehat{\mathbb{X}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + nC_g = \|\widehat{\mathbb{X}}\widehat{\boldsymbol{\beta}} - \widehat{\mathbb{X}}\boldsymbol{\beta}^*\|_2 \frac{2\boldsymbol{\epsilon}^T\widehat{\mathbb{X}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)}{\|\widehat{\mathbb{X}}\widehat{\boldsymbol{\beta}} - \widehat{\mathbb{X}}\boldsymbol{\beta}^*\|_2} + nC_g. \tag{B.3.2}$$

By the assumptions that $s \leq K$ and $2K \leq dd_0$, letting $\mathcal{B}_0(2K) = \{\mathbf{v} \in \mathbb{R}^{dd_0} : \|\mathbf{v}\|_0 \leq 2K\}$, we have that $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \in \mathcal{B}_0(2K)$. Denote by $\widehat{\mathcal{K}} = \text{supp}(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})$, and let the $nd \times |\widehat{\mathcal{K}}|$ submatrix of $\widehat{\mathbb{X}}$ be $\widehat{\mathbb{X}}_{\widehat{\mathcal{K}}}$, where $\widehat{\mathcal{K}} \subset \{1, 2, ..., d_0d\}$. Assume that $\widehat{\mathbb{X}}_{\widehat{\mathcal{K}}}$ is of rank r. Let $\Psi_r = [\psi_1, ..., \psi_r] = \mathbb{R}^{nd \times r}$ be an orthonormal basis for the column space of $\widehat{\mathbb{X}}_{\widehat{\mathcal{K}}}$. We have that there exists a vector $\boldsymbol{\nu} \in \mathbb{R}^r$,

$$\widehat{\mathbb{X}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) = \widehat{\mathbb{X}}_{\widehat{\mathcal{K}}}\big\{\widehat{\boldsymbol{\beta}}(\widehat{\mathcal{K}}) - \boldsymbol{\beta}^*(\widehat{\mathcal{K}})\big\} = \Psi_r\boldsymbol{\nu}.$$

Plugging the above equation into (B.3.2), we have

$$\frac{\boldsymbol{\epsilon}^T\widehat{\mathbb{X}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)}{\|\widehat{\mathbb{X}}\widehat{\boldsymbol{\beta}} - \widehat{\mathbb{X}}\boldsymbol{\beta}^*\|_2} = \frac{\boldsymbol{\epsilon}^T\Psi_r\boldsymbol{\nu}}{\|\boldsymbol{\nu}\|_2} \leq \max_{r \leq 2K} \sup_{\mathbf{v} \in \mathbb{S}^{r-1}} [\boldsymbol{\epsilon}^T\Psi_r]\mathbf{v},$$

where $\mathbb{S}^{r-1} = \{\mathbf{z} \in \mathbb{R}^r : \|\mathbf{z}\|_2 \leq 1\}$. Then, by (B.3.2) and some algebraic manipulation, we have

$$\|\widehat{\mathbb{X}}\widehat{\boldsymbol{\beta}} - \widehat{\mathbb{X}}\boldsymbol{\beta}^*\|_2^2 \leq 8 \max_{r \leq 2K} \sup_{\mathbf{v} \in \mathbb{S}^{r-1}} (\widetilde{\boldsymbol{\epsilon}}_r^T\mathbf{v})^2 + 2nC_g,$$

where $\widetilde{\boldsymbol{\epsilon}}_r = \Psi_r^T\boldsymbol{\epsilon}$ is a Gaussian random vector with marginal variance at most σ^2.

Taking a union bound, we have that for any $t > 0$,

$$\mathbb{P}\big\{\max_{r \leq 2K} \sup_{\mathbf{v} \in \mathbb{S}^{r-1}} (\widetilde{\boldsymbol{\epsilon}}_r^T\mathbf{v})^2 > t\big\} \leq \sum_{r \leq 2K} \mathbb{P}\big\{\sup_{\mathbf{v} \in \mathbb{S}^{r-1}} (\widetilde{\boldsymbol{\epsilon}}_r^T\mathbf{v})^2 > t\big\}.$$

By Lemma B.4.1, we have

$$\mathbb{P}\big\{\max_{r \leq 2K} \sup_{\mathbf{v} \in \mathbb{S}^{r-1}} (\widetilde{\boldsymbol{\epsilon}}_r^T\mathbf{v})^2 > t\big\} \leq 6^r \exp\big\{-t/(8\sigma^2)\big\} \leq 6^{2K} \exp\big\{-t/(8\sigma^2)\big\}.$$

Combining the above three inequalities, we have, for any $t > 0$,

$$\mathbb{P}\big(\|\widehat{\mathbb{X}}\widehat{\boldsymbol{\beta}} - \widehat{\mathbb{X}}\boldsymbol{\beta}^*\|_2^2 > 8t + 2nC_g\big) \leq \sum_{j=1}^{2K} \binom{dd_0}{j} 6^{2K} e^{-\frac{t}{8\sigma^2}}.$$

To ensure that the RHS of the above inequality is bounded up by δ, we need to ensure that

$$t \geq 8\sigma^2 \log\Big(\sum_{j=1}^{2K} \binom{dd_0}{j}\Big) + 16K\sigma^2 \log(6) + 8\sigma^2 \log(1/\delta),$$

and the claim (3.3.3) holds as desired. □

B.3.2 Proof of Lemma 3.3.3

Proof. For each j, since the loss function $\mathcal{L}_j(\cdot)$ is nonnegative, we have,

$$\mathcal{L}_j(\mathbf{0}) - \min_{\boldsymbol{\beta}_j} \mathcal{L}_j(\boldsymbol{\beta}_j) \leq \mathcal{L}_j(\mathbf{0}) = \sum_{i=1}^{n} x_{ij}^2.$$

Since each X_j follows a normal distribution with variance at most σ^2, we have $\nu_j^{-1} \sum_{i\in[n]} x_{ij}^2$ follows a chi-squared distribution with n degrees of freedom, where $\nu_j < \sigma^2$ is a positive constant. Denote by $Z_j = \nu_j^{-1} \sum_{i\in[n]} x_{ij}^2$. By Vempala (2005), we have,

$$\mathbb{P}\big(\big|n^{-1}Z_j - 1\big| \geq t\big) \leq 2\exp\big\{-4^{-1}n(t^2 - t^3)\big\}, \text{ for any } t > 0.$$

Using Bonferroni's method, we have

$$\mathbb{P}\big(\big|n^{-1}\max_{j\leq d} Z_j - 1\big| \geq t\big) \leq 2d\exp\big\{-4^{-1}n(t^2 - t^3)\big\}, \text{ for any } t > 0.$$

Our claim follows by performing some algebraic manipulations. □

B.3.3 Proof of Corollary 3.3.4

Proof. By (B.2.1), the estimator obtained by the SPICA algorithm is an optimal solution under the constraint $\sum_{j\in[d]} \|\boldsymbol{\beta}_j\|_0 = s - s'$, for some $s' \in \{0, ..., d_0\}$. We prove that the global minimizer of problem (3.1.1) recovers the true support with high probability. Using similar arguments, It is easy to generalize the proof to show that the estimator recovers $s - s'$ components of true support if $s' \in \{1, ..., d_0\}$, and we omit details to avoid repetition.

Adopt the notations in (B.3.1). For any subset $\mathcal{T} \subset \{1, ..., d_0 d\}$, let

$$f(\mathcal{T}) = \min_{\boldsymbol{\beta}\in\mathbb{R}^{|\mathcal{T}|}} \|\mathbf{y} - \widehat{\mathbb{X}}_{\mathcal{T}}\boldsymbol{\beta}_{\mathcal{T}}\|^2.$$

For a fixed set $\mathcal{S} = \cup_{j\in[d]}\text{supp}(\boldsymbol{\beta}_j^*)$, let $\Delta(\mathcal{K}) = f(\mathcal{K}) - f(\mathcal{S})$ and $\mathcal{G} = \{\mathcal{K} \subset \{1, ..., d_0 d\} : |\mathcal{K}| = s, \mathcal{K} \neq \mathcal{S}\}$. Then

$$\mathbb{P}(\mathcal{K} \neq \mathcal{S}) = \mathbb{P}\big(\cup_{\mathcal{K}\in\mathcal{G}} \{\Delta(\mathcal{K}) < 0\}\big) \leq \sum_{\mathcal{K}\in\mathcal{G}} \mathbb{P}\big(\Delta(\mathcal{K}) < 0\big).$$

Denote the least square estimator restricted to the support $\mathcal{K}$ by

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\mathcal{K}}^{\text{OLS}} &= (\widehat{\mathbb{X}}_{\mathcal{K}}^T\widehat{\mathbb{X}}_{\mathcal{K}})^{-1}\widehat{\mathbb{X}}_{\mathcal{K}}^T\mathbf{y} = (\widehat{\mathbb{X}}_{\mathcal{K}}^T\widehat{\mathbb{X}}_{\mathcal{K}})^{-1}\widehat{\mathbb{X}}_{\mathcal{K}}^T(\widehat{\mathbb{X}}_{\mathcal{K}}\boldsymbol{\beta}_{\mathcal{K}}^* + \widehat{\mathbb{X}}_{\mathcal{D}}\boldsymbol{\beta}_{\mathcal{D}}^* + \boldsymbol{\epsilon}) \\ &= \boldsymbol{\beta}_{\mathcal{K}}^* + (\widehat{\mathbb{X}}_{\mathcal{K}}^T\widehat{\mathbb{X}}_{\mathcal{K}})^{-1}\widehat{\mathbb{X}}_{\mathcal{K}}(\widehat{\mathbb{X}}_{\mathcal{D}}\boldsymbol{\beta}_{\mathcal{D}}^* + \boldsymbol{\epsilon}).\end{aligned}$$

Let the difference between $\mathcal{S}$ and $\mathcal{K}$ be $\mathcal{D} = \mathcal{S}\backslash\mathcal{K}$. We have

$$\begin{aligned}&\|\mathbf{y} - \widehat{\mathbb{X}}_{\mathcal{K}}\widehat{\boldsymbol{\beta}}_{\mathcal{K}}^{\text{OLS}}\|_2^2 \\ =&\|\boldsymbol{\epsilon} + \widehat{\mathbb{X}}_{\mathcal{K}}\boldsymbol{\beta}_{\mathcal{K}}^* + \widehat{\mathbb{X}}_{\mathcal{D}}\boldsymbol{\beta}_{\mathcal{D}}^* - \widehat{\mathbb{X}}_{\mathcal{K}}\boldsymbol{\beta}_{\mathcal{K}}^* - \widehat{\mathbb{X}}_{\mathcal{K}}(\widehat{\mathbb{X}}_{\mathcal{K}}^T\widehat{\mathbb{X}}_{\mathcal{K}})^{-1}\widehat{\mathbb{X}}_{\mathcal{K}}^T(\widehat{\mathbb{X}}_{\mathcal{D}}\boldsymbol{\beta}_{\mathcal{D}}^* + \boldsymbol{\epsilon})\|_2^2 \\ =&\|\big\{\mathbf{I}_{nd} - \widehat{\mathbb{X}}_{\mathcal{K}}(\widehat{\mathbb{X}}_{\mathcal{K}}^T\widehat{\mathbb{X}}_{\mathcal{K}})^{-1}\widehat{\mathbb{X}}_{\mathcal{K}}^T\big\}\boldsymbol{\epsilon} + \big\{\mathbf{I}_{nd} - \widehat{\mathbb{X}}_{\mathcal{K}}(\widehat{\mathbb{X}}_{\mathcal{K}}^T\widehat{\mathbb{X}}_{\mathcal{K}})^{-1}\widehat{\mathbb{X}}_{\mathcal{K}}^T\big\}\widehat{\mathbb{X}}_{\mathcal{D}}\boldsymbol{\beta}_{\mathcal{D}}^*\|_2^2 \\ =&\boldsymbol{\epsilon}^T\mathbf{P}_{\mathcal{K}}^{\perp}\boldsymbol{\epsilon} + 2\boldsymbol{\epsilon}^T\mathbf{P}_{\mathcal{K}}^{\perp}\widehat{\mathbb{X}}_{\mathcal{D}}\boldsymbol{\beta}_{\mathcal{D}}^* + \|\mathbf{P}_{\mathcal{K}}^{\perp}\widehat{\mathbb{X}}_{\mathcal{D}}\boldsymbol{\beta}_{\mathcal{D}}^*\|_2^2,\end{aligned}$$

where $\mathbf{P}_{\mathcal{K}}^{\perp} = \mathbf{I}_{nd} - \mathbf{P}_{\mathcal{K}}$, and $\mathbf{P}_{\mathcal{K}} = \widehat{\mathbb{X}}_{\mathcal{K}}(\widehat{\mathbb{X}}_{\mathcal{K}}^T\widehat{\mathbb{X}}_{\mathcal{K}})^{-1}\widehat{\mathbb{X}}_{\mathcal{K}}^T$ is a projection matrix, i.e., $\mathbf{P}_{\mathcal{K}}^2 = \mathbf{P}_{\mathcal{K}}$, which implies that $\mathbf{P}_{\mathcal{K}}^{\perp}$ is also a projection matrix.

By the same argument, we have

$$f(\mathcal{S}) = \boldsymbol{\epsilon}^T\mathbf{P}_{\mathcal{S}}^{\perp}\boldsymbol{\epsilon}.$$

We have, when $\Delta(\mathcal{K}) < 0$,

$$\boldsymbol{\epsilon}^T\mathbf{P}_{\mathcal{K}}^{\perp}\boldsymbol{\epsilon} + 2\boldsymbol{\epsilon}^T\mathbf{P}_{\mathcal{K}}^{\perp}\widehat{\mathbb{X}}_{\mathcal{D}}\boldsymbol{\beta}_{\mathcal{D}}^* + \|\mathbf{P}_{\mathcal{K}}^{\perp}\widehat{\mathbb{X}}_{\mathcal{D}}\boldsymbol{\beta}_{\mathcal{D}}^*\|_2^2 < \boldsymbol{\epsilon}^T\mathbf{P}_{\mathcal{S}}^{\perp}\boldsymbol{\epsilon}.$$

Let $\mathcal{S}_j = \mathrm{supp}(\boldsymbol{\beta}_j^*)$ and $\mathcal{K}_j = \mathrm{supp}(\widehat{\boldsymbol{\beta}}_j^K)$, and let $k' > 0$ be the number of $\mathcal{S}_j \neq \mathcal{K}_j$. By Lemma B.4.3, we have

$$\Delta(\mathcal{K}) \geq \widehat{\Delta}(\mathcal{K}) = 2\boldsymbol{\epsilon}\mathbf{P}_{\mathcal{K}}^{\perp}\widehat{\mathbb{X}}_{\mathcal{D}}\boldsymbol{\beta}_{\mathcal{D}}^* + \|\mathbf{P}_{\mathcal{K}}^{\perp}\widehat{\mathbb{X}}_{\mathcal{D}}\boldsymbol{\beta}_{\mathcal{D}}^*\|_2^2 - \sigma^2 R,$$

where R follows a chi-square distribution with $k'd_0$ degrees of freedom.

We have, when $\widehat{\Delta}(\mathcal{K}) < 0$ holds,

$$-2\boldsymbol{\epsilon}^T\mathbf{P}_{\mathcal{K}}^{\perp}\widehat{\mathbb{X}}_{\mathcal{D}}\boldsymbol{\beta}_{\mathcal{D}}^* \geq \|\mathbf{P}_{\mathcal{K}}^{\perp}\widehat{\mathbb{X}}_{\mathcal{D}}\boldsymbol{\beta}_{\mathcal{D}}^*\|_2^2 - \sigma^2 R.$$

Using the techniques in the proof of Theorem 3.3.1, we get,

$$\|\mathbf{P}_{\mathcal{K}}^{\perp}\widehat{\mathbb{X}}_{\mathcal{D}}\boldsymbol{\beta}_{\mathcal{D}}^*\|_2^2 \leq 8 \sup_{\mathbf{u}\in\mathbb{R}^{|\mathcal{D}|}} (\widetilde{\boldsymbol{\epsilon}}\mathbf{u})^2 + 2\sigma^2 R.$$

With probability at least $1 - 2\delta$, it follows that, by Lemma B.4.1 and Lemma 10 of Kolar and Liu (2013),

$$\|\mathbf{P}_{\mathcal{K}}^{\perp}\widehat{\mathbb{X}}_{\mathcal{D}}\boldsymbol{\beta}_{\mathcal{D}}^*\|_2^2 \leq 64\sigma^2|\mathcal{D}|\log 6 + 64\sigma^2\log\delta^{-1} + 2\big(k'd_0\sigma^2 + C\sqrt{k'd_0\log\delta^{-1}}\big),$$

for some constant C.

By Lemma B.4.2 and the sparse eigenvalue assumption, $\Lambda_{\max}(\widehat{\mathbb{X}}_{\mathcal{K}}^T\widehat{\mathbb{X}}_{\mathcal{K}}) \geq \Lambda_{\max}(\widehat{\mathbb{X}}_{\mathcal{D}}^T\mathbf{P}_{\mathcal{K}}^{\perp}\widehat{\mathbb{X}}_{\mathcal{D}}) \geq \Lambda_{\min}(\widehat{\mathbb{X}}_{\mathcal{D}}^T\mathbf{P}_{\mathcal{K}}^{\perp}\widehat{\mathbb{X}}_{\mathcal{D}}) \geq \Lambda_{\min}(\widehat{\mathbb{X}}_{\mathcal{K}}^T\widehat{\mathbb{X}}_{\mathcal{K}}) \geq \rho$ with probability at least $1 - \mathcal{O}(d^{-1})$. Let $s_1 = |\mathcal{D}|$. Taking $\delta = \big\{\binom{dd_0-s}{s_1}\binom{s}{s-s_1}d\big\}^{-1}$, as $s > s_1$, it is not difficult to see that the dominating term is $64\sigma^2 \log \delta^{-1}$, where

$$\begin{aligned}\log \delta^{-1} &= \log\left\{\binom{dd_0 - s}{s_1}\right\} + \log\left\{\binom{s}{s_1}\right\} + \log d \\ &\leq C_1 s_1 \log dd_0 + C_2 s_1 \log s + \log d,\end{aligned}$$

where the dominating term is $s_1 \log dd_0$.

Consequently, we have that, with probability at least $1 - \mathcal{O}\Big(\big\{\binom{dd_0-s}{s_1}\binom{s}{s-s_1}d\big\}^{-1}\Big)$,

$$s_1\|\boldsymbol{\beta}_{\mathcal{D}}^*\|_{\min}^2 \leq \frac{64 s_1 \sigma^2 \log d}{nC},$$

for some constant C, which violates our assumption (3.3.4). This implies that $\Delta(\mathcal{K}) < 0$ holds with probability at most $\mathcal{O}\Big(\big\{\binom{dd_0-s}{s_1}\binom{s}{s-s_1}d\big\}^{-1}\Big)$. Taking a union bound, we have

$$\mathbb{P}(\mathcal{K} \neq \mathcal{S}) = \mathcal{O}(d^{-1}),$$

as desired. □

B.3.4 Proof of Theorem 3.3.6

Proof. Recall that the logistic loss of the j-th variable is

$$\mathcal{L}_j(\boldsymbol{\beta}_j) = \frac{1}{n}\sum_{i=1}^{n} \log\big\{1 + \exp(-x_{ij}\mathbf{x}_{i,\mathcal{N}_j}^T\boldsymbol{\beta}_j)\big\},$$

and its gradient at $\boldsymbol{\beta}_j^*$ is

$$\nabla \mathcal{L}_j(\boldsymbol{\beta}_j^*) = -\frac{1}{n}\sum_{i=1}^{n} x_{ij}\mathbf{x}_{i,N_j}\big\{1+\exp(x_{ij}\mathbf{x}_{i,\mathcal{N}_j}^T\boldsymbol{\beta}^*)\big\}^{-1}.$$

Taking expectation, by tower rule, we have

$$\begin{aligned}\mathbb{E}[\nabla \mathcal{L}_j(\boldsymbol{\beta}_j^*)] &= \mathbb{E}\Big[-\mathbb{E}\big[\frac{\boldsymbol{X}_{\mathcal{N}_j}}{1+\exp(\boldsymbol{X}_{\mathcal{N}_j}\boldsymbol{\beta}_j^*)}\mathbb{P}(X_j=1|\boldsymbol{X}_{\mathcal{N}_j}) \\ &\qquad + \frac{\boldsymbol{X}_{\mathcal{N}_j}}{1+\exp(-\boldsymbol{X}_{\mathcal{N}_j}\boldsymbol{\beta}_j^*)}\mathbb{P}(X_j=-1|\boldsymbol{X}_{\mathcal{N}_j})\big|\boldsymbol{X}_{\mathcal{N}_j}\big]\Big] \\ &= \mathbf{0}.\end{aligned}$$

Note that each component of $\nabla \mathcal{L}_j(\boldsymbol{\beta}^*) \in \mathbb{R}^{d_0}$ is bounded in $[-1,1]$. By Hoeffding's inequality, we have that, for any $k \in \mathcal{N}_j$, and any $t>0$, $\nabla_k \mathcal{L}_j(\boldsymbol{\beta}_j^*)$ satisfies

$$\mathbb{P}(|\nabla_k \mathcal{L}_j(\boldsymbol{\beta}_j^*)| \geq t) \leq 2\exp(-nt^2/2).$$

Using Bonferroni's method, we have, with probability at least $1-\mathcal{O}(d^{-1})$, for some constant $C>0$

$$\max_j \|\nabla \mathcal{L}_j(\boldsymbol{\beta}_j^*)\|_\infty \leq C\cdot\sqrt{\frac{\log d}{n}}. \tag{B.3.3}$$

Next, the Hessian of $\mathcal{L}(\boldsymbol{\beta})$ is a block diagonal matrix of $\nabla^2\mathcal{L}_1(\boldsymbol{\beta}_1),...,\nabla^2\mathcal{L}_d(\boldsymbol{\beta}_d)$. By assumptions (A.1) and (A.2), it has been shown by Lemma 5 of Ravikumar et al. (2010) that $\mathcal{L}(\boldsymbol{\beta}^*)$ is ρ-strongly convex with respect to the support $\mathcal{S}\cup\mathcal{K}$ with probability at least $1-\mathcal{O}(d^{-1})$, where $\mathcal{S}$ is the support of $\{\boldsymbol{\beta}_j^*\}_{j=1}^d$, and $\mathcal{K}$ is the support of $\{\widehat{\boldsymbol{\beta}}_j\}_{j=1}^d$.

