CONSTRUCTING CONFIDENCE INTERVALS IN HIGH-DIMENSIONAL MODELS AND DEALING WITH PLEIOTROPY IN MENDELIAN RANDOMIZATION

BY SAI LI

A dissertation submitted to the
School of Graduate Studies
Rutgers, The State University of New Jersey
In partial fulfillment of the requirements
For the degree of
Doctor of Philosophy
Graduate Program in Statistics
Written under the direction of
Cun-Hui Zhang and Steven Buyske
And approved by

New Brunswick, New Jersey
MAY, 2018

ABSTRACT OF THE DISSERTATION

Constructing Confidence Intervals in High-Dimensional Models and Dealing with Pleiotropy in Mendelian Randomization

by Sai Li

Dissertation Director: Cun-Hui Zhang and Steven Buyske

Constructing confidence intervals in high-dimensional models is a challenging task due to the lack of knowledge on the distribution of many regularized estimators. The debiased Lasso approach (Zhang and Zhang, 2014) has been proposed for constructing confidence intervals of low-dimensional parameters in high-dimensional linear models. This thesis generalizes the idea of "debiasing" to make inference in high-dimensional Cox models with time-dependent covariates. A quadratic optimization algorithm is proposed for computing the debiased Lasso estimator and its benefits are demonstrated. This thesis also studies the sample size conditions for inference in high-dimensional linear models with bootstrapped debiased Lasso. It is proved that bootstrap can further correct the bias of debiased Lasso and new sample size conditions involving the number of weak signals are obtained.

In many economical and biological applications, estimating the causal effect of an exposure on an outcome is an important task. Mendelian Randomization, in particular, uses genetic variants as instruments to estimate causal effects in epidemiological studies. However, when there exist pleiotropic effects, conventional instrumental variable methods can be biased. Theoretical properties of Bayes estimators induced by single and mixture Gaussian priors are studied in the existence of pleiotropy. The methods under consideration are generalized to deal with summarized data and demonstrated in various simulation settings and on two real datasets.

Acknowledgements

I am deeply indebted to my advisors, Professor Cun-Hui Zhang and Professor Steven Buyske, for their help, guidance and care. As I started my research, Professor Zhang has taught me many valuable research experiences, from technical details to generating new ideas. His seriousness and high standards on research has immense influence on me. Without his sharing his time and ideas, I cannot complete my PhD study in such a smooth way. Professor Buyske always helps me broaden the horizon, encourages me to explore new problems and provide instructions on my writing and presentations patiently. His encouragement and support gets me through many difficulties. I cannot be more fortunate to have Professor Zhang and Professor Buyske as my advisors. Their passion, intelligence and commitment to statistics are what I would pursue in my future academic career.

I really appreciate the support and help I received from the whole statistics department at Rutgers. Especially, I would like to thank Professors Zijian Guo, John Kolassa, Regina Y. Liu, Harold B. Sackrowitz, William Edward Strawderman and Zhiqiang Tan for their constructive advices and kind help on various aspects of my PhD study.

I cannot be more grateful to my parents, who always care about me, support me and love me deeply. During my study in U.S., my mother has suffered through a serious health problem. I felt so sorry that I cannot go back to stay with her and take care of her during that terrible period of time. Fortunately, she has survived it with her strong will and positivity as well as with the best care from my family. As always, my family are my advancing motivations.

Table of Contents

Abstract ii

Acknowledgements iv

List of Figures viii

List of Tables x

1. Introduction 1

1.1. High-Dimensional inference 2

1.1.1. Debiased Lasso 3

1.1.2. Bootstrap 5

1.2. Causal effect estimation in Mendelian Randomization 6

1.2.1. Mendelian Randomization 6

1.2.2. Instrumental variable assumptions and two-stage least square estimator 7

1.2.3. Dealing with pleiotropy 8

1.3. Notations 11

1.4. Organization of the thesis 11

2. Confidence intervals in high-dimensional Cox models with time-dependent covariates 12

2.1. Model set-up 12

2.2. Methdology 14

2.2.1. LDPE 14

2.2.2. The proposed approach 14

2.2.3. Outline of the proof 16

2.3. Theoretical properties 17

2.3.1. Main theorems 19

2.3.2. Supporting Lemmas 20

2.3.3. Feasibility 21

2.4. Discussion 22

3. Debiasing the debiased Lasso with bootstrap 23

3.1. Methodology 23

3.2. Main ideas 25

3.3. Main results: deterministic designs 27

3.4. Main results: Gaussian designs 31

3.5. Simulations 34

3.6. Discussions 37

4. Dealing with pleiotropy in Mendelian Randomization 38

4.1. Motivation of Bayesian methods 38

4.2. Hierarchical models 39

4.2.1. Single Gaussian prior 39

4.2.2. A mixture Gaussian prior 42

4.3. Dealing with unknown hyper parameters 45

4.4. Implementation with summary data 46

4.5. Simulations and real studies 46

4.5.1. Synthetic data experiments 46

4.5.2. Case study (i): HDL and type 2 diabetes 50

4.5.3. Case study (ii): LDL and type 2 diabetes 51

4.6. Discussions 52

A. Appendices 62

A.1. Proofs in Chapter 2 62

A.1.1. Supportive lemmas 62

A.1.2. Proof of Lemmas in Section 2.3.2 68

A.1.3. Proof of theorems in Section 2.3.1 71

A.2. Proofs for theorems and lemmas in Section 3 75

A.2.1. Some technical lemmas 75

A.2.2. Proof of lemmas and theorems in Section 3 81

A.3. Proofs for Chapter 4 94

A.3.1. Proof of theorems and lemmas 94

A.3.2. Implementation details 101

A.3.3. Implementation with summary statistics 102

List of Figures

1.1. Illustrative diagram of conventional instrumental variable assumptions. $Z_1, \ldots, Z_J$ are J genetic variants as instruments. D and Y are the exposure and the outcome under consideration. U represents common confounders of D and Y. Crosses indicate violations of assumptions. . . 7

1.2. Illustrative diagram of relaxing ER assumption as a result of the existence of pleiotropic effects. Dashed arrows indicate the effects allowed to exist in this paper. Parameters in the parentheses correspond to the notations in model (1.5). 9

3.1. Boxplots of the double debiased Lasso (DDB) (3.10), the debiased Lasso (DB) (3.2) and the Lasso (Las) (3.1) with the identity covariance matrix in setting (ii). First row consists of estimates for strong signals: $\beta_6 = \beta_7 = \beta_8 = 2$. Second row consists of estimates for strong signals: $\beta_1 = \beta_2 = \beta_3 = 1$. Third row consists of estimates for zeros: $\beta_{21} = \beta_{22} = \beta_{23} = 0$. Each Boxplot is based on 1000 independent replications. . . . 36

4.1. $\beta = 0$ and InSIDE assumption is satisfied. The x-axis of each plot is the number of nonzero α_j. Each point represents the MSE of 200 experiments for $\mu_\alpha = -0.2, 0$ and 0.2 from left to right, respectively. 48

4.2. $\beta = 0.2$ and InSIDE assumption is satisfied. The x-axis of each plot is the number of nonzero α_j. Each point represents the MSE of 200 experiments for $\mu_\alpha = -0.2, 0$ and 0.2 from left to right, respectively. . . 48

4.3. $\beta = 0.2$ and InSIDE assumption is not satisfied. The x-axis of each plot is the number of nonzero α_j. Each point represents the MSE of 200 experiments for $\mu_\alpha = -0.2, 0$ and 0.2 from left to right, respectively. . . 49

4.4. $\beta = 0.2$ and InSIDE assumption is not satisfied. The x-axis of each plot is the number of nonzero α_j. Each point represents the MSE of 200 experiments for $\mu_\alpha = -0.2, 0$ and 0.2 from left to right, respectively. . . 49

List of Tables

Coverage probability and length of confidence interval for debiased Lasso, bootstrapped debiased Lasso and bootstrapped Adaptive Lasso. 35

Chapter 1

Introduction

Regression models are fundamental tools to study the association between predictors and response variables. In low-dimensional or fixed-dimensional settings, the least square estimator (LSE) and maximum likelihood estimator (MLE) have been well-studied under some classical assumptions. However, those assumptions can be too restrictive in modern applications. Consider a low-dimensional Gaussian linear model:

$$y = X\beta + \epsilon, \tag{1.1}$$

where the regression coefficient vector $\beta \in \mathbb{R}^p$, the design $X \in \mathbb{R}^{n\times p}$ can be either deterministic or random and the random error $\epsilon_i \sim_{i.i.d.} N(0, \sigma^2)$ conditioning on X for some positive constant σ^2. In the low-dimensional case ($p < n$), the least square estimator for β is

$$\widehat{\beta}^{(LS)} = (X^T X)^{-1} X^T y$$

and its consistency and asymptotic normality can be proved without much efforts. When some of the classical assumptions are violated, the LSE may have undesirable performance and new techniques and methodologies need to be developed. The topics in this thesis arise from following generalizations and relaxations of regression model (1.1), which are motivated by broad applications in modern science.

(a) We consider the high-dimensional setting, where the dimension p of the model is allowed to be larger or much larger than the sample size n. In this case, the sample Gram matrix $X^T X/n$ is not invertible and the least square estimator is undefined. This motivates new methodologies, say regularized least square estimators, for estimation and inference in high-dimensional models.

(b) We consider the effect of unknown confounders, which results in the correlation between some predictors and the random errors. In this case, the least square estimator can be biased. This motivates the use and study of instrumental variables.

1.1 High-Dimensional inference

In many areas of applications, including genomics, machine learning and astronomy, many opportunities and challenges have been posed by large-scale data. It is common to observe large numbers of parameters and/or large numbers of observations. Hence, it is important to study the methodologies and theory in high-dimensional scenarios.

When the dimension p of the model is larger than the sample size n, regularized least square estimators are typically used when the signal is believed to be sparse. Popular approaches include, but are not restricted to, the Lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001), Adaptive Lasso (Zou, 2006), Dantzig Selector (Candes and Tao, 2007), Lasso and Dantzig (Bickel et al., 2009), MCP (Zhang, 2010) and scaled Lasso (Sun and Zhang, 2012). Properties of regularized least square estimators in prediction, estimation and variable selection have been extensively studied in high-dimensional linear models. Among the regularized regression procedures, the Lasso (Tibshirani, 1996) is one of the most popular methods as it is computationally manageable and theoretically well-understood (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006; Bunea et al., 2007; Zhang and Huang, 2008; Meinshausen and Yu, 2009; Wainwright, 2009; Ye and Zhang, 2010).

Some approaches have been generalized beyond linear models. For example, van de Geer (2008) and Huang and Zhang (2012) studied the oracle inequality of weighted Lasso in high-dimensional generalized linear models. Tibshirani (1997) and Fan and Li (2002) extended the Lasso and SCAD to the Cox model, respectively. Bradic et al. (2011) established strong oracle properties of nonconcave penalized methods for non-polynomial dimensional Cox model. Huang et al. (2013) and Kong and Nan (2014) studied the non-asymptotic oracle inequalities of the Lasso in the high-dimensional Cox regression. Lin and Lv (2013) studied a class of penalized methods for variable selection and estimation in the high-dimensional additive hazards models. Belloni et al. (2012) and Spindler (2016) studied the Lasso and Post-Lasso in linear instrumental variable models with high-dimensional instruments.

However, the distributions of many regularized estimators are not tractable. For example, the limiting distribution of the Lasso estimator (Knight and Fu, 2000) depends on unknown parameters in low-dimensional settings and is not available in high-dimensional settings. Thus, there is substantial difficulty in drawing valid inference based on the Lasso estimates directly.

1.1.1 Debiased Lasso

In the $p \gg n$ scenario, Zhang and Zhang (2014) proposed to construct confidence intervals for low-dimensional regression coefficients by "debiasing" an initial Lasso estimator. Such estimators enjoy asymptotic normality under certain conditions and are known as the low-dimensional projection estimator (LDPE) or "debiased Lasso". It is worth mentioning that unlike the approaches based on selection consistency, the debiased Lasso approach does not assume the "beta-min" condition, which requires all the signals to be stronger than certain threshold level.

Along this line of research, many recent papers study computational algorithms and theoretical guarantees for the debiased Lasso and its extensions beyond linear models. van de Geer et al. (2014) proved asymptotic efficiency of the debiased Lasso estimator in linear models and for convex loss functions. Javanmard and Montanari (2014a) carefully studied quadratic programming in Zhang and Zhang (2014) to generate a direction for debiasing the Lasso in high-dimensional linear models. The asymptotic normality of the resulting estimator does not rely on the condition on the sparsity of the precision matrix of the design. Belloni et al. (2013, 2014) proposed to construct confidence regions for the quantile regression and instrumental median regression estimator, respectively, based on Neyman's orthogonalization, which is first-order equivalent to the bias correction. Jankova and van de Geer (2015) and Ren et al. (2015) proved asymptotic efficiency of the debiased Lasso in estimating individual entries of a precision matrix. Mitra and

Zhang (2016) proposed to debias a scaled group Lasso for chi-squared-based statistical inference for large variable groups. Fang et al. (2016) considered statistical inference for a single parameter with a decorrelated Wald statistic based on debiased Lasso in high-dimensional Cox model. Chernozhukov et al. (2017) studied debiased method in a semiparametric model with machine learning approaches.

In Chapter 2, we consider a low-dimensional projection estimator (LDPE) for a linear combination of regression coefficients in high-dimensional Cox models with time-dependent covariates. The computation is via a quadratic optimization algorithm, which can be viewed as the extension of the quadratic programming in linear models (Zhang and Zhang, 2014; Javanmard and Montanari, 2014a) to the Cox model with time-dependent covariates. We prove the asymptotic normality of LDPE under proper conditions.

The sample size requirement for asymptotic normality of the debiased Lasso in aforementioned papers is typically $n \gg (s \log p)^2$, where s is the number of nonzero regression coefficients. However, it is known that estimation consistency (in ℓ_1-norm) of the Lasso estimator holds with $n \gg s \log p$. Therefore, it becomes an intriguing question whether it is possible to conduct statistical inference of individual coefficients in the range $s \log p \ll n \lesssim (s \log p)^2$.

Very little work has been done in this direction. For the standard Gaussian design, Javanmard and Montanari (2014b) studied that the debiased estimator is asymptotically Gaussian in an average sense if $s = O(n/\log p)$ with s/p, n/p constant, but they did not provide theoretical results when the covariance of the design is unknown. Let s_j be the number of nonzero elements in j-th column of the precision matrix. Javanmard and Montanari (2015) proved that asymptotic normality for the debiased Lasso holds when $s \ll n/(\log p)^2$, $s_j \ll n/\log p$ and $\min\{s, s_j\} \ll \sqrt{n}/\log p$ under Gaussian design with unknown covariance matrix and other technical conditions. Cai and Guo (2017) proved that adaptivity in s is infeasible for statistical inference with random design when $n \lesssim (s \log p)^2$ in a minimax sense.

1.1.2 Bootstrap

Bootstrap has been widely studied for conducting inference in both low-dimensional and high-dimensional models. Mammen (1993) considered estimating the distribution of linear contrasts and of F-test statistics when p increases with n. Chatterjee and Lahiri (2010) showed the inconsistency of residual bootstrap for the Lasso if at least one true coefficient is zero in fixed-dimensional settings. For fixed number of covariates p, Chatterjee and Lahiri (2011) proposed to apply bootstrap to a modified Lasso estimator as well as to the Adaptive Lasso estimator. Chatterjee and Lahiri (2013) showed the consistency of bootstrap for Adaptive Lasso when p increases with n under some conditions which guarantee sign consistency. They also proved the second-order correctness for a studentized pivot with a bias-correction term. It is worth mentioning that a beta-min condition is required in their theorems as sign consistency is used to prove bootstrap consistency. In the high-dimensional setting, multiplier bootstrap has been studied to approximate the maximum of a sum of high-dimensional random vectors (Chernozhukov et al., 2013; Deng and Zhang, 2017).

For the debiased Lasso procedure, Zhang and Cheng (2017) proposed a Gaussian bootstrap method to conduct simultaneous inference with non-Gaussian errors. Dezeure et al. (2016) proposed residual, paired and wild multiplier bootstrap for debiased Lasso estimators, which demonstrates the benefits of bootstrap for heteroscedastic errors as well as simultaneous inference. However, the aforementioned papers do not provide improvement on the sample size conditions.

In Chapter 3, we prove the consistency of bootstrap approximation for the distribution of debiased Lasso under proper regularity conditions. We prove that the required sample size condition can be weaker than the typical condition for debiased Lasso, which involves the number of weak signals.

1.2 Causal effect estimation in Mendelian Randomization

1.2.1 Mendelian Randomization

Studying the causality between exposures and outcomes is a crucial task in social science and epidemiology. Mendelian randomization (MR) uses genetic variants as instruments to measure the causal effect of a specific exposure on an outcome (Lawlor et al., 2008; Davey Smith and Hemani, 2014). As a counterpart to the randomized controlled trial (RCT), MR can address areas where an RCT would be impossible or unethical. With more and more available genome-wide association studies (GWAS), researchers are able to find genetic variants which are robustly associated with target exposures and infer the causality between exposures and outcomes via the variation of genetic variants.

For instance, some recent studies raised a puzzling question whether there exists a causal relationship between low-density lipoprotein (LDL) cholesterol and type 2 diabetes. Statin therapy has been shown to reduce cardiovascular disease by lowering LDL (Baigent et al., 2005). However, it is associated with a 9% increased risk for incident diabetes in RCT studies (Sattar et al., 2010). On the other hand, another LDL lowering drug, Evolocumab, which uses a different bological pathway, has not been shown to have a significant effect on the incident diabetes in RCTs (Sabatine et al., 2017). Thus, it is of interest to study whether the increased risk of diabetes is caused by lowering LDL as opposed to medication-specific effects. This problem is analyzed in Chapter 4 as a case study with MR methods applied on summary data from GWAS.

There are many advantages of genetic variants serving as instruments. Firstly, in genetic associations, the direction of causation is always from the genetic polymorphism to the phenotype of interest, and not vice versa. Secondly, genetic variants are subject to relatively small measurement error or confoundedness, as opposed to conventionally measured environmental exposures, which are often associated with a wide range of behavioral, social and physiological confounding factors. Thirdly, MR is more cost-effective compared with RCTs.

1.2.2 Instrumental variable assumptions and two-stage least square estimator

By conventional instrumental variable literatures, typical assumptions for genetic variants to be valid instruments (Figure 1.1) are (van Kippersluis and Rietveld, 2017)

(i) Relevance: The genetic variants have an effect on exposure.

(ii) Independence: The genetic variants are uncorrelated with any confounders of the exposure-outcome relationship.

(iii) Exclusion restriction (ER): The genetic variants affect the outcome only through exposure.

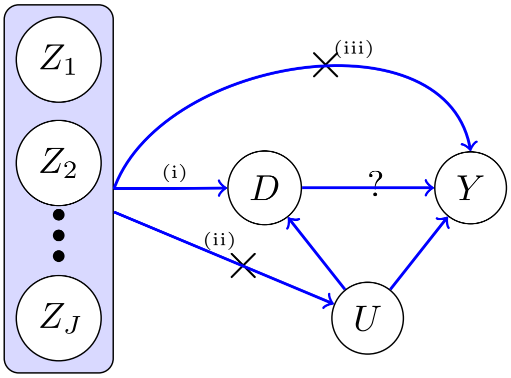

Figure 1.1: Illustrative diagram of conventional instrumental variable assumptions. $Z_1, \ldots, Z_J$ are J genetic variants as instruments. D and Y are the exposure and the outcome under consideration. U represents common confounders of D and Y. Crosses indicate violations of assumptions.

Observed genotypes are usually coded as the number of minor alleles, 0, 1 or 2. Without loss of generality, we consider the case where the instruments are continuous. Let Z be an $n \times J$ matrix of genetic variants whose i-th row consists of the i-th observation Z_i. Let $D = (D_1, \ldots, D_n)^T \in \mathbb{R}^n$, where $D_i \in \mathbb{R}$ is the i-th observation of the exposure. Let $Y = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$, where $Y_i \in \mathbb{R}$ is the i-th observation of the outcome. Suppose that we observe independent copies of $(Z_i, D_i, Y_i), i = 1, \ldots, n$.

We adopt the Neyman-Rubin's potential outcome framework (Rubin, 1974; Splawa-Neyman, 1990) and set up the model for observed data under IV assumptions (i) - (iii). For $i = 1, \ldots, n$,

$$\begin{cases} D_i &= Z_i\gamma + v_i \\ Y_i &= \beta D_i + \epsilon_i, \end{cases} \tag{1.2}$$

where $\gamma = (\gamma_1, \ldots, \gamma_J)^T \in \mathbb{R}^J$ with γ_j the strength of the j-th instrumental variable, $\beta \in \mathbb{R}$ is the causal effect of interest, and (v_i, ϵ_i) has mean zero and covariance matrix $\begin{pmatrix} \sigma_v^2 & \sigma_{v\epsilon}^2 \\ \sigma_{v\epsilon}^2 & \sigma_\epsilon^2 \end{pmatrix}$ conditioning on Z, $\beta \in \mathbb{R}$ is the causal effect of interest. v and ϵ are correlated due to the effect of the common confounders U. In model (1.2), D is correlated with the error term ϵ and hence the LSE based on the exposure-outcome model in (1.2) can be biased. The Two-Stage Least Square (TSLS) estimator has been proposed to solve this issue.

Specifically, one can construct a proxy of D, namely the least square estimate $\hat{D}$, such that

$$\hat{D} = Z\hat{\gamma}, \text{ with } \hat{\gamma} = (Z^T Z)^{-1} Z^T D. \tag{1.3}$$

Then we use $\hat{D}$ as a proxy of D in the exposure-outcome model:

$$Y_i = \beta \hat{D}_i + \hat{\eta}_i, \tag{1.4}$$

where $\hat{\eta}_i = \epsilon_i + \beta(D_i - \hat{D}_i)$. The moment condition $\mathbb{E}[Z_i^T \hat{\eta}_i] = 0$, resulting from the model assumption, sheds light on the TSLS estimator:

$$\hat{\beta}^{(TSLS)} = \arg\min_{\beta \in \mathbb{R}} \|Y - \hat{D}\beta\|_2^2.$$

It is easy to see that $\hat{\beta}^{(TSLS)}$ is an asymptotically unbiased estimator of β assuming that $\|\hat{D}\|_2^2/n \to K_1 > 0$ as $n \to \infty$.

1.2.3 Dealing with pleiotropy

Some concerns, such as weak instruments, the confoundness of genotype and canalization, have been raised about applying the MR methods. Using multiple instruments can increase the power of genotype-exposure and genotype-outcome association, but may also introduce issues with linkage disequilibrium and pleiotropy (Davey Smith and Ebrahim, 2008; VanderWeele et al., 2014). Genetic variants with pleiotropy, which means that one gene can influence two or more seemingly unrelated phenotypic traits, may fail the ER assumption, because they may have direct effects on the outcome. To deal with this issue, we consider causal effect estimation in the scenario of Figure 1.2 in comparison to Figure 1.1.

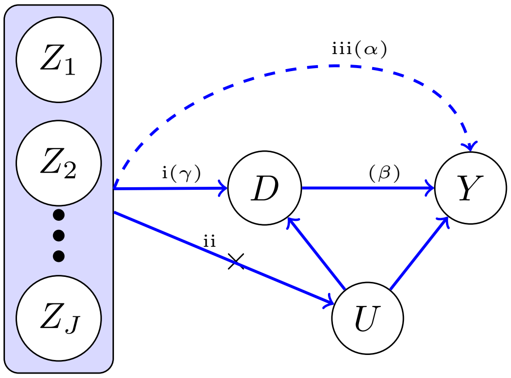

Figure 1.2: Illustrative diagram of relaxing ER assumption as a result of the existence of pleiotropic effects. Dashed arrows indicate the effects allowed to exist in this paper. Parameters in the parentheses correspond to the notations in model (1.5).

The model for observed data corresponding to Figure 1.2 can be formulated as follows. For $i = 1, \ldots, n$,

$$\begin{cases} D_i &= Z_i\gamma + v_i \\ Y_i &= \beta D_i + Z_i\alpha + \epsilon_i, \end{cases} \tag{1.5}$$

where α_j the effect of j-th genetic variant Z_j on the outcome Y not via the exposure D. For simplicity, we refer to α_j as the pleiotropic effect of Z_j. As in (1.4), we can rewrite the exposure-outcome model in (1.5) as

$$Y_i = \beta\hat{D}_i + Z_i\alpha + \hat{\eta}_i, \tag{1.6}$$

for $\hat{D}$ defined via (1.3). Since $\hat{D}$ is a linear combination of Z, the column space of matrix $(\hat{D}, Z) \in \mathbb{R}^{n\times(J+1)}$ is rank deficient. If it is known the subset of genetic variants with pleiotropic effects, i.e. the support of α is known and $\|\alpha\|_0 < J$, one can perform a multivariate least square regarding (1.6) (Donald and Newey, 2001). However, in many realistic cases, the support of α is always unknown.

Many recent works have studied causal effect estimation in existence of pleiotropic effects from various perspectives. Bowden et al. (2015) introduced Egger's regression method under the InSIDE assumption (instrument strength independent of direct effect). Kang et al. (2016) showed that the causal effect can be identified if at least 50% of instruments are valid and also developed a Lasso-type estimator under some sparsity conditions and other regularity conditions. Bowden et al. (2016) developed a weighted median estimator which is consistent when at least 50% of instruments are valid. A pleiotropy-robust MR method was introduced by van Kippersluis and Rietveld (2017) using a subsample which is independent of the exposure to estimate the pleiotropic effects. However, these types of assumptions can be hard to check in reality and hence restrict the applicability of such estimators.

The Bayesian methods have been used and studied in order to deal with pleiotropic effects. However, theoretical properties of Bayesian estimators are not well understood. Feller and Gelman (2015) considered a hierarchical model to account for the randomness in data collection, unmeasured covariates and treatment effect variation. However, their approach does not incorporate instrumental variables, while an MR problem is intrinsically equipped with genetic variants as instruments. Berzuini et al. (2017) considered horseshoe prior on pleiotropic effects and demonstrated its performance through simulations. Thompson et al. (2017) considered Bayesian model averaging among three pleiotropy models. Schmidt and Dudbridge (2017) studied a joint normal prior on causal effect and direct effects via simulation. However, there have been no theoretical justifications on the non-asymptotic or asymptotic performance of Bayesian estimators in MR problem with pleiotropy. It is important to study theoretical performance of Bayesian methods, which can provide some guidance in applications. The purpose of Chapter 4 is to study the theoretical performance of Bayesian estimators under some well-adopted priors.

1.3 Notations

We use following general notations in this thesis. For vectors u and v, let $\|u\|_q$ denote the ℓ_q norm of u, $\|u\|_0$ the number of nonzero entries of u, $\langle v, u\rangle = u^T v$ the inner product. For a set $\mathcal{T}$, let T^c denote its complement, $|\mathcal{T}|$ the cardinality of $\mathcal{T}$ and $u_{\mathcal{T}}$ the subvector of u with components in $\mathcal{T}$. Define $u^{\otimes 0} = 1 \in \mathbb{R}$, $u^{\otimes 1} = u$ and $u^{\otimes 2} = uu^T$. We use e_j to refer to the j-th standard basis element, for example, $e_1 = (1, 0, \ldots, 0)$. For a matrix $A \in \mathbb{R}^{k_1 \times k_2}$, let $\|A\|_q$ denote the ℓ_q operator norm of A. Specially, let $\|A\|_\infty = \max_{j \le k_1} \|A_{j,.}\|_1$. let P_A be the $k_1 \times k_1$ orthonormal projection matrix onto the column space of A, i.e. $P_A = A(A^T A)^{-1} A^T$ and $P_A^\perp = I_{k_1 \times k_1} - P_A$. Let $\Lambda_{\max}(A)$ and $\Lambda_{\min}(A)$ be the largest and smallest singular values of A, $A_{\mathcal{T}_1, \mathcal{T}_2}$ the submatrix of A consisting of rows in $\mathcal{T}_1$ and columns in $\mathcal{T}_2$. We say $A' \preceq A$ iff $A - A'$ is a positive definite matrix.

Let $\Phi(\cdot)$ and $\phi(\cdot)$ be the cdf and pdf of a standard Gaussian random variable, respectively. Let $\mathrm{U}[a, b]$ be the uniform distribution on $[a, b]$ for $a < b$. Let $\xrightarrow{\mathcal{D}}$ denote convergence in distribution.

1.4 Organization of the thesis

In Chapter 2, we study constructing confidence intervals for a linear combination of coefficients in high-dimensional Cox models with time-dependent covariates with a quadratic optimization scheme. In Chapter 3, we prove that bootstrap can further correct the bias of debiased Lasso for both deterministic and Gaussian designs. In Chapter 4, we study Bayesian estimators of causal effect in the existence of pleiotropy.

Chapter 2

Confidence intervals in high-dimensional Cox models with time-dependent covariates

In this chapter, we consider interval estimates and hypothesis testing in high-dimensional Cox models with time-dependent covariates. We first formally define a low-dimensional projection estimator (LDPE) for a linear combination of regression coefficients and then develop a one-step estimator (OSE) as a computationally efficient alternative of LDPE. The OSE in the high-dimensional linear model is equivalent to the debiased Lasso approach (Zhang and Zhang, 2014). The computation of LDPE and OSE is via a quadratic optimization algorithm, which can be viewed as the extension of Zhang and Zhang (2014) and Javanmard and Montanari (2014a) to the Cox model with time-dependent covariates. We prove the asymptotic normality of LDPE and OSE under proper conditions.

Fang et al. (2017) considered a similar problem as ours. They developed a decorrelated Wald test for a single parameter. It can be seen from latter sections that our analysis applies to a general problem with different set of conditions. Detailed discussion is in Section 2.3.

2.1 Model set-up

Let $\mathbf{N}^{(n)}(t) = (N_1(t), \cdots, N_n(t))'$, $t > 0$ be an n-dimensional counting process on a time interval $[0, \tau]$ with $\tau > 0$, where $N_i(t)$ counts the number of observed events for the i-th individual in the time interval $[0, t]$. For $t > 0$, let $\mathcal{F}_t$ be the filtration representing all the information available up to time t. Following the definitions in Andersen and Gill (1982), assume that for $\{\mathcal{F}_t, t \geq 0\}$, $\mathbf{N}^{(n)}$ has a predictable compensator $\Lambda^{(n)} =$

$(\Lambda_1, \cdots, \Lambda_n)$ with

$$d\Lambda_i(t) = Y_i(t) \exp\{X_i(t)\beta_0\} d\Lambda_0(t), \quad 1 \le i \le n, \tag{2.1}$$

where $\beta_0 \in \mathbb{R}^p$ is the true unknown parameter, $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ is an unknown baseline cumulative hazard function and for each i, $Y_i(t) \in \{0,1\}$ and $X_i(t) = (X_{i1}(t), X_{i2}(t), \ldots, X_{ip}(t))'$ are both $\mathcal{F}_t$ predictable. Here p is large and possibly much larger than n.

To ease our notation, define

$$\gamma_{ni}(t,\beta) = \frac{Y_i(t)\exp\{X_i(t)\beta\}}{nS^{(0)}(t,\beta)} \quad \text{and} \quad S^{(r)}(t,\beta) = \frac{1}{n}\sum_{i=1}^n X_i(t)^{\otimes r} Y_i(t) e^{X_i(t)\beta}, \quad r = 0,1,2.$$

For two time-dependent vectors $f(t), g(t) \in \mathbb{R}^n$, $t \in [\tau_1, \tau_2]$, define

$$\begin{aligned}
\bar{f}_n(t,\beta) &= \sum_{i=1}^n \gamma_{ni}(t,\beta) f_i(t,\beta) \\
\overline{\mathrm{Cov}(f,g)}(t,\beta) &= \sum_{i=1}^n \gamma_{ni}(t,\beta)(f_i(t) - \bar{f}_n(t,\beta))(g_i(t) - \bar{g}_n(t,\beta))^T \\
\overline{\mathrm{Var}(f)}(t,\beta) &= \overline{\mathrm{Cov}(f,f)}(t,\beta).
\end{aligned}$$

As in Andersen and Gill (1982), the log-partial likelihood function is defined as:

$$C(\beta,\tau) = \sum_{i=1}^n \int_0^\tau \{X_i(t)\beta\} dN_i(t) - \int_0^\tau \log\left[\sum_{i=1}^n Y_i(t)\exp\{X_i(t)\beta\}\right] d\bar{N}(t),$$

where $\bar{N} = \sum_{i=1}^n N_i$. Let $\ell([0,\tau];\beta) = -C(\beta,\tau)/n$ for some $\tau > 0$. By differentiation and rearrangement of terms, it can be shown that the gradient of $\ell([0,\tau],\beta)$ is

$$\dot{\ell}([0,\tau];\beta) = \frac{\partial \ell([0,\tau];\beta)}{\partial \beta} = \frac{1}{n}\sum_{i=1}^n \int_0^\tau [X_i(t) - \bar{X}_n(t,\beta)] dN_i(t) \tag{2.2}$$

and the Hessian matrix of $\ell([0,\tau];\beta)$ is

$$\ddot{\ell}([0,\tau];\beta) = \frac{\partial^2 \ell([0,\tau];\beta)}{\partial\beta\partial\beta^T} = \frac{1}{n}\int_0^\tau \overline{\mathrm{Var}(X)}(t,\beta) d\bar{N}(t). \tag{2.3}$$

Different from linear models and generalized linear models, the score function and Hessian matrix of the Cox model do not have the structure of sum of independent variables. In the low-dimensional case, the central limit theory of MLE is based on the martingale structure of the score function. In the high-dimensional case, we need to keep the martingale property of the score function regarding the proposed estimator and prove the consistency of Hessian matrix by a careful decomposition based on empirical process theory.