Let $\mathcal{L}^{(\mathcal{K}')}:\mathbb{R}^{|\mathcal{S}\cup\mathcal{K}|}\to\mathbb{R}$ be the empirical loss function restricted to the support $\mathcal{K}\cup\mathcal{S}$. By the definition of ρ-strongly convexity, we have, with probability at least $1-O(d^{-1})$,

$$\sum_{j=1}^d \mathcal{L}_j^{(\mathcal{K}')}(\widehat{\boldsymbol{\beta}}_j) \geq \sum_{j=1}^d \mathcal{L}_j^{(\mathcal{K}')}(\boldsymbol{\beta}_j^*) + \sum_{j=1}^d \nabla\mathcal{L}^{(\mathcal{K}')}(\boldsymbol{\beta}_j^*)^T(\widehat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j^*) + \frac{\rho}{2}\sum_{j=1}^d\|\widehat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j^*\|_2^2.$$

Next, by Theorem 3.2.1, we have

$$\sum_{j=1}^{d} \mathcal{L}_j(\widehat{\boldsymbol{\beta}}) \leq \sum_{j=1}^{d} \mathcal{L}_j(\boldsymbol{\beta}^*) + C_g,$$

where $C_g = \max_j \big\{\mathcal{L}_j(\mathbf{0}) - \min_{\boldsymbol{\beta}_j} \mathcal{L}_j(\boldsymbol{\beta}_j)\big\}$. Under assumption (A.1), we have that the logistic loss is bounded, which implies that C_g is a bounded constant. Thus, together with (B.3.3), we have, with probability at least $1 - \mathcal{O}(d^{-1})$,

$$\sum_{j=1}^{d} \|\boldsymbol{\beta}_j^* - \widehat{\boldsymbol{\beta}}_j\|_2^2 \leq C\Big(\frac{(s+K)\log d}{n\rho^2} + \frac{C_g}{\rho}\Big),$$

for some constant $C > 0$, as desired. □

B.4 Technical Lemmas

Lemma B.4.1. *Let $\boldsymbol{X} \in \mathbb{R}^d$ be a sub-Gaussian random vector with mean 0 and variance proxy σ^2. Then,*

$$\mathbb{P}(\max_{\boldsymbol{\theta} \in \mathbb{S}^{d-1}} \boldsymbol{\theta}^T \boldsymbol{X} > t) \leq 6^d \exp(-t^2/8\sigma^2).$$

Proof. Let $\mathcal{N}$ be a 1/2-net of $\mathbb{S}^{d-1}$ with respect to the ℓ_2-norm that satisfies $|\mathcal{N}| \leq 6^d$, where such a net exists by Boucheron et al. (2013). For any $\boldsymbol{\theta} \in \mathbb{S}^{d-1}$, there exists $\mathbf{z}_1 \in \mathcal{N}$ and $\|\mathbf{z}_2\|_2 \leq 1/2$ such that $\boldsymbol{\theta} = \mathbf{z}_1 + \mathbf{z}_2$. We have

$$\max_{\boldsymbol{\theta} \in \mathbb{S}^{d-1}} \boldsymbol{\theta}^T \boldsymbol{X} \leq \max_{\mathbf{z}_1 \in \mathcal{N}} \mathbf{z}_1^T \boldsymbol{X} + \max_{\mathbf{z}_2 : \|\mathbf{z}_2\|_2 \leq 1/2} \mathbf{z}_2^T \boldsymbol{X} = \max_{\mathbf{z}_1 \in \mathcal{N}} \mathbf{z}_1^T \boldsymbol{X} + \frac{1}{2} \max_{\mathbf{z}_2 \in \mathbb{S}^{d-1}} \mathbf{z}_2^T \boldsymbol{X}.$$

Therefore, we have

$$\max_{\boldsymbol{\theta} \in \mathbb{S}^{d-1}} \boldsymbol{\theta}^T \boldsymbol{X} \leq 2 \max_{\mathbf{z} \in \mathcal{N}} \mathbf{z}^T \boldsymbol{X}.$$

Consequently, it holds that

$$\mathbb{P}(\max_{\boldsymbol{\theta}\in\mathbb{S}^{d-1}} \boldsymbol{\theta}^T\boldsymbol{X} > t) \leq \mathbb{P}(2\max_{\mathbf{z}\in\mathcal{N}} \mathbf{z}^T\boldsymbol{X} > t) \leq |\mathcal{N}|e^{-t^2/8\sigma^2} \leq 6^d e^{-t^2/8\sigma^2},$$

where the second inequality follows by taking a union bound and the fact that any sub-Gaussian random variable X with variance proxy σ^2, we have $\mathbb{P}(X > t) \leq \exp(-t/8\sigma^2)$. □

Lemma B.4.2. *Given any $\mathbb{Z}_1 \in \mathbb{R}^{n\times d_1}$ and $\mathbb{Z}_2 \in \mathbb{R}^{n\times d_2}$, let $\mathbb{Z} = (\mathbb{Z}_1, \mathbb{Z}_2) \in \mathbb{R}^{n\times(d_1+d_2)}$. We have*

$$\Lambda_{\max}(\mathbb{Z}^T\mathbb{Z}) \geq \Lambda_{\max}(\mathbb{Z}_1^T\mathbf{P}_2^{\perp}\mathbb{Z}_1) \geq \Lambda_{\min}(\mathbb{Z}_1^T\mathbf{P}_2^{\perp}\mathbb{Z}_1) \geq \Lambda_{\min}(\mathbb{Z}^T\mathbb{Z}),$$

where $\mathbf{P}_2 = (\mathbf{I}_n - \mathbb{Z}_2(\mathbb{Z}_2^T\mathbb{Z}_2)^{-1}\mathbb{Z}_2^T)$.

Proof. Observe that

$$\mathbb{Z}^T\mathbb{Z} = \begin{pmatrix} \mathbb{Z}_1^T\mathbb{Z}_1 & \mathbb{Z}_1^T\mathbb{Z}_2 \\ \mathbb{Z}_2^T\mathbb{Z}_1 & \mathbb{Z}_2^T\mathbb{Z}_2 \end{pmatrix},$$

and its the inverse of $\mathbb{Z}^T\mathbb{Z}$ is

$$\begin{aligned} &\begin{pmatrix} \mathbb{Z}_1^T\mathbb{Z}_1 & \mathbb{Z}_1^T\mathbb{Z}_2 \\ \mathbb{Z}_2^T\mathbb{Z}_1 & \mathbb{Z}_2^T\mathbb{Z}_2 \end{pmatrix}^{-1} \\ &= \begin{pmatrix} (\mathbb{Z}_{11} - \mathbb{Z}_{12}\mathbb{Z}_{22}^{-1}\mathbb{Z}_{21})^{-1} & -\mathbb{Z}_{11}^{-1}\mathbb{Z}_{12}(\mathbb{Z}_{22} - \mathbb{Z}_{21}\mathbb{Z}_{11}^{-1}\mathbb{Z}_{12})^{-1} \\ -\mathbb{Z}_{22}^{-1}\mathbb{Z}_{21}(\mathbb{Z}_{11} - \mathbb{Z}_{12}\mathbb{Z}_{22}^{-1}\mathbb{Z}_{21})^{-1} & (\mathbb{Z}_{22} - \mathbb{Z}_{21}\mathbb{Z}_{11}^{-1}\mathbb{Z}_{12})^{-1} \end{pmatrix}, \end{aligned}$$

where $\mathbb{Z}_{jk} = \mathbb{Z}_j^T\mathbb{Z}_k$ for $j, k = 1, 2$.

It is seen that $(\mathbb{Z}_1^T\mathbb{Z}_1 - \mathbb{Z}_1^T\mathbb{Z}_2(\mathbb{Z}_2^T\mathbb{Z}_2)^{-1}\mathbb{Z}_2^T\mathbb{Z}_1)^{-1}$ is a submatrix of $(\mathbb{Z}^T\mathbb{Z})^{-1}$. Our claim follows immediately that the eigenvalues of the matrix $(\mathbb{Z}_1^T\mathbb{Z}_1 - \mathbb{Z}_1^T\mathbb{Z}_2(\mathbb{Z}_2^T\mathbb{Z}_2)^{-1}\mathbb{Z}_2^T\mathbb{Z}_1)^{-1}$ is in the range $[\Lambda_{\min}(\mathbb{Z}^T\mathbb{Z}), \Lambda_{\max}(\mathbb{Z}^T\mathbb{Z})]$. □

Lemma B.4.3. *Suppose each $\boldsymbol{\epsilon}_j = (\epsilon_{j1}, ..., \boldsymbol{\epsilon}_{jd})^T \in \mathbb{R}^d$ where $\epsilon_{ii'}$'s are i.i.d normally distributed with mean 0 and variance σ_j^2 for $j = 1, ..., d$. Let $\sigma^2 = \max_{j=1,...,d}\sigma_j^2$. Using*

notations used in Section B.3.3, let $\mathbf{P}_{\mathcal{K}} = \mathbb{X}_{\mathcal{K}}(\mathbb{X}_{\mathcal{K}}^T\mathbb{X}_{\mathcal{K}})^{-1}\mathbb{X}_{\mathcal{K}}$ *and* $\mathbf{P}_{\mathcal{S}} = \mathbb{X}_{\mathcal{S}}(\mathbb{X}_{\mathcal{S}}^T\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}$*. Let* $\mathcal{K}_j = supp(\widehat{\boldsymbol{\beta}}_j)$ *and* $\mathcal{S}_j = supp(\boldsymbol{\beta}_j^*)$ *for* $j = 1, ..., d$*, and let* k' *be the number of* $\mathcal{K}_j \neq \mathcal{S}_j$*. We have,*

$$\boldsymbol{\epsilon}^T\mathbf{P}_{\mathcal{S}}\boldsymbol{\epsilon} - \boldsymbol{\epsilon}^T\mathbf{P}_{\mathcal{K}}\boldsymbol{\epsilon} \geq -\sigma^2 Z,$$

where Z *is a random variable follows a chi-square distribution with* d_0k' *degrees of freedom.*

Proof. Let $\widehat{\mathcal{D}} = \{j_1', j_2', ..., j_{k'}'\}$ be the indices that $\mathcal{S}_{j_k'} \neq \mathcal{K}_{j_k'}$ for $k = 1, ..., k'$. It is readily seen that

$$\begin{aligned}
&\boldsymbol{\epsilon}^T\mathbf{P}_{\mathcal{S}}\boldsymbol{\epsilon} - \boldsymbol{\epsilon}^T\mathbf{P}_{\mathcal{K}}\boldsymbol{\epsilon}^T \\
=&\sum_{j=1}^{d} \boldsymbol{\epsilon}_j^T\mathbb{X}_{\mathcal{S}_j}(\mathbb{X}_{\mathcal{S}_j}^T\mathbb{X}_{\mathcal{S}_j})^{-1}\mathbb{X}_{\mathcal{S}_j}\boldsymbol{\epsilon}_j - \boldsymbol{\epsilon}_j^T\mathbb{X}_{\mathcal{K}_j}(\mathbb{X}_{\mathcal{K}_j}^T\mathbb{X}_{\mathcal{K}_j})^{-1}\mathbb{X}_{\mathcal{K}_j}\boldsymbol{\epsilon}_j \\
=&\sum_{j\in\{1,...,d\}\setminus\widehat{\mathcal{D}}} \left(\boldsymbol{\epsilon}_j^T\mathbb{X}_{\mathcal{S}_j}(\mathbb{X}_{\mathcal{S}_j}^T\mathbb{X}_{\mathcal{S}_j})^{-1}\mathbb{X}_{\mathcal{S}_j}\boldsymbol{\epsilon}_j - \boldsymbol{\epsilon}_j^T\mathbb{X}_{\mathcal{K}_j}(\mathbb{X}_{\mathcal{K}_j}^T\mathbb{X}_{\mathcal{K}_j})^{-1}\mathbb{X}_{\mathcal{K}_j}\boldsymbol{\epsilon}_j\right) \\
&\quad + \sum_{j\in\widehat{\mathcal{D}}} \left(\boldsymbol{\epsilon}_j^T\mathbb{X}_{\mathcal{S}_j}(\mathbb{X}_{\mathcal{S}_j}^T\mathbb{X}_{\mathcal{S}_j})^{-1}\mathbb{X}_{\mathcal{S}_j}\boldsymbol{\epsilon}_j - \boldsymbol{\epsilon}_j^T\mathbb{X}_{\mathcal{K}_j}(\mathbb{X}_{\mathcal{K}_j}^T\mathbb{X}_{\mathcal{K}_j})^{-1}\mathbb{X}_{\mathcal{K}_j}\boldsymbol{\epsilon}_j\right) \\
\geq& -\sum_{j\in\widehat{\mathcal{D}}}^{d} \boldsymbol{\epsilon}_j^T\mathbb{X}_{\mathcal{K}_j}(\mathbb{X}_{\mathcal{K}_j}^T\mathbb{X}_{\mathcal{K}_j})^{-1}\mathbb{X}_{\mathcal{K}_j}\boldsymbol{\epsilon}_j
\end{aligned}$$

Since $\boldsymbol{\epsilon}_j^T\mathbb{X}_{\mathcal{K}_j}(\mathbb{X}_{\mathcal{K}_j}^T\mathbb{X}_{\mathcal{K}_j})^{-1}\mathbb{X}_{\mathcal{K}_j}\boldsymbol{\epsilon}_j$ is a projection matrix, its eigenvalues are 1's and 0's. In addition, the matrix's rank is $|\mathcal{K}_j| \leq d_0$ as $n > d_0$. We have

$$-\sum_{j\in\widehat{\mathcal{D}}} \boldsymbol{\epsilon}_j^T\mathbb{X}_{\mathcal{K}_j}(\mathbb{X}_{\mathcal{K}_j}^T\mathbb{X}_{\mathcal{K}_j})^{-1}\mathbb{X}_{\mathcal{K}_j}\boldsymbol{\epsilon}_j = -\sum_{j\in\widehat{\mathcal{D}}} \sigma_j^2 Z_j \geq -\sigma^2 Z,$$

where $Z_j \sim \chi^2_{|\mathcal{K}_j|}$ and $Z \sim \chi^2_{k'd_0}$, and the claim holds as desired. □

Lemma B.4.4. *For any integers* $k, d > 0$*, such that* $k \in [1, d]$*, we have*

$$\sum_{j=0}^{k} \binom{d}{j} \leq \left(\frac{ed}{k}\right)^k \tag{B.4.1}$$

Proof. Since the function $f(x) = (x^{-1}ed)^x$ is increasing when $x \geq 1$, if $k \geq d/2$, we have

$$\sum_{j=0}^{k} \binom{d}{j} \leq 2^d \leq (2e)^{d/2} = f(d/2) \leq f(k).$$

When $k < d/2$ and let $Z \sim \mathrm{Bin}(d, 0.5)$. We have

$$\mathbb{P}(Z \leq k) = \sum_{j=0}^{k} \binom{d}{j} 2^{-d}.$$

Thus, we have

$$\sum_{j=0}^{k} \binom{d}{j} = 2^d \mathbb{P}(Z - \mathbb{E}(Z) \leq k - d/2).$$

Taking a Chernoff bound, we have that for any $k' > 0$

$$\begin{aligned} 2^n \mathbb{P}(Z - \mathbb{E}(Z) \leq k - d/2) &\leq \exp\Big\{ d\phi(k') + k'(k - d/2) + d\log 2 \Big\} \\ &= \exp\big\{ kk' + d\log(1 + e^{-k'}) \big\}, \end{aligned} \tag{B.4.2}$$

where we let $U \sim \mathrm{Ber}(0.5)$ and

$$\phi(k') = \log \mathbb{E}\Big(e^{k'(1/2 - U)} \Big) = \frac{k'}{2} + \log(1 + e^{-k'}) - \log 2.$$

Next, we bound the term $kk' + d\log(1 + e^{-k'}$. Taking $k^* = \log(\frac{d+k}{k})$ and $z = k/d < 1/2$, we have

$$k^* k + d\log(1 + e^{-k^*}) = d\Big\{ z\log\big(\frac{1+z}{z}\big) + \log\big(1 + \frac{z}{1+z}\big) \Big\} \leq d\big\{ z - z\log z \big\},$$

where the inequality follows by Lemma B.4.5. Plugging this into (B.4.2), we have

$$\sum_{j=0}^{k} \binom{d}{j} \leq \exp\big\{ k + k\log(d/k) \big\} = \Big(\frac{en}{k} \Big)^k,$$

which finishes the proof. □

Lemma B.4.5. *The function*

$$\psi(z) = z \log\big(\frac{1+z}{z}\big) + \log\big(1 + \frac{z}{1+z}\big)$$

satisfies

$$\psi(z) \leq \phi(z) = z - z\log(z), \text{ for any } z \in (0, 1/2].$$

Proof. It is not difficult to verify that the function $\psi(z) - z\log(z)$ is convex. Thus, we only need to prove that $\psi(z) - z\log(z) \leq z$ for $z = 0$ and $1/2$. By L'Hospital's rule, we have

$$\lim_{z \to 0^+} z\log(1+z) + \log\big(1 + \frac{z}{1+z}\big) = 0.$$

Next, by some computation, we have $\phi(1/2) - \log(1/2)/2 < 1/2$, and our claim follows as desired. □

B.5 Some Definitions in Computational Complexity

In this section, we introduce some basic definitions in the computational complexity theory. More detailed explanation and discussion can be found in iterature such as Arora and Barak (2009).

Definition B.5.1 (Class of P)**.** The complexity class P is defined as all problems that can be solved by a deterministic Turing machine using a polynomial time.

Definition B.5.2 (Class of NP)**.** The complexity class NP (Nondeterministic Polynomial time) is defined as all problems whose solutions can be verified by a deterministic Turing machine using a polynomial time.

Definition B.5.3 (Class of NP-hard)**.** A problem H is NP-hard if for any problem L in NP, there is a polynomial-time reduction from L to H, where a polynomial-time reduction is a method of solving one problem by means of a hypothetical subroutine for solving a different problem, that uses polynomial time excluding the time within the subroutine.

Remark B.5.4. Roughly speaking, a polynomial-time reduction from L to H is that, given a problem L of size d, we can construct a problem H in a polynomial-time $\mathcal{O}(d^k)$ for some $k \in \mathbb{N}$, and solving the problem H will provide a solution to the original problem L. Consequently, finding an algorithm to solve any one of NP-hard problems within a polynomial-time would provide a universal algorithm to solve all the problems in NP.

Definition B.5.5 (Class of NP-Complete)**.** A problem H is NP-complete if it satisfies: (1) H is in the class of NP, and (2) For every problem L in NP, there exists a polynomial-time reduction from L to H.

B.6 Proof of Theorem 3.4.2

Proof. To prove the claim, we construct a two way polynomial-time reduction (i) from a problem instance of knapsack problem to a problem instance of (3.4.4) and (ii) from a problem instance of (3.4.4) to a problem instance of knapsack problem.

(i) We first prove that given a knapsack problem instance (3.4.2) with input $(\mathbf{c}, \mathbf{b}, b_0)$, we can construct a problem instance of the form (3.4.4), and by solving the new problem instance with a ϵ-optimal solution, we can recover the optimal solution to the knapsack problem instance. We consider the general case that we assume all the coefficients b_0, b_j's and c_j's are rational numbers. This assumption is general in the sense that computers can only take input as rational numbers. In addition, our analysis can be generalized to irrational case using numerical analysis techniques, see Trefethen and Bau III (1997) for example. To facilitate our discussion, denote by c the least common multiple of the denominators of b_0, all

b_j's and c_j's. Since there are only finitely many feasible solutions to the knapsack problem, we have that there exists a positive gap δ between the optimal value and other feasible solutions' corresponding objective values. In addition, since we assume all the components are rational numbers, we have $\delta \geq 1/c$. Choosing $\epsilon = 1/(2c)$, by (B.1), within a polynomial time, we can construct $\mathbb{X}_j$, $\mathbb{Y}_j$, such that

$$\mathcal{L}(0; \mathbb{X}_j, \mathbb{Y}_j) = 0, \text{ and } \min_{\beta_j} \mathcal{L}(\beta_j; \mathbb{X}_j, \mathbb{Y}_j) = -c_j - \epsilon/d, \text{ for all } j.$$

Here we let $\mathcal{L}(0; \mathbb{X}_j, \mathbb{Y}_j) = 0$ instead of c_0 in (B.1) for ease of presentation, which does not lose generality. We essentially need to construct $\mathbb{X}_j$'s and $\mathbb{Y}_j$'s such that the difference between $\mathcal{L}(0, \mathbb{X}_j, \mathbb{Y}_j)$ and $\min_{\beta_j} \mathcal{L}(\beta_j; \mathbb{X}_j, \mathbb{Y}_j)$ is $-c_j - \epsilon/d$ for all j, and all $\mathcal{L}(0, \mathbb{X}_j, \mathbb{Y}_j)$ are identical. In addition, we let $\mathcal{R}_j(\beta_j) = b_j \|\beta_j\|_0$. This constructs an instance of problem (3.4.4), and the optimal solution to this instance lies in a compact set that $\beta_j^* \in [-r, r]$ for all j, where $r = \max_j \|\beta_j'\|_\infty$, and $\beta_j' = \operatorname{argmin}_{\beta_j} \mathcal{L}(\beta_j; \mathbb{X}_j, \mathbb{Y}_j)$. Also, since $\mathcal{L}$ is convex, there exists a constant $g > 0$, such that $|\mathcal{L}(\beta_j; \mathbb{X}_j, \mathbb{Y}_j) - \mathcal{L}(\widehat{\beta}_j; \mathbb{X}_j, \mathbb{Y}_j)| \leq g\|\beta_j - \widehat{\beta}_j\|_2$ for any $\beta_j, \widehat{\beta}_j \in [-r, r]$ for all j.

Denote by f_K^* and $-f_L^*$ the optimal objective values of the knapsack problem instance and the new problem instance, where $f_K^*, f_L^* \geq 0$. First, we show that $f_L^* \geq f_K^* + \epsilon$. Denote by $\mathbf{x}^*$ an optimal solution to the instance of multiple-choice knapsack problem. Looking at the solution

$$\widehat{\beta}_j = \begin{cases} \operatorname{argmin}_{\beta_j} \mathcal{L}(\beta_j; \mathbb{X}_j, \mathbb{Y}_j), & \text{if } x_j^* = 1, \\ 0 & \text{otherwise,} \end{cases}$$

we have $\{\widehat{\boldsymbol{\beta}}_j\}_{j \in [d]}$ is a feasible solution to the instance of problem (3.4.4), and the corresponding objective value equals $-f_K^* - \epsilon$. Thus, we conclude $f_L^* \geq f_K^* + \epsilon$. Consequently, letting an ϵ-optimal solution's corresponding value be $-f_L'$, we have $f_L' \geq f_K^*$.