2.2 Methdology

2.2.1 LDPE

Consider the problem of estimating a linear functional $\theta = \langle a_0, \beta \rangle$ with respect to the negative partial likelihood $\ell([0,\tau];\beta)$. As proposed in Zhang (2011), the low-dimensional projection estimator (LDPE) of θ can be constructed as a univariate MLE

$$\widehat{\theta}^{(LDP)} = \langle a_0, \widehat{\beta}^{(init)} \rangle + \arg\min_{\phi \in \mathbb{R}} \ell\big([0,\tau]; \widehat{\beta}^{(init)} + u\phi\big), \tag{2.4}$$

where $\widehat{\beta}^{(init)}$ is an initial estimator of β with desired rate of convergence and u is the least favorable submodel which has proper bias correction effect. Realization of $\widehat{\beta}^{(init)}$ and u will be specified in next section. The formulation of LDPE in (2.4) is equivalent to

$$\hat{\theta}^{(LDP)} = \arg\min_{\theta \in \mathbb{R}} \ell\Big([0,\tau]; \widehat{\beta}^{(init)} + u\big(\theta - \langle a_0, \widehat{\beta}^{(init)} \rangle\big)\Big). \tag{2.5}$$

The negative partial score function regarding (2.5) is

$$\begin{aligned} D(\theta) &= u^T \dot{\ell}\left([0,\tau]; \widehat{\beta}^{(init)} + u(\theta - \langle a_0, \widehat{\beta}^{(init)} \rangle)\right) \\ &= \frac{1}{n} \sum_{i=1}^{n} \int_0^{\tau} \left[z_i(t) - \frac{\sum_{i=1}^{n} z_i(t) Y_i(t) e^{X_i(t)\widehat{\beta}^{(init)} + z_i(t)(\theta - \widehat{\theta}^{(init)})}}{\sum_{i=1}^{n} Y_i(t) e^{X_i(t)\widehat{\beta}^{(init)} + z_i(t)(\theta - \widehat{\theta}^{(init)})}} \right] dN_i(t), \end{aligned} \tag{2.6}$$

where $z_i(t) = X_i(t)u$. Hence, $\hat{\theta}^{(LDP)}$ can be solved by finding the root of $D(\theta) = 0$ for given $z(t)$.

2.2.2 The proposed approach

The asymptotic normality of $\hat{\theta}^{(LDP)}$ defined in (2.5) requires $\widehat{\beta}^{(init)}$ to satisfy certain rate of estimation consistency and $z(t)$ to be a $\mathcal{F}_t$-predictable process with large probability. Since the computation of $z(t)$ is based on $\widehat{\beta}^{(init)}$, we propose to compute $\widehat{\beta}^{(init)}$ based on the history information prior to the time interval where $z(t)$ is calculated from. Specifically, we assume that the time interval we observed is $[-\tau_0, \tau]$ for the sake of notational simplicity. We split $[-\tau_0, \tau]$ into two parts and estimate $\widehat{\beta}^{(init)}$ with the observed events in $[-\tau_0, 0]$. Based on this $\widehat{\beta}^{(init)}$, we compute $z(t),\ t \in [0, \tau]$ with the events in $[0, \tau]$.

For high-dimensional Cox model, the Lasso (Tibshirani, 1997) and SCAD (Fan and Li, 2002) approaches have been developed. We use the Lasso method to construct $\widehat{\beta}^{(init)}$. That is, for a tuning parameter $\lambda > 0$,

$$\widehat{\beta}^{(init)} = \arg\min_{\beta \in \mathbb{R}^p} \{\ell([-\tau_0, 0]; \beta) + \lambda \|\beta\|_1\}. \tag{2.7}$$

Oracle inequalities and rate of convergence of $\widehat{\beta}^{(init)}$ have been studied in Huang et al. (2013) under certain conditions. Another possibility is to estimate $\widehat{\beta}^{(init)}$ from a set of independent observations, which can be achieved by randomly splitting the data into two parts, computing $\widehat{\beta}^{(init)}$ from one part of the data and computing $z(t)$ from the other part of data.

To compute the correction score $z(t)$, we first compute $z'(t),\ t \in [0, \tau]$ via the following optimization scheme. Suppose there are q ordered failure times in $[0, \tau]$: $0 \le T_{(1)} \le T_{(2)} \le \cdots \le T_{(q)} \le \tau$ (setting $T_{(0)} = 0$ and $T_{(q+1)} = \tau$). At each $T_{(l)}$ for $l = 1, \ldots, q$, consider

$$\min_{z'(T_{(l)}) \in \mathbb{R}^{(n)}} \overline{\mathrm{Var}(z')}(T_{(l)}, \widehat{\beta}^{(init)}) \tag{2.8}$$

$$\text{subject to } \begin{cases} \max_{i \in R(T_{(l)})} |z'_i(T_{(l)})| \le K_4 \\ \left\| \overline{\mathrm{Cov}(z', X)}(T_{(l)}, \widehat{\beta}^{(init)}) - a_0 \right\|_\infty \le \lambda' \asymp \sqrt{\frac{\log p}{n}}. \end{cases}$$

Let $z'(t) = z'(T_{(l)})$ for $t \in [T_{(l)}, T_{(l+1)})$ and $z'(t) = 0$ for $t \in [T_{(0)}, T_{(1)})$. Then we obtain $z(t)$ by

$$z(t) = z'(t) / \overline{\mathrm{Var}(z')}(t, \widehat{\beta}^{(init)}) \text{ for } t \in [0, \tau]. \tag{2.9}$$

The construction of such optimization scheme (2.8) clearly demonstrates the necessary conditions for the success of the debiased Lasso: achieving smallest variance (target function) with sufficient bias reduction guaranteed (second constraint). Specifically, the boundedness restriction of $z'(t)$ is used to justify that the dominating term in the error of $\widehat{\theta}^{(LDP)}$ is a locally bounded martingale. The second restriction imposes the bias correction effect of $z'(t)$ and will guarantee the remainder terms in the estimation error of $\widehat{\theta}^{(LDP)}$ are sufficiently small under some regularity conditions.

Moreover, an important advantage of optimization over the Lasso approach is that any realized correction score $z(t)$ via (2.8) and (2.9), which is not required to be close to the true least favorable direction, can successfully debias the initial estimator. Hence, the conditions regarding the existence of the true least favorable direction and the consistency of the estimated one are not required. We will specify our conditions in section 2.3.

The computation of $z'(t)$ via (2.8) is manageable but more demanding than the Lasso approach proposed in Fang et al. (2017) for debiasing a single parameter. However, (2.8) a more natural condition and it requires weaker regularity conditions. Especially, the "debiasing" efect of $z(t)$ does not rely on the sparsity of Fisher information at β_0 and time t (2.14).

Remark 2.2.1. *Actually, to get sufficient bias correction effect, the second constraint in (2.8) can be replaced with*

$$\left\| \frac{1}{n}\int_0^\tau \overline{Cov(z', X)}(t, \widehat{\beta}^{(init)})d\bar{N}(t) - a_0 \right\|_\infty \leq \lambda' \asymp \sqrt{\frac{\log p}{n}}.$$

However, the $z'(t)$ calculated under the above constraint does not have the martingale property and this will bring difficulty to the justification for the central limit theory.

Based on $\widehat{\beta}^{(init)}$ and $z(t)$ computed as above, $\widehat{\theta}^{(LDP)}$ can be computed as the solution to $D(\theta) = 0$ for $D(\theta)$ defined in (2.6). For a faster computation, we can construct a one-step estimator (OSE) based on the first Newton-Raphson iteration as an approximation of $\widehat{\theta}^{(LDP)}$:

$$\widehat{\theta}^{(OS)} = \left\langle a_0, \widehat{\beta}^{(init)} \right\rangle - \left\{ \frac{dD(\theta)}{d\theta}\Big|_{\theta=\widehat{\theta}^{(init)}} \right\}^{-1} D(\widehat{\theta}^{(init)}). \tag{2.10}$$

2.2.3 Outline of the proof

We will show that with $\widehat{\beta}^{(init)}$ defined in (2.7) and $z(t)$ computed via (2.8) and (2.9), for any constant δ and $D(\theta)$ defined in (2.6),

$$\sqrt{n}D\big(\theta_0 + \delta/\sqrt{n}\big) = \sqrt{n}\xi(0, \beta_0) + \delta F_z([0, \tau]; \beta_0) + o_P(1), \tag{2.11}$$

where

$$\xi(0,\beta_0)=\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}[z_i(t)-\bar{z}_n(t,\beta_0)]dN_i(t) \tag{2.12}$$

is asymptotically normal and $F_z([0,\tau];\beta_0)$ can be viewed as the Fisher information when $z(t)$ is given the least favorable direction for the estimation of θ (See (2.16) below).

As a consequence, we can prove that under certain regularity conditions, for the estimators defined in (2.5) and (2.10) respectively, we have

$$\sqrt{n}(\widehat{\theta}^{(LDP)}-\theta_0)=-\sqrt{n}F_z([0,\tau];\beta_0)^{-1}\xi(0,\beta_0)+o_P(1)\xrightarrow{\text{D}}N(0,F_z^{-1}([0,\tau];\beta_0)) \tag{2.13a}$$

$$\sqrt{n}(\widehat{\theta}^{(OS)}-\theta_0)=-\sqrt{n}F_z([0,\tau];\beta_0)^{-1}\xi(0,\beta_0)+o_P(1)\xrightarrow{\text{D}}N(0,F_z^{-1}([0,\tau];\beta_0)). \tag{2.13b}$$

Hence, confidence intervals and hypothesis testing procedures can be performed based on $\widehat{\theta}^{(LDP)}$ or $\widehat{\theta}^{(OS)}$. Detailed analysis will be carried out in the next section.

2.3 Theoretical properties

Define the following population version of quantities

$$s^{(r)}(t,\beta_0)=\mathbb{E}[S^{(r)}(t,\beta_0)],\quad r=0,1,2,\quad \mu(t,\beta_0)=\frac{s^{(1)}(t,\beta_0)}{s^{(0)}(t,\beta_0)}.$$

The Fisher information of β at β_0 and time t is defined as

$$F(t,\beta_0)=\frac{s^{(2)}(t,\beta_0)}{s^{(0)}(t,\beta_0)}-\mu^{\otimes 2}(t,\beta_0). \tag{2.14}$$

The Fisher information of β at β_0 over $[0,\tau]$ is defined as

$$F([0,\tau];\beta_0)=\int_0^{\tau}F(t,\beta_0)s^{(0)}(t,\beta_0)d\Lambda_0(t). \tag{2.15}$$

To establish the asymptotic properties of the LDPE (2.5) and OSE (2.10) for Cox model (2.1), we require the following conditions.

Condition 2.3.1. *For any vector v belonging to the cone $\mathscr{C}(\xi,S)=\{v\in\mathbb{R}^p:\|v_{S^c}\|_1\le \xi\|v_S\|_1\}$, there is a constant K_0 such that*

$$RE(\xi,S)=\inf_{0\ne v\in\mathscr{C}(\xi,S)}\frac{\{v^T\ddot{\ell}([-\tau_0,0];\beta_0)v\}^{1/2}}{\|v\|_2}\ge K_0>0.$$

Condition 2.3.2. *$\{Y_i(t), X_i(t), i = 1, \ldots, n\}$ are i.i.d. processes from $\{Y(t), X(t), t \in [-\tau_0, \tau]\}$, $\mathbb{P}\left\{\max_{1\le i\le n} N_i(\tau) \le 1\right\} = 1$ and*

$$\mathbb{P}\left\{\sup_{t\in[-\tau_0,\tau]} \max_{1\le i\le n} \|X_i(t) - \mu(t, \beta_0)\|_\infty > K_1\right\} = o(1).$$

Condition 2.3.3. *It holds that*

$$\mathbb{P}\left\{\sup_{t\in[-\tau_0,\tau]} \max_{1\le i\le n} |e^{X_i(t)\beta_0}/s^{(0)}(t, \beta_0)| > K_2\right\} = o(1).$$

Condition 2.3.4. *For $F(t, \beta_0)$ in (2.14) and $F([0, \tau]; \beta_0)$ in (2.15), it holds that*

$$\Lambda_{\max}(F([0, \tau]; \beta_0)) \le c_1^* \text{ and } \sup_{t\in[0,\tau]} \max_{1\le j\le p} (F_{j,j}(t, \beta_0)) \le c_2^*.$$

Moreover,

$$\int_0^\tau \overline{Var(z)}(t, \beta_0) S^{(0)}(t, \beta_0) d\Lambda_0(t) \to F_z([0, \tau]; \beta_0) > \epsilon > 0 \text{ in probability.} \tag{2.16}$$

Condition 2.3.5. *The parameter of interest $\theta = \langle a_0, \beta\rangle$ is given with a constant vector $a_0 \in \mathbb{R}^p$ satisfying $\|a_0\|_2 = O_P(1)$ and $\|a_0\|_\infty \ge a_*$.*

Condition 2.3.1 assumes that the restricted eigenvalue of $\ddot{\ell}([-\tau_0, 0]; \beta_0)$ is lower bounded, which was used to prove desired rate of consistency for $\widehat{\beta}^{(init)}$ in Huang et al. (2013) together with Condition 2.3.2. Condition 2.3.3 is used to prove desirable rate of convergence of the Hessian matrix (2.3). In Condition 2.3.4, the diagonal elements of the Fisher information at each t for $t \in [0, \tau]$ are required to be upper bounded. This is due to our construction of the time-dependent correction score $z(t)$. Condition 2.3.4 also assumes that (2.16) holds true, which can be justified under some other conditions (see Lemma 2.3.7). Condition 2.3.5 puts some conditions on the linear coefficients a_0, which can be easily satisfied in many situations.

Note that Conditions 2.3.2 and 2.3.3 are normalized and weaker versions of Assumptions 1 and 3 in Fang et al. (2017). Especially, neither the low bound on the eigenvalues of $F([0, \tau]; \beta_0)$ nor the sparsity of the least favorable direction $F^{-1}([0, \tau]; \beta_0)a_0$ are required.

2.3.1 Main theorems

Now we state the main theorems characterizing the asymptotic normality of the LDPE (2.5) and OSE (2.10) of θ_0, respectively. Define

$$\sqrt{n}\xi(0,\beta_0) = n^{-1/2}\sum_{i=1}^{n}\int_0^{\tau}[z_i(t) - \bar{z}_n(t,\beta_0)]dN_i(t). \tag{2.17}$$

Theorem 2.3.1. *Assume Conditions 2.3.1-2.3.5 hold true, $\lambda \asymp \lambda' \asymp \sqrt{\log p/n}$ and $s = o(\sqrt{n}/\log p)$. Suppose $z(t)$ is a correction score computed via (2.8) and (2.8). For any constant δ,*

$$\sqrt{n}D\big(\theta_0 + \delta n^{-1/2}\big) = \sqrt{n}\xi(0,\beta_0) + \delta F_z([0,\tau];\beta_0) + o_P(1).$$

The proof of Theorem 2.3.1 is in the Appendix. The result of Theorem 2.3.1 is based on the second order expansion of $D(\theta)$ (2.6). The remainder terms are controlled by the second constraint in (2.8) and continuity of $\dot{D}(\theta_0)$ in its exponent (see Lemma A.1.3). With the above theorem, we are able to establish the asymptotic normality for $\widehat{\theta}^{(LDP)}$ and $\widehat{\theta}^{(OS)}$ as desired. The inverse of variance estimator can be computed via

$$\hat{F}_z(\widehat{\beta}^{(init)}) = \frac{1}{n}\int_0^{\tau}\overline{\mathrm{Var}(z)}(t,\widehat{\beta}^{(init)})d\bar{N}(t). \tag{2.18}$$

Corollary 2.3.2. *Under the conditions of Theorem 2.3.1. The estimators $\hat{\theta}^{(LDP)}$ in (2.5) and $\hat{\theta}^{(OS)}$ in (2.10) satisfy, respectively,*

$$\sqrt{n\hat{F}_z(\widehat{\beta}^{(init)})}(\widehat{\theta}^{(LDP)} - \theta_0) \xrightarrow{D} N(0,1)$$
$$\sqrt{n\hat{F}_z(\widehat{\beta}^{(init)})}(\widehat{\theta}^{(OS)} - \theta_0) \xrightarrow{D} N(0,1),$$

for $\hat{F}_z(\widehat{\beta}^{(init)})$ defined in (2.18).

The confidence interval and hypothesis testing for parameter θ can be constructed based on Corollary 2.3.2. Here we omit the proof details.

Corollary 2.3.3. *Under the conditions of Theorem 2.3.1, an asymptotic two-sided confidence interval for θ_0, with significance $0 < \alpha < 1$ is given by*

$$\left(\widehat{\theta} - (n\hat{F}_z(\widehat{\beta}^{(init)}))^{-1/2}\Phi^{-1}(1-\alpha/2),\ \ \widehat{\theta} + (n\hat{F}_z(\widehat{\beta}^{(init)}))^{-1/2}\Phi^{-1}(1-\alpha/2)\right).$$

Moreover, the asymptotic p-value for testing the hypothesis $H_0: \ \theta_0 = 0$ *versus* $H_A:$ $\theta_0 \neq 0$ *is*

$$\mathcal{P} = 2\left\{1 - \Phi\left(\sqrt{n\hat{F}_z(\widehat{\beta}^{(init)})}|\widehat{\theta}|\right)\right\},$$

where $\hat{F}_z(\widehat{\beta}^{(init)})$ *is defined in (2.18), and* $\widehat{\theta}$ *can be the LDPE or OSE, which are given in (2.5) and (2.10), respectively.*

2.3.2 Supporting Lemmas

We state several preliminary lemmas which are important for proving the asymptotic normality of $\widehat{\theta}^{(LDP)}$ and $\widehat{\theta}^{(OS)}$. We postpone the proofs to the Appendix.

Lemma 2.3.4 (Properties of the initial Lasso estimator)**.** *Assume Conditions 2.3.1 - 2.3.2 hold. For* $\lambda \asymp \sqrt{\log p/n}$ *and* $\widehat{\beta}^{(init)}$ *defined in (2.7), it holds that*

$$\|\widehat{\beta}^{(init)} - \beta_0\|_1 = O_P(s\lambda) \quad \textit{and} \quad \|\widehat{\beta}^{(init)} - \beta_0\|_2^2 = O_P(s\lambda^2).$$

Lemma 2.3.4 states the rate of convergence of the initial Lasso estimator $\widehat{\beta}^{(init)}$. The results are directly from Theorem 4.1 of Huang et al. (2013).