Next, we show that given an ϵ-optimal solution $\{\beta_j^*\}_{j\in[d]}$ to the instance of problem (3.4.4), we have that there exists an optimal solution $\mathbf{x}^*$ to the knapsack problem instance that $x_j^* = 1$ if $\beta_j^* \neq 0$, and $x_j^* = 0$ otherwise. Then, we can recover an optimal solution $\mathbf{x}^*$ to the knapsack problem by setting $x_j^* = 1$ if $\bar{\beta}_j^* \neq 0$ and 0 otherwise.

For any feasible solution $\{\bar{\beta}_j\}_{j\in[d]}$, suppose that there does not exist an optimal solution $\mathbf{x}^*$ to the multiple-choice knapsack problem such that $\mathcal{R}_j(\bar{\beta}_j) > 0$ for $x_j^* = 1$ and $\mathcal{R}_j(\bar{\beta}_j) = 0$ for $x_j^* = 0$. We have $\{\bar{\beta}_j\}_{j\in[d]}$'s corresponding objective value's absolute value is upper bounded by $\sum_{j:\bar{\beta}_j\neq 0} c_j + \epsilon + \sum_{j:\bar{\beta}_j=0} \mathcal{L}(0; \mathbb{X}_j, \mathbb{Y}_j)$, where the term $\sum_{j:\bar{\beta}_j\neq 0} c_j$ is upper bounded by $f_K^* - \delta$, and the term $\sum_{j:\bar{\beta}_j=0} \mathcal{L}(0; \mathbb{X}_j, \mathbb{Y}_j)$ is 0 by construction. By our choices of ϵ, the objective is no greater than f_K^*. This proves $\{\bar{\beta}_j\}_{j\in[d]}$ cannot be an ϵ-optimal solution to the problem (3.4.4). Meanwhile, for any ϵ-optimal solution $\{\beta_j^*\}_{j\in[d]}$, we have its corresponding objective value's absolute value is lower-bounded by $\sum_{j:\beta_j^*\neq 0} c_j + \epsilon + \sum_{j:\beta_j=0} \mathcal{L}(0; \mathbb{X}_j, \mathbb{Y}_j) \geq f_K^*$. Since the term $\sum_{j:\beta_j=0} \mathcal{L}(0; \mathbb{X}_j, \mathbb{Y}_j)$ is 0, we have the term $\sum_{j:\beta_j^*\neq 0} c_j + \epsilon$ is strictly larger than $f_K^* - \delta$, which implies that $\sum_{j:\beta_j^*\neq 0} c_j$ is strictly larger than $f_K^* - \delta$ since $\epsilon < \delta$ and $\sum_{j:\beta_j^*\neq 0} c_j$ can by difference at least δ. Meanwhile, we have $\sum_{j:\beta_j^*\neq 0} b_j \leq b_0$. This proves the feasibility of $\mathbf{x}^*$. Thus, we have $\mathbf{x}^*$ is feasible, and its corresponding objective value is strictly greater than $f_K^* - \delta$, which implies its corresponding objective can only be the optimal value f_K^*. We have that an ϵ-optimal solution $\{\beta^*\}_{j\in[d]}$ to the instance of problem (3.4.4) recovers an optimal solution to the knapsack problem by setting $x_j^* = 1$ if $\beta_j^* \neq 0$, and $x_j^* = 0$ otherwise.

(ii) For the other direction, we first introduce the multiple-choice knapsack problem for ease of presentation. Denote by $\mathbf{C} = (\mathbf{c}_1, ..., \mathbf{c}_d)^T \in \mathbb{R}^{d\times d_0}$ and $\mathbf{B} = (\mathbf{b}_1, ..., \mathbf{b}_d)^T \in \mathbb{R}^{d\times d_0}$, where $\mathbf{c}_j = (c_{j1}, ..., c_{jd_0})^T, \mathbf{b}_j = (b_{j1}, ..., b_{jd_0})^T$. Consider the multiple-choice knapsack problem with input $(\mathbf{C}, \mathbf{B}, b_0)$, where all the coefficients are positive rational numbers:

$$\max_{x_{jk}} \sum_{j=1}^{d}\sum_{k=1}^{d_0} c_{jk}x_{jk}, \text{ subject to } \sum_{j=1}^{d}\sum_{k=1}^{d_0} b_{jk}x_{jk} \leq b_0, \sum_{k=1}^{d_0} x_{jk} \leq 1, \ x_{jk} \in \{0, 1\},$$

for all $j \in [d]$ and all $k \in [d_0]$. It is well known that there exists a two-way polynomial-time reduction between the 0-1 knapsack problem (3.4.2) and the multiple-choice knapsack problem (Williamson and Shmoys, 2011). Thus, given a problem of the form (3.4.2), we only need to find a polynomial time reduction to a problem instance of the multiple-choice knapsack problem.

Without loss of generality, we assume all c_{jk}, b_{jk}'s and b_0 are positive rational numbers. Given an instance of problem (3.4.4) with input $(\{\mathbb{X}_j, \mathbb{Y}_j\}_{j\in[d]}, b_0)$, the solution belongs to a bounded region that $\boldsymbol{\beta}_j^* \in [-r, r]^{d_0}$ for all j. Since $\mathcal{L}$ is convex, we have that there exists a constant $g > 0$, such that $|\mathcal{L}(\boldsymbol{\beta}_j; \mathbb{X}_j, \mathbb{Y}_j) - \mathcal{L}(\widehat{\boldsymbol{\beta}}_j; \mathbb{X}_j, \mathbb{Y}_j)| \leq g\|\boldsymbol{\beta}_j - \widehat{\boldsymbol{\beta}}_j\|_2$ for any $\boldsymbol{\beta}_j, \widehat{\boldsymbol{\beta}}_j \in [-r, r]^{d_0}$ for all j. For each j, we first discretize $[-r, r]^{d_0}$ into a set of d_1 points $\{\mathbf{p}_1^{(j)}, ..., \mathbf{p}_{d_1}^{(j)}\}$ denoted by $\mathcal{B}_j$, where we discretize the set sufficiently finely that for any $\boldsymbol{\beta}_j \in \mathcal{B}_j$, there exists a $\mathbf{p}^{(j)} \in \mathcal{B}_j$ such that $\|\mathbf{p}^{(j)} - \boldsymbol{\beta}_j\|_\infty < \delta'$ and $\mathcal{R}(\mathbf{p}^{(j)}) \leq \mathcal{R}(\boldsymbol{\beta}_j)$, and $\delta' < (dg)^{-1}\epsilon$. By assumption (B.2), this can be done within a polynomial time. We further compute corresponding objective and constraint values for all points $\mathbf{p}_k^{(j)} \in \mathcal{B}_j$. Specifically, for any $\mathbf{p}_k^{(j)} \in \mathcal{B}_j$, we compute $b'_{jk} = \mathcal{R}_j(\mathbf{p}_k^{(j)})$ and $c'_{jk} = -\mathcal{L}(\mathbf{p}_k^{(j)}; \mathbb{X}_j, \mathbb{Y}_j)$ for $j = 1, ..., d$ and $k = 1, ..., d_1$. Also, letting $b'_0 = b_0$, we have an instance of multiple-choice knapsack problem with input $(\mathbf{C}', \mathbf{B}', b'_0)$. It is not difficult to see that an optimal solution to the instance of multiple-choice knapsack problem $\mathbf{x}^*$ gives a feasible solution of the original problem (3.4.4) by assigning $\widehat{\boldsymbol{\beta}}_j = \mathbf{p}_k^{(j)}$ if $x_{jk}^* = 1$. Next, consider an optimal solution $\{\boldsymbol{\beta}_j^*\}_{j=1}^d$ to the instance of problem (3.4.4), we have that by our discretization, there exists some feasible point $\{\widehat{\mathbf{p}}^{(j)}\}_{j\in[d]}$ belongs to the discretized set, and $\|\widehat{\mathbf{p}}^{(j)} - \boldsymbol{\beta}_j^*\|_2 < \delta'$ for all j. Consequently, we have $\{\widehat{\mathbf{p}}^{(j)}\}$'s corresponding objective's absolute value is lower bounded by $f_L^* - dg\delta' \geq f_L^* - \epsilon$, where $-f_L^*$ denotes the objective value of the instance of problem (3.4.4). Thus, we have that $\{\widehat{\mathbf{p}}^{(j)}\}_{j\in[d]}$ is an ϵ-optimal solution to the problem (3.4.4). Consequently, the instance of multiple-choice knapsack problem's optimal value is lower-bounded by $f_L^* - \epsilon$.

Thus, solving the instance of multiple-choice knapsack problem gives an ϵ-optimal solution to the instance of problem (3.4.4), which finishes the proof. □

B.7 Computational Complexity

We briefly discuss the computational complexities of the dynamic programming approach and the SPICA algorithm in this section. We denote by d_m the average degree per-vertex, so we have

$$d^{-1}K = \Theta(d_m).$$

We are interested in the computational complexity for solving the problem (3.1.1). The dual method proposed in this work involves iteratively calling a solver to solve the subproblems

$$\max_{\beta_j} \mathcal{L}_j(\beta_j) + \lambda\|\beta_j\|_0,$$

for every $j = 1, \ldots, d$. Each subproblem is a combinatorial problem with dimension d_0. We assume that each call to this subproblem solver incurs an identical time complexity, which is a function of d_0, and we denote by $\mathcal{T}(d_0)$.

In the SPICA algorithm, we use the golden section method for maximizing the dual function. To achieve an $\epsilon-$optimal numerical solution, the golden section method requires $\mathcal{O}\left(\log \frac{1}{\epsilon}\right)$ iterations. Each iteration involves calling the subproblem solver for d times. The total computational complexity is

$$\mathcal{O}\left(d \cdot \log \frac{1}{\epsilon} \cdot \mathcal{T}(d_0)\right).$$

Suppose the optimal statistical error is of the order $\mathcal{O}(\frac{K \log d}{dn})$. Consider the regime where the duality gap is no larger than the statistical error, i.e., $\frac{1}{d} = \mathcal{O}(\frac{K \log d}{dn})$, we need $n = \mathcal{O}(K \log d) = \mathcal{O}(d_m d \log d)$. We also require that the optimization error ϵ to be bounded

by the statistical error, i.e., $\epsilon = \Theta(\frac{K \log d}{dn})$, which means that the computational complexity becomes

$$\begin{aligned} d \cdot \log \frac{1}{\epsilon} \cdot \mathcal{T}(d_0) &= d \cdot \log \frac{dn}{K \log d} \cdot \mathcal{T}(d_0) \\ &= d \cdot (\log n - \log d_m - \log \log d) \cdot \mathcal{T}(d_m) \\ &= d \cdot (\log(d_m d \log d) - \log d_m - \log \log d) \cdot \mathcal{T}(d_m) \\ &= \mathcal{O}\Big(d \cdot \log d \cdot \mathcal{T}(d_m)\Big). \end{aligned}$$

This is the time complexity to obtain an overall error rate $\mathcal{O}\left(\frac{K \log d}{dn}\right)$.

For comparison, let us consider the scheme in which we enumerate all possible subproblem solutions and solves the problem using dynamic programming as discussed in Section 3.4. This approach provides an exact solution, and its computational complexity is dd_0K. However, enumerating all possible subproblem solutions is costly. For each node, there are d_0 subproblems. The total computational complexity is

$$\mathcal{O}\Big(dd_0 \cdot \mathcal{T}(d_0) + dd_0K\Big).$$

Now we compare the time complexity for these two methods if they both achieve the optimal statistical rates. Suppose that the sample size n is fixed. We consider the relation between the graph's property and the algorithms' computational complexity. Clearly, the SPICA algorithm is more time efficient for graphs with more potential local neighbors such that $d_0 \gg \log(d)$, and the dynamic programming approach is more efficient if $d_0 \ll \log(d)$. Meanwhile, suppose that d_0 is fixed. The SPICA algorithm is faster if $\log n - \log d_m - \log \log d \ll d_m$, i.e.,

$$n \leq \mathcal{O}\Big(d_m \log d \cdot \min\{e^{d_m}, d\}\Big),$$

and the dynamic programming approach is faster otherwise. To summarize, there always exists an algorithm that achieves an statistical error

$$\mathcal{O}\Big(\frac{K\log d}{dn}\Big) = \mathcal{O}\Big(\frac{d_m \log d}{n}\Big)$$

within computational complexity

$$\mathcal{O}\Big(d \cdot \min\{\log d, d_m\} \cdot \mathcal{T}(d_0)\Big).$$

Appendix C

Appendix to Chapter 4

C.1 Proof of Theorem 4.3.4

We first provide a key lemma which characterizes the asymptotic normality of $\nabla\mathcal{L}(\boldsymbol{\beta}^*)$. This lemma is essential in our later proofs to derive the asymptotic distributions of the test statistics.

Lemma C.1.1. *Under Assumptions 4.1.1, 4.3.2 and 4.3.3, for any vector* $\mathbf{v} \in \mathbb{R}^d$*, if* $\|\mathbf{v}\|_0 \leq s'$ *and* $n^{-1/2}\sqrt{s'^3 \log d} = o(1)$*, it holds that*

$$\frac{\sqrt{n}\mathbf{v}^T\nabla\mathcal{L}(\boldsymbol{\beta}^*)}{\sqrt{\mathbf{v}^T\mathbf{H}^*\mathbf{v}}} \xrightarrow{d} N(0,1), \quad \textit{where } \mathbf{H}^* \textit{ is defined in } (4.1.7).$$

Proof. Let $M_i(t) = N_i(t) - \int_0^t Y_i(u)\lambda_0(u)du$. By the definition of $\nabla\mathcal{L}(\boldsymbol{\beta}^*)$ in (4.1.4), we have

$$\begin{aligned}\nabla\mathcal{L}(\boldsymbol{\beta}^*) &= -\frac{1}{n}\sum_{i=1}^n \int_0^\tau \{\boldsymbol{X}_i(u) - \bar{\boldsymbol{Z}}(u,\boldsymbol{\beta}^*)\}dM_i(u) \\ &= -\frac{1}{n}\sum_{i=1}^n \int_0^\tau \{\boldsymbol{X}_i(u) - \mathbf{e}(u,\boldsymbol{\beta}^*)\}dM_i(u) - \frac{1}{n}\sum_{i=1}^n \int_0^\tau \{\mathbf{e}(u,\boldsymbol{\beta}^*) - \bar{\boldsymbol{Z}}(u,\boldsymbol{\beta}^*)\}dM_i(u),\end{aligned}$$

Thus, by the identity $\mathbf{H}^* = \sqrt{n}\,\mathrm{Var}\{\nabla\mathcal{L}(\boldsymbol{\beta}^*)\}$, we have

$$\frac{\sqrt{n}\mathbf{v}^T\nabla\mathcal{L}(\boldsymbol{\beta}^*)}{\sqrt{\mathbf{v}^T\mathbf{H}^*\mathbf{v}}} = -\frac{1}{\sqrt{n}}\underbrace{\frac{\mathbf{v}^T}{\sqrt{\mathbf{v}^T\mathbf{H}^*\mathbf{v}}}\sum_{i=1}^{n}\int_0^\tau\{\boldsymbol{X}_i(u)-\mathbf{e}(u,\boldsymbol{\beta}^*)\}dM_i(u)}_{S}$$
$$-\underbrace{\frac{1}{\sqrt{n}}\frac{\mathbf{v}^T}{\sqrt{\mathbf{v}^T\mathbf{H}^*\mathbf{v}}}\sum_{i=1}^{n}\int_0^\tau\{\mathbf{e}(u,\boldsymbol{\beta}^*)-\bar{\boldsymbol{Z}}(u,\boldsymbol{\beta}^*)\}dM_i(u)}_{E}.$$

For the first term S, denote by

$$\xi_i = \frac{\mathbf{v}^T}{\sqrt{\mathbf{v}^T\mathbf{H}^*\mathbf{v}}}\int_0^\tau\big\{\boldsymbol{X}_i(u)-\mathbf{e}(u,\boldsymbol{\beta}^*)\big\}dM_i(u).$$

We have $\mathbb{E}(\xi_i)=0$, and $\mathrm{Var}(n^{-1/2}S)=1$. Thus S is a sum of n independent random variables with mean 0. To get the asymptotic distribution of $n^{-1/2}S$, we verify the Lyapunov condition. Indeed, we have

$$\frac{1}{n^{3/2}}\sum_{i=1}^{n}\mathbb{E}\Big|\frac{\mathbf{v}^T}{\sqrt{\mathbf{v}^T\mathbf{H}^*\mathbf{v}}}\int_0^\tau\{\boldsymbol{X}_i(u)-\mathbf{e}(u,\boldsymbol{\beta}^*)\}dM_i(u)\Big|^3$$
$$\le \frac{C}{C_h^{3/2}n^{3/2}}\sum_{i=1}^{n}s'^{3/2}\sup_{u\in[0,\tau]}\|\boldsymbol{X}_i(u)-\mathbf{e}(u,\boldsymbol{\beta}^*)\|_\infty^3 = \mathcal{O}(s'^{3/2}n^{-1/2}),$$

where the inequality follows by Assumption 4.3.3 for some constant C, and the equality holds by Lemma C.6.1 and Assumption 4.1.1. Thus, the Lyapunov condition holds by our scaling assumption that $s'^{3/2}n^{-1/2}=o(1)$. Applying Lindeberg Feller Central Limit Theorem, we have $n^{-1/2}S\xrightarrow{d}N(0,1)$.

Next, we prove that the second term $E=o_{\mathbb{P}}(1)$. Since

$$E = \frac{1}{\sqrt{n}}\frac{\mathbf{v}^T}{\sqrt{\mathbf{v}^T\mathbf{H}^*\mathbf{v}}}\sum_{i=1}^{n}\int_0^\tau\Big[\{\mathbf{e}(u,\boldsymbol{\beta}^*)-\bar{\boldsymbol{Z}}(u,\boldsymbol{\beta}^*)\}1dM_i(u)\Big]$$
$$\le \frac{1}{\sqrt{n}}\frac{s'^{1/2}}{\lambda_{\min}}\sup_{u\in[0,\tau]}\|\mathbf{e}(u,\boldsymbol{\beta}^*)-\bar{\boldsymbol{Z}}(u,\boldsymbol{\beta}^*)\|_\infty\int_0^\tau\Big|\sum_{i=1}^{n}1dM_i(u)\Big|.$$

By Lemma C.6.1, it holds that $\sup_{u\in[0,\tau]} \|\mathbf{e}(u,\boldsymbol{\beta}^*) - \bar{\boldsymbol{Z}}(u,\boldsymbol{\beta}^*)\|_\infty = \mathcal{O}_{\mathbb{P}}(\sqrt{n^{-1}\log d})$. It holds that, for some constant $C > 0$,

$$E \leq \frac{C}{\sqrt{n}} \frac{1}{\lambda_{\min}} \sqrt{\frac{s' \log d}{n}} \int_0^\tau \Big| \sum_{i=1}^n 1 dM_i(u) \Big|.$$

It remains to bound the term $\int_0^\tau \Big| \sum_{i=1}^n 1 dM_i(u) \Big|$. By Theorem 2.11.9 and Example of 2.11.16 of van der Vaart and Wellner (1996), $\bar{G}(t) := n^{-1/2} \sum_{i=1}^n M_i(t)$ converges weakly to a tight Gaussian process $G(t)$. Furthermore, by Strong Embedding Theorem of Shorack and Wellner (2009), there exists another probability space such that $(S^{*(0)}(\boldsymbol{\beta},t), S^{*(1)}(\boldsymbol{\beta},t), \bar{G}^*(t))$ converges almost surely to $(s^{*(0)}(\boldsymbol{\beta},t), \mathbf{s}^{*(1)}(\boldsymbol{\beta},t), G^*(t))$, where $*$ indicates the existences in a new probability space. This implies that $\int_0^\tau |dG(t)| = \int_0^\tau |dG^*(t)| + o_{\mathbb{P}}(1)$. We have, by our assumption $n^{-1}\sqrt{s' \log d} = o_{\mathbb{P}}(1)$, the term E satisfies that

$$E = \mathcal{O}_{\mathbb{P}}\Big(\sqrt{\frac{s' \log d}{n}} \cdot \frac{1}{\sqrt{n}} \Big) = o_{\mathbb{P}}(1).$$

Combining this with the result that $n^{-1/2} S \xrightarrow{d} N(0,1)$ concludes the proof. □

Before proving Theorem 4.3.4, we need some additional technical lemmas which characterize several concentration results. We present them here and defer detailed proofs to Section C.7.

Lemma C.1.2. *Under Assumptions 4.1.1, 4.1.2, 4.3.1, 4.3.2 and 4.3.3, there exists a positive constant C, such that with probability at least $1 - \mathcal{O}(d^{-3})$,*

$$\|\nabla \mathcal{L}(\boldsymbol{\beta}^*)\|_\infty \leq C \sqrt{\frac{\log d}{n}}.$$

Lemma C.1.3. *Under Assumptions 4.1.1, 4.1.2, 4.3.1, 4.3.2 and 4.3.3, let $\widehat{\boldsymbol{\beta}}$ be the estimator for $\boldsymbol{\beta}^*$ estimated by* (4.1.1) *satisfying the result in* (4.1.2) *that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = \mathcal{O}_{\mathbb{P}}(s\lambda)$*

with $\lambda \asymp \mathcal{O}(\sqrt{n^{-1}\log d})$. *Then, we have, for any* $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta}^* + u(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$ *with* $u \in [0,1]$,

$$\|\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}})\|_\infty = \mathcal{O}_{\mathbb{P}}(1), \text{ and } \|\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}) - \mathbf{H}^*\|_\infty = \mathcal{O}_{\mathbb{P}}\Big(s\sqrt{\frac{\log d}{n}}\Big).$$

Lemma C.1.4. *Under Assumptions 4.1.1, 4.1.2, 4.3.1, 4.3.2 and 4.3.3, If* $\lambda' \asymp s'\sqrt{n^{-1}\log d}$, *we have*

$$\|\widehat{\mathbf{w}} - \mathbf{w}^*\|_1 = \mathcal{O}_{\mathbb{P}}(s's\sqrt{n^{-1}\log d}), \tag{C.1.1}$$

where $\mathbf{w}^*$ *and* $\widehat{\mathbf{w}}$ *are defined in* (4.2.1) *and* (4.2.2).

Now, we are ready to prove Theorem 4.3.4.