Lemma 2.3.5 (Asymptotic normality of the dominant term)**.** *Suppose* $z'(t)$ *is a solution to (2.8) and* $z(t)$ *is computed via (2.9). Under Conditions 2.3.1-2.3.5, for* $\xi(0,\beta_0)$ *defined in (2.12) and* $F_z([0,\tau];\beta_0)$ *defined in (2.16), we have*

$$\sqrt{n}\xi(0,\beta_0) \xrightarrow{D} N(0, F_z([0,\tau];\beta_0)), \ \textit{as } n \to \infty.$$

In the proof of Lemma 2.3.5, we mainly use the martingale central limit theorem.

Lemma 2.3.6 (Consistency of variance estimator)**.** *Under Conditions 2.3.1 - 2.3.5,* $\hat{F}_z(\widehat{\beta}^{(init)})$ *(2.18) is a consistent estimator of* $F_z([0,\tau];\beta_0)$ *(2.16), i.e.*

$$\hat{F}_z(\widehat{\beta}^{(init)}) = F_z([0,\tau];\beta_0) + o_P(1).$$

In the next Lemma, we provide sufficient conditions guaranteeing (2.16) in Condition 2.3.4.

Lemma 2.3.7. *Suppose that* $\sup_{t\in[0,\tau]}\lambda\|u_0(t,\beta_0)\|_1 = o(1)$ *and* $z'(T_{(l)})$ *is a solution of (2.8) for* $l = 1,\ldots,q$. *Then we have*

$$\overline{Var(z')}(T_{(l)},\beta_0) = a_0^T F^{-1}(T_{(l)},\beta_0)a_0 + o_P(1).$$

This Lemma implies that $F_z([0,\tau],\beta_0)$ satisfies, specifically,

$$\begin{aligned}\lim_{n\to\infty}\int_0^\tau \overline{\mathrm{Var}(z)}(T_{(l)},\beta_0)S^{(0)}(t,\beta_0)d\Lambda_0(t) &= \lim_{n\to\infty}\int_0^\tau (a_0^T F^{-1}(t,\beta_0)a_0)^{-1}S^{(0)}(t,\beta_0)d\Lambda_0(t)\\ &= \int_0^\tau (a_0^T F^{-1}(t,\beta_0)a_0)^{-1}s^{(0)}(t,\beta_0)d\Lambda_0(t).\end{aligned}$$

Note that condition $\sup_{t\in[0,\tau]}\lambda\|u_0(t,\beta_0)\|_1 = o(1)$ implies $\sup_{t\in[0,\tau]}\|u_0(t,\beta_0)\|_1 = o(\sqrt{n/\log p})$.

2.3.3 Feasibility

As discussed before, the asymptotic normality holds with any realized $z(t)$ via (2.8) and (2.9). To provide theoretical guarantees for the existence of a solution, we further assume the following conditions.

Condition 2.3.6. $\inf_{t\in[0,\tau]}\Lambda_{\min}(F(t,\beta_0)) \geq c_*$.

Condition 2.3.7. *Let* $u_0(t,\beta_0) = F^{-1}(t,\beta_0)a_0$. *It holds that*

$$\mathbb{P}\left\{\sup_{t\in[0,\tau]}\max_{1\leq i\leq n}|(X_i(t)-\mu(t,\beta_0))u_0(t,\beta_0)| > K_3\right\} = o(1).$$

Condition 2.3.6 guarantees the existence of $u_0(t,\beta_0)$ in Condition 2.3.7. Condition 2.3.7 is a weaker version of Assumption 4 in Fang et al. (2017).

Lemma 2.3.8 (Feasibility of the optimization scheme (2.8)). *Assume Conditions 2.3.1-2.3.7. If* $\lambda' \asymp \lambda \asymp \sqrt{\log p/n}$ *and* $s = o(\sqrt{n}/\log p)$, *then the optimization problem defined in (2.8) is feasible with large probability.*

The feasibility is proved by showing that a modification of $X(t)u_0(t,\beta_0)$ defined in Condition 2.3.7 satisfies the constraints in (2.8) under our conditions with large probability.

2.4 Discussion

The quadratic programming for computing debiased Lasso was first proposed in equation (46) of Zhang and Zhang (2014) for high-dimensional linear models and its theoretical properties were carefully studied in Javanmard and Montanari (2014a). In this chapter, we extend the quadratic optimization scheme to debias the Lasso estimator in high-dimensional Cox models. By proposing a time-dependent correction score, the martingale property of the score function regarding θ is maintained.

We emphasize that there is no need to check the conditions which are only used to prove the feasibility of the optimization algorithm (2.8) in application. This is because once a solution from (2.8) is obtained, the feasibility is automatically satisfied. Moreover, for any realization from the optimization algorithm, its bias-correction effect is automatically guaranteed. Hence, the regularity conditions regarding the true least favorable direction are not required for proving the asymptotic normality.

Chapter 3

Debiasing the debiased Lasso with bootstrap

As introduced in Section 1.1.1, there is a gap of required sample size between estimation consistency ($n \gg s\log p$) and inference with debiased Lasso ($n \gg (s\log p)^2$) in high-dimensional linear models. In this chapter, we prove that the bias of the debiased Lasso estimator (Zhang and Zhang, 2014) can be further removed by bootstrap without assuming the beta-min condition for both deterministic and Gaussian designs. We provide a refined analysis to distinguish the effects of small and large coefficients and show that bootstrap can remove the bias caused by strong coefficients. Our results demonstrate that if a majority of signals are strong, our sample size condition is weaker than the usual $n \gg (s\log p)^2$.

3.1 Methodology

Consider a linear regression model

$$y_i = x_i\beta + \epsilon_i,$$

where $\beta \in \mathbb{R}^p$ is the true unknown parameter and $\epsilon_1, \ldots, \epsilon_n$ are *i.i.d.* random variables with mean 0 and variance σ^2. We assume the true β is sparse in the sense that the number of nonzero entries of β is relatively small compared with $\min\{n, p\}$. For simplicity, we also assume that x_j's are normalized, s.t. $\|x_j\|_2^2 = n$, for $j = 1, \ldots, p$.

The Lasso estimator (Tibshirani, 1996) is defined as

$$\hat\beta = \arg\min_{b\in\mathbb{R}^p}\left\{\frac{1}{2n}\|y - Xb\|_2^2 + \lambda\|b\|_1\right\}, \tag{3.1}$$

where $\lambda > 0$ is a tuning parameter.

Suppose that we are interested in making inference of a single coordinate β_j, $j = 1, \ldots, p$. The debiased Lasso (Zhang and Zhang, 2014) corrects the Lasso estimator by a term calculated from residuals. Specifically, it takes the form

$$\hat{\beta}_j^{(DB)} = \hat{\beta}_j + \frac{z_j^T(y - X\hat{\beta})}{z_j^T x_j}, \tag{3.2}$$

where z_j is an estimate of the least favorable direction (Zhang, 2011). For the construction of z_j, it can be computed either as the residual of another ℓ_1-penalized regression of x_j on X_{-j} (Zhang and Zhang, 2014; van de Geer et al., 2014) or by a quadratic optimization (Zhang and Zhang, 2014; Javanmard and Montanari, 2014a). We adopt the first procedure in this paper. Formally,

$$z_j = x_j - X_{-j}\hat{\gamma}_{-j}, \text{ with} \tag{3.3}$$

$$\hat{\gamma}_{-j} = \underset{\gamma_{-j}\in\mathbb{R}^{p-1}}{\arg\min} \left\{ \frac{1}{2n}\|x_j - X_{-j}\gamma_{-j}\|_2^2 + \lambda_j\|\gamma_{-j}\|_1 \right\}, \ \lambda_j > 0. \tag{3.4}$$

While it is also possible to debias other regularized estimators of β, such as Dantzig selector (Candes and Tao, 2007), SCAD (Fan and Li, 2001) and MCP (Zhang, 2010), we restrict our attention to bootstrapping the debiased Lasso.

We consider Gaussian bootstrap although the noise ϵ_i are not necessarily assumed to be normally distributed. We generate the bootstrapped response vector as

$$y_i^* = x_i\hat{\beta} + \hat{\epsilon}_i^*, \ i = 1, \dots, n, \tag{3.5}$$

where x_i are unchanged and $\hat{\epsilon}_i^*$ are *i.i.d.* standard Gaussian random variables multiplied by an estimated standard deviation $\hat{\sigma}$. Namely,

$$\hat{\epsilon}_i^* = \hat{\sigma}\xi_i, \ i = 1, \dots, n, \tag{3.6}$$

where ξ_i are *i.i.d* standard normal. For the choice of variance estimator, we use

$$\hat{\sigma}^2 = \frac{1}{n - \|\hat{\beta}\|_0}\|y - X\hat{\beta}\|_2^2 \tag{3.7}$$

(Sun and Zhang, 2012; Reid et al., 2016; Zhang and Cheng, 2017; Dezeure et al., 2016). This is the same proposal of bootstrapping the residuals as in Zhang and Cheng (2017). However, we do not directly use $\hat{\epsilon}_i^*$ in (3.6) to simulate the distribution of the debiased estimator. Instead, we recompute the debiased Lasso based on (X, y^*) as follows:

$$\hat{\beta}_j^{(*,DB)} = \hat{\beta}_j^* + \frac{z_j^T(y^* - X\hat{\beta}^*)}{z_j^T x_j}, \tag{3.8}$$

where z_j is the same as the sample version in (3.3) and $\hat{\beta}^*$ is the bootstrap version of the Lasso estimator computed via (3.1) with (X, y^*) instead of the original sample.

We construct the confidence interval for β_j as

$$\left(\hat{\beta}_j^{(DB)} - q_{1-\alpha/2}(\hat{\beta}_j^{(*,DB)} - \hat{\beta}_j),\ \hat{\beta}_j^{(DB)} - q_{\alpha/2}(\hat{\beta}_j^{(*,DB)} - \hat{\beta}_j)\right), \tag{3.9}$$

where $q_c(u)$ is the c-quantile of the distribution of u.

We prove that under proper conditions, the approximation error of the debiased Lasso estimator $\hat{\beta}_j^{(DB)}$ in (3.2) is dominated by a constant term. We propose to estimate this dominating constant bias by the median of the bootstrapped approximation errors and construct a double debiased Lasso (DDB) estimator

$$\hat{\beta}_j^{(DDB)} = \hat{\beta}_j^{(DB)} - median\left(\hat{\beta}_j^{(*,DB)} - \hat{\beta}_j\right), \tag{3.10}$$

which is asymptotically normal under proper conditions.

3.2 Main ideas

Our analysis is based on a different error decomposition for the debiased Lasso from the one originally introduced. In Zhang and Zhang (2014), the error of the debiased Lasso is decomposed into two terms, a noise term and a remainder term:

$$\hat{\beta}_j^{(DB)} - \beta_j = \underbrace{\frac{z_j^T \epsilon}{z_j^T x_j}}_{Orig.noise} \underbrace{- \left(e_j^T - \frac{z_j^T X}{z_j^T x_j}\right)(\hat{\beta} - \beta)}_{Orig.remainder}. \tag{3.11}$$

This is the starting point of many existing analysis of the debiased Lasso (van de Geer et al., 2014; Javanmard and Montanari, 2014a; Dezeure et al., 2016). Typically, the *Orig.remainder* is bounded by $O_P(s\lambda\lambda_j)$ through an ℓ_∞-ℓ_1 splitting with λ_j in (3.4).

Our analysis is motivated by the following observations. Let S and $\hat{S}$ be the support of β and $\hat{\beta}$ respectively. For a vector $b \in \mathbb{R}^p$, let $sgn(b)$ be an element of the subdifferential of the ℓ_1 norm of b. It follows from the KKT condition of the Lasso (3.1) that

$$\hat{\beta}_{\hat{S}} = \beta_{\hat{S}} + \left(\frac{1}{n}X_{\hat{S}}^T X_{\hat{S}}\right)^{-1}\left(\frac{1}{n}X_{\hat{S}}^T \epsilon\right) - \lambda\left(\frac{1}{n}X_{\hat{S}}^T X_{\hat{S}}\right)^{-1} sgn(\hat{\beta}_{\hat{S}}) \ \text{ and } \ \hat{\beta}_{\hat{S}^c} = 0,$$

assuming that $X_{\hat{S}}^T X_{\hat{S}}/n$ is invertible. Our idea is to approximate $\hat{\beta}$ by an oracle estimator $\hat{\beta}^o$ in the analysis, where

$$\hat{\beta}_S^o = \beta_S + \left(\frac{1}{n}X_S^T X_S\right)^{-1}\left(\frac{1}{n}X_S^T\epsilon\right) - \lambda\left(\frac{1}{n}X_S^T X_S\right)^{-1} sgn(\beta_S) \text{ and } \hat{\beta}_{S^c}^o = 0, \quad (3.12)$$

when $X_S^T X_S/n$ is invertible. This estimator $\hat{\beta}^o$ is oracle as it requires the knowledge of the true support of β. However, it is different from the oracle least square estimator as the last term in $\hat{\beta}_S^o$ is added to mimic the Lasso estimator. In fact, $\hat{\beta}^o = \hat{\beta}$ when the Lasso estimator is sign consistent.

Inference based on the oracle estimator $\hat{\beta}^o$ (3.12) is relatively easy, because its approximation error does not involve random support selection. In fact, its approximation error is linear in ϵ with an unknown intercept. Our idea is that when the difference between the oracle estimator $\hat{\beta}^o$ and the Lasso estimator $\hat{\beta}$ is small, the approximation error of the debiased Lasso in (3.11) is dominated by a bias term associated with this intercept. Therefore, bootstrap can be used to remove this main bias term. Specifically, we decompose the error of the debiased Lasso in (3.2) as

$$\begin{aligned}
\hat{\beta}_j^{(DB)} - \beta_j &= \frac{z_j^T\epsilon}{z_j^T x_j} - \left(e_j^T - \frac{z_j^T X}{z_j^T x_j}\right)_S (\hat{\beta}^o - \beta) - \left(e_j^T - \frac{z_j^T X}{z_j^T x_j}\right)(\hat{\beta} - \hat{\beta}^o) \\
&= \underbrace{\frac{z_j^T\epsilon}{z_j^T x_j} - \left(e_j^T - \frac{z_j^T X}{z_j^T x_j}\right)_S \left(\frac{1}{n}X_S^T X_S\right)^{-1}\left(\frac{1}{n}X_S^T\epsilon\right)}_{Noise} \\
&\quad + \underbrace{\lambda\left(e_j^T - \frac{z_j^T X}{z_j^T x_j}\right)_S \left(\frac{1}{n}X_S^T X_S\right)^{-1} sgn(\beta_S)}_{Bias} + \underbrace{\left(e_j^T - \frac{z_j^T X}{z_j^T x_j}\right)(\hat{\beta}^o - \hat{\beta})}_{Remainder}.
\end{aligned} \quad (3.13)$$

Here the $Noise$ term is the sum of the $Orig.noise$ in (3.11) and a noise term associated with the oracle estimator $\hat{\beta}^o$ in (3.12). The $Bias$ term is from the intercept of the oracle estimator $\hat{\beta}^o$ in (3.12), which is a constant of order $O_P(s\lambda\lambda_j)$ (Remark 3.3.1 in Section 3.3). The $Remainder$ term arises from the difference between the oracle estimator $\hat{\beta}^o$ and the Lasso estimator $\hat{\beta}$ in (3.1). We prove the consistency of bootstrap when the $Remainder$ term is of order $o(n^{-1/2})$, even if the $Bias$ term is of larger order than $n^{-1/2}$. The error decomposition in (3.13) will demonstrate benefits over the decomposition in

(3.11) when the *Remainder* term in (3.13) is of smaller order than the *Orig.remainder* term in (3.11).

One way to bound the *Remainder* term in (3.13) is by considering the event that the selected support by Lasso is inside the true support and the an ℓ_∞-bound exists for its estimation error:

$$\Omega_0 = \left\{ \hat{S} \subseteq S \;\text{ and }\; \|\hat{\beta} - \beta\|_\infty \leq C_{n,p}\lambda,\ C_{n,p} > 0 \right\}. \tag{3.14}$$

Recall that when the Lasso estimator $\hat{\beta}$ is sign consistent, $\hat{\beta}^o = \hat{\beta}$ and *Remainder* in (3.13) is zero. Let $\tilde{S}$ be a set of "small" coefficients, such that $\tilde{S} = \{j : 0 < |\beta_j| \leq C_{n,p}\lambda\}$. In Ω_0, we can get $sgn(\beta_j) = sgn(\hat{\beta}_j)$, for $j \in S \backslash \tilde{S}$. And hence the sign inconsistency only occurs on $\tilde{S}$. Formally,

$$\|sgn(\hat{\beta}_S) - sgn(\beta_S)\|_1 \leq 2|\tilde{S}|. \tag{3.15}$$

We show that the *Remainder* term in (3.13) is associated with the order of $|\tilde{S}|$. This leads to the improvement in sample size requirement when $|\tilde{S}|$ is of smaller order than $|S|$.

3.3 Main results: deterministic designs

In this section, we carry out detailed analysis for deterministic designs. We first provide sufficient conditions for our theorems. For ease of notation, let $\Sigma^n = X^T X/n$.