Proof of Theorem 4.3.4. To derive the asymptotic distribution of $\sqrt{n}\widehat{U}(0,\widehat{\boldsymbol{\theta}})$, we start with decomposing $\widehat{U}(0,\widehat{\boldsymbol{\theta}})$ into several terms.

$$\begin{aligned}
&\widehat{U}(0,\widehat{\boldsymbol{\theta}}) = \nabla_\alpha \mathcal{L}(0,\widehat{\boldsymbol{\theta}}) - \widehat{\mathbf{w}}^T \nabla_{\boldsymbol{\theta}} \mathcal{L}(0,\widehat{\boldsymbol{\theta}}) \\
=&\nabla_\alpha \mathcal{L}(0,\boldsymbol{\theta}^*) + \nabla^2_{\alpha\boldsymbol{\theta}} \mathcal{L}(0,\bar{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \big\{\widehat{\mathbf{w}}^T \nabla_{\boldsymbol{\theta}} \mathcal{L}(0,\boldsymbol{\theta}^*) + \widehat{\mathbf{w}}^T \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(0,\widetilde{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\big\} \qquad \text{(C.1.2)} \\
=&\underbrace{\nabla_\alpha \mathcal{L}(0,\boldsymbol{\theta}^*) - \mathbf{w}^{*T} \nabla_{\boldsymbol{\theta}} \mathcal{L}(0,\boldsymbol{\theta}^*)}_{S} + \underbrace{(\mathbf{w}^* - \widehat{\mathbf{w}})^T \nabla_{\boldsymbol{\theta}} \mathcal{L}(0,\boldsymbol{\theta}^*)}_{E_1} + \underbrace{\big\{\nabla^2_{\alpha\boldsymbol{\theta}} \mathcal{L}(0,\bar{\boldsymbol{\theta}}) - \widehat{\mathbf{w}}^T \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(0,\widetilde{\boldsymbol{\theta}})\big\}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)}_{E_2},
\end{aligned}$$

where the second equality holds by the mean value theorem for some $\bar{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + u(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$, $\widetilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + u'(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ and $u, u' \in [0,1]$.

We consider the terms S, E_1 and E_2 separately. For the first term S, by Lemma C.1.1, taking $\mathbf{v} = (1, -\mathbf{w}^{*T})^T$. we have

$$\sqrt{n}S \xrightarrow{d} Z, \text{ where } Z \sim N(0, H_{\alpha|\boldsymbol{\theta}}). \tag{C.1.3}$$

For the term E_1, we have,

$$E_1 \le \|\widehat{\mathbf{w}} - \mathbf{w}^*\|_1 \|\nabla_{\boldsymbol{\theta}} \mathcal{L}(0,\boldsymbol{\theta}^*)\|_\infty = \mathcal{O}_{\mathbb{P}}(s'\lambda'\sqrt{n^{-1}\log d}), \tag{C.1.4}$$

where $\|\widehat{\mathbf{w}} - \mathbf{w}^*\|_1 = \mathcal{O}_{\mathbb{P}}(s'\lambda')$ by Lemma C.1.4, and $\|\nabla_{\boldsymbol{\theta}}\mathcal{L}(0, \boldsymbol{\theta}^*)\|_\infty = \mathcal{O}_{\mathbb{P}}\big(\sqrt{n^{-1}\log d}\big)$ by Lemma C.1.2.

For the term E_2, we have,

$$E_2 = \underbrace{\big\{\nabla^2_{\alpha\boldsymbol{\theta}}\mathcal{L}(0, \bar{\boldsymbol{\theta}}) - \mathbf{H}^*_{\alpha\boldsymbol{\theta}}\mathbf{H}^{*-1}_{\boldsymbol{\theta\theta}}\nabla^2_{\boldsymbol{\theta\theta}}\mathcal{L}(0, \widetilde{\boldsymbol{\theta}})\big\}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)}_{E_{21}} + \underbrace{(\mathbf{w}^* - \widehat{\mathbf{w}})^T\nabla^2_{\boldsymbol{\theta\theta}}\mathcal{L}(0, \widetilde{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)}_{E_{22}}. \tag{C.1.5}$$

Considering the terms E_{21} and E_{22} separately, first, we have,

$$\begin{aligned} E_{21} =& \big\{\nabla^2_{\alpha\boldsymbol{\theta}}\mathcal{L}(0, \bar{\boldsymbol{\theta}}) - \mathbf{H}^*_{\alpha\boldsymbol{\theta}} + \mathbf{H}^*_{\alpha\boldsymbol{\theta}} - \mathbf{H}^*_{\alpha\boldsymbol{\theta}}\mathbf{H}^{*-1}_{\boldsymbol{\theta\theta}}\nabla^2_{\boldsymbol{\theta\theta}}\mathcal{L}(0, \widetilde{\boldsymbol{\theta}})\big\}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \\ \leq& \|\nabla^2_{\alpha\boldsymbol{\theta}}\mathcal{L}(0, \bar{\boldsymbol{\theta}}) - \mathbf{H}^*_{\alpha\boldsymbol{\theta}}\|_\infty\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 + |\mathbf{H}^*_{\alpha\boldsymbol{\theta}}(\mathbf{I}_{d-1} - \mathbf{H}^{*-1}_{\boldsymbol{\theta\theta}}\nabla^2_{\boldsymbol{\theta\theta}}\mathcal{L}(0, \widetilde{\boldsymbol{\theta}}))(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)|, \end{aligned} \tag{C.1.6}$$

where the inequality holds by Hölder's inequality. For the first term in the above inequality, we have

$$\|\nabla^2_{\alpha\boldsymbol{\theta}}\mathcal{L}(0, \bar{\boldsymbol{\theta}}) - \mathbf{H}^*_{\alpha\boldsymbol{\theta}}\|_\infty\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 = \mathcal{O}_{\mathbb{P}}(s^2\lambda^2), \tag{C.1.7}$$

since $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 = \mathcal{O}_{\mathbb{P}}(s\lambda)$ by (4.1.2), and $\|\nabla_{\alpha\boldsymbol{\theta}}\mathcal{L}(0, \bar{\boldsymbol{\theta}}) - \mathbf{H}^*_{\alpha\boldsymbol{\theta}}\|_\infty = \mathcal{O}_{\mathbb{P}}(s\lambda)$ by Lemma C.1.3.

For the second term in (C.1.6), by Hölder's inequality, we have

$$\begin{aligned} &|\mathbf{H}^*_{\alpha\boldsymbol{\theta}}(\mathbf{I}_{d-1} - \mathbf{H}^{*-1}_{\boldsymbol{\theta\theta}}\nabla^2_{\boldsymbol{\theta\theta}}\mathcal{L}(0, \widetilde{\boldsymbol{\theta}}))(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)| = |\mathbf{H}^*_{\alpha\boldsymbol{\theta}}\mathbf{H}^{*-1}_{\boldsymbol{\theta\theta}}\big(\mathbf{H}^*_{\boldsymbol{\theta\theta}} - \nabla^2_{\boldsymbol{\theta\theta}}\mathcal{L}(0, \widetilde{\boldsymbol{\theta}})\big)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)| \\ \leq& \|\mathbf{w}^*\|_1\|\mathbf{H}^*_{\boldsymbol{\theta\theta}} - \nabla^2_{\boldsymbol{\theta\theta}}\mathcal{L}(0, \widetilde{\boldsymbol{\theta}})\|_\infty\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 = \mathcal{O}_{\mathbb{P}}(s's^2\lambda^2), \end{aligned} \tag{C.1.8}$$

where the last equality holds since $\|\mathbf{w}^*\|_1 = \mathcal{O}(s')$ by Assumption 4.3.2, $\|\mathbf{H}^*_{\boldsymbol{\theta\theta}} - \nabla^2_{\boldsymbol{\theta\theta}}\mathcal{L}(0, \widetilde{\boldsymbol{\theta}})\|_\infty = \mathcal{O}_{\mathbb{P}}(s\lambda)$ by Lemma C.1.3, and $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 = \mathcal{O}_{\mathbb{P}}(s\lambda)$ by (4.1.2). Plugging (C.1.7) and (C.1.8) into (C.1.6), we have

$$|E_{21}| = \mathcal{O}_{\mathbb{P}}(s's^2\lambda^2). \tag{C.1.9}$$

For the second term E_{22} in (C.1.5), we have,

$$|E_{22}| \leq \|\widehat{\mathbf{w}} - \mathbf{w}^*\|_1 \|\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(0, \widetilde{\boldsymbol{\theta}})\|_\infty \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 = \mathcal{O}_{\mathbb{P}}(s' s \lambda' \lambda), \tag{C.1.10}$$

where we use the results that $\|\widehat{\mathbf{w}} - \mathbf{w}^*\|_1 = \mathcal{O}_{\mathbb{P}}(s'\lambda')$ by Lemma C.1.4, $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 \leq \mathcal{O}_{\mathbb{P}}(s\lambda)$ by (4.1.2), and $\|\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(0, \widetilde{\boldsymbol{\theta}})\|_\infty = \mathcal{O}_{\mathbb{P}}(1)$ by Lemma C.1.3.

Plugging (C.1.6) and (C.1.10) into (C.1.5), we have $E_2 = \mathcal{O}_{\mathbb{P}}(n^{-1} s' s^2 \log d)$. Combining it with (C.1.4), we have

$$|E_1| + |E_2| = \mathcal{O}_{\mathbb{P}}\Big(\frac{s' s^2 \log d}{n}\Big) = o_{\mathbb{P}}\Big(\frac{1}{\sqrt{n}}\Big), \tag{C.1.11}$$

where the last equality holds by the assumption that $n^{-1/2} s^3 \log d = o(1)$ and $s \asymp s'$. Combining (C.1.11), (C.1.3) and (C.1.2), our claim (4.3.1) holds as desired. □

C.2 Proof of Theorem 4.3.9

Proof of Theorem 4.3.9. We have

$$\begin{aligned}
&\mathcal{L}(\widetilde{\alpha}, \widehat{\boldsymbol{\theta}} - \widetilde{\alpha}\widehat{\mathbf{w}}) - \mathcal{L}(0, \widehat{\boldsymbol{\theta}}) \\
=&\widetilde{\alpha} \nabla_\alpha \mathcal{L}(0, \widehat{\boldsymbol{\theta}}) - \widetilde{\alpha} \widehat{\mathbf{w}}^T \nabla_{\boldsymbol{\theta}} \mathcal{L}(0, \widehat{\boldsymbol{\theta}}) + \frac{\widetilde{\alpha}^2}{2} \nabla^2_{\alpha\alpha} \mathcal{L}(\bar{\alpha}, \widehat{\boldsymbol{\theta}}) + \frac{\widetilde{\alpha}^2}{2} \widehat{\mathbf{w}}^T \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(0, \bar{\boldsymbol{\theta}}) \widehat{\mathbf{w}} - \widetilde{\alpha}^2 \widehat{\mathbf{w}}^T \nabla_{\boldsymbol{\theta}} \mathcal{L}(\bar{\alpha}', \widehat{\boldsymbol{\theta}}) \\
=&\underbrace{\widetilde{\alpha} \widehat{U}(0, \widehat{\boldsymbol{\theta}})}_{T_1} + \underbrace{\frac{\widetilde{\alpha}^2}{2} \Big\{ \nabla^2_{\alpha\alpha} \mathcal{L}(\bar{\alpha}, \widehat{\boldsymbol{\theta}}) + \widehat{\mathbf{w}}^T \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(0, \bar{\boldsymbol{\theta}}) \widehat{\mathbf{w}} - 2 \widetilde{\mathbf{w}}^T \nabla^2_{\boldsymbol{\theta}\alpha} \mathcal{L}(\bar{\alpha}', \bar{\boldsymbol{\theta}}') \Big\}}_{T_2},
\end{aligned} \tag{C.2.1}$$

where the first equality follows by the mean value theorem with $\bar{\alpha} = u_1 \widehat{\alpha}$, $\bar{\alpha}' = u_2 \widehat{\alpha}$, $\bar{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + u_3(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$, and $\bar{\boldsymbol{\theta}}' = \boldsymbol{\theta}^* + u_4(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ for some $0 \leq u_1, u_2, u_3, u_4 \leq 1$.

We first look at the term T_1. Under the null hypothesis $\alpha^* = 0$, $\sqrt{n}\widehat{U}(0, \widehat{\boldsymbol{\theta}}) \xrightarrow{d} Z$ and $\sqrt{n}\widetilde{\alpha} = -H_{\alpha|\boldsymbol{\theta}}^{-1}\widehat{U}(0, \widehat{\boldsymbol{\theta}}) + o_{\mathbb{P}}(1)$ by Theorems 4.3.4 and 4.3.7, where $Z \sim N(0, H_{\alpha|\boldsymbol{\theta}})$. We have,

$$2nT_1 = -2\widehat{U}^2(0, \widehat{\boldsymbol{\theta}}) + o_{\mathbb{P}}(1) \xrightarrow{d} -2Z^2 H_{\alpha|\boldsymbol{\theta}}^{-1}. \tag{C.2.2}$$

Next, we look at the term T_2,

$$\begin{aligned} T_2 = &\underbrace{\frac{\widetilde{\alpha}^2}{2}\big(\mathbf{H}^*_{\alpha\alpha} + \mathbf{H}_{\alpha\boldsymbol{\theta}}\mathbf{H}^{*-1}_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathbf{H}^*_{\boldsymbol{\theta}\alpha} - 2\mathbf{H}^*_{\alpha\boldsymbol{\theta}}\mathbf{H}^{*-1}_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathbf{H}^*_{\boldsymbol{\theta}\alpha}\big)}_{T_{21}} \\ &+ \underbrace{\frac{\widetilde{\alpha}^2}{2}\Big[\big\{\nabla^2_{\alpha\alpha}\mathcal{L}(\bar{\alpha}, \widehat{\boldsymbol{\theta}}) - \mathbf{H}^*_{\alpha\alpha}\big\} + \big\{\widehat{\mathbf{w}}^T\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(0, \bar{\boldsymbol{\theta}})\widehat{\mathbf{w}} - \mathbf{w}^*\mathbf{H}^*_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathbf{w}^*\big\} - 2\big\{\widetilde{\mathbf{w}}^T\nabla^2_{\boldsymbol{\theta}\alpha}\mathcal{L}(\bar{\alpha}', \bar{\boldsymbol{\theta}}') - \mathbf{H}^*_{\alpha\boldsymbol{\theta}}\mathbf{w}^*\big\}\Big]}_{T_{22}} \end{aligned} \tag{C.2.3}$$

It holds by Theorem 4.3.7 that $\sqrt{n}\widetilde{\alpha} \xrightarrow{d} H^{-1}_{\alpha\boldsymbol{\theta}}Z$. Together with the regularity condition $H_{\alpha|\boldsymbol{\theta}} = \mathcal{O}(1)$, we have,

$$2nT_{21} = n\widetilde{\alpha}^2 H_{\alpha|\boldsymbol{\theta}} \xrightarrow{d} H^{-1}_{\alpha|\boldsymbol{\theta}}Z^2. \tag{C.2.4}$$

Considering the term T_{22}, we have

$$\begin{aligned} T_{22} = &\frac{\widetilde{\alpha}^2}{2}\Big[\underbrace{\big\{\nabla^2_{\alpha\alpha}\mathcal{L}(\bar{\alpha}, \widehat{\boldsymbol{\theta}}) - \mathbf{H}^*_{\alpha\alpha}\big\}}_{R_1} + \underbrace{\big\{\widehat{\mathbf{w}}^T\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(0, \bar{\boldsymbol{\theta}})\widehat{\mathbf{w}} - \mathbf{w}^*\mathbf{H}^*_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathbf{w}^*\big\}}_{R_2} \\ &- \underbrace{2\big\{\widetilde{\mathbf{w}}^T\nabla^2_{\alpha\boldsymbol{\theta}}\mathcal{L}(\bar{\alpha}', \bar{\boldsymbol{\theta}}') - \mathbf{w}^{*T}\mathbf{H}^*_{\alpha\boldsymbol{\theta}}\big\}}_{R_3}\Big]. \end{aligned} \tag{C.2.5}$$

For the first term $|R_1|$, we have, by Lemma C.1.3, $|R_1| = |\nabla^2_{\alpha\alpha}\mathcal{L}(\bar{\alpha}, \widehat{\boldsymbol{\theta}}) - \mathbf{H}^*_{\alpha\alpha}| = \mathcal{O}_{\mathbb{P}}(s\lambda)$. For the second term,

$$
\begin{aligned}
|R_2| =& |\widehat{\mathbf{w}}^T \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(0, \bar{\boldsymbol{\theta}}) \widehat{\mathbf{w}} - \mathbf{w}^* \mathbf{H}^*_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathbf{w}^*| \\
\leq& |(\widehat{\mathbf{w}} - \mathbf{w}^*)^T \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(0, \bar{\boldsymbol{\theta}})(\widehat{\mathbf{w}} - \mathbf{w}^*)| + 2|\mathbf{w}^* \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(0, \bar{\boldsymbol{\theta}})(\widehat{\mathbf{w}} - \mathbf{w}^*)| \\
&+ |\mathbf{w}^{*T}(\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(0, \bar{\boldsymbol{\theta}}) - \mathbf{H}^*_{\boldsymbol{\theta}\boldsymbol{\theta}})\mathbf{w}^*| \\
\leq& \|\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(0, \bar{\boldsymbol{\theta}})\|_\infty \|\widehat{\mathbf{w}} - \mathbf{w}^*\|_1^2 + 2\|\mathbf{w}^*\|_1 \|\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(0, \bar{\boldsymbol{\theta}})\|_\infty \|\widehat{\mathbf{w}} - \mathbf{w}^*\|_1 \\
&+ \|\mathbf{w}^*\|_1^2 \|\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(0, \bar{\boldsymbol{\theta}}) - \mathbf{H}^*_{\boldsymbol{\theta}\boldsymbol{\theta}}\|_\infty \\
=& \mathcal{O}_{\mathbb{P}}(s'^2 \lambda'^2) + \mathcal{O}_{\mathbb{P}}(s'^2 \lambda') + \mathcal{O}_{\mathbb{P}}(s'^2 s\lambda),
\end{aligned}
\tag{C.2.6}
$$

where the last equality follows by (4.1.2), Lemma C.6.3, Lemma C.1.4 and the sparsity Assumption 4.3.1 of $\mathbf{w}^*$.

For the third term $|R_3|$, we have

$$
\begin{aligned}
|R_3| \leq& 2\Big[|\big\{\nabla^2_{\alpha\boldsymbol{\theta}} \mathcal{L}(\bar{\alpha}', \bar{\boldsymbol{\theta}}') - \mathbf{H}^*_{\alpha\boldsymbol{\theta}}\big\}\widehat{\mathbf{w}}| + |\mathbf{H}^*_{\alpha\boldsymbol{\theta}}(\widehat{\mathbf{w}} - \mathbf{w}^*)|\Big] \\
\leq& 2\Big[|\big\{\nabla^2_{\alpha\boldsymbol{\theta}} \mathcal{L}(\bar{\alpha}', \bar{\boldsymbol{\theta}}') - \mathbf{H}^*_{\alpha\boldsymbol{\theta}}\big\}(\widehat{\mathbf{w}} - \mathbf{w}^*)| + |\big\{\nabla^2_{\alpha\boldsymbol{\theta}} \mathcal{L}(\bar{\alpha}', \bar{\boldsymbol{\theta}}') - \mathbf{H}^*_{\alpha\boldsymbol{\theta}}\big\}\mathbf{w}^*| + |\mathbf{H}^*_{\alpha\boldsymbol{\theta}}(\widehat{\mathbf{w}} - \mathbf{w}^*)|\Big] \\
\leq& 2\|\nabla^2_{\alpha\boldsymbol{\theta}} \mathcal{L}(\bar{\alpha}', \bar{\boldsymbol{\theta}}') - \mathbf{H}^*_{\alpha\boldsymbol{\theta}}\|_\infty \|\widehat{\mathbf{w}} - \mathbf{w}^*\|_1 + 2\|\nabla^2_{\alpha\boldsymbol{\theta}} \mathcal{L}(\bar{\alpha}', \bar{\boldsymbol{\theta}}') - \mathbf{H}^*_{\alpha\boldsymbol{\theta}}\|_\infty \|\mathbf{w}^*\|_1 \\
&+ 2\|\mathbf{H}^*_{\alpha\boldsymbol{\theta}}\|_\infty \|\widehat{\mathbf{w}} - \mathbf{w}^*\|_1.
\end{aligned}
$$

Note that $\|\nabla^2_{\alpha\boldsymbol{\theta}} \mathcal{L}(\bar{\alpha}', \bar{\boldsymbol{\theta}}') - \mathbf{H}^*_{\alpha\boldsymbol{\theta}}\|_\infty \|\widehat{\mathbf{w}} - \mathbf{w}^*\|_1 = \mathcal{O}_{\mathbb{P}}(s' s\lambda' \lambda)$ by Lemma C.1.4 and Lemma C.6.3, $\|\nabla_{\alpha\boldsymbol{\theta}} \mathcal{L}(\bar{\alpha}', \bar{\boldsymbol{\theta}}') - \mathbf{H}^*_{\alpha\boldsymbol{\theta}}\|_\infty \|\mathbf{w}^*\|_1 = \mathcal{O}_{\mathbb{P}}(s' s\lambda)$ by Lemma C.6.3 and Assumption 4.3.2, and $\|\mathbf{H}^*_{\alpha\boldsymbol{\theta}}\|_\infty \|\widehat{\mathbf{w}} - \mathbf{w}^*\|_1 = \mathcal{O}_{\mathbb{P}}(s'\lambda')$ by Assumption 4.3.3 and Lemma C.1.4. We have $|R_3| = \mathcal{O}_{\mathbb{P}}(s' s\lambda)$.

Combining the results above, we have,

$$
T_{22} = \frac{\widetilde{\alpha}^2}{2} \cdot \mathcal{O}_{\mathbb{P}}(s'^2 s\lambda) = \mathcal{O}_{\mathbb{P}}\Big(\frac{s'^2 s\sqrt{\log d}}{n^{3/2}}\Big) = o_{\mathbb{P}}(n^{-1}), \tag{C.2.7}
$$

where the second equality follows by Theorem 4.3.7 that $\widetilde{\alpha} = \mathcal{O}_{\mathbb{P}}(n^{-1/2})$ under the null hypothesis, and the last equality follows by the assumption that $n^{-1/2}s's^2 \log d = o(1)$.

Combining (C.2.4) and (C.2.7) with (C.2.3), we have

$$2nT_2 \xrightarrow{d} H_{\alpha|\boldsymbol{\theta}}^{-1} Z^2, \text{ where } Z \sim N(0, H_{\alpha|\boldsymbol{\theta}}). \tag{C.2.8}$$

Plugging (C.2.2) and (C.2.8) into (C.2.1), by Theorem 4.3.4,

$$-2n\big\{\mathcal{L}(\widetilde{\alpha}, \widehat{\boldsymbol{\theta}} - \widetilde{\alpha}\widehat{\mathbf{w}}) - \mathcal{L}(0, \widehat{\boldsymbol{\theta}})\big\} \xrightarrow{d} Z_\chi^2, \text{where } Z_\chi \sim \chi_1^2,$$

which concludes the proof. □

C.3 Proofs in Section 4.3

In this section, we provide the detailed proofs in Section 4.3.