Condition 3.3.1. *The design matrix X is deterministic with*

$$\Lambda_{\min}(\Sigma^n_{S,S}) = C_{\min} > 0 \text{ and } \max_{i \leq n, j \leq p} |X_{i,j}| \leq K_0.$$

Condition 3.3.2.

$$\left\| \Sigma^n_{S^c,S} (\Sigma^n_{S,S})^{-1} \right\|_\infty \leq \kappa < 1.$$

Condition 3.3.3.

$$\left\| (\Sigma^n_{S,S})^{-1} \right\|_\infty \leq K_1 < \infty.$$

Condition 3.3.4. *ϵ_i, $i = 1, \dots, n$, are i.i.d. random variables from a distribution with $\mathbb{E}[\epsilon_1] = 0, \mathbb{E}[\epsilon_1^2] = \sigma^2$ and $\mathbb{E}[|\epsilon_1|^4] \leq M_0$.*

Condition 3.3.5. *For any $j \leq p$, $\|z_j\|_4^4 = o(\|z_j\|_2^4)$ and $\|z_j\|_2^2/n \geq K_2 > 0$.*

As K_1 is assumed to be a constant in Condition 3.3.3, the eigenvalue condition in Condition 3.3.1 is redundant in the sense that $\Lambda_{\min}(\Sigma_{S,S}^n) \geq 1/K_1$. Note that the eigenvalue condition and Condition 3.3.3 are only required on a block of the Gram matrix consisting of rows and columns in the true support. The quantity in Condition 3.3.2 is called incoherence parameter (Wainwright, 2009). This condition is equivalent to the uniformity of the strong irrepresentable condition (Zhao and Yu, 2006) over all sign vectors. Another related condition, the neighborhood stability condition (Meinshausen and Bühlmann, 2006), has been studied for model selection in Gaussian graphical models. Condition 3.3.3 is required for establishing an ℓ_∞-bound of estimation error of the Lasso estimator. Condition 3.3.4 involves only first four moments of ϵ allowing some heavy-tailed distributions. Condition 3.3.5 contains some regularity conditions on z_j, which are verifiable after the calculation of z_j.

We first prove that event Ω_0 in (3.14) holds true with large probability for deterministic designs.

Lemma 3.3.1. *Suppose that Conditions 3.3.1 - 3.3.4 are satisfied and (n, p, s, λ) satisfies that*

$$n \geq \frac{32\sigma^2}{\lambda^2(1-\kappa)^2} \quad \textit{and} \quad \lambda > \frac{16\sigma}{1-\kappa}\sqrt{\frac{2\log p}{n}}. \tag{3.16}$$

Then it holds that

$$\hat{S} \subseteq S \quad \textit{and} \quad \|\hat{\beta}_S - \beta_S\|_\infty \leq \underbrace{K_1\lambda + 8\sigma\sqrt{\frac{2\log p}{C_{\min} n}}}_{g_1(\lambda)}, \tag{3.17}$$

with probability greater than $1 - 4\exp(-c_1 \log p) - c_2/n$ for some $c_1, c_2 > 0$.

Lemma 3.3.1 is proved in Appendix. Lemma 3.3.1 asserts that the Lasso estimator does not have false positive selection with large probability under Conditions 3.3.1 - 3.3.4. It is known that Condition 3.3.3 and beta-min condition together imply the selection consistency of the Lasso estimator. However, we do not impose the beta-min condition but distinguish the effects of small and large signals. Note that $g_1(\lambda) \asymp \lambda$ for $\lambda \asymp \sqrt{\log p/n}$.

Next we show that analogous results of Lemma 3.3.1 hold for the bootstrap version of the Lasso estimator $\hat{\beta}^*$.

Lemma 3.3.2. *Assume that Conditions 3.3.1 - 3.3.4 are satisfied. If* (n, p, s, λ) *satisfies (3.16),* $n \gg s \log p$ *and*

$$\frac{4\sigma}{1-\kappa}\sqrt{\frac{2\log p}{n}} \leq \lambda \asymp \sqrt{\log p/n}, \tag{3.18}$$

then with probability going to 1,

$$\hat{S}^* \subseteq S \text{ and } \|\hat{\beta}^*_S - \hat{\beta}_S\|_\infty \leq \underbrace{K_1\lambda + 2\sigma\sqrt{\frac{2\log p}{C_{\min} n}}}_{g_1'(\lambda)}. \tag{3.19}$$

Lemma 3.3.2 is proved in Appendix. We mention that the condition $n \gg s \log p$ is required for the consistency of $\hat{\sigma}^2$ (see Lemma A.2.4 for details). Note that it is $\hat{S}$ instead of S that is the true support under the bootstrap resampling proposal and $\hat{S} \subseteq S$ with large probability by Lemma 3.3.1. In fact, as will be proved in the next lemma, (3.19) is sufficient for achieving the following decomposition.

$$\begin{aligned}\hat{\beta}_j^{(*,DB)} - \hat{\beta}_j = & \underbrace{\frac{z_j^T \hat{\epsilon}^*}{z_j^T x_j} - \left(e_j^T - \frac{z_j^T X}{z_j^T x_j}\right)_S \left(\frac{1}{n}X_S^T X_S\right)^{-1} \left(\frac{1}{n}X_S^T \hat{\epsilon}^*\right)}_{Noise^*} + Bias \\ & + \underbrace{\lambda \left(e_j^T - \frac{z_j^T X}{z_j^T x_j}\right)_S \left(\frac{1}{n}X_S^T X_S\right)^{-1} \Big(sgn(\hat{\beta}^*_S) - sgn(\beta_S)\Big)}_{Remainder^*},\end{aligned} \tag{3.20}$$

for $Bias$ defined in (3.13). Thus, bootstrapping the debiased Lasso is consistent when $Remainder$ and $Remiander^*$ are sufficiently small and the bootstrap approximation of $Noise^*$ to $Noise$ is consistent.

To bound the remainder terms, let $\tilde{s}$ be the number of small coefficients, such that

$$\tilde{s} = \left|\left\{j : 0 < |\beta_j| < g_1(\lambda) + g_1'(\lambda)\right\}\right|, \tag{3.21}$$

for $g_1(\lambda)$ and $g_1'(\lambda)$ defined in (3.17) and (3.19) respectively.

Lemma 3.3.3. *Suppose that Conditions 3.3.1 - 3.3.4 hold true,* $\lambda \asymp \sqrt{\log p/n}$ *satisfies (3.16) and (3.18) and* $n \gg s \log p$*. For* $\hat{\beta}_j^{(DB)}$ *and* $\hat{\beta}_j^{(*,DB)}$ *defined in (3.2) and (3.8)* *respectively, we have*

$$\mathbb{P}\left(\left|\hat{\beta}_j^{(DB)} - \beta_j - Noise - Bias\right| > 2K_1 \frac{\tilde{s}\lambda\lambda_j}{z_j^T x_j/n}\right) = o(1), \tag{3.22}$$

$$\mathbb{P}\left(\left|\hat{\beta}_j^{(*,DB)} - \hat{\beta}_j - Noise^* - Bias\right| > 2K_1 \frac{\tilde{s}\lambda\lambda_j}{z_j^T x_j/n}\right) = o(1), \tag{3.23}$$

where $Noise$ and $Bias$ are defined in (3.13), $Noise^*$ is defined in (3.20) and $\tilde{s}$ is defined in (3.21).

Lemma 3.3.3 is proved in Appendix. The factor $z_j^T x_j/n$ is calculable and can be treated as a positive constant typically. In fact, this factor is proportional to the standard deviation of $Noise$ and $Noise^*$, so that it will be cancelled in the analysis of the asymptotic normality. Therefore, we have proved that the $Remainder$ term in (3.13) and the $Remainder^*$ term in (3.20) are of order $O_P(\tilde{s}\lambda\lambda_j)$.

Remark 3.3.1. *Under Conditions 3.3.1 - 3.3.4 and $\lambda \asymp \lambda_j \asymp \sqrt{\log p/n}$, we can get a natural upper bound on $Bias$ in (3.13):*

$$Bias = O_P\left(\frac{s\lambda\lambda_j}{(z_j^T x_j/n)C_{\min}}\right) = O_P\left(\frac{s\log p}{n}\right) = o_P(1).$$

Note that the order of $Bias$ is not guaranteed to be $o(n^{-1/2})$ under the sample size conditions of Lemma 3.3.3. There will be no guarantee of improvement on the sample size requirement if we do not remove the $Bias$ term.

Inference for β_j is based on the following pivotal statistics

$$R_j = \frac{z_j^T x_j}{\|z_j\|_2}(\hat{\beta}_j^{(DB)} - \beta_j) \quad \text{and} \quad R_j^* = \frac{z_j^T x_j}{\|z_j\|_2}(\hat{\beta}_j^{(*,DB)} - \hat{\beta}_j), \tag{3.24}$$

where $\hat{\beta}_j^{(DB)}$ and $\hat{\beta}_j^{(*,DB)}$ are defined in (3.2) and (3.8) respectively. We show the consistency of bootstrap approximation of R_j^* to R_j as well as the asymptotic normality of a pivot based on the double debiased Lasso estimator $\hat{\beta}_j^{(DDB)}$ in (3.10):

$$R_j^{(DDB)} = \frac{z_j^T x_j}{\hat{\sigma}\|z_j\|_2}(\hat{\beta}_j^{(DDB)} - \beta_j). \tag{3.25}$$

We specify the sample size conditions as following:

$$\begin{aligned}\mathcal{A}_1 = \Big\{&(n,p,s,\tilde{s},\lambda,\lambda_j) : (n,p,s,\lambda) \text{ satisfies (3.16) and (3.18)},\ \lambda \asymp \lambda_j \asymp \sqrt{\log p/n} \\ &\text{and}\ \ n \gg \max\{s\log p, (\tilde{s}\log p)^2\} \text{ for } \tilde{s} \text{ in (3.21)}\Big\}.\end{aligned} \tag{3.26}$$

As mentioned in Section 1, the condition on the overall sparsity recovers the rate of point estimation. If $\tilde{s} \ll s$, our sample size condition is weaker than the typical one $n \gg (s \log p)^2$.

Theorem 3.3.4. *Assume that Conditions 3.3.1 - 3.3.5 hold true and* $(n, p, s, \tilde{s}, \lambda, \lambda_j) \in \mathcal{A}_1$*. Then for* R_j *and* R_j^* *defined in (3.24), it holds that*

$$\sup_{\alpha \in (0,1)} \left| \mathbb{P}\{R_j \le q_\alpha(R_j^*)\} - \alpha \right| = o_P(1),$$

where $q_\alpha(R_j^*)$ *is the* α*-quantile of the distribution of* R_j^**. Moreover, for* $R_j^{(DDB)}$ *defined in (3.25), we have*

$$\sup_{\alpha \in (0,1)} \left| \mathbb{P}(R_j^{(DDB)} \le z_\alpha) - \alpha \right| = o_P(1).$$

Theorem 3.3.4 is proved in Appendix. Based on Theorem 3.3.4, a two-sided $100 \times (1-\alpha)\%$ confidence interval for β_j can be constructed as in (3.9).

For the double debiased estimator (3.10), the *Bias* in (3.20) is estimated by the median of the distribution of $\hat{\beta}_j^{(*,DB)} - \hat{\beta}_j$. In practice, the *median* $\left(\hat{\beta}_j^{(*,DB)} - \hat{\beta}_j\right)$ can be approximated by the sample median of bootstrap realizations.

Remark 3.3.2. *Suppose we are interested in making inference for a linear combination of regression coefficients* $\langle a_0, \beta \rangle$ *for* $a_0 \in \mathbb{R}^p$*. It is not hard to see that Gaussian bootstrap remains consistent under the conditions of Theorem 3.3.4 if* $\|a_0\|_1 / \|a_0\|_2$ *is bounded.*

3.4 Main results: Gaussian designs

This section includes main results in the case of Gaussian designs. The proof follows similar steps as for deterministic designs. We first describe conditions we impose in our theorems.

Condition 3.4.1. *The design matrix* X *has independent Gaussian rows with mean 0 and covariance* Σ*.*

Condition 3.4.2.

$$\left\| \Sigma_{S^c,S} \Sigma_{S,S}^{-1} \right\|_\infty \le \kappa < 1.$$

Condition 3.4.3.

$$\left\|\Sigma_{S,S}^{-1/2}\right\|_{\infty}^{2} \le K_1 < \infty.$$

Condition 3.4.4.

$$\Lambda_{\min}(\Sigma_{S,S}) \ge C_{\min},\ \max_{j\le p}(\Sigma^{-1})_{j,j} \le 1/C_* \ \textit{and}\ \max_{j\le p}\Sigma_{j,j} \le C^* < \infty.$$

Condition 3.4.5. *ϵ_i, $i = 1, \ldots, n$, are i.i.d from Gaussian distribution with mean 0 and variance σ^2.*

Condition 3.4.2 - Condition 3.4.3 are population versions of Condition 3.3.2 - Condition 3.3.3. In Condition 3.4.4, we require that the largest diagonal element of Σ^{-1} is upper bounded, in order to lower bound $\|z_j\|_2^2/n$ asymptotically. It is also worth mentioning that Condition 3.4.3 is related to condition (iii) in Javanmard and Montanari (2015) (denoted as [JM15]). Specifically, [JM15] required that

$$\rho(\Sigma, C_0 s_0) = \max_{T\subseteq[p],|T|\le C_0 s}\left\|\Sigma_{T,T}^{-1}\right\|_{\infty} < \rho, \quad \text{for } C_0 \ge 33.$$

This condition is on a set T, which is actually the support of the estimation error of a perturbed Lasso estimator, while Condition 3.4.3 is assumed on the true support S.

Lemma 3.4.1. *Assume that Conditions 3.4.1 - 3.4.5 are satisfied and (n, p, s, λ) satisfies that*

$$s\lambda^2 \le \frac{\sigma^2 C_{\min}}{2} \quad \textit{and}\ \lambda \ge \frac{8\sigma}{1-\kappa}\sqrt{\frac{C^*\log p}{n}}. \tag{3.27}$$

Then for

$$g_2(\lambda) = (1 + C_n)K_1\lambda + 4\sigma\sqrt{\frac{\log p}{C_{\min} n}} \quad \textit{with}\ \ C_n = O\left(\frac{s \vee \sqrt{s}\log p}{\sqrt{n}}\right), \tag{3.28}$$

it holds that

$$\hat{S} \subseteq S \quad \textit{and}\ \ \|\hat{\beta}_S - \beta_S\|_\infty \le g_2(\lambda), \tag{3.29}$$

with probability greater than $1 - c_1/n - 2\exp(-c_2\log p) - 2\exp(-c_3 n)$ for some $c_1, c_2, c_3 > 0$.

Lemma 3.4.1 is proved in Appendix. Note that $g_2(\lambda) = o(1)$ if $n \gg s \log p$. If $s^2 \vee s \log p = O(n)$, then $g_2(\lambda) = O(\lambda)$. In fact, Theorem 3 in Wainwright (2009) considers the same scenario, but their results require $s \to \infty$ and their upper bound on $\|\hat{\beta} - \beta\|_\infty$ only holds for sign consistency case.

In the next Lemma, we prove a bootstrap analogue of Lemma 3.4.1.

Lemma 3.4.2. *Assume that Conditions 3.4.1 - 3.4.5 hold true. If (n, p, s, λ) satisfies (3.27), $n \gg s \log p$ and*

$$\frac{4\sigma}{1-\kappa}\sqrt{\frac{\log p}{n}} \leq \lambda \asymp \sqrt{\frac{\log p}{n}}, \tag{3.30}$$

then with probability going to 1,

$$\hat{S}^* \subseteq S \text{ and } \|\hat{\beta}^*_S - \hat{\beta}_S\|_\infty \leq g_2(\lambda), \text{ for } g_2(\lambda) \text{ in (3.28)}. \tag{3.31}$$

Lemma 3.4.2 in proved in Appendix. Same as the deterministic design case, the condition $n \gg s \log p$ is required for the consistency of $\hat{\sigma}^2$.

Under Conditions 3.4.1 - 3.4.5, we prove the consistency of Gaussian bootstrap under Gaussian designs.

For $g_2(\lambda)$ defined in (3.28), define

$$\tilde{s} = |\{j : 0 < |\beta_j| < 2g_2(\lambda)\}|. \tag{3.32}$$

We specify the required sample size condition as following:

$$\mathcal{A}_2 = \Big\{(n, p, s, \tilde{s}, s_j, \lambda, \lambda_j) : (n, p, s, \lambda) \text{ satisfies (3.27) and (3.30)}, \ \lambda \asymp \lambda_j \asymp \sqrt{\log p/n} \\ \text{and } \ n \gg \max\{s \log p, s\tilde{s} \log p, (\tilde{s} \log p)^2, s_j \log p\} \text{ for } \tilde{s} \text{ in (3.32)}\Big\}. \tag{3.33}$$

Theorem 3.4.3. *Suppose that Conditions 3.4.1 - 3.4.5 are satisfied and $(n, p, s, \tilde{s}, s_j, \lambda, \lambda_j) \in \mathcal{A}_2$. Then it holds that*

$$\sup_{\alpha \in (0,1)} \left|\mathbb{P}\{R_j \leq q_\alpha(R_j^*)\} - \alpha\right| = o_P(1).$$

Moreover, for $R_j^{(DDB)}$ defined in (3.25), we have

$$\sup_{\alpha \in (0,1)} \left|\mathbb{P}(R_j^{(DDB)} \leq z_\alpha) - \alpha\right| = o_P(1).$$

It can be seen from the proof that condition $n \gg s\tilde{s}\log p$ in (3.33) is used to achieve desired rates of $|Remainder|$ and $|Remainder^*|$, such that $\|(\Sigma^n_{S,S})^{-1}\|_\infty \tilde{s}\lambda^2 = o_P(n^{-1/2})$. The condition $n \gg s_j \log p$ is required to prove that $\|z_j\|_2^2/n$ is asymptotically bounded away from zero.

In terms of the sparsity requirements, $\mathcal{A}_2$ (3.33) implies that it is sufficient to require $s = O(\sqrt{n})$ and $\tilde{s} = o(\sqrt{n}/\log p)$. Compared with the typical condition, $s = o(\sqrt{n}/\log p)$, our condition allows at least an extra order of $\log p$. Moreover, if $\tilde{s}$ is constant, our requirement on s is $s \ll n/\log p$, which recovers the rate of point estimation. Comparing with the sparsity condition assumed in [JM15] for unknown Gaussian design case, our analysis still benefits when $\tilde{s}$ is sufficiently small:

- If the sparsity of the j-th column of precision matrix is much larger than the sparsity of β, i.e. $\tilde{s} \leq s \ll s_j$, [JM15] required $n \gg \max\{(s\log p)^2, s_j \log p\}$, which is no better than the rate in $\mathcal{A}_2$ (3.33) as discussed above. If $\tilde{s} \ll s$, $\mathcal{A}_2$ is weaker than the sparsity conditions assumed in [JM15].

- If the j-th column of the precision matrix is much sparser, i.e. $s \gg s_j$, [JM15] required that $n \gg \max\{s(\log p)^2,\ (s_j \log p)^2\}$. If $\tilde{s} \ll \log p$, then $s(\tilde{s} \vee 1)\log p \ll s(\log p)^2$ and hence the sample size condition in $\mathcal{A}_2$ is weaker. If $\tilde{s} \gg \log p$, [JM15] required weaker condition on s but stronger condition on s_j.

3.5 Simulations

In this section, we report the performance of the debiased Lasso with Gaussian bootstrap and other comparable methods in simulation experiments.

Consider deterministic design case with $n = 100$, $p = 500$, $X_i \sim N(0, I_p)$ and $\epsilon_i \sim N(0,1)$. We consider a relatively large sparsity level, $s = 20$, and two levels of true regression coefficients as following.

(i) All the signals are strong: $\beta_1 = \cdots = \beta_{20} = 2$.

(ii) A large proportion of signals are strong: $\beta_1 = \cdots = \beta_5 = 1$, $\beta_6 = \cdots = \beta_{20} = 2$.

We compare the performance of bootstrapping the debiased Lasso (BS-DB), the debiased Lasso without bootstrap (DB) and the Adaptive Lasso with residual bootstrap (BS-ADP). For BS-DB, we generate $(1-\alpha)\%$ confidence interval (CI) according to (3.9) with 500 bootstrap resamples. We take $\lambda = \lambda_j$ at the universal level for the Lasso procedures. For DB, we estimate the noise level by (3.7) and take $\lambda = \lambda_j$ at the universal level for the Lasso procedures. $(1-\alpha)\%$ confidence intervals are generated according to

$$\left(\hat{\beta}_j^{(DB)} + \hat{\sigma} z_{\alpha/2} \frac{\|z_j\|_2}{z_j^T x_j},\ \hat{\beta}_j^{(DB)} + \hat{\sigma} z_{1-\alpha/2} \frac{\|z_j\|_2}{z_j^T x_j}\right).$$

For BS-ADP, we consider the pivot defined in (4.2) of Chatterjee and Lahiri (2013), which can achieve second-order correctness under some conditions. Such estimators also have a bias-correction term, which can be explicitly calculated assuming sign consistency. The choices of $\lambda_{1,n}$ and $\lambda_{2,n}$ are according to Section 6 of Chatterjee and Lahiri (2013). Each confidence interval is generated with 500 bootstrap resamples.

We construct two-sided 95% confidence intervals using each of the aforementioned methods. Each setting is replicated with 1000 independent realizations. In the following table, we report the average coverage probability on S and S^c ($\widehat{cov}_S$ and $\widehat{cov}_{S^c}$, respectively) as well as the average length of CIs on S and S^c (ℓ_S and ℓ_{S^c}, respectively) for identity covariance matrix and equicorrelated covariance matrix with $\Sigma_{j,j} = 1$ and $\Sigma_{j,k} = 0.2$ $(j \neq k)$.

β	Methods	$\Sigma_{j,k} = 0$ $(j \neq k)$				$\Sigma_{j,k} = 0.2$ $(j \neq k)$			
		$\widehat{cov}_S$	$\widehat{cov}_{S^c}$	ℓ_S	ℓ_{S^c}	$\widehat{cov}_S$	$\widehat{cov}_{S^c}$	ℓ_S	ℓ_{S^c}
(i)	BS-DB	0.997	0.999	1.127	0.539	0.893	0.997	1.074	0.436
	DB	0.940	0.982	0.886	0.885	0.627	0.934	0.760	0.778
	BS-ADP	0.274	0.950	0.181	0.201	0.241	0.945	0.432	0.319
(ii)	BS-DB	0.974	0.998	1.057	0.554	0.820	0.987	0.963	0.424
	DB	0.939	0.982	0.887	0.886	0.649	0.925	0.716	0.733
	BS-ADP	0.279	0.951	0.195	0.187	0.280	0.943	0.638	0.280

One can see that BS-DB always gives larger coverage probabilities than DB across different settings. We mention that noise level is overestimated. For example, in setting (i) and (ii) with the identity covariance matrix, the average of $\hat{\sigma}$ is 2.240 and 2.244, respectively. The CIs given by BS-DB are longer than those computed with DB on S, but on S^c the CIs given by BS-DB are shorter than the ones given by DB. On the other hand, BS-ADP exhibits the overconfidence phenomenon: the average lengths of CIs are small, which results in low coverage probabilities on S. In the presence of equicorrelation, which is a harder case, BS-DB is significantly better than DB and BS-ADP in terms of coverage probability.

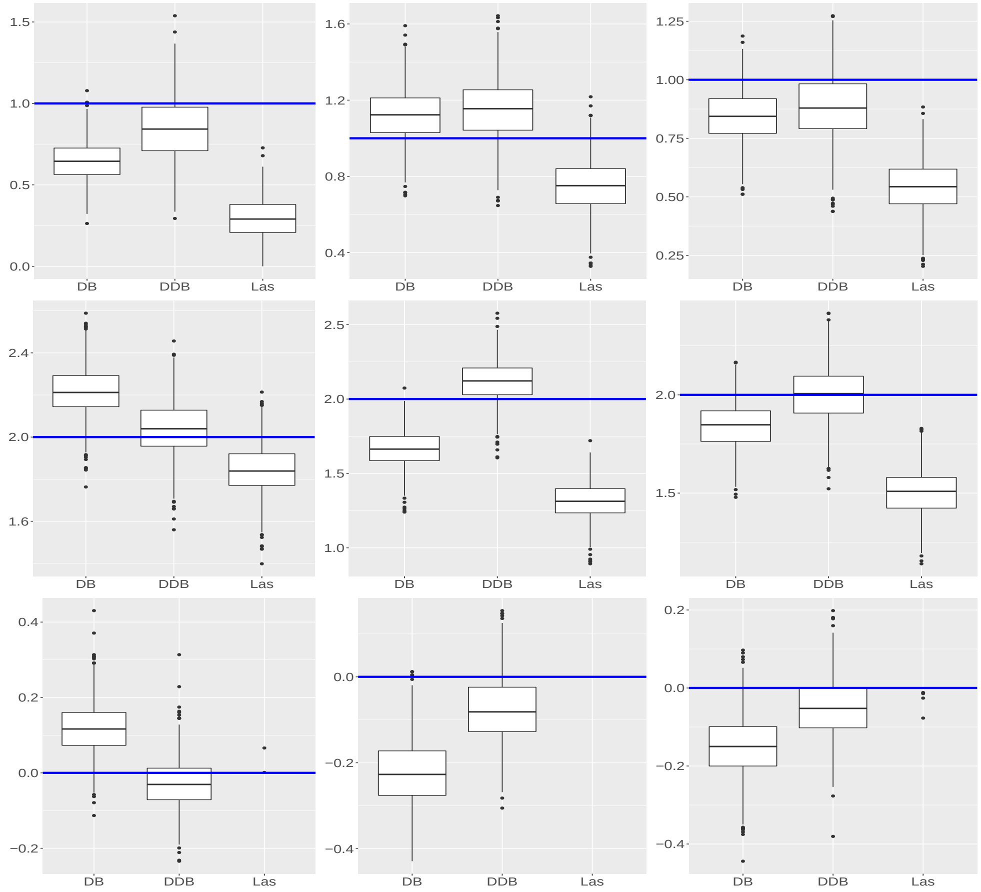

Figure 3.1: Boxplots of the double debiased Lasso (DDB) (3.10), the debiased Lasso (DB) (3.2) and the Lasso (Las) (3.1) with the identity covariance matrix in setting (ii). First row consists of estimates for strong signals: $\beta_6 = \beta_7 = \beta_8 = 2$. Second row consists of estimates for strong signals: $\beta_1 = \beta_2 = \beta_3 = 1$. Third row consists of estimates for zeros: $\beta_{21} = \beta_{22} = \beta_{23} = 0$. Each Boxplot is based on 1000 independent replications.

Figure 3.1 demonstrates the bias-correction effects of debiasing and bootstrap across different levels of signal strengths. Concerning the overall performance, DDB is better than DB in terms of bias-correction, which is in line with our theoretical results. For $j \in S$, DDB and DB are less biased than the Lasso estimators. On S^c, the Lasso estimates the regression coefficients as zero with a large probability. Thus, the Boxplot degenerates to a point at zero with a few outliers. Comparing row-wise, one can see that bootstrap has more significant correction effects on strong signals (first row) than on weak signals (second row). When true coefficients are zeros, DDB is also less biased than DB.

3.6 Discussions

We consider the bias-correction effect of bootstrap for statistical inference with debiased Lasso under proper conditions. Our analysis on the approximation error of debiased Lasso admits sample size conditions in terms of the number of weak signals. Our results contribute to the inference problem in the regime $s \log p \ll n \lesssim (s \log p)^2$, but also demonstrate the benefits of having strong signals for the debiased Lasso procedure. We establish the consistency of Gaussian bootstrap and show that confidence intervals can be constructed based on bootstrap samples.

Besides Gaussian bootstrap, we also considered residual bootstrap, which is robust in the presence of heteroscedastic errors. However, the proof involves a more technical analysis and may impair the sample size conditions. To focus on the main idea, this is omitted from the paper. We also considered the proof techniques in [JM15], which construct a perturbed version of the Lasso estimator assuming β_j is known and utilize its independence of x_j. However, these techniques cannot be directly applied to the bootstrapped debiased Lasso, since the "true" parameters $\hat{\beta}$ and $\hat{\sigma}$ under the bootstrap resampling plan are not independent with x_j for $j \in S$.

Chapter 4

Dealing with pleiotropy in Mendelian Randomization

In this chapter, we consider causal effect estimation with instruments which may violate the ER assumption (Figure 1.1). Mathematically, our goal is to identify β given possibly nonzero α in model (1.5). We first consider a single Gaussian prior on α and characterize the estimation error of the Bayes rule in terms of the ratio of the variation within α and the instrument strength. We also study a joint estimation of unknown hyper parameters and parameters of interests. Secondly, we consider a mixture Gaussian prior to deal with sparse α. We study the estimation error of posterior mean and propose a computation algorithm to deal with unknown hyper parameters. The proposed method is generalized to fit summarized data under some conditions. Simulations and real studies are demonstrated in Section 4.5.

4.1 Motivation of Bayesian methods

Bayesian methods have been used to deal with pleiotropic effects in MR for its robustness to model misspecification (Thompson et al., 2017; Schmidt and Dudbridge, 2017). Our analysis of Bayesian methods is motivated by following observations from a frequentist point of view. Suppose that pleiotropic effects α_j are all equal but not necessarily zero, i.e. $\alpha_1 = \cdots = \alpha_J = \mu_\alpha$ and μ_α is unknown. Then the exposure-outcome model (1.6) can be written as

$$Y_i = \hat{D}_i\beta + \tilde{Z}_i\mu_\alpha + \hat{\eta}_i, \tag{4.1}$$

where $\tilde{Z}_i = \sum_{j=1}^{J} Z_{i,j}$. A multiple least square estimator with respect to (4.1) is consistent under typical assumptions as long as $\hat{D}$ is linearly independent of $\tilde{Z}$. This motivates us to develop an estimator of β, which is consistent when the variation within pleiotropic effects α is “sufficiently” small. Such estimators can be very useful when we have prior knowledge of the approximate variation of α. Let us consider a regularized least square estimator with ℓ_2-regularization:

$$(\hat{\beta}, \hat{\alpha}) = \underset{b \in \mathbb{R}, a \in \mathbb{R}^J}{\arg\min} \left\{ \|Y - \hat{D}b - Za\|_2^2 + \lambda \sum_{j=1}^{J} (a_j - \mu_\alpha)^2 \right\}, \tag{4.2}$$

for some $\mu_\alpha \in \mathbb{R}$ and tuning parameter $\lambda > 0$. The penalty term in (4.2) plays two roles: Firstly, it regularizes the rank deficient Gram matrix formed by $(\hat{D}, Z)$. Secondly, it favors small variation in pleiotropic effects α. Specially, if $\alpha_1 = \cdots = \alpha_J = \mu_\alpha$, then the penalty part of (4.2) is zero. It is worth mentioning that different from a typical Ridge-regression, the causal effect β is not penalized in (4.2). We can equivalently formulate (4.2) in a Bayesian framework by specifying some proper priors on α, in order to get some generalization on the penalty term in later sections. We mention that we focus on the low-dimensional scenario of α, i.e. $J < n$.

4.2 Hierarchical models

Hierarchical models are useful tools for pooling information and simultaneous inference. In Genetics, effect sizes are often modeled under a Gaussian prior (Stephens and Balding, 2009). We first consider a single Gaussian prior where the hyper parameters can be known or unknown. Then we consider utilizing a mixture Gaussian prior to deal with possibly sparse pleiotropic efffects.

4.2.1 Single Gaussian prior

It is not hard to see that $(\hat{\beta}, \hat{\alpha})$ defined in (4.2) can be viewed as the posterior mode of the following hierarchical model.

$$Y_i | \hat{D}, Z, \beta, \alpha, \sigma_\eta^2 \sim_{ind} N(\hat{D}_i \beta + Z_i \alpha, \sigma_\eta^2), \tag{4.3}$$

$$\alpha_j | \sigma_\eta^2 \sim_{iid} N(\mu_\alpha, \tau_n^2 \sigma_\eta^2), \tag{4.4}$$

$$\sigma_\eta^{-2} \sim \text{Gamma}(\nu_1, \nu_2), \tag{4.5}$$

where $\text{Gamma}(a, b)$ is Gamma distribution with shape parameter a and rate parameter b and μ_α, τ_n^2, ν_1, ν_2 are some prespecified constants based on our prior knowledge. We emphasize that (4.3) - (4.5) are not assumptions on the unknown parameters but used to induced proper regularization from its log-likelihood. Note that the target function in (4.2) is proportional to its negative log likelihood with $\lambda = \tau_n^{-2}$. We set ν_1 and ν_2 to be small in order to make the priors noninfluential. In (4.4), we allow the variance of α_j, τ_n^2, to depend on sample size n. The next theorem justifies a non-asymptotic error bound for the Bayes rule $\hat{\beta}^{(SG)}$ in the model (4.3) - (4.5) with the single Gaussian prior. Let $\Sigma_Z^n = Z^T Z/n$.

Condition 4.2.1. *The eigenvalues of Σ_Z^n are bounded from above and below.*

Theorem 4.2.1. *If Condition 4.2.1 holds, then estimation error of $\hat{\beta}$ satisfies*

$$|\hat{\beta}^{(SG)} - \beta| \leq \frac{\Lambda_{\max}^{1/2}(\Sigma_Z^n)}{\Lambda_{\min}^{1/2}(\Sigma_Z^n)} \frac{\|\alpha - \mu_\alpha \mathbb{1}_J\|_2}{\|\hat{\gamma}\|_2} + \frac{\tau_n^2 \sqrt{n} \Lambda_{\max}^{1/2}(\Sigma_Z^n) \|Z^T P_{\hat{D}}^{\perp} \hat{\eta}\|_2}{\|\hat{D}\|_2} + \frac{|\hat{D}^T \hat{\eta}|}{\|\hat{D}\|_2^2}. \tag{4.6}$$

The proof of Theorem 4.2.1 can be found in the Appendix. Inequality (4.6) provides an empirical bound for the estimation error of $\hat{\beta}^{(SG)}$. The first term arises from the bias of $\hat{\beta}^{(SG)}$, which is caused by the regularization on α. It is quantified by the ratio of variation of pleiotropic effects from μ_α, $\|\alpha - \mu_\alpha \mathbb{1}_J\|_2$, and overall instrument strength $\|\hat{\gamma}\|_2$. The eigenvalues of the sample Gram matrix Σ_Z^n are positive finite numbers when Σ_Z^n is positive definite, which is not hard to satisfy in low-dimensional setting. The last two terms in (4.6) arise from the noise. Specifically, the variance parameter in the prior distribution τ_n^2 plays a role in regularizing the second term. We need small enough τ_n^2 for the second term to converge to zero. In the next Corollary, we study the order of the noise-related terms, which sheds lights on the selection of τ_n^2.