C.3.1 Proofs in Section 4.3.1

Proof of Lemma 4.3.5. By the definition of $H_{\alpha|\boldsymbol{\theta}}$ and $\widehat{H}_{\alpha|\boldsymbol{\theta}}$, we have

$$|H_{\alpha|\boldsymbol{\theta}} - \widehat{H}_{\alpha|\boldsymbol{\theta}}| \leq \underbrace{|\mathbf{H}_{\alpha\alpha}^* - \nabla_{\alpha\alpha}^2 \mathcal{L}(\widehat{\alpha}, \widehat{\boldsymbol{\theta}})|}_{E_1} + \underbrace{|\mathbf{H}_{\alpha\boldsymbol{\theta}}^* \mathbf{H}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{*-1} \mathbf{H}_{\boldsymbol{\theta}\alpha}^* - \widehat{\mathbf{w}}^T \nabla_{\boldsymbol{\theta}\alpha}^2 \mathcal{L}(\widehat{\alpha}, \widehat{\boldsymbol{\theta}})|}_{E_2}. \tag{C.3.1}$$

We consider the two terms separately. For the first term E_1, we have by Lemma C.1.3, $E_1 = \mathcal{O}_{\mathbb{P}}(s\lambda)$. For the second term E_2, we have,

$$\begin{aligned} E_2 &= |\mathbf{H}_{\alpha\boldsymbol{\theta}}^* \mathbf{H}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{*-1} \mathbf{H}_{\boldsymbol{\theta}\alpha}^* - \widehat{\mathbf{w}}^T \nabla_{\boldsymbol{\theta}\alpha}^2 \mathcal{L}(\widehat{\alpha}, \widehat{\boldsymbol{\theta}})| = |\mathbf{H}_{\alpha\boldsymbol{\theta}}^* \mathbf{H}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{*-1} \mathbf{H}_{\boldsymbol{\theta}\alpha}^* - \widehat{\mathbf{w}}^T \mathbf{H}_{\boldsymbol{\theta}\alpha}^* + \widehat{\mathbf{w}}^T \mathbf{H}_{\boldsymbol{\theta}\alpha}^* - \widehat{\mathbf{w}}^T \nabla_{\boldsymbol{\theta}\alpha}^2 \mathcal{L}(\widehat{\alpha}, \widehat{\boldsymbol{\theta}})| \\ &\leq \underbrace{|\mathbf{H}_{\alpha\boldsymbol{\theta}}^* \mathbf{H}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{*-1} \mathbf{H}_{\boldsymbol{\theta}\alpha}^* - \widehat{\mathbf{w}}^T \mathbf{H}_{\boldsymbol{\theta}\alpha}^*|}_{E_{21}} + \underbrace{|\widehat{\mathbf{w}}^T \mathbf{H}_{\boldsymbol{\theta}\alpha}^* - \widehat{\mathbf{w}}^T \nabla_{\boldsymbol{\theta}\alpha}^2 \mathcal{L}(\widehat{\alpha}, \widehat{\boldsymbol{\theta}})|}_{E_{22}}. \end{aligned}$$

For the term E_{21}, we have, by Hölder's inequality,

$$E_{21} \leq \|\mathbf{H}^*_{\alpha\boldsymbol{\theta}}\mathbf{H}^{*-1}_{\boldsymbol{\theta}\boldsymbol{\theta}} - \widehat{\mathbf{w}}^T\|_1 \|\mathbf{H}^*_{\boldsymbol{\theta}\alpha}\|_\infty = \mathcal{O}_{\mathbb{P}}(s'\lambda'), \tag{C.3.2}$$

where the last inequality holds by the fact that $\|\mathbf{H}^*_{\alpha\boldsymbol{\theta}}\mathbf{H}^{*-1}_{\boldsymbol{\theta}\boldsymbol{\theta}} - \widehat{\mathbf{w}}^T\|_1 = \mathcal{O}_{\mathbb{P}}(s'\lambda')$, and $\|\mathbf{H}^*_{\boldsymbol{\theta}\alpha}\|_\infty = \mathcal{O}(1)$ by Assumption 4.3.3.

For the second term E_{22}, we have, by Hölder's inequality,

$$E_{22} \leq \|\widehat{\mathbf{w}}\|_1 \|\mathbf{H}^*_{\boldsymbol{\theta}\alpha} - \nabla^2_{\boldsymbol{\theta}\alpha}\mathcal{L}(\widehat{\alpha}, \widehat{\boldsymbol{\theta}})\|_\infty = \mathcal{O}_{\mathbb{P}}(s's\lambda), \tag{C.3.3}$$

where the last equality holds by the assumption that $\|\mathbf{w}^*\|_1 = \mathcal{O}(s')$, the result $\|\widehat{\mathbf{w}} - \mathbf{w}^*\| = \mathcal{O}_{\mathbb{P}}(s'\lambda')$ by (C.1.1) and by Lemma C.1.3 that $\|\mathbf{H}^*_{\boldsymbol{\theta}\alpha} - \nabla^2_{\boldsymbol{\theta}\alpha}\mathcal{L}(\widehat{\alpha}, \widehat{\boldsymbol{\theta}})\|_\infty = \mathcal{O}_{\mathbb{P}}(s\lambda)$.

Combining (C.3.2) and (C.3.3), we have, $E_2 \leq E_{21} + E_{22} = \mathcal{O}_{\mathbb{P}}(s'\lambda')$. Together with the result that $E_1 = \mathcal{O}_{\mathbb{P}}(s^2\lambda)$, the claim holds as desired. □

Proof of Corollary 4.3.6. The claim follows immediately from Theorem 4.3.4 and Lemma 4.3.5. □

Proof of Theorem 4.3.7. Based on our construction of $\widetilde{\alpha}$ in (4.2.7), we have

$$\begin{aligned}
\widetilde{\alpha} &= \widehat{\alpha} - \Big\{\frac{\partial \widehat{U}(\widehat{\alpha}, \widehat{\boldsymbol{\theta}})}{\partial \alpha}\Big\}^{-1} \widehat{U}(\widehat{\alpha}, \widehat{\boldsymbol{\theta}}) = \widehat{\alpha} - H^{-1}_{\alpha|\boldsymbol{\theta}} \widehat{U}(\widehat{\alpha}, \widehat{\boldsymbol{\theta}}) + \underbrace{\widehat{U}(\widehat{\alpha}, \widehat{\boldsymbol{\theta}})\Big[H^{-1}_{\alpha|\boldsymbol{\theta}} - \Big\{\frac{\partial \widehat{U}(\widehat{\alpha}, \widehat{\boldsymbol{\theta}})}{\partial \alpha}\Big\}^{-1}\Big]}_{R_1} \\
&= \widehat{\alpha} - H^{-1}_{\alpha|\boldsymbol{\theta}}\Big\{\widehat{U}(0, \widehat{\boldsymbol{\theta}}) + \frac{(\widehat{\alpha} - 0)\partial \widehat{U}(\bar{\alpha}, \widehat{\boldsymbol{\theta}})}{\partial \alpha}\Big\} + R_1 \qquad \text{(C.3.4)} \\
&= \widehat{\alpha} - H^{-1}_{\alpha|\boldsymbol{\theta}} \widehat{U}(0, \widehat{\boldsymbol{\theta}}) - \widehat{\alpha} H^{-1}_{\alpha|\boldsymbol{\theta}} H_{\alpha|\boldsymbol{\theta}} + \underbrace{\widehat{\alpha} H^{-1}_{\alpha|\boldsymbol{\theta}}\Big\{H_{\alpha|\boldsymbol{\theta}} - \frac{\partial \widehat{U}(\bar{\alpha}, \widehat{\boldsymbol{\theta}})}{\partial \alpha}\Big\}}_{R_2} + R_1 = -H^{-1}_{\alpha|\boldsymbol{\theta}} \widehat{U}(0, \widehat{\boldsymbol{\theta}}) + R_1 + R_2,
\end{aligned}$$

where (C.3.4) holds by the mean value theorem for some $\bar{\alpha} = u\widehat{\alpha}$ and $u \in [0, 1]$. For the term R_1, note that

$$|\widehat{U}(\widehat{\alpha}, \widehat{\boldsymbol{\theta}}) - \widehat{U}(0, \widehat{\boldsymbol{\theta}})| = |\widehat{\alpha}| \cdot \Big|\frac{\partial \widehat{U}(\bar{\alpha}', \widehat{\boldsymbol{\theta}})}{\partial \alpha}\Big|$$

where the equality holds by mean value theorem with $\bar{\alpha}' = u\widehat{\alpha}$ for some $u \in [0,1]$. Under the null hypothesis $\alpha^* = 0$, by Theorem 3.2 of Huang et al. (2013), $|\widehat{\alpha} - \alpha^*| \leq \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = \mathcal{O}_{\mathbb{P}}(s\lambda)$. By regularity condition $H_{\alpha|\boldsymbol{\theta}} = \mathcal{O}(1)$ and Lemma 4.3.5, it also holds that $|\partial \widehat{U}(\bar{\alpha}', \widehat{\boldsymbol{\theta}})/\partial\alpha| = \mathcal{O}_{\mathbb{P}}(1)$. Thus, we have

$$|\widehat{U}(\widehat{\alpha}, \widehat{\boldsymbol{\theta}}) - \widehat{U}(0, \widehat{\boldsymbol{\theta}})| = \mathcal{O}_{\mathbb{P}}(s\lambda), \text{ and } |\widehat{U}(0, \widehat{\boldsymbol{\theta}})| = \mathcal{O}_{\mathbb{P}}(n^{-1/2}), \tag{C.3.5}$$

where the second equality holds by Theorem 4.3.4. Thus, by triangle inequality, we have

$$|R_1| \leq |\widehat{U}(\widehat{\alpha}, \widehat{\boldsymbol{\theta}}) - \widehat{U}(0, \widehat{\boldsymbol{\theta}})| \cdot \Big| H_{\alpha|\boldsymbol{\theta}}^{-1} - \Big\{ \frac{\partial \widehat{U}(\widehat{\alpha}, \widehat{\boldsymbol{\theta}})}{\partial \alpha} \Big\}^{-1} \Big| + |\widehat{U}(0, \widehat{\boldsymbol{\theta}})| \cdot \Big| H_{\alpha|\boldsymbol{\theta}}^{-1} - \Big\{ \frac{\partial \widehat{U}(\widehat{\alpha}, \widehat{\boldsymbol{\theta}})}{\partial \alpha} \Big\}^{-1} \Big| = \mathcal{O}_{\mathbb{P}}\Big(s^3 \frac{\log d}{n} \Big),$$

where the last equality holds by (C.3.5) and Lemma 4.3.5.

For the term R_2, we have,

$$|R_2| \leq |\widehat{\alpha} H_{\alpha|\boldsymbol{\theta}}^{-1}| \cdot \Big| H_{\alpha|\boldsymbol{\theta}} - \frac{\partial \widehat{U}(\bar{\alpha}, \widehat{\boldsymbol{\theta}})}{\partial \alpha} \Big| = \mathcal{O}_{\mathbb{P}}\Big(s^3 \frac{\log d}{n} \Big),$$

where the last inequality holds by the fact that $|\widehat{\alpha}| = \mathcal{O}_{\mathbb{P}}(s\lambda)$, $|H_{\alpha|\boldsymbol{\theta}}| = \mathcal{O}(1)$ and Lemma 4.3.5.

Consequently, it holds that,

$$\sqrt{n}\widetilde{\alpha} \xrightarrow{d} Z, \text{ where } Z \sim N(0, H_{\alpha|\boldsymbol{\theta}}^{-1}),$$

and the last equality follows by Theorem 4.3.4 and our the assumption that $n^{-1/2} s^3 \log d = o(1)$. The claim follows as desired. □

Proof of Corollary 4.3.8. The claim follows from the argument Theorem 4.3.7 by replacing 0 by α^*. □

Proof of Corollary 4.3.10. The claim follows from Theorem 4.3.9 directly. □

Proof. The claim follows from the proofs of Theorems 4.3.4, 4.3.7 and 4.3.9 by looking at the asymptotic distributions of the test statistics. □

C.3.2 Proofs in Section 4.3.2

Proof of Theorem 4.3.14. Under the alternative hypothesis that $\alpha^* = n^{-1/2}c$, we look at the decorrelated score function $\widehat{U}(0, \widehat{\boldsymbol{\theta}})$. By the same derivation as in (C.1.2) and mean value theorem, it holds that

$$\begin{aligned}\widehat{U}(0, \widehat{\boldsymbol{\theta}}) = &\underbrace{\nabla_\alpha \mathcal{L}(0, \boldsymbol{\theta}^*) - \mathbf{w}^{*T} \nabla_{\boldsymbol{\theta}} \mathcal{L}(0, \boldsymbol{\theta}^*)}_{S} + \underbrace{(\mathbf{w}^* - \widehat{\mathbf{w}})^T \nabla_{\boldsymbol{\theta}} \mathcal{L}(0, \boldsymbol{\theta}^*)}_{E_1} \\ &+ \underbrace{\big\{\nabla^2_{\alpha\boldsymbol{\theta}} \mathcal{L}(0, \bar{\boldsymbol{\theta}}) - \widehat{\mathbf{w}}^T \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(0, \widetilde{\boldsymbol{\theta}})\big\}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)}_{E_2},\end{aligned}$$

where $\bar{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + u_1(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ and $\widetilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + u_2(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ for some $0 \le u_1, u_2 \le 1$.

The proof of Theorem 4.3.4 cannot be directly applied to characterize the asymptotic distribution of the first term S. This is because the vector $(0, \boldsymbol{\theta}^{*T})^T \neq \boldsymbol{\beta}^*$ under the alternative hypothesis, and thus Lemma C.1.1 cannot be applied. To derive the asymptotic distribution of S, we have

$$\begin{aligned}S &= \nabla_\alpha \mathcal{L}(0, \boldsymbol{\theta}^*) - \mathbf{w}^{*T} \nabla_{\boldsymbol{\theta}} \mathcal{L}(0, \boldsymbol{\theta}^*) \\ &= \underbrace{\nabla_\alpha \mathcal{L}(\alpha^*, \boldsymbol{\theta}^*) - \mathbf{w}^{*T} \nabla_{\boldsymbol{\theta}} \mathcal{L}(\alpha^*, \boldsymbol{\theta}^*)}_{S_1} + \underbrace{\alpha^* \mathbf{w}^{*T} \nabla^2_{\boldsymbol{\theta}\alpha} \mathcal{L}(\bar{\alpha}', \boldsymbol{\theta}^*) - \alpha^* \nabla^2_{\alpha\alpha} \mathcal{L}(\bar{\alpha}, \boldsymbol{\theta}^*)}_{R},\end{aligned}$$

where the second equality holds by mean value theorem for some $\bar{\alpha} = v_1\alpha^*, \bar{\alpha}' = v_2\alpha^*$ and $0 \le v_1, v_2 \le 1$.

By Lemma C.1.1, taking $\mathbf{v} = (1, -\mathbf{w}^{*T})^T$, under the alternative hypothesis, it holds that the first term

$$S_1 \xrightarrow{d} Z, \text{ where } Z \sim N(0, H_{\alpha|\boldsymbol{\theta}}). \tag{C.3.6}$$

For the second term R, we have

$$R = -\alpha^*(\mathbf{H}^*_{\alpha\alpha} - \mathbf{w}^{*T}\mathbf{H}^*_{\boldsymbol{\theta}\alpha}) + \underbrace{\alpha^*\{\mathbf{H}^*_{\alpha\alpha} - \nabla^2_{\alpha\alpha}\mathcal{L}(\bar{\alpha}, \boldsymbol{\theta}^*)\}}_{R_1} + \underbrace{\alpha^*\mathbf{w}^{*T}\{\nabla^2_{\boldsymbol{\theta}\alpha}\mathcal{L}(\bar{\alpha}', \boldsymbol{\theta}^*) - \mathbf{H}^*_{\boldsymbol{\theta}\alpha}\}}_{R_2}. \tag{C.3.7}$$

For the term R_1, we have, under the alternative hypothesis $\alpha^* = n^{-1/2}c$,

$$|R_1| = \alpha^*|\mathbf{H}^*_{\alpha\alpha} - \nabla^2_{\alpha\alpha}\mathcal{L}(\bar{\alpha}, \boldsymbol{\theta}^*)| \leq cn^{-1/2}\|\mathbf{H}^* - \nabla^2\mathcal{L}(\bar{\alpha}, \boldsymbol{\theta}^*)\|_\infty = \mathcal{O}_{\mathbb{P}}(n^{-1}\sqrt{\log d}), \tag{C.3.8}$$

where the last equality holds by Lemma C.1.3.

Next, we look at the term R_2. It holds that

$$|R_2| \leq \|\alpha^*\mathbf{w}^*\|_1\|\nabla^2_{\boldsymbol{\theta}\alpha}\mathcal{L}(\bar{\alpha}', \boldsymbol{\theta}^*) - \mathbf{H}^*_{\boldsymbol{\theta}\alpha}\|_\infty = \mathcal{O}_{\mathbb{P}}(n^{-1}s's\sqrt{\log d}), \tag{C.3.9}$$

where the first inequality holds by Hölder's inequality, and the equality holds by Lemma C.1.3.

Plugging (C.3.8) and (C.3.9) into (C.3.7), and by the identity $H_{\alpha|\boldsymbol{\theta}} = \mathbf{H}^*_{\alpha\alpha} - \mathbf{w}^{*T}\mathbf{H}^*_{\boldsymbol{\theta}\alpha}$, we have $R = -\alpha^* H_{\alpha|\boldsymbol{\theta}} + \mathcal{O}_{\mathbb{P}}(n^{-1}s's\sqrt{\log d})$. Combining this result with (C.3.6), we have, $S \xrightarrow{d} Z - \alpha^* H_{\alpha|\boldsymbol{\theta}}$.

Meanwhile, by the similar argument as in the proof of Theorem 4.3.4, it is not difficult to get that the two latter terms $E_1 = o_{\mathbb{P}}(n^{-1/2})$ and $E_2 = o_{\mathbb{P}}(n^{-1/2})$.

To summarize, under the alternative hypothesis that $\alpha^* = cn^{-1/2}$, the decorrelated score function satisfies

$$\sqrt{n}\widehat{U}(0, \widehat{\boldsymbol{\theta}}) \xrightarrow{d} Z', \text{ where } Z' \sim N(-cH_{\alpha|\boldsymbol{\theta}}, H_{\alpha|\boldsymbol{\theta}}),$$

which concludes the proof. □

Proof of Theorem 4.3.16. (a). By the definition of $\widetilde{\alpha}$ in (4.2.7), we have

$$\widetilde{\alpha} = -H_{\alpha|\boldsymbol{\theta}}^{-1}\widehat{U}(0,\widehat{\boldsymbol{\theta}}) + \underbrace{\widehat{U}(\widehat{\alpha},\widehat{\boldsymbol{\theta}})\Big[H_{\alpha|\boldsymbol{\theta}}^{-1} - \Big\{\frac{\partial\widehat{U}(\widehat{\alpha},\widehat{\boldsymbol{\theta}})}{\partial\alpha}\Big\}^{-1}\Big]}_{R_1} + \underbrace{\widehat{\alpha}H_{\alpha|\boldsymbol{\theta}}^{-1}\Big\{H_{\alpha|\boldsymbol{\theta}} - \frac{\partial\widehat{U}(\bar{\alpha},\widehat{\boldsymbol{\theta}})}{\partial\alpha}\Big\}}_{R_2}.$$

By Theorem 4.3.14, we have

$$-\sqrt{n}H_{\alpha|\boldsymbol{\theta}}^{-1}\widehat{U}(0,\widehat{\boldsymbol{\theta}}) \xrightarrow{d} Z, \text{ where } Z \sim N(\alpha^*, H_{\alpha|\boldsymbol{\theta}}^{-1}).$$

In addition, by the similar argument as in Theorem 4.3.7, we have $R_1 = o_{\mathbb{P}}(n^{-1/2})$ and $R_2 = o_{\mathbb{P}}(n^{-1/2})$. Under the null hypothesis $\alpha^* = n^{-1/2}c$, we have

$$\sqrt{n}\widetilde{\alpha} \xrightarrow{d} Z', \text{ where } Z' \sim N(c, H_{\alpha|\boldsymbol{\theta}}^{-1}),$$

and our claim holds as desired.

(b). By the definition of the test statistic of the decorrelated partial likelihood ratio test (4.2.10), we have

$$\mathcal{L}(\widetilde{\alpha},\widehat{\boldsymbol{\theta}}-\widetilde{\alpha}\widehat{\mathbf{w}}) - \mathcal{L}(0,\widehat{\boldsymbol{\theta}}) = \underbrace{\widetilde{\alpha}\widehat{U}(0,\widehat{\boldsymbol{\theta}})}_{T_1} + \underbrace{\frac{\widetilde{\alpha}^2}{2}\Big\{\nabla_{\alpha\alpha}^2\mathcal{L}(\bar{\alpha},\widehat{\boldsymbol{\theta}}) + \widehat{\mathbf{w}}^T\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2\mathcal{L}(0,\bar{\boldsymbol{\theta}})\widehat{\mathbf{w}} - 2\widetilde{\mathbf{w}}^T\nabla_{\boldsymbol{\theta}\alpha}^2\mathcal{L}(\bar{\alpha}',\bar{\boldsymbol{\theta}}')\Big\}}_{T_2},$$

where the equality holds by mean value theorem with $\bar{\alpha} = \widehat{\alpha} + u_1(0-\widehat{\alpha})$, $\bar{\alpha}' = \widehat{\alpha} + u_2(0-\widehat{\alpha})$ and $\bar{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}} + u_3(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}})$ for some $0 \le u_1, u_2, u_3 \le 1$.

By Theorem 4.3.14 and 4.3.16, we have $T_1 = -\big\{\widehat{U}(0,\widehat{\boldsymbol{\theta}})\big\}^2 H_{\alpha|\boldsymbol{\theta}}^{-1} + o_{\mathbb{P}}(n^{-1})$. In addition, by the similar argument as in Theorem 4.3.9, we have $T_2 = \frac{1}{2}\{\widehat{U}(0,\widehat{\boldsymbol{\theta}})\}^2 H_{\alpha|\boldsymbol{\theta}}^{-1} + o_{\mathbb{P}}(n^{-1})$.