Corollary 4.2.2. *If Condition 4.2.1 holds true, then*

$$|\hat{\beta}^{(SG)} - \beta| \leq \frac{\|\alpha - \mu_\alpha \mathbb{1}_J\|_2}{\|\hat{\gamma}\|_2} \frac{\Lambda_{\max}^{1/2}(\Sigma_Z^n)}{\Lambda_{\min}^{1/2}(\Sigma_Z^n)} + O_P\left(\frac{\tau_n^2 \sqrt{J} n \sigma_\epsilon \Lambda_{\max}(\Sigma_Z^n)}{\|\hat{D}\|_2} + \frac{\sigma_\epsilon}{\|\hat{D}\|_2} \right). \tag{4.7}$$

Corollary 4.2.2 implies that for $\hat{\beta}$ to be consistent, it is sufficient for the variation of α from μ_α to be of smaller order than the instrument strength, i.e. $\|\alpha - \mu_\alpha\|_2 = o(\|\hat{\gamma}\|_2)$, and parameter $\tau_n^2 = o(\|\hat{D}\|_2/(\sqrt{J}n))$.

Note that if we fix $\mu_\alpha = 0$, then the first term in (4.7) is proportional to the ratio of the strength of pleiotropic effects, $\|\alpha\|_2$, and instrument strength, $\|\hat{\gamma}\|_2$. In many applications, the hyper parameter μ_α is unknown. We propose to treat it as another nuisance parameter and estimate $(\beta, \alpha, \mu_\alpha)$ via the profile likelihood method with respect to (4.2). This can be considered as an empirical Bayes method, where hyper parameters are estimated from the sample. Specifically, consider estimating unknown μ_α in the following way.

$$(\hat{\beta}^{(SG*)}, \hat{\alpha}^{(SG*)}, \hat{\mu}_\alpha^{(SG*)}) = \underset{b\in\mathbb{R}, a\in\mathbb{R}^J, \mu_a\in\mathbb{R}}{\arg\min} \left\{ \|Y - \hat{D}b - Za\|_2^2 + \tau_n^{-2} \sum_{j=1}^{J} (a_j - \mu_a)^2 \right\}, \tag{4.8}$$

for some $\tau_n^2 > 0$.

For two vectors $a_1 \in \mathbb{R}^k, a_2 \in \mathbb{R}^k$, let $\mathrm{Cor}(a_1, a_2)$ be the correlation between a_1 and a_2, i.e.

$$\mathrm{Cor}(a_1, a_2) = \langle a_1, a_2 \rangle / \|a_1\|_2 \|a_2\|_2.$$

Condition 4.2.2. *For* $\tilde{Z} = Z\mathbb{1}_J$,

$$|Cor(\hat{D}, \tilde{Z})| < \sqrt{\frac{\Lambda_{\min}(\Sigma_Z^n)}{\Lambda_{\max}(\Sigma_Z^n)} + 1} - 1.$$

Theorem 4.2.3. *Assume that Conditions 4.2.1 and 4.2.2 hold true. If* $n\tau_n^2 \leq \Lambda_{\max}^{-1}(\Sigma_Z^n)$, *then* $\hat{\beta}^{(SG*)}$ *defined in (4.8) satisfies*

$$|\hat{\beta}^{(SG*)} - \beta| \leq \frac{\Lambda_{\max}^{1/2}(\Sigma_Z^n)}{\Lambda_{\min}^{1/2}(\Sigma_Z^n)} \frac{\|\alpha - \bar{\alpha}\mathbb{1}_J\|_2}{n\tau_n^2 r_n^* \|\hat{\gamma}\|_2} + O_P\left(\frac{\sqrt{J}\sigma_\epsilon \Lambda_{\max}(\Sigma_Z^n)}{r_n^* \|\hat{D}\|_2} + \frac{\sigma_\epsilon}{\|\hat{D}\|_2} \right), \tag{4.9}$$

where $r_n^* = \Lambda_{\min}(\Sigma_Z^n) + \Lambda_{\max}(\Sigma_Z^n) - \Lambda_{\max}(\Sigma_Z^n)(|Cor(\hat{D}, \tilde{Z})| + 1)^2$.

The proof of Theorem 4.2.3 is in the Appendix. Theorem 4.2.3 implies that if the correlation between $\hat{D}$ and $\tilde{Z}$ is sufficiently small, then the bias of $\hat{\beta}^{(SG*)}$ caused by regularization is proportional to the ratio of the variation of α (from its mean) and the overall instrument strength. We can see that the first term on the right hand side of (4.9) favors large τ_n^2. Together with our conditions, it suggests taking $n\tau_n^2 \asymp 1$.

Remark 4.2.1. *The Egger's regression method is also related to the variation of pleiotropic effects. The causal effect estimate via Egger's regression is*

$$\hat{\beta}^{(Egger)} = \frac{(\hat{\gamma} - \bar{\hat{\gamma}}\mathbb{1}_J)^T (\hat{\Gamma} - \bar{\hat{\Gamma}}\mathbb{1}_J)}{\|\hat{\gamma} - \bar{\hat{\gamma}}\mathbb{1}_J\|_2^2},$$

where $\hat{\Gamma}$ *is the least square estimate of* Y *on* Z*. It can be easily shown that*

$$|\hat{\beta}^{(Egger)} - \beta| \leq \frac{\|\alpha - \bar{\alpha}\mathbb{1}_J\|_2}{\|\hat{\gamma} - \bar{\hat{\gamma}}\mathbb{1}_J\|_2} + O_P\left(\frac{\sigma_\epsilon}{\sqrt{n}\Lambda_{\min}^{1/2}(\Sigma_Z^n)\|\gamma - \bar{\gamma}\mathbb{1}_J\|_2}\right).$$

The bias term of Egger's estimator is bounded above by the ratio of the variation in α *and* $\hat{\gamma}$*, while the first term of* $\hat{\beta}$ *in (4.7) has the order of the ratio of the variation within* α *and the size of* $\hat{\gamma}$*. Note that* $\|\hat{\gamma}\|_2$ *is always no smaller than* $\|\hat{\gamma} - \bar{\hat{\gamma}}\mathbb{1}_J\|_2$*. A scenario favoring the Bayes rule over the Egger's method is when* $\hat{\gamma}_j$*'s are nonzero but close to each other.*

It is also important to consider the case where some genetic variants do not have pleiotropic effects, i.e. $\alpha_j = 0$ for some $j \in \{1, \ldots, J\}$. On one hand, a direct application of a single Gaussian prior with nonzero μ_α is not desirable because the penalty term in (4.2) cannot be zero in this case. On the other hand, sparsity of α may also bring some benefits on the invertibility of the Gram matrix. In the next section, we consider a proper model for sparse α.

4.2.2 A mixture Gaussian prior

In this section, we consider a hierarchical model with a mixture Gaussian prior.

$$\alpha_j|\mu_\alpha, \xi_j, \sigma_\eta^2 \sim_{ind} N(\mu_\alpha\xi_j, \tau_{0n}^2\sigma_\eta^2 + (\tau_{1n}^2 - \tau_{0n}^2)\xi_j\sigma_\eta^2) \tag{4.10}$$

$$\xi_j|p_1 \sim_{iid} \text{Bernoulli}(p_1) \tag{4.11}$$

$$\sigma_\eta^{-2} \sim \text{Gamma}(\nu_1, \nu_2), \tag{4.12}$$

where μ_α, p_1, τ_{0n}^2 and τ_{1n}^2 are prespecified constants. If $\xi_j = 1$, the prior distribution of α_j is $N(\mu_\alpha, \tau_{1n}^2\sigma_\eta^2)$; if $\xi_j = 0$, the prior distribution of α_j is $N(0, \tau_{0n}^2\sigma_\eta^2)$ with a small constant τ_{0n}^2 ($\tau_{0n}^2 = o(1)$).

The above model is closely related to the "Spike-and-Slab" (George and McCulloch, 1993, 1997; Ročková and George, 2014; Narisetty and He, 2014), which is a well-established Bayesian variable selection procedure. The "Spike-and-Slab" prior consists of a spike component and a slab component both centered at 0. Especially, the estimation consistency of ξ with Spike-and-Slab prior has been carefully studied in

Narisetty and He (2014) under the same hierarchical prior as (4.10) - (4.12) for high-dimensional regression, in which case an initial LSE estimator is not attainable. This is analogous to the situation under consideration with $\mu_\alpha = 0$.

Nevertheless, a nonzero μ_α is of practical importance when unbalanced pleiotropic effect is of interest and our goal is estimation rather than variable selection. Hence, we will first provide theoretical justifications of the posterior mean $\hat{\beta}^{(MG)}$ (under mixture Gaussian prior) for any given hyper-parameters μ_α and p_0. Then we will illustrate the realizations when hyper parameters are unknown.

The regularization property of the mixture Gaussian prior with nonzero center (4.10) - (4.12) can be analyzed similarly as for the Spike-and-Slab Lasso penalty in Ročková and George (2016). Consider the marginal prior of α given σ_η^{-2}, which is

$$\alpha_j | p_1, \tau_{0n}^2, \tau_{1n}^2, \sigma_\eta^2 \sim_{ind} p_1 N(\mu_\alpha, \tau_{1n}^2 \sigma_\eta^2) + (1-p_1) N(0, \tau_{0n}^2 \sigma_\eta^2).$$

The estimation of $\hat{\beta}^{(MG)}$ with respect to (4.10)-(4.12) and (4.3) can be formulated as

$$(\hat{\beta}^{(MG)}, \hat{\alpha}^{(MG)}) = \underset{b\in\mathbb{R}, a\in\mathbb{R}^J}{\arg\min} \left\{ \|Y - \hat{D}b - Za\|_2^2 + Pen(a) \right\}, \tag{4.13}$$

where

$$Pen(a) = -2\sigma_\eta^2 \sum_{j=1}^{J} \log\left(p_1 \phi(a_j; \mu_\alpha, \tau_{1n}^2 \sigma_\eta^2) + (1-p_1)\phi(a_j; 0, \tau_{0n}^2 \sigma_\eta^2) \right). \tag{4.14}$$

By studying the derivative of $Pen(a)$, the following Lemma characterizes the regularization property of prior (4.10)-(4.12). Let $\pi_1(a_j)$ be the posterior probability of $\xi_j = 1$ conditioning on $\alpha_j = a_j$, i.e.

$$\pi_1(a_j) = p(\xi_j = 1 | \alpha_j = a_j, \mu_\alpha, \tau_{0n}^2, \tau_{1n}^2, \sigma_\eta^2) = \frac{p_1 \phi(a_j; \mu_\alpha, \tau_{1n}^2 \sigma_\eta^2)}{p_1 \phi(a_j; \mu_\alpha, \tau_{1n}^2 \sigma_\eta^2) + (1-p_1)\phi(a_j; 0, \tau_{0n}^2 \sigma_\eta^2)}.$$

Lemma 4.2.4. *The derivative of $Pen(\alpha)$ in (4.14) satisfies*

$$\frac{\partial Pen(\alpha)}{\partial \alpha_j} = 2\left[\frac{1-\pi_1(\alpha_j)}{\tau_{0n}^2} + \frac{\pi_1(\alpha_j)}{\tau_{1n}^2} \right] (\alpha_j - \omega(\alpha_j)\mu_\alpha), \tag{4.15}$$

where

$$\omega(\alpha_j) = \frac{\pi_1(\alpha_j)/\tau_{1n}^2}{\pi_1(\alpha_j)/\tau_{1n}^2 + (1-\pi_1(\alpha_j))/\tau_{0n}^2}.$$

Equation (4.15) shows that $Pen(\alpha)$ has adaptive shrinkage effects. As expected, larger $\pi_1(\alpha_j)$ favors smaller regularization τ_{1n}^{-2} and a larger center parameter $\omega(\alpha_j)\mu_\alpha$. Especially, the center $\omega(\alpha_j)\mu_\alpha = o(1)$ if $\pi_1(\alpha_j)/(1-\pi_1(\alpha)) = o(\tau_{1n}^2/\tau_{0n}^2)$. In comparison, the penalty specified by prior (4.4) - (4.5) has a constant level of regularization (τ_n^{-2}) and a fixed center μ_α.

The format of penalty obtained in Lemma 4.2.4 suggests that the results in Theorem 4.2.1 can be extended to quantify the estimation error of $\hat{\beta}^{(MG)}$ in (4.13). However, in the next Theorem, we study the estimation error of $\hat{\beta}^{(MG)}$ by considering the effect of different amount of shrinkage imposed on vector α.

Let $Q^n = Z^T P_{\hat{D}}^{\perp} Z/n$. For prior parameters, we take $\tau_{0n}^2 = n^{-1}$ for simplicity and τ_{1n}^2 is a positive constant. We split the set of pleiotropic effects into two sets according to the posterior distribution of ξ_j, $\pi_1(\hat{\alpha}_j^{(MG)})$. Secifically, let $\hat{S}_\alpha$ be $\hat{S}_\alpha = \{j : \lim_{n\to\infty} \pi_1(\hat{\alpha}_j^{(MG)}) = 1\}$. The complement of $\hat{S}_\alpha$ can be written as $\hat{S}_\alpha^c = \{j : \pi_1(\hat{\alpha}_j^{(MG)}) \leq 1 - c_0,\ \text{for some constant } c_0 > 0\}$.

Condition 4.2.3. *We assume that* $\hat{S}_\alpha \neq [J]$ *and* $\Lambda_{\min}(Q_{\hat{S}_\alpha,\hat{S}_\alpha}^n) > 0$.

In Condition 4.2.3, it is assumed that not all estimated α_j has large probability to be selected to the nonzero component.

Theorem 4.2.5. *Under Conditions 4.2.1 and 4.2.3, for given* p_1 *and* μ_α*, the estimation error of* $\hat{\beta}^{(MG)}$ *in (4.13) satisfies*

$$\begin{aligned} |\hat{\beta}^{(MG)} - \beta| \leq & \frac{2c_n^* \Lambda_{\max}^{1/2}(\Sigma_Z^n)}{\Lambda_{\min}^{1/2}(\Sigma_Z^n)} \left[\frac{c_{1n} \|(\alpha - \omega(\hat{\alpha}^{(MG)})\mu_\alpha)_{\hat{S}_\alpha}\|_2}{\|\hat{\gamma}\|_2} + \frac{\|\alpha_{\hat{S}_\alpha^c} - c_{2n}\|_2}{\|\hat{\gamma}\|_2} \right] \\ & + O_P\left(\frac{\sqrt{J}\Lambda_{\max}(\Sigma_Z^n)\sigma_\epsilon}{\|\hat{D}\|_2} + \frac{\sigma_\epsilon}{\|\hat{D}\|_2} \right), \end{aligned} \tag{4.16}$$

where $c_n^* \leq \max\left\{ \frac{1+c_0^{-1}\Lambda_{\max}(\Sigma_Z^n)}{\Lambda_{\min}(Q_{\hat{S}_\alpha,\hat{S}_\alpha}^n)}, c_0^{-1} \right\}$, $c_{1n} = o(1)$ *and* $c_{2n} = O(n^{-1})$.

The proof is in the Appendix. In Theorem 4.2.5, we consider the shrinkage effects of α_j in $\hat{S}_\alpha$ and $\hat{S}_\alpha^c$ separately. As discussed before, adaptive mixing weights provide different amount of regularization on pleiotropic effects. It is possible to use large regularization $(H_{j,j},\ j \in \hat{S}_\alpha^c)$ on a subset of α to sufficiently stabilize the Gram matrix.

Specifically, in the proof we show that if $\Lambda_{\min}(Q^n_{\hat{S}_\alpha,\hat{S}_\alpha})$ is a positive constant and $\tau^2_{0n} = n^{-1}$, then c^*_n can be treated as a constant. The eigenvalue condition on Q^n is restricted to a submatrix and it can be viewed as a restricted isometry condition. As a result, this allows for small regularizations (of order $o(1)$) on $\alpha_j,\ j \in \hat{S}_\alpha$, which can reduce the bias caused by the variation of $\alpha_{\hat{S}_\alpha}$. For $j \in \hat{S}^c_\alpha$, the benefit of adaptivity is shown by shrinking the weighted center to zero at the rate of n^{-1}.

4.3 Dealing with unknown hyper parameters

In this section, we mainly discuss the computation algorithm of the posterior mean considered in section 4.2.2.

For $\hat{\beta}^{(MG)}$ given by model (4.3) and (4.10) - (4.12), if hyper parameters μ_α and p_1 are known, Gibbs sampler can be applied to draw samples from the posterior distribution of $\hat{\beta}^{(MG)}$. Implementation details can be found in the Appendix. In many applications, however, the hyper parameters are always unknown.

When μ_α and p_1 are unknown, we propose an empirical Bayes method to estimate them from the data, which can be viewed as an approximation of a fully hierarchical Bayes analysis (Carlin and Gelfand, 1990, 1991). The estimation can be realized by a variation of expectation-maximization (EM) algorithm, the Monte Carlo EM (MCEM) algorithm (Meng and Schilling, 1996; Levine and Casella, 2001). After m rounds of Gibb's sampling conditioning on hyper parameters, we compute hyper parameters as the maximizer of the marginal likelihood approximated by Gibbs samples. The MCEM algorithm iteratively estimates the marginal parameters and samples the middle-layered parameters until it converges. (See Casella (2001) for a general description.) Implementation details can be found in the Appendix.