Consequently, the test statistic $\widehat{L}_n$ in (4.2.10) satisfies

$$2n\big\{\mathcal{L}(0,\widehat{\boldsymbol{\theta}}) - \mathcal{L}(\widetilde{\alpha},\widehat{\boldsymbol{\theta}} - \widetilde{\alpha}\widehat{\mathbf{w}})\big\} = n\widehat{U}(0,\widehat{\boldsymbol{\theta}})^2 H_{\alpha|\boldsymbol{\theta}}^{-1} + o_{\mathbb{P}}(1) \xrightarrow{d} Z_\chi' + o_{\mathbb{P}}(1),$$

where $Z'_\chi \sim NC_{\chi_1}(c^2 H_{\alpha|\boldsymbol{\theta}})$ by Theorem 4.3.14. Our claim follows as desired. □

C.4 Proofs in Section 4.4

In this section, we provide detailed proofs in Section 4.4.

Lemma C.4.1. *Under Assumptions 4.1.1, 4.1.2, 4.3.2, 4.3.3 and 4.4.1,* $\|\nabla\widehat{\Lambda}_0(t,\widehat{\boldsymbol{\beta}}) - \nabla\Lambda_0(t,\boldsymbol{\beta}^*)\|_\infty = \mathcal{O}_{\mathbb{P}}(s\sqrt{n^{-1}\log d})$.

Proof. By the definition of $\widehat{\Lambda}_0(t,\widehat{\boldsymbol{\beta}})$ in (4.4.1), we have,

$$\|\nabla\widehat{\Lambda}_0(t,\widehat{\boldsymbol{\beta}})-\nabla\Lambda_0(t,\boldsymbol{\beta}^*)\|_\infty = \Big\|\frac{1}{n}\int_0^t \frac{S^{(1)}(u,\widehat{\boldsymbol{\beta}})d\bar{N}(u)}{\{S^{(0)}(u,\widehat{\boldsymbol{\beta}})\}^2}+\mathbb{E}\int_0^t \frac{\mathbf{s}^{(1)}(u,\boldsymbol{\beta}^*)dN(u)}{\{\mathbf{s}^{(0)}(u,\boldsymbol{\beta}^*)\}^2}\Big\|_\infty = \mathcal{O}_{\mathbb{P}}\Big(s\sqrt{\frac{\log d}{n}}\Big),$$

where the last inequality follows by the same argument in Lemma C.1.3. □

A corollary of Lemma C.4.1 and Lemma C.1.4 follows immediately which characterizes the rate of convergence of $\widehat{\mathbf{u}}(t)$.

Corollary C.4.2. Under Assumptions 4.1.1, 4.1.2, 4.3.2, 4.3.3 and 4.4.1, if $\delta \asymp s'\sqrt{n^{-1}\log d}$ we have,

$$\|\widehat{\mathbf{u}}(t) - \mathbf{u}^*(t)\|_1 = \mathcal{O}_{\mathbb{P}}\Big(ss'\sqrt{\frac{\log d}{n}}\Big).$$

Now, we are ready to prove Theorem 4.4.2.

Proof of Theorem 4.4.2. We first decompose $\sqrt{n}\big\{\Lambda_0(t) - \widetilde{\Lambda}_0(t,\widehat{\boldsymbol{\beta}})\big\}$ into two terms that

$$\sqrt{n}\big\{\Lambda_0(t) - \widetilde{\Lambda}_0(t,\widehat{\boldsymbol{\beta}})\big\} = \sqrt{n}\underbrace{\big\{\Lambda_0(t) - \widehat{\Lambda}_0(t,\boldsymbol{\beta}^*)\big\}}_{I_1(t)} + \sqrt{n}\underbrace{\big\{\widehat{\Lambda}_0(t,\boldsymbol{\beta}^*) - \widetilde{\Lambda}_0(t,\widehat{\boldsymbol{\beta}})\big\}}_{I_2(t)}.$$

Let $M_i(t) = N_i(t) - \int_0^t Y_i(u)\lambda_0(u)du$. For the first term $\sqrt{n}I_1(t)$, we have

$$\sqrt{n}I_1(t) = \int_0^t \frac{\sqrt{n}\sum_{i=1}^n dM_i(u)}{\sum_{i=1}^n Y_i(u)\exp\{\boldsymbol{X}_i^T(u)\boldsymbol{\beta}^*\}}.$$

Since $M_i(t)$ is a martingale, $\sqrt{n}I_1(t)$ becomes a sum of martingale residuals. By Andersen and Gill (1982), we have, as $n \to \infty$, $\sqrt{n}I_1(t) \xrightarrow{d} N\big(0, \sigma_1^2(t)\big)$, where

$$\sigma_1^2(t) = \int_0^t \frac{\lambda_0(u)du}{\mathbb{E}\big[\exp\{\mathbf{X}^T(u)\boldsymbol{\beta}^*\}Y(u)\big]}.$$

For the second term $I_2(t)$, we have, by mean value theorem, for some $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta}^* + t(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$, $\widetilde{\boldsymbol{\beta}}' = \boldsymbol{\beta}^* + t'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$ and $0 \le t, t' \le 1$,

$$\begin{aligned}
I_2(t) &= \widehat{\Lambda}_0(t, \boldsymbol{\beta}^*) - \widehat{\Lambda}_0(t, \widehat{\boldsymbol{\beta}}) + \{\widehat{\mathbf{u}}(t)\}^T \nabla \mathcal{L}(\widehat{\boldsymbol{\beta}}) \\
&= (\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})^T \nabla \widehat{\Lambda}_0(t, \widetilde{\boldsymbol{\beta}}) + \{\widehat{\mathbf{u}}(t)\}^T \{\nabla \mathcal{L}(\boldsymbol{\beta}^*) + \nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}')(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\} \\
&= \{\mathbf{u}^*(t)\}^T \nabla \mathcal{L}(\boldsymbol{\beta}^*) + \underbrace{(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})^T \nabla \widehat{\Lambda}_0(t, \widetilde{\boldsymbol{\beta}}) + \{\mathbf{u}^*(t)\}^T \nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}')(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)}_{R_1} \\
&\quad + \underbrace{\{\widehat{\mathbf{u}}(t) - \mathbf{u}^*(t)\}^T \{\nabla \mathcal{L}(\boldsymbol{\beta}^*) + \nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}')(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\}}_{R_2}.
\end{aligned}$$

Next, we consider the two terms R_1 and R_2. For the term R_1, we have

$$\begin{aligned}
R_1 &= (\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})^T \nabla \widehat{\Lambda}_0(t, \widetilde{\boldsymbol{\beta}}) + \{\mathbf{u}^*(t)\}^T \nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}')(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \\
&= (\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})^T \big[\mathbf{H}^* \mathbf{H}^{*-1} \nabla \widehat{\Lambda}_0(t, \widetilde{\boldsymbol{\beta}}) - \nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}') \mathbf{H}^{*-1} \nabla \widehat{\Lambda}_0(t, \boldsymbol{\beta}^*)\big] \\
&= \underbrace{(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})^T \{\nabla \widehat{\Lambda}_0(t, \widetilde{\boldsymbol{\beta}}) - \nabla \Lambda_0(t, \boldsymbol{\beta}^*)\}}_{R_{11}} + \underbrace{(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})^T \big[\mathbf{H}^* - \nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}')\big] \mathbf{H}^{*-1} \nabla \Lambda_0(t, \boldsymbol{\beta}^*)}_{R_{12}}.
\end{aligned}$$

It holds that $|R_{11}| \le \|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}\|_1 \|\nabla \Lambda_0(t, \widetilde{\boldsymbol{\beta}}) - \nabla \widehat{\Lambda}_0(t, \boldsymbol{\beta}^*)\|_\infty = \mathcal{O}_{\mathbb{P}}(s^2 n^{-1} \log d)$ by (4.1.2) and Lemma C.4.1, and $|R_{12}| \le \|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}\|_1 \|\mathbf{H}^* - \nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}')\|_\infty \|\mathbf{u}^*(t)\|_1 = \mathcal{O}_{\mathbb{P}}(s' s^2 n^{-1} \log d)$. Summing them up, by triangle inequality, we have $|R_1| = \mathcal{O}_{\mathbb{P}}(s' s^2 n^{-1} \log d)$.

For the term R_2, we have

$$\begin{aligned}
|R_2| &\le \|\widehat{\mathbf{u}}(t) - \mathbf{u}^*(t)\|_1 \|\nabla \mathcal{L}(\boldsymbol{\beta}^*)\|_\infty + \|\widehat{\mathbf{u}}(t) - \mathbf{u}^*(t)\|_1 \|\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\beta}}')\|_\infty \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \\
&= \mathcal{O}_{\mathbb{P}}(s' s n^{-1} \log d) + \mathcal{O}_{\mathbb{P}}(s' s^2 n^{-1} \log d),
\end{aligned}$$

where the last inequality holds by Lemma C.1.2 and C.1.3.

Meanwhile, by Lemma C.1.1, taking $\mathbf{v} = \mathbf{u}^*(t)$, we have the term $\sqrt{n}\mathbf{u}^{*T}(t)\nabla\mathcal{L}(\boldsymbol{\beta}^*) \xrightarrow{d} N(0, \sigma_2^2(t))$, where $\sigma_2^2(t) = \nabla\Lambda_0(t, \boldsymbol{\beta}^*)^T\mathbf{H}^{*-1}\nabla\Lambda_0(t, \boldsymbol{\beta}^*)$. Thus, we have,

$$\sqrt{n}I_2(t) \xrightarrow{d} Z, \text{ where } Z \sim N(0, \sigma_2^2(t)),$$

and $\sigma_2^2(t) = \nabla\Lambda_0(t, \boldsymbol{\beta}^*)^T\mathbf{H}^{*-1}\nabla\Lambda_0(t, \boldsymbol{\beta}^*)$.

Following the standard martingale theory, the covariance between $I_1(t)$ and $I_2(t)$ is 0. Our claim holds as desired. □

Proof of Corollary 4.4.3. This follows immediately from Theorem 4.4.2. □

C.5 Extension to Conditional Hazard Function Inference

In this section, we extend the procedure proposed in Section 4.4 to conduct conditional hazard function inference given the covariate. For ease of presentation, we assume that the covariates are fixed through time t. Given the i-th sample's covariate $\boldsymbol{X}_i$, the conditional hazard rate function and the cumulative conditional hazard function at time t are

$$\lambda_0(t, \boldsymbol{X}_i) = \lambda_0(t)\exp(\boldsymbol{X}_i^T\boldsymbol{\beta}^*), \text{ and } \Lambda_0(t, \boldsymbol{X}_i) = \int_0^t \lambda_0(u, \boldsymbol{X}_i)du = \int_0^t \lambda_0(u)du \cdot \exp(\boldsymbol{X}_i^T\boldsymbol{\beta}^*).$$

Similar to Section 4.4, we adopt a Breslow-type estimator for the conditional hazard function. Given the initial penalized estimator $\widehat{\boldsymbol{\beta}}$, we use the direct plug-in estimator for the conditional hazard function at time t as

$$\widehat{\Lambda}_0(t, \boldsymbol{X}_i) = \int_0^t \frac{dN_i(u) \cdot \exp\left(\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}\right)}{\sum_{i'=1}^n \exp\left(\boldsymbol{X}_{i'}^T\widehat{\boldsymbol{\beta}}\right)Y_{i'}(u)}.$$

Due to the intractable distribution of $\widehat{\boldsymbol{\beta}}$, we cannot directly conduct inference based on $\widehat{\Lambda}_0(t, \boldsymbol{X}_i)$. Using the decorrelation approach, we propose to estimate the conditional hazard function by the sample version of $\widehat{\Lambda}_0(t, \boldsymbol{X}_i) - \big\{\nabla\Lambda_0(t, \boldsymbol{X}_i)\big\}\mathbf{H}^{*-1}\nabla\mathcal{L}(\widehat{\boldsymbol{\beta}})$, where the gradient $\nabla\Lambda_0(t, \boldsymbol{X}_i)$ is taken with respect to $\boldsymbol{\beta}$ at $\boldsymbol{\beta}^*$. Similar to (4.4.2), we directly estimate the product $\mathbf{H}^{*-1}\nabla\Lambda_0(t, \boldsymbol{X}_i, \boldsymbol{\beta}^*)$ by the following Dantzig type estimator

$$\widehat{\mathbf{u}}(t) = \operatorname{argmin} \|\mathbf{u}(t)\|_1, \text{ subject to } \|\nabla\widehat{\Lambda}_0(t, \boldsymbol{X}_i) - \nabla^2\mathcal{L}(\widehat{\boldsymbol{\beta}})\mathbf{u}(t)\|_\infty \leq \delta, \tag{C.5.1}$$

where δ is a tuning parameter. By the following assumption, which is analogous to Assumption 4.4.1 $\widehat{\mathbf{u}}(t)$ converges to $\mathbf{u}^*(t) = \mathbf{H}^{*-1}\nabla\Lambda_0(t, \boldsymbol{X}_i)$ at a fast rate.

Assumption C.5.1. It holds that $\|\mathbf{u}^*(t)\|_0 = s' \asymp s$ for all $0 \leq t \leq \tau$.

Hence, the decorrelated conditional hazard function estimator at time t is

$$\widetilde{\Lambda}_0(t, \boldsymbol{X}_i) = \widehat{\Lambda}_0(t, \boldsymbol{X}_i) - \widehat{\mathbf{u}}(t)^T\nabla\mathcal{L}(\widehat{\boldsymbol{\beta}}), \quad \text{where } \widehat{\mathbf{u}}(t) \text{ is defined in (C.5.1).} \tag{C.5.2}$$

Consequently, the conditional survival function can be estimated by $\widetilde{S}(t, \boldsymbol{X}_i) = \exp\{-\widetilde{\Lambda}_0(t, \boldsymbol{X}_i)\}$. The next theorem characterizes the asymptotic normality of $\widetilde{\Lambda}_0(t, \boldsymbol{X}_i)$ and $\widetilde{S}(t, \boldsymbol{X}_i)$. The proof is analogous to the proof to Theorem 4.4.2, which we omit here to avoid repetitions.

Theorem C.5.2. *Suppose Assumptions 4.1.1, 4.1.2, 4.3.1, 4.3.3 and C.5.1 hold,* $\lambda \asymp \sqrt{n^{-1}\log d}$, $\delta \asymp s'\sqrt{n^{-1}\log d}$ *and* $n^{-1/2}s^3\log d = o(1)$. *We have that for any* $t \in [0, \tau]$, *the decorrelated conditional hazard function estimator* $\widetilde{\Lambda}_0(t, \boldsymbol{X}_i)$ *in* (C.5.2) *satisfies*

$$\sqrt{n}\big\{\Lambda_0(t, \boldsymbol{X}_i) - \widetilde{\Lambda}_0(t, \boldsymbol{X}_i)\big\} \xrightarrow{d} Z, \text{ where } Z \sim N\big(0, \sigma_1^2(t) + \sigma_2^2(t)\big),$$

where

$$\sigma_1^2(t) = \int_0^t \frac{\lambda_0(u, \boldsymbol{X}_i)du \cdot \exp(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)}{\mathbb{E}\big\{\exp(\boldsymbol{X}^T\boldsymbol{\beta}^*)Y(u)\big\}}, \quad \text{and } \sigma_2^2(t) = \nabla\Lambda_0(t, \boldsymbol{X}_i)^T\mathbf{H}^{*-1}\nabla\Lambda_0(t, \boldsymbol{X}_i). \tag{C.5.3}$$

The estimated survival function $\widetilde{S}(t, \boldsymbol{X}_i)$ satisfies

$$\sqrt{n}\big\{\widetilde{S}(t, \boldsymbol{X}_i) - S_0(t, \boldsymbol{X}_i)\big\} \xrightarrow{d} Z', \text{ where } Z' \sim N\bigg(0, \frac{\sigma_1^2(t) + \sigma_2^2(t)}{\exp\big\{2\Lambda_0(t, \boldsymbol{X}_i)\big\}}\bigg).$$

Note that, the limiting variance can be estimated by plug-in estimators that

$$\widehat{\sigma}_1^2(t) = \int_0^t \frac{d\widehat{\Lambda}_0(u, \boldsymbol{X}_i)}{n^{-1}\sum_{i'=1}^n \exp\big(\boldsymbol{X}_{i'}^T\widehat{\boldsymbol{\beta}}\big)Y_{i'}(u)} \text{ and } \widehat{\sigma}_2^2(t) = \big\{\nabla\widehat{\Lambda}_0(t, \boldsymbol{X}_i)\big\}^T\widehat{\mathbf{u}}(t).$$

To conclude, based on above Theorem C.5.2, we can conduct valid inference and construct confidence intervals for the conditional hazard function and survival function.

C.6 Technical Lemmas

In this section, we prove some concentration results of the sample gradient $\nabla\mathcal{L}(\boldsymbol{\beta}^*)$ and sample Hessian matrix $\nabla^2\mathcal{L}(\boldsymbol{\beta}^*)$. The mathematical tools we use are mainly from empirical process theory.

We start from introducing the following notations. Let $\|\cdot\|_{\mathbb{P},r}$ denote the $L_r(\mathbb{P})$-norm. For any given $\epsilon > 0$ and the function class $\mathcal{F}$, let $N_{[\,]}\big(\epsilon, \mathcal{F}, L_r(\mathbb{P})\big)$ and $N\big(\epsilon, \mathcal{F}, L_2(\mathbb{Q})\big)$ denote the bracketing number and the covering number, respectively. The quantifies $\log N_{[\,]}\big(\epsilon, \mathcal{F}, L_r(\mathbb{P})\big)$ and $\log N\big(\epsilon, \mathcal{F}, L_2(\mathbb{Q})\big)$ are called entropy with bracketing and entropy. In addition, let F be an envelope of $\mathcal{F}$ where $|f| \leq F$ for all $f \in \mathcal{F}$. The bracketing integral and uniform entropy integral are defined as

$$J_{[\,]}\big(\delta, \mathcal{F}, L_r(\mathbb{P})\big) = \int_0^\delta \sqrt{\log N_{[\,]}\big(\epsilon, \mathcal{F}, L_r(\mathbb{P})\big)}d\epsilon,$$

and

$$J\big(\delta, \mathcal{F}, L_2\big) = \int_0^\delta \sqrt{\log \sup_{\mathbb{Q}} N\big(\epsilon\|F\|_{\mathbb{Q},2}, \mathcal{F}, L_2(\mathbb{Q})\big)}d\epsilon,$$

respectively, where the supremum is taken over all probability measures $\mathbb{Q}$ with $\|F\|_{\mathbb{Q},2} > 0$. Denote the empirical process by $\mathbb{G}_n(f) = n^{1/2}(\mathbb{P}_n - \mathbb{P})(f)$, where $\mathbb{P}_n(f) = n^{-1}\sum_{i=1}^n f(X_i)$ and $\mathbb{P}(f) = \mathbb{E}(f(X_i))$. The following three Lemmas characterize the bounds for the expected maximal empirical processes and the concentration of the maximal empirical processes.

Lemma C.6.1. *Under Assumptions 4.1.1, 4.1.2, 4.3.1, 4.3.2 and 4.3.3, there exist some constant $C > 0$, such that, for $r = 0, 1, 2$, with probability at least $1 - \mathcal{O}(d^{-3})$,*

$$\sup_{t\in[0,\tau]} \|\mathbf{s}^{(r)}(t, \boldsymbol{\beta}^*) - S^{(r)}(t, \boldsymbol{\beta}^*)\|_\infty \leq C\sqrt{\frac{\log d}{n}},$$

where $\mathbf{s}^{(r)}(t, \boldsymbol{\beta}^*)$ and $S^{(r)}(t, \boldsymbol{\beta}^*)$ are defined in (4.1.6) *and* (4.1.3).

Proof. We will only prove the case for $r = 1$, and the cases for $r = 0$ and 2 follow by the similar argument. For $j = 1, ..., d$, let

$$E_j = \sup_{t\in[0,\tau]} |S_j^{(1)}(t, \boldsymbol{\beta}^*) - s_j^{(1)}(t, \boldsymbol{\beta}^*)|,$$

where $S_j^{(1)}(t, \boldsymbol{\beta}^*)$ and $s_j^{(1)}(t, \boldsymbol{\beta}^*)$ denote the j-th component of $S^{(1)}(t, \boldsymbol{\beta}^*)$ and $s^{(1)}(t, \boldsymbol{\beta}^*)$, respectively. We will prove a concentration result of E_j.

First, we show the class of functions $\{X_j(t)Y(t)\exp\left(\boldsymbol{X}^T(t)\boldsymbol{\beta}^*\right) : \ t \in [0, \tau]\}$ has bounded uniform entropy integral. By Lemma 9.10 of Kosorok (2007), the class $\mathcal{F} = \{X_j(t) : \ t \in [0, \tau]\}$ is a VC-hull class associated with a VC class of index 2. By Corollary 2.6.12 of van der Vaart and Wellner (1996), the entropy of the class $\mathcal{F}$ satisfies $\log N(\epsilon\|F\|_{Q,2}, \mathcal{F}, L_2(\mathbb{Q})) \leq C'(1/\epsilon)$ for some constant $C' > 0$, and hence $\mathcal{F}$ has the uniform entropy integral $J(1, \mathcal{F}, L_2) \leq \int_0^1 \sqrt{K(1/\epsilon)}d\epsilon < \infty$. By the same argument, we have that $\left\{\exp\{\boldsymbol{X}(t)^T\boldsymbol{\beta}^*\} : t \in [0, \tau]\right\}$ also has a uniform entropy integral. Meanwhile, by example 19.16 of van der Vaart and Wellner (1996), $\{Y(t) : \ t \in [0, \tau]\}$ is a VC class and hence has bounded uniform entropy integral. Thus, by Theorem 9.15 of Kosorok (2007), we have $\left\{X_j(t)Y(t)\exp\{\boldsymbol{X}(t)^T\boldsymbol{\beta}^*\} : \ t \in [0, \tau]\right\}$ has bounded uniform entropy integral.

Next, taking the envelop F as $\sup_{t\in[0,\tau]}|X_j(t)Y(t)\exp\{\boldsymbol{X}^T(t)\boldsymbol{\beta}^*\}|$, by Lemma 19.38 of van der Vaart (2000),

$$\mathbb{E}(E_j) \le C_1 n^{-1/2} J(1, \mathcal{F}, L_2)\|F\|_{\mathbb{P},2} \le C n^{-1/2},$$

for some positive constants C_1 and C. By McDiarmid's inequality, we have, for any $\Delta > 0$,

$$\mathbb{P}\big(E_j \ge Cn^{-1/2}(1+\Delta)\big) \le \mathbb{P}\big(E_j \ge \mathbb{E}(E_j) + n^{-1/2}C\Delta\big) \le \exp(-C_2\Delta^2 L^{-2}),$$

for some positive constant C_2 and L, and the desired result follows by taking $\Delta = \sqrt{n^{-1}\log d}$ a union bound over $j = 1, ..., d$. □

Lemma C.6.2. *Suppose the Assumptions 4.1.1, 4.1.2, 4.3.1, 4.3.2 and 4.3.3 hold, and* $\lambda \asymp \sqrt{n^{-1}\log d}$. *We have, for* $r = 0, 1, 2$ *and* $t \in [0, \tau]$,

$$\|S^{(r)}(t, \widehat{\boldsymbol{\beta}}) - S^{(r)}(t, \boldsymbol{\beta}^*)\|_\infty = \mathcal{O}_{\mathbb{P}}\Big(s\sqrt{\frac{\log d}{n}}\Big).$$

Proof. Similar to the previous Lemma, we only prove the case for $r = 1$, and the other two cases follow by the similar argument. For the case $r = 1$, we have

$$\begin{aligned}
\|S^{(1)}(t, \widehat{\boldsymbol{\beta}}) - S^{(1)}(t, \boldsymbol{\beta}^*)\|_\infty &= \Big\|\frac{1}{n}\sum_{i=1}^n Y_i(t)\big[\exp\{\boldsymbol{X}_i^T(t)\widehat{\boldsymbol{\beta}}\} - \exp\{\boldsymbol{X}_i^T(t)\boldsymbol{\beta}^*\}\big]\boldsymbol{X}_i(t)\Big\|_\infty \\
&\le \max_i \big\{Y_i(t)\|\boldsymbol{X}_i(t)\|_\infty\big|\exp\{\boldsymbol{X}_i^T(t)\widehat{\boldsymbol{\beta}}\} - \exp\{\boldsymbol{X}_i^T(t)\boldsymbol{\beta}^*\}\big|\big\} \\
&\le C_X \cdot \max_i \big|\exp\{\boldsymbol{X}_i^T(t)\boldsymbol{\beta}^*\}\big[\exp\{\boldsymbol{X}_i^T(t)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\} - 1\big]\big| \qquad \text{(C.6.1)} \\
&\le C_X \cdot C_1 \cdot \max_i \|\boldsymbol{X}_i(t)\|_\infty \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \qquad \text{(C.6.2)} \\
&= \mathcal{O}_{\mathbb{P}}\Big(s\sqrt{\frac{\log d}{n}}\Big),
\end{aligned}$$

where (C.6.1) holds by the Assumption 4.1.1 for some constant $C_X > 0$; (C.6.2) holds by Assumption 4.3.1 that $\boldsymbol{X}_i^T(t)\boldsymbol{\beta}^* = \mathcal{O}(1)$ and $\exp(|x|) \leq 1+2|x|$ for any $|x|$ sufficiently small, and the last equality holds by (4.1.2). Our claim holds as desired. □

Lemma C.6.3. *Under Assumptions 4.1.1, 4.1.2, 4.3.1, 4.3.2 and 4.3.3, for any $1 \leq j, k \leq d$, there exists a positive constant C, such that with probability at least $1 - \mathcal{O}(d^{-1})$,*

$$\max_{j,k=1,\ldots,d} |\nabla^2_{jk}\mathcal{L}(\boldsymbol{\beta}^*) - \mathbf{H}^*_{jk}| \leq C\sqrt{\frac{\log d}{n}}. \tag{C.6.3}$$

Proof. By the definitions of $\nabla^2\mathcal{L}(\boldsymbol{\beta}^*)$ and $\mathbf{H}^*$ in (4.1.5) and (4.1.7), we have

$$\begin{aligned}
\nabla^2\mathcal{L}(\boldsymbol{\beta}^*) - \mathbf{H}^* = & \underbrace{\frac{1}{n}\int_0^\tau \left\{\frac{S^{(2)}(t,\boldsymbol{\beta}^*)}{S^{(0)}(t,\boldsymbol{\beta}^*)} - \frac{\mathbf{s}^{(2)}(t,\boldsymbol{\beta}^*)}{\mathbf{s}^{(0)}(t,\boldsymbol{\beta}^*)}\right\} d\bar{N}(t)}_{T_1} \\
& + \underbrace{\frac{1}{n}\int_0^\tau \frac{\mathbf{s}^{(2)}(t,\boldsymbol{\beta}^*)}{\mathbf{s}^{(0)}(t,\boldsymbol{\beta}^*)} d\bar{N}(t) - \mathbb{E}\Big[\int_0^\tau \frac{\mathbf{s}^{(2)}(t,\boldsymbol{\beta}^*)}{\mathbf{s}^{(0)}(t,\boldsymbol{\beta}^*)} dN(t)\Big]}_{T_2} \\
& + \underbrace{\frac{1}{n}\int_0^\tau \left\{\mathbf{e}(t,\boldsymbol{\beta}^*)^{\otimes 2} - \bar{\boldsymbol{Z}}(t,\boldsymbol{\beta}^*)^{\otimes 2}\right\} d\bar{N}(t)}_{T_3} \\
& + \underbrace{\mathbb{E}\Big[\int_0^\tau \mathbf{e}(t,\boldsymbol{\beta}^*)^{\otimes 2} dN(t)\Big] - \frac{1}{n}\int_0^\tau \mathbf{e}(t,\boldsymbol{\beta}^*)^{\otimes 2} d\bar{N}(t)}_{T_4}.
\end{aligned}$$

For the term T_1, we have, with probability at least $1 - \mathcal{O}(d^{-1})$,

$$\|T_1\|_\infty \leq \sup_{t\in[0,\tau]} \Big\|\frac{S^{(2)}(t,\boldsymbol{\beta}^*)}{S^{(0)}(t,\boldsymbol{\beta}^*)} - \frac{\mathbf{s}^{(2)}(t,\boldsymbol{\beta}^*)}{\mathbf{s}^{(0)}(t,\boldsymbol{\beta}^*)}\Big\|_\infty \cdot \frac{1}{n}\int_0^\tau d\bar{N}(t) \leq C_1\sqrt{\frac{\log d}{n}},$$

where the last inequality follows by Lemma C.6.1. Next, by Assumption 4.1.1, we have

$$\Big\|\frac{\mathbf{s}^{(2)}(t,\boldsymbol{\beta}^*)}{\mathbf{s}^{(0)}(t,\boldsymbol{\beta}^*)}\Big\|_\infty < \infty.$$

Consequently, T_2 becomes an i.i.d. sum of mean 0 bounded random variables. Hoeffding's inequality gives that with probability at least $1-\mathcal{O}(d^{-1})$, $\|T_2\|_\infty \leq C_2\sqrt{n^{-1}\log d}$. Meanwhile, the terms T_3 and T_4 can be bounded similarly. Our claim holds as desired. □

Lemma C.6.4. *Under Assumptions 4.1.1, 4.1.2 4.3.1, 4.3.2 and 4.3.3, it holds that*

$$\|\nabla^2_{\alpha\boldsymbol{\theta}}\mathcal{L}(\widehat{\boldsymbol{\beta}}) - \mathbf{w}^{*T}\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\widehat{\boldsymbol{\beta}})\|_\infty = \mathcal{O}_{\mathbb{P}}\Big(s\sqrt{\frac{\log d}{n}}\Big).$$

Proof. By triangle inequality, we have

$$\begin{aligned}
&\|\nabla^2_{\alpha\boldsymbol{\theta}}\mathcal{L}(\widehat{\boldsymbol{\beta}}) - \mathbf{w}^{*T}\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\widehat{\boldsymbol{\beta}})\|_\infty \\
\leq\ &\underbrace{\|\mathbf{H}^*_{\alpha\boldsymbol{\theta}} - \mathbf{w}^{*T}\mathbf{H}^*_{\boldsymbol{\theta}\boldsymbol{\theta}}\|_\infty}_{E_1} + + \underbrace{\|\nabla^2_{\boldsymbol{\theta}\alpha}\mathcal{L}(\widehat{\boldsymbol{\beta}}) - \mathbf{H}^*_{\boldsymbol{\theta}\alpha}\|_\infty}_{E_2} + \underbrace{\|\mathbf{w}^{*T}\big\{\mathbf{H}^*_{\boldsymbol{\theta}\boldsymbol{\theta}} - \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\widehat{\boldsymbol{\beta}})\big\}\|_\infty}_{E_3}.
\end{aligned}$$

It is seen that $E_1 = 0$ by the definition of $\mathbf{w}^* = \mathbf{H}^{*-1}_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathbf{H}^*_{\boldsymbol{\theta}\alpha}$ in (4.2.1). In addition, $E_2 = \mathcal{O}_{\mathbb{P}}(s\sqrt{n^{-1}\log d})$ by Lemma C.1.3. For the term E_3, we have

$$E_3 \leq \underbrace{\|\mathbf{w}^{*T}\big\{\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\widehat{\boldsymbol{\beta}}) - \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\beta}^*)\big\}\|_\infty}_{E_{31}} + \underbrace{\|\mathbf{w}^{*T}\big\{\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\beta}^*) - \mathbf{H}^*_{\boldsymbol{\theta}\boldsymbol{\theta}}\big\}\|_\infty}_{E_{32}}.$$

For the term E_{31}, by the definition of $\nabla^2\mathcal{L}(\cdot)$ in (4.1.5), we have

$$\begin{aligned}
\mathbf{w}^{*T}\big\{\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\widehat{\boldsymbol{\beta}}) - \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\beta}^*)\big) =\ &\underbrace{\mathbf{w}^{*T}\Big\{\frac{1}{n}\sum_{i=1}^n\int_0^\tau \frac{S^{(2)}\big(t,\widehat{\boldsymbol{\beta}}\big)}{S^{(0)}\big(t,\widehat{\boldsymbol{\beta}}\big)} - \frac{S^{(2)}\big(t,\boldsymbol{\beta}^*\big)}{S^{(0)}\big(t,\boldsymbol{\beta}^*\big)}dN_i(t)\Big\}_{\boldsymbol{\theta}\boldsymbol{\theta}}}_{T_1} \\
&+ \underbrace{\mathbf{w}^{*T}\Big\{\frac{1}{n}\sum_{i=1}^n\int_0^\tau \bar{\boldsymbol{Z}}\big(t,\widehat{\boldsymbol{\beta}}\big)^{\otimes 2} - \bar{\boldsymbol{Z}}\big(t,\boldsymbol{\beta}^*\big)^{\otimes 2}\Big\}_{\boldsymbol{\theta}\boldsymbol{\theta}}}_{T_2}.
\end{aligned}$$

For the term T_1, we have

$$T_1 = \frac{1}{n}\sum_{i=1}^n\int_0^\tau \frac{S^{(0)}\big(t,\boldsymbol{\beta}^*\big)\mathbf{w}^{*T}S^{(2)}_{\boldsymbol{\theta}\boldsymbol{\theta}}\big(t,\widehat{\boldsymbol{\beta}}\big) - S^{(0)}\big(t,\widehat{\boldsymbol{\beta}}\big)\mathbf{w}^{*T}S^{(2)}_{\boldsymbol{\theta}\boldsymbol{\theta}}\big(t,\boldsymbol{\beta}^*\big)}{S^{(0)}\big(t,\widehat{\boldsymbol{\beta}}\big)S^{(0)}\big(t,\boldsymbol{\beta}^*\big)}$$

For ease of notation, in the rest of the proof, let $\widehat{S}^{(r)}(t) := S^{(r)}\big(t, \widehat{\boldsymbol{\beta}}\big)$ and $S^{*(r)}(t) := S^{(r)}\big(t, \boldsymbol{\beta}^*\big)$ for $r = 0, 1, 2$. We have, for the k-th component of T_1,

$$
\begin{aligned}
T_{1,k} = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \frac{S^{*(0)}(t)\frac{1}{n}\sum_{i'=1}^{n} y_{i'}(t)\exp\{\boldsymbol{X}_{i'}^T(t)\widehat{\boldsymbol{\beta}}\}\mathbf{w}^{*T}\boldsymbol{X}_{i',\boldsymbol{\theta}}(t)X_{i',k}(t)}{\widehat{S}^{(0)}(t)S^{*(0)}(t)}dN_i(t) \\
- \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \frac{\widehat{S}^{(0)}(t)\sum_{i'=1}^{n} y_{i'}(t)\exp\{\boldsymbol{X}_{i'}^T(t)\boldsymbol{\beta}^*\}\mathbf{w}^{*T}\boldsymbol{X}_{i',\boldsymbol{\theta}}(t)X_{i',k}(t)}{\widehat{S}^{(0)}(t)S^{*(0)}(t)}dN_i(t).
\end{aligned}
$$

Consequently, it holds that

$$
\begin{aligned}
&|T_{1,k}| \\
&\le \Big|\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \frac{\big\{S^{*(0)}(t) - \widehat{S}^{(0)}(t)\big\}\frac{1}{n}\sum_{i'=1}^{n} Y_{i'}(t)\exp\{\boldsymbol{X}_{i'}^T(t)\widehat{\boldsymbol{\beta}}\}\mathbf{w}^{*T}\boldsymbol{X}_{i',\boldsymbol{\theta}}(t)X_{i',k}(t)}{\widehat{S}^{(0)}(t)S^{*(0)}(t)}dN_i(t)\Big| \\
&+ \Big|\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \frac{\widehat{S}^{(0)}(t)\frac{1}{n}\sum_{i'=1}^{n} Y_{i'}(t)\big[\exp\{\boldsymbol{X}_{i'}^T(t)\widehat{\boldsymbol{\beta}}\} - \exp\{\boldsymbol{X}_{i'}^T(t)\boldsymbol{\beta}^*\}\big]\mathbf{w}^{*T}\boldsymbol{X}_{i',\boldsymbol{\theta}}(t)X_{i',k}(t)}{\widehat{S}^{(0)}(t)S^{*(0)}(t)}dN_i(t)\Big| \\
&\le \sup_{t\in[0,\tau]}\Big|\frac{1}{n}\sum_{i=1}^{n}\frac{\big\{S^{*(0)}(t) - \widehat{S}^{(0)}(t)\big\}\big[\frac{1}{n}\sum_{i'=1}^{n} Y_{i'}(t)\exp\{\boldsymbol{X}_{i'}^T(t)\boldsymbol{\beta}^*\}\mathbf{w}^{*T}\boldsymbol{X}_{i',\boldsymbol{\theta}}(t)X_{i',k}(t)\big]}{\widehat{S}^{(0)}(t)S^{*(0)}(t)}\Big|\cdot\tau \\
&+ \Big|\frac{1}{n}\sum_{i=1}^{n}\frac{\big\{S^{*(0)}(t) - \widehat{S}^{(0)}(t)\big\}\Big[\frac{1}{n}\sum_{i'=1}^{n} Y_{i'}(t)\big[\exp\{\boldsymbol{X}_{i'}^T(t)\widehat{\boldsymbol{\beta}}\} - \exp\{\boldsymbol{X}_{i'}^T(t)\boldsymbol{\beta}^*\}\big]\mathbf{w}^{*T}\boldsymbol{X}_{i',\boldsymbol{\theta}}(t)X_{i',k}(t)\Big]}{\widehat{S}^{(0)}(t)S^{*(0)}(t)}\Big|\cdot\tau \\
&+ \frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{S}^{(0)}(t)\frac{1}{n}\sum_{i'=1}^{n} Y_{i'}(t)\big[\exp\{\boldsymbol{X}_{i'}^T(t)\widehat{\boldsymbol{\beta}}\} - \exp\{\boldsymbol{X}_{i'}^T(t)\boldsymbol{\beta}^*\}\big]\mathbf{w}^{*T}\boldsymbol{X}_{i',\boldsymbol{\theta}}(t)X_{i',k}(t)}{\widehat{S}^{(0)}(t)S^{*(0)}(t)}\cdot\tau \\
&= \mathcal{O}_{\mathbb{P}}(s\sqrt{n^{-1}\log d}),
\end{aligned}
$$

where the last equality holds by Assumptions 4.1.1 and 4.3.1 that $\boldsymbol{X}_i^T(t)\boldsymbol{\beta}^*$ is bounded, $S^{*(0)}(t)$ is bounded away from 0, and by Lemma C.6.2 that $|\widehat{S}^{(r)}(t) - S^{*(r)}(t)| = \mathcal{O}_{\mathbb{P}}(s\sqrt{n^{-1}\log d})$.

The term T_2 can be bounded by the similar argument, and our claim holds as desired. □

Lemma C.6.5. *Under Assumptions 4.1.1 and 4.1.2, and if $n^{-1/2}s^3\log d = o(1)$, the RE condition holds for the sample Hessian matrix $\nabla^2\mathcal{L}(\widehat{\boldsymbol{\beta}})$. Specifically, for the vectors in the*

cone $\mathcal{C} = \{\mathbf{v} | \|\mathbf{v}_{\mathcal{S}}\|_1 \leq \xi \|\mathbf{v}_{\mathcal{S}^C}\|_1\}$*, we have*

$$\frac{\mathbf{v}^T \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\beta}}) \mathbf{v}}{\|\mathbf{v}\|_2} \geq \frac{1}{2} \kappa^2 \big(\xi, |\mathcal{S}|; \nabla^2 \mathcal{L}(\boldsymbol{\beta}^*)\big), \text{ for all } \mathbf{v} \in \mathcal{C}.$$

Proof. By Lemma 3.2 of Huang et al. (2013), we have $\exp(-2\xi_{\mathbf{b}}) \nabla^2 \mathcal{L}(\boldsymbol{\beta}) \preceq \nabla^2 \mathcal{L}(\boldsymbol{\beta} + \mathbf{b})$, where $\xi_{\mathbf{b}} = \max_{u \geq 0} \max_{i,i',k,k'} |\mathbf{b}^T \{\boldsymbol{X}_{ik}(u) - \boldsymbol{X}_{i'k'}(u)\}|$. Let $\mathbf{b} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$. By Assumption 4.1.1 that $\|\{\boldsymbol{X}_{ik}(u) - \boldsymbol{X}_{i'k'}(u)\}\|_\infty \leq C_X$, we have $\xi_{\mathbf{b}} = \mathcal{O}_{\mathbb{P}}(s\sqrt{n^{-1} \log d})$ by (4.1.2), we have $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = \mathcal{O}_{\mathbb{P}}(s\lambda)$. By the scaling assumption that $n^{-1/2} s^3 \log d = o(1)$, we have $\xi_{\mathbf{b}} \leq \frac{1}{2} \log 2$. Consequently, $\exp(-2\xi_{\mathbf{b}}) \geq 1/2$. We have $\nabla^2 \mathcal{L}(\widehat{\boldsymbol{\beta}}) \succeq \frac{1}{2} \cdot \nabla^2 \mathcal{L}(\boldsymbol{\beta}^*)$. Since the cone $\mathcal{C}$ is a subset of $\mathbb{R}^d$, our claim follows as desired. □

Lemma C.6.6. *Under Assumptions 4.1.1, 4.1.2, 4.3.1, 4.3.2 and 4.3.3, if*

$$\|\nabla^2_{\boldsymbol{\theta}\alpha} \mathcal{L}(\widehat{\boldsymbol{\beta}}) - \mathbf{w}^{*T} \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(\widehat{\boldsymbol{\beta}})\|_\infty \leq \lambda', \tag{C.6.4}$$

we have, the Dantzig selector $\widehat{\mathbf{w}}$ *defined in* (4.2.2) *satisfies*

$$\|\widehat{\mathbf{w}} - \mathbf{w}^*\|_1 \leq \frac{16 \lambda' s'}{\kappa^2(1, s'; \nabla^2 \mathcal{L}(\boldsymbol{\beta}^*))}.$$

Proof. We first derive the result that the vector $\widehat{\Delta} = \widehat{\mathbf{w}} - \mathbf{w}^*$ belongs to the cone $\mathcal{C} = \{\mathbf{v} | \|\mathbf{v}_{\mathcal{S}^C}\|_1 \leq \|\mathbf{v}_{\mathcal{S}}\|_1\}$. By our assumption (C.6.4), and since $\|\widehat{\mathbf{w}}\|_1 \leq \|\mathbf{w}^*\|_1$ by the optimality condition of Dantzig selector, we have

$$\|\widehat{\mathbf{w}}_{\mathcal{S}}\|_1 + \|\widehat{\mathbf{w}}_{\mathcal{S}^C}\|_1 \leq \|\mathbf{w}^*_{\mathcal{S}}\|_1,$$

where we use the fact that $\|\mathbf{w}^*_{\mathcal{S}^C}\|_1 = 0$.