Another realistic concern is the inverse of the sum of Σ^n_Z and a diagonal matrix, which can be computationally expensive. This issue can be efficiently solved by block updates (Ishwaran and Rao, 2005).

4.4 Implementation with summary data

Many public datasets, especially GWAS datasets, are available only up to summary statistics for the association studies between individual genetic variants and traits. Moreover, in many cases the data on the interested exposure and that on the interested outcome are available in independent samples. Developing methodology for this type of data can broaden the applicability of MR and is of great relevance.

We generalize the methods under consideration to the case where only $(\hat{\gamma}^{(2)}, \hat{\Gamma}, \hat{\sigma}^2_\Gamma)$ are available, where $\hat{\Gamma}_j$ $(j = 1, \dots, J)$ is the estimated regression coefficient between the interested outcome Y and the j-th genetic variant Z_j, $\hat{\sigma}^2_{\Gamma,j}$ $(j = 1, \dots, J)$ is the estimated variance of $\hat{\Gamma}_j$ and $\hat{\gamma}^{(2)}_j$ $(j = 1, \dots, J)$ is the estimated regression coefficient between the interested exposure D and Z_j from an independent sample. To apply the methods under consideration, we assume that $Z^T Z$ is a diagonal matrix, i.e. there is no linkage disequilibrium. Then we can replace (4.3) by

$$\hat{\Gamma}_j | \hat{\gamma}, \beta, \alpha, \sigma^2_\Gamma \sim_{ind} N(\hat{\gamma}_j \beta + \alpha_j, \sigma^2_{\Gamma,j}). \tag{4.17}$$

A conjugate prior on α, say mixture Gaussian prior, can be

$$\alpha_j | \mu_\alpha, \xi_j, \sigma^2_\Gamma \sim_{ind} N(\mu_\alpha \xi_j, \tau^2_{0n} \sigma^2_{\Gamma,j} + (\tau^2_{1n} - \tau^2_{0n}) \xi_j \sigma^2_{\Gamma,j}) \tag{4.18}$$

$$\xi_j \sim_{i.i.d} \text{Bernoulli}(p_1). \tag{4.19}$$

Implementation details of the MCEM algorithm can be found in the Appendix.

4.5 Simulations and real studies

4.5.1 Synthetic data experiments

In this section, we evaluate the performance of Bayesian estimators considered in Section 4.2. We consider (1) the model under single Gaussian prior (4.3) - (4.5) with a data-driven center (SG*); (2) the model under mixture Gaussian prior (4.3), (4.10) - (4.12) with $\mu_\alpha = 0$ and $p_1 = 0.5$ (MG); (3) the model under mixture Gaussian prior (4.3), (4.10) - (4.12) with data-driven μ_α and p_0 (MG*). These methods are numerically studied in comparison to the TSLS and the Lasso estimators in various simulation settings. The TSLS estimator is computed as a benchmark from classical instrumental variable literatures. The Lasso estimator, which is essentially the sisVIVE estimator in Kang et al. (2016), is proposed to deal with sparse α and hence is added in comparison. The threshold parameter is chosen by 10-fold cross validation as suggested in the paper.

In all the experiments presented in this section, each sample consists of $n = 1000$ observations and $J = 30$ candidate genetic variants. The genetic variants Z_i, $i = 1, \ldots, n$, are drawn from a multivariate normal distribution with mean zero and identity covariance matrix. The phenotypes (D_i, Y_i), $i = 1, \ldots, n$, are generated according to model (1.5), where each (v_i, ϵ_i) is generated from a bivariate normal distribution with mean zero and covariance matrix $\begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}$.

With and without the InSIDE assumption, we allow the following parameters to vary: the strength of causal effect, the distribution of pleiotropic effects and the proportion of invalid instruments. Specifically, we consider two levels of signal strength $\beta \in \{0, 0.2\}$, three levels of the mean of pleiotropic effects $\mu_\alpha \in \{-0.2, 0, 0.2\}$, and five levels of sparsity ($p_1^* \in \{0, 0.25, 0.5, 0.75, 1\}$) of α. In each of these settings, we generate γ_j from U$[0.1, 0.3]$, ξ_j^* from Bernoulli$(1, p_1^*)$, and u_j from U$[\mu_\alpha - 0.2, \mu_\alpha + 0.2]$ in an *i.i.d.* fashion for $j = 1, \ldots, J$. The pleiotropic effects $\alpha_j = \xi_j^* u_j$ if the InSIDE assumption is satisfied and $\alpha_j = (0.2\gamma_j + u_j)\xi_j^*$ if the InSIDE assumption is not satisfied, for $j = 1, \ldots, J$. In each setting, the experiment is independently replicated for 100 times and the mean square error (MSE) is reported.

As explained before, we set ν_1 and ν_2 to be small numbers, $\nu_1 = \nu_2 = 0.001$, and $\tau_n^2 = \tau_{0n}^2 = 1/n$ and $\tau_{1n}^2 = 1$.

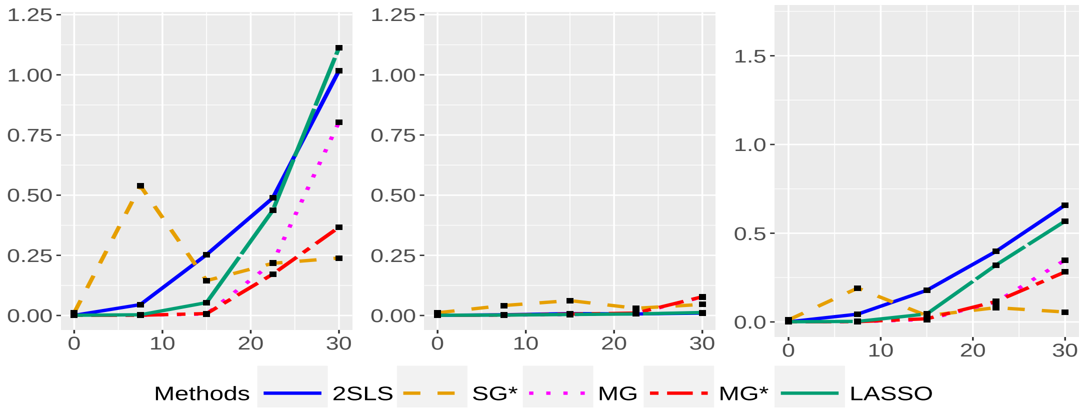

Figure 4.1: $\beta = 0$ and InSIDE assumption is satisfied. The x-axis of each plot is the number of nonzero α_j. Each point represents the MSE of 200 experiments for $\mu_\alpha = -0.2, 0$ and 0.2 from left to right, respectively.

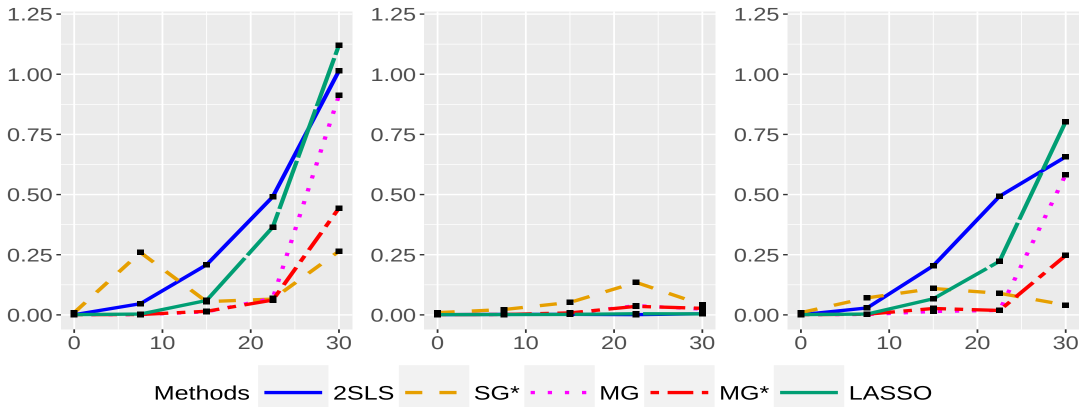

Figure 4.2: $\beta = 0.2$ and InSIDE assumption is satisfied. The x-axis of each plot is the number of nonzero α_j. Each point represents the MSE of 200 experiments for $\mu_\alpha = -0.2, 0$ and 0.2 from left to right, respectively.

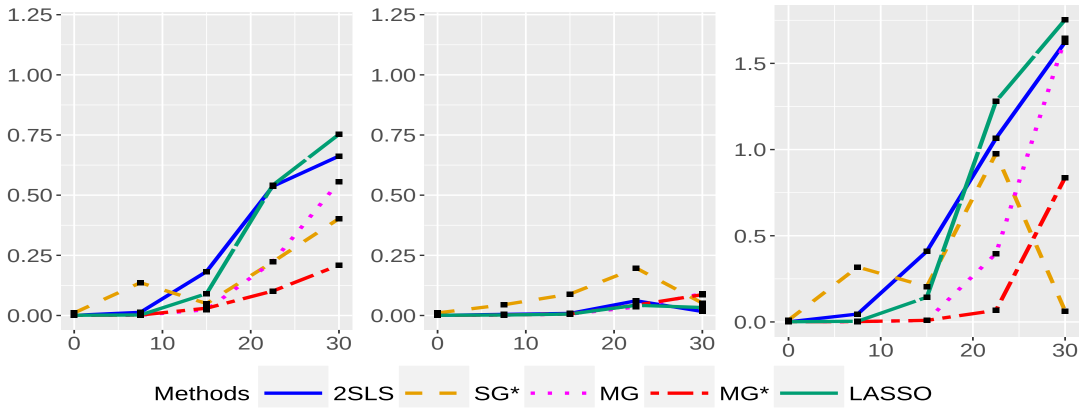

Figure 4.3: $\beta = 0.2$ and InSIDE assumption is not satisfied. The x-axis of each plot is the number of nonzero α_j. Each point represents the MSE of 200 experiments for $\mu_\alpha = -0.2, 0$ and 0.2 from left to right, respectively.

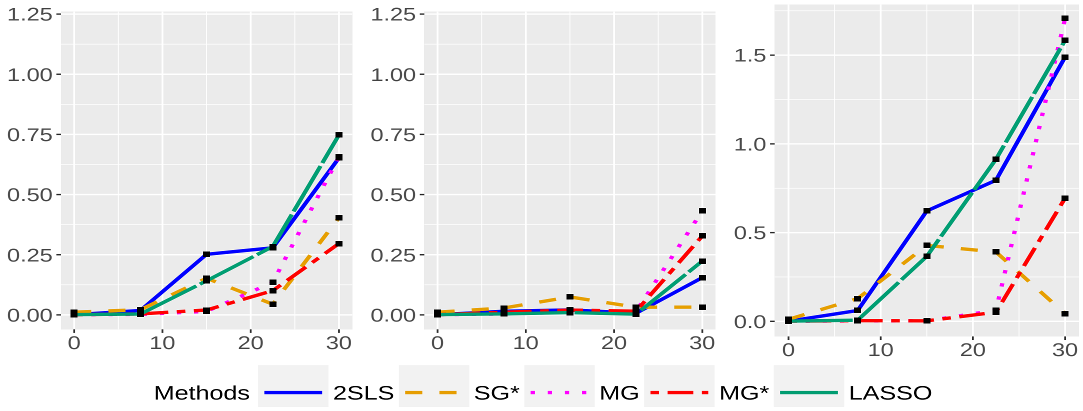

Figure 4.4: $\beta = 0.2$ and InSIDE assumption is not satisfied. The x-axis of each plot is the number of nonzero α_j. Each point represents the MSE of 200 experiments for $\mu_\alpha = -0.2, 0$ and 0.2 from left to right, respectively.

From the above plots, one can see that Bayesian methods with mixture Gaussian prior (MG and MG*) have reliable performance at many levels of sparsity. MG* has superior performance in the unbalanced pleiotropic effects scenario, while MG is slightly more accurate in the balanced pleiotropic effects case. On the other hand, the method with single Gaussian prior (SG*) has smaller MSE comparing with other methods when all the pleiotropic effects are nonzero, especially when the pleiotropic effects are unbalanced. This is due to the appropriateness of the prior as well as the data-dependent selection of μ_α.

The benefits of the Lasso approach under sparsity conditions have been studied in Kang et al. (2016). In comparison to 2SLS and the Lasso approach, MG and MG* have smaller MSE remains stable in a broader range of sparsity levels in comparison to existing methods. Moreover, the performances of 2SLS and the Lasso are sensitive to the unbalanced pleiotropic effects. As explained before, the Bayesian methods with data-dependent priors (SG* and MG*) have demonstrated more reliable performance.

When InSIDE assumption is not satisfied (Figure 4.3 and 4.4), which is a harder scenario, we observe similar pattern as when InSIDE assumption is satisfied. Especially, MG* has most reliable performance as long as not all pleiotropic effects are nonzero.

4.5.2 Case study (i): HDL and type 2 diabetes

The high density lipoprotein (HDL) cholesterol has the reputation as a "good" cholesterol, since it is negatively associated in observational studies with the risk of many diseases, for example, myocardial infarction and type 2 diabetes. However, the supporting studies have been unable to control various potential confounders, while the negative association with HDL has lacked convincing biological mechanisms. Hence, the association does not necessarily imply a causal effect.

Haase et al. (2012) use the traditional MR method to estimate the causality between HDL and the risk of type 2 diabetes. Their results suggest that there is no causal effect of HDL on type 2 diabetes. We access a different set of summary data from MRbase (Hemani et al., 2016) and arrive at a similar conclusion for a related trait. The MRbase is a database and an analytical platform for MR studies, which provides summary data of many published GWAS and some basic analytic tools.

The exposure data is measured plasma HDL cholesterol (unit: mg/dL) from the Global Lipids Genetics Consortium (Willer et al., 2013) with a sample size 187167. The outcome data is measured fasting glucose (unit: mmol/L) from the Meta-Analyses of Glucose and Insulin-related traits Consortium (Dupuis et al., 2010) with a sample size 46186. Hyperglycemia in the fasting state is one of the criteria that defines type

2 diabetes (Kahn, 2003). Thus, fasting glucose is an important indicator of Type 2 diabetes.

For the analysis, 83 genetic variants are selected and harmonized automatically by the MRbase, which excludes linkage disequilibrium and selects variants which are robustly associated with the target traits with the genome-wide significance threshold 5×10^{-8}.

Four method are applied on this dataset. The estimate given by TSLS is -0.0282 mmol/L per mg/dL; the estimate given by the Egger's regression (Bowden et al., 2015) is -0.0345 mmol/L per mg/dL; the estimate given by the inverse-variance weighted median estimator (Bowden et al., 2016) is -0.0290 mmol/L per mg/dL; the estimate given by the MG* estimator is -0.0312 mmol/L per mg/dL, where τ_{0n}^2 and τ_{1n}^2 are specified as 0.0001 and 0.1, respectively. One can see that these methods generate similar estimates for this dataset.

4.5.3 Case study (ii): LDL and type 2 diabetes

As introduced at the beginning of the paper, we study the causal effect of LDL cholesterol on the type 2 diabetes in this section.

The exposure data is measured plasma LDL cholesterol (unit: mg/dL) from the Global Lipids Genetics Consortium (Willer et al., 2013) with a sample size 173082. The outcome data is the fasting glucose (unit: mmol/dL) which is from the same source of data as in case study (i). For the analysis, 72 genetic variants are selected and harmonized.

For this dataset, the estimate given by TSLS is -0.0157 mmol/L per mg/dL; the estimate given by the Egger's regression (Bowden et al., 2015) is -0.0248 mmol/L per mg/dL; the estimate given by the inverse-variance weighted median estimator (Bowden et al., 2016) is -0.0121 mmol/L per mg/dL; the estimate given by the MG* estimator is -0.0038 mmol/L per mg/dL, for which τ_{0n}^2 and τ_{1n}^2 are specified as 0.0001 and 0.1, respectively.

4.6 Discussions

In this chapter, we have studied Bayesian hierarchical models for estimating the causal effect in the presence of invalid instruments for MR studies. Due to the confoundedness of pleiotropic effects, the causal effect cannot be identified with the traditional TSLS estimator. To deal with this problem, we consider single Gaussian priors and mixture Gaussian priors to incorporate the pleiotropic effects. The estimation errors of the considered methods are studied under proper conditions. In order to deal with unknown hyper parameters, computational algorithm are proposed to estimate them from data, which demonstrate superior performance in unbalanced pleiotropic effects scenario through our simulation. Moreover, the Bayesian estimators under mixture Gaussian prior are more stable in a broader range of sparsity levels of pleiotropic effects in comparison to some other comparable methods, say the Lasso.

There are still interesting and open problems in the scope of current topic. Firstly, many epidemiological studies are interested in the causal effect of exposures on the risk of certain diseases. An important generalization of the proposed method is to estimate the causal effect for the binary outcome data, or equivalently the probability of occurrence of an event. Secondly, this paper together with many previous works have been focusing on the estimation procedure, while generating valid interval estimates with mild conditions and cheap computation remains to be a challenging and worthwhile topic.

Bibliography

Per Kragh Andersen and Richard David Gill. Cox's regression model for counting processes: a large sample study. *The annals of statistics*, 10:1100–1120, 1982.

C. Baigent