By triangle inequality, we have

$$\|\mathbf{w}^*_{\mathcal{S}}\|_1 \leq \|\widehat{\mathbf{w}}_{\mathcal{S}}\|_1 + \|\widehat{\Delta}_{\mathcal{S}}\|_1.$$

Summing up the above two inequalities, we have

$$\|\widehat{\Delta}_{\mathcal{S}^C}\|_1 \leq \|\widehat{\Delta}_{\mathcal{S}}\|_1. \tag{C.6.5}$$

Meanwhile, by the feasibility conditions of the Dantzig selector $\widehat{\mathbf{w}}$ and $\mathbf{w}^*$, we have

$$\|\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\widehat{\boldsymbol{\beta}})\widehat{\Delta}\|_\infty \leq \|\nabla^2_{\boldsymbol{\theta}\alpha}\mathcal{L}(\widehat{\boldsymbol{\beta}}) - \mathbf{w}^{*T}\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\widehat{\boldsymbol{\beta}})\|_\infty + \|\nabla^2_{\boldsymbol{\theta}\alpha}\mathcal{L}(\widehat{\boldsymbol{\beta}}) - \widehat{\mathbf{w}}\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\widehat{\boldsymbol{\beta}})\|_\infty \leq 2\lambda'. \tag{C.6.6}$$

By (C.6.5) and (C.6.6), we have

$$\widehat{\Delta}^T\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\widehat{\boldsymbol{\beta}})\widehat{\Delta} \leq \|\widehat{\Delta}\|_1\|\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\widehat{\boldsymbol{\beta}})\widehat{\Delta}\|_\infty \leq 2\lambda'\|\widehat{\Delta}\|_1 \leq 4\lambda'\|\widehat{\Delta}_{\mathcal{S}}\|_1.$$

By Lemma C.6.5, it holds that

$$\widehat{\Delta}^T\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\widehat{\boldsymbol{\beta}})\widehat{\Delta} \geq \frac{1}{2}\kappa^2(1, s'; \nabla^2\mathcal{L}(\boldsymbol{\beta}^*))\|\widehat{\Delta}_{\mathcal{S}}\|_2^2,$$

which implies that

$$\widehat{\Delta}^T\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\widehat{\boldsymbol{\beta}})\widehat{\Delta} \geq \frac{1}{2}\kappa^2(1, s'; \nabla^2\mathcal{L}(\boldsymbol{\beta}^*))s'^{-1}\|\widehat{\Delta}_{\mathcal{S}}\|_2^1.$$

Consequently, we have

$$\|\widehat{\Delta}_{\mathcal{S}}\|_1 \leq \frac{8\lambda' s'}{\kappa^2(1, s'; \nabla^2\mathcal{L}(\boldsymbol{\beta}^*))}.$$

By (C.6.5), it holds that

$$\|\widehat{\Delta}\|_1 \leq 2\|\widehat{\Delta}_{\mathcal{S}}\|_1 \leq \frac{16\lambda' s'}{\kappa^2(1, s'; \nabla^2\mathcal{L}(\boldsymbol{\beta}^*))}$$

as desired. □

C.7 Proof of Some Technical Lemmas

Proof of Lemma C.1.4. As shown in Lemma C.6.4, under Assumptions 4.1.1, 4.1.2, 4.3.1, 4.3.2 and 4.3.3, the feasibility condition (C.6.4), $\|\nabla^2_{\boldsymbol{\theta}\alpha}\mathcal{L}(\widehat{\boldsymbol{\beta}}) - \mathbf{w}^{*T}\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\widehat{\boldsymbol{\beta}})\|_\infty \leq \lambda'$, is satisfied for $\lambda' \asymp s'\sqrt{n^{-1}\log d}$. By Lemma C.6.6, we have

$$\|\widehat{\mathbf{w}} - \mathbf{w}^*\|_1 = \mathcal{O}_{\mathbb{P}}(s's\sqrt{n^{-1}\log d}),$$

which concludes the proof. □

Proof of Lemma C.1.2. By definition, we have, for all $j = 1, ..., d$,

$$\begin{aligned}\nabla_j\mathcal{L}(\boldsymbol{\beta}^*) &= -\frac{1}{n}\sum_{i=1}^n\int_0^\tau \big\{X_{ij}(u,\boldsymbol{\beta}^*) - \bar{Z}_j(u,\boldsymbol{\beta}^*)\big\}dM_i(u)\\ &= \frac{1}{n}\sum_{i=1}^n\int_0^\tau \bar{Z}_j(u,\boldsymbol{\beta}^*)dM_i(u) - \frac{1}{n}\sum_{i=1}^n\int_0^\tau X_{ij}(u,\boldsymbol{\beta}^*)dM_i(u). \end{aligned} \tag{C.7.1}$$

For the first term, we have for all $t \in [0,\tau]$,

$$\bar{Z}_j(t,\boldsymbol{\beta}^*) - e_j(t,\boldsymbol{\beta}^*) = \frac{S_j^{(1)}(t,\boldsymbol{\beta}^*) - s_j^{(1)}(t,\boldsymbol{\beta}^*)}{S^{(0)}(t,\boldsymbol{\beta}^*)} - \frac{s_j^{(1)}(t,\boldsymbol{\beta}^*)\big\{S^{(0)}(t,\boldsymbol{\beta}^*) - s^{(0)}(t,\boldsymbol{\beta}^*)\big\}}{S^{(0)}(t,\boldsymbol{\beta}^*)s^{(0)}(t,\boldsymbol{\beta}^*)}. \tag{C.7.2}$$

By Assumption 4.1.1 and the fact that $\mathbb{P}(y(\tau) > 0) > 0$, we have that $\sup_{t\in[0,\tau]}|\bar{Z}_j(t,\boldsymbol{\beta}^*) - e_j(t)| \leq C_1$ for some constant $C_1 > 0$. In addition,

$$\frac{1}{n}\sum_{i=1}^n\int_0^\tau \bar{Z}_j(u,\boldsymbol{\beta}^*)dM_i(u) \leq \sup_{f\in\mathcal{F}_j}\frac{1}{n}\sum_{i=1}^n\int_0^\tau f(u)dM_i(u),$$

where $\mathcal{F}_j$ denotes the class of functions $f : [0,\tau] \to \mathbb{R}$ which have uniformly bounded variation and satisfy $\sup_{t\in[0,\tau]}|f(t) - e_j(t)| \leq \delta_1$ for some δ_1. By constructing ℓ_∞ balls centered at piecewise constant functions on a regular grid, one can show that the covering number of

the class $\mathcal{F}_j$ satisfies $N(\epsilon, \mathcal{F}_j, \ell_\infty) \leq (C_2\epsilon^{-1})^{C_3\epsilon^{-1}}$ for some positive constants C_2, C_3. Let $\mathcal{G}_j = \{\int_0^\infty f(t)dM(t) : f \in \mathcal{F}_j\}$. Note that for any two $f_1,\ f_2 \in \mathcal{F}_j$,

$$\Big| \int_0^\tau f_1(t) - f_2(t)dM(t) \Big| \leq \sup_{u\in[0,\tau]} |f_1(u) - f_2(u)| \int_0^\tau |dM(t)|.$$

By Theorem 2.7.11 of van der Vaart and Wellner (1996), the bracketing number of the class $\mathcal{G}_j$ satisfies $N_{[\,]}\big(2\epsilon\|F\|_{\mathbb{P},2}, \mathcal{G}_j, \ell_2(\mathbb{P})\big) \leq N(\epsilon, \mathcal{F}_j, \|\cdot\|_\infty) \leq (C_2\epsilon^{-1})^{C_3\epsilon^{-1}}$, where $F = \int_0^\tau |dM(t)|$. Hence, $\mathcal{G}_j$ has bounded bracketing integral. An application of Corollary 19.35 of van der Vaart (2000) yields that

$$\mathbb{E}\bigg(\sup_{f\in\mathcal{F}_j} \frac{1}{n} \sum_{i=1}^n \int_0^\tau f(u)dM_i(u) \bigg) \leq n^{-1/2} C_4$$

for some constant $C_4 > 0$. Then, by McDiarmid's inequality, for any $c > 0$

$$\mathbb{P}\bigg(\frac{1}{n} \sum_{i=1}^n \int_0^\tau \bar{Z}_j(u, \boldsymbol{\beta}^*)dM_i(u) > c \bigg) \leq \mathbb{P}\bigg(\sup_{f\in\mathcal{F}_j} \frac{1}{n} \sum_{i=1}^n \int_0^\tau f(u)dM_i(u) > c \bigg) \leq \exp\bigg(-\frac{nc^2}{C_5} \bigg),$$

for some constant C_5. Following by the union bound, we have with probability at least $1 - \mathcal{O}(d^{-3})$,

$$\bigg\| \frac{1}{n} \sum_{i=1}^n \int_0^\tau \bar{Z}_j(u, \boldsymbol{\beta}^*)dM_i(u) \bigg\|_\infty \leq C\sqrt{\frac{\log d}{n}}.$$

Note that the second term of (C.7.1) is a sum of i.i.d. mean-zero bounded random variables. Following by the Hoeffding inequality and the union bound, we have with probability at least $1 - \mathcal{O}(d^{-3})$,

$$\bigg\| \frac{1}{n} \sum_{i=1}^n \int_0^\infty X_{ij}(u, \boldsymbol{\beta}^*)dM_i(u) \bigg\|_\infty \leq C\sqrt{\frac{\log d}{n}},$$

for some constant C. The claim follows as desired. □

Proof of Lemma C.1.3. Let $\xi = \max_{u\geq 0}\max_{i,i'}|\boldsymbol{\Delta}^T\{\boldsymbol{X}_i(u) - \boldsymbol{X}_{i'}(u)\}|$, where $\boldsymbol{\Delta} = \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$. By Lemma 3.2 of Huang et al. (2013), it holds that,

$$\exp(-2\xi)\nabla^2\mathcal{L}(\boldsymbol{\beta}^*) \preceq \nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}) \preceq \exp(2\xi)\nabla^2\mathcal{L}(\boldsymbol{\beta}^*), \tag{C.7.3}$$

where $\mathbf{A} \preceq \mathbf{B}$ means that the matrix $\mathbf{B} - \mathbf{A}$ is a positive semidefinite matrix.

Note that the diagonal elements of a positive semidefinite matrix can only be nonnegative. In addition, for a positive semidefinite matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, it is easy to see that $\|\mathbf{A}\|_\infty = \max_j\{a_{jj}\}_{j=1}^d$. We have,

$$\exp(-2\xi)\|\nabla^2\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty \leq \|\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}})\|_\infty \leq \exp(2\xi)\|\nabla^2\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty.$$

By (4.1.2) that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = \mathcal{O}_{\mathbb{P}}(s\lambda)$, which implies that $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = \mathcal{O}(s\lambda)$ as $\widetilde{\boldsymbol{\beta}}$ is on the line segment connecting $\boldsymbol{\beta}^*$ and $\widehat{\boldsymbol{\beta}}$. Hence, $\xi = \mathcal{O}_{\mathbb{P}}(s\lambda)$. By triangle inequality,

$$\|\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}) - \mathbf{H}^*\|_\infty \leq \underbrace{\|\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}) - \nabla^2\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty}_{E_1} + \underbrace{\|\nabla^2\mathcal{L}(\boldsymbol{\beta}^*) - \mathbf{H}^*\|_\infty}_{E_2}.$$

We consider the two terms separately, for the first term E_1, we have, by (C.7.3) and taking the Taylor's expansion of $\exp(2\xi)$,

$$\|\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}) - \nabla^2\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty \leq 2\ \|\xi\nabla^2\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty + o_{\mathbb{P}}(\xi).$$

Since $\xi = \mathcal{O}_{\mathbb{P}}(s\lambda)$, and by Assumption 4.3.3, we have,

$$\|\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}) - \nabla^2\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty = \mathcal{O}_{\mathbb{P}}(s\lambda),$$

and $E_1 = \mathcal{O}_{\mathbb{P}}(s\sqrt{n^{-1}\log d})$ as $\lambda \asymp \sqrt{n^{-1}\log d}$. In addition, $E_2 = \mathcal{O}_{\mathbb{P}}(\sqrt{n^{-1}\log d})$ by Lemma C.6.3. It further implies that $\|\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}})\|_\infty = \mathcal{O}_{\mathbb{P}}(1)$. □

References

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511.

Alizadeh, A. A., Gentles, A. J., Alencar, A. J., Liu, C. L., Kohrt, H. E., Houot, R., Goldstein, M. J., Zhao, S., Natkunam, Y., Advani, R. H., et al. (2011). Prediction of survival in diffuse large b-cell lymphoma based on the expression of 2 genes reflecting tumor and microenvironment. *Blood*, 118(5):1350–1358.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.*, pages 1100–1120.

Antoniadis, A., Fryzlewicz, P., and Letué, F. (2010). The Dantzig selector in Cox's proportional hazards model. *Scand. J. Stat.*, 37(4):531–552.

Arora, S. and Barak, B. (2009). *Computational Complexity: A Modern Approach*. Cambridge University Press.

Bauer, P. (1989). Multistage testing with adaptive designs(with discussion). *Biometrie und Informatik in Medizin und Biologie*, 20:130148.

Bauer, P. and Köhne, K. (1994). Evaluations of experiment s with adaptive interim analyses. *Biometrics*, 50:10291041.

Belloni, A., Chernozhukov, V., and Wei, Y. (2013). Honest confidence regions for logistic regression with a large number of controls. *arXiv preprint arXiv:1304.3969*.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. Springer.

Bergmann, B. and Hommel, G. (1988). Improvements of general multiple test procedures for redundant systems of hypotheses. In Bauer, P., Hommel, G., and Sonnemann, E., editors, *Multiple Hypothesenprüfung/Multiple Hypotheses Testing*. Springer, Berlin.

Bertsekas, D. P. (1999). *Nonlinear Programming*. Athena Scientific.

Boessen, R., van der Baan, F., Groenwold, R., Egberts, A., Klungel, O., Grobbee, D., Knol, M., and Roes, K. (2013a). Optimizing trial design in pharmacogenetics research: comparing a fixed parallel group, group sequential, and adaptive selection design on sample size requirements. *Pharmaceutical Statistics*.

Boessen, R., van der Baan, F., Groenwold, R., Egberts, A., Klungel, O., Grobbee, D., Knol, M., and Roes, K. (2013b). Optimizing trial design in pharmacogenetics research: comparing a fixed parallel group, group sequential, and adaptive selection design on sample size requirements. *Pharmaceutical Statistics*.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford.

Bradic, J., Fan, J., and Jiang, J. (2011). Regularization for Cox's proportional hazards model with NP-dimensionality. *Ann. Statist.*, 39(6):3092–3120.

Brannath, W., Zuber, E., Branson, M., Bretz, F., Gallo, P., Posch, M., and Racine-Poon, A. (2009a). Confirmatory adaptive designs with bayesian decision tools for a targeted therapy in oncology. *Statistics in Medicine*, 28(10):1445–1463.

Brannath, W., Zuber, E., Branson, M., Bretz, F., Gallo, P., Posch, M., and Racine-Poon, A. (2009b). Confirmatory adaptive designs with bayesian decision tools for a targeted therapy in oncology. *Statistics in Medicine*, 28(10):1445–1463.

Bretz, F., Schmidli, H., König, F., Racine, A., and Maurer, W. (2006). Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: General concepts. *Biometrical Journal*, 48(4):623–634.

Bullmore, E. and Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198.

Bullmore, E. T. and Bassett, D. S. (2011). Brain graphs: graphical models of the human brain connectome. *Annual Review of Clinical Psychology*, 7:113–140.

Cai, J., Fan, J., Li, R., and Zhou, H. (2005). Variable selection for multivariate failure time data. *Biometrika*, 92(2):303–316.

Cai, T., Liu, W., and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Stat. Assoc.*, 106(494):594–607.

Cao, L. and Fei-Fei, L. (2007). Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE.

Cox, D. R. (1972). Regression models and life-tables. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 34(2):187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2):269–276.

Di Gaetano, N., Cittera, E., Nota, R., Vecchi, A., Grieco, V., Scanziani, E., Botto, M., Introna, M., and Golay, J. (2003). Complement activation determines the therapeutic activity of rituximab in vivo. *J. Immunol.*, 171(3):1581–1587.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50(272):1096–1121.

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.*, 30(1):74–99.

FDA (2010). Draft guidance for industry. Adaptive design clinical trials for drugs and biologics. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM201790.pdf.

Follmann, D. (1997). Adaptively changing subgroup proportions in clinical trials. *Statistica Sinica*, 7:1085–1102.

Freidlin, B. and Simon, R. (2005). Adaptive signature design: An adaptive clinicaltrial design for generating and prospectivelytesting agene expression signature for sensitive patients. *Clin Cancer Res*, 11:7872–7878.

Friede, T., Parsons, N., and Stallard, N. (2012). A conditional error function approach for subgroup selection in adaptive clinical trials. *Statistics in medicine*, 31(30):4309–4320.

Hampson, L. V. and Jennison, C. (2015). Optimizing the data combination rule for seamless phase II/III clinical trials. *Statistics in Medicine*, 34(1):39–58.

Hiai, H., Tsuruyama, T., and Yamada, Y. (2003). Pre-B lymphomas in SL/Kh mice: A multifactorial disease model. *Cancer Science*, 94(10):847–850.

Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*. Wiley Interscience, New York.

Howard, A., Matarić, M. J., and Sukhatme, G. S. (2002). Mobile sensor network deployment using potential fields: A distributed, scalable solution to the area coverage problem. In *Distributed Autonomous Robotic Systems 5*, pages 299–308. Springer.

Huang, J., Sun, T., Ying, Z., Yu, Y., and Zhang, C.-H. (2013). Oracle inequalities for the Lasso in the Cox model. *Ann. Statist.*, 41(3):1142–1165.

Javanmard, A. and Montanari, A. (2013). Confidence intervals and hypothesis testing for high-dimensional statistical models. In *NIPS*, pages 1187–1195.

Jenkins, M., Stone, A., and Jennison, C. (2011a). An adaptive seamless phase ii/iii design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical Statistics*, 10(4):347–356.

Jenkins, M., Stone, A., and Jennison, C. (2011b). An adaptive seamless phase ii/iii design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical Statistics*, 10(4):347–356.

Jennison, C. and Turnbull, B. W. (2007). Adaptive seamless designs: Selection and prospective testing of hypotheses. *J. Biopharmaceutical Statistics*, pages 1135–1161, doi: 10.1080/10543400701645215.

Kalbfleisch, J. D. and Prentice, R. L. (2011). *The statistical analysis of failure time data*, volume 360. John Wiley & Sons.

Kolar, M. and Liu, H. (2013). Optimal feature selection in high-dimensional discriminant analysis. *arXiv preprint arXiv:1306.6557*.

Kong, S. and Nan, B. (2014). Non-asymptotic oracle inequalities for the high-dimensional Cox regression via Lasso. *Stat. Sinica*, 24:25–42.

Kosorok, M. R. (2007). *Introduction to Empirical Processes and Semiparametric Inference*. Springer.

Lai, T. L., Lavori, P. W., and Liao, O. Y.-W. (2014). Adaptive choice of patient subgroup for comparing two treatments. *Contemporary clinical trials*.

Langendoen, K., Baggio, A., and Visser, O. (2006). Murphy loves potatoes: Experiences from a pilot sensor network deployment in precision agriculture. In — *Proceedings 20th IEEE International Parallel & Distributed Processing Symposium*, page 155. IEEE.

Lee, S. H., Lee, S., Song, H., and Lee, H. S. (2009). Wireless sensor network design for tactical military applications: remote large-scale environments. In *Military Communications Conference, 2009. MILCOM 2009. IEEE*, pages 1–7. IEEE.

Lehmacher, W. and Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics*, 55(4):1286–1290.

Lehmann, E. L. and Scheffé, H. (1950). Completeness, similar regions, and unbiased estimation. i. *Sankhyā*, 10(4):305–340.

Liu, H. and Wang, L. (2012). Tiger: A tuning-insensitive approach for optimally estimating gaussian graphical models. *arXiv preprint arXiv:1209.2437*.

Lockhart, R., Taylor, J., Tibshirani, R. J., Tibshirani, R., et al. (2014). A significance test for the Lasso. *Ann. Statist.*, 42(2):413–468.

Magazine, M. J. and Chern, M.-S. (1984). A note on approximation schemes for multidimensional knapsack problems. *Math. Oper. Res.*, 9(2):244–247.

Marcus, R., Peritz, E., and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63:655–660.

Meierhoff, G., Dehmel, U., Gruss, H., Rosnet, O., Birnbaum, D., Quentmeier, H., Dirks, W., and Drexler, H. (1995). Expression of FLT3 receptor and FLT3-ligand in human leukemia-lymphoma cell lines. *Leukemia*, 9(8):1368–1372.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.*, 34(3):1436–1462.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 72(4):417–473.

Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.*, 37(1):246–270.

Nielsen, K. R., Steffensen, R., Bendtsen, M. D., Rodrigo-Domingo, M., Baech, J., Haunstrup, T. M., Bergkvist, K. S., Schmitz, A., Boedker, J. S., Johansen, P., et al. (2015). Inherited inflammatory response genes are associated with b-cell non-hodgkinÕs lymphoma risk and survival. *PLOS ONE*, 10(10):e0139329.

Ning, Y. and Liu, H. (2014). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *arXiv preprint arXiv:1412.8765*.

Nishiu, M., Yanagawa, R., Nakatsuka, S.-i., Yao, M., Tsunoda, T., Nakamura, Y., and Aozasa, K. (2002). Microarray analysis of gene-expression profiles in diffuse large b-cell lymphoma: Identification of genes related to disease progression. *Cancer Science*, 93(8):894–901.

O'Brien, P. and Fleming, T. (1979). A multiple testing procedure for clinical trials. *Biometrics*, 35:549–556.

Pisinger, D. (1995). A minimal algorithm for the multiple-choice knapsack problem. *Eur. J. Oper. Res.*, 83(2):394–410.

Ravikumar, P., Wainwright, M. J., and Lafferty, J. D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.*, 38(3):1287–1319.

Rosenblum, M., Liu, H., and Yen, E.-H. (2014). Optimal tests of treatment effects for the overall population and two subpopulations in randomized trials, using sparse linear programming. *Journal of the American Statistical Association*, 109(507):1216–1228.

Rosenblum, M. and van der Laan, M. J. (2011). Optimizing randomized trial designs to distinguish which subpopulations benefit from treatment. *Biometrika*, 98(4):845–860.

Russek-Cohen, E. and Simon, R. M. (1997). Evaluating treatments when a gender by treatment interaction may exist. *Statistics in Medicine*, 16:455–464.

Schmidli, H., Bretz, F., Racine, A., and Maurer, W. (2006). Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: applications and practical considerations. *Biometrical Journal*, 48(4):635–643.

Shorack, G. R. and Wellner, J. A. (2009). *Empirical processes with applications to statistics*, volume 59. SIAM.

Stallard, N., Hamborg, T., Parsons, N., and Friede, T. (2014). Adaptive designs for confirmatory clinical trials with subgroup selection. *Journal of biopharmaceutical statistics*, 24(1):168–187.

Starr, R. M. (1969). Quasi-equilibria in markets with non-convex preferences. *Econometrica*, pages 25–38.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, pages 267–288.

Tibshirani, R. (1997). The Lasso method for variable selection in the Cox model. *Stat. Med.*, 16(4):385–395.

Trefethen, L. N. and Bau III, D. (1997). *Numerical Linear Algebra*. Number 50. SIAM.

Tsiatis, A. A. (1981). A large sample study of Cox's regression model. *Ann. Statist.*, 9(1):93–108.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42(3):1166–1202.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer.

Vempala, S. S. (2005). *The Random Projection Method*, volume 65. American Mathematical Society.

Wang, S., Nan, B., Zhu, N., and Zhu, J. (2009a). Hierarchically penalized Cox regression with grouped variables. *Biometrika*, 96(2):307–322.

Wang, S. J., Hung, H., and O'Neill, R. T. (2009b). Adaptive patient enrichment designs in therapeutic trials. *Biometrical Journal*, 51:358–374.

Wang, S. J., O'Neill, R. T., and Hung, H. (2007). Approaches to evaluation of treatment effect in randomized clinical trials with genomic subsets. *Pharmaceut. Statist.*, 6:227–244.

Williamson, D. P. and Shmoys, D. B. (2011). *The Design of Approximation Algorithms*. Cambridge University Press.

Xue, L., Zou, H., and Cai, T. (2012). Nonconcave penalized composite conditional likelihood estimation of sparse ising models. *Ann. Statist.*, 40(3):1403–1429.

Yick, J., Mukherjee, B., and Ghosal, D. (2008). Wireless sensor network survey. *Computer Networks*, 52(12):2292–2330.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 76(1):217–242.

Zhang, H. H. and Lu, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika*, 94(3):691–703.

Zhang, T. (2009). On the consistency of feature selection using greedy least squares regression. *J. Mach. Learn. Res.*, 10:555–568.

Zhang, Y., Wainwright, M. J., and Jordan, M. I. (2014). Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *Proceedings of Annual Conference on Learning Theory*.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7:2541–2563.

Zhao, S. D. and Li, Y. (2012). Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *J. Multivariate. Anal.*, 105(1):397–411.

Zhong, P.-S., Hu, T., and Li, J. (2015). Tests for coefficients in high-dimensional additive hazard models. *Scandinavian Journal of Statistics*, 42:649–664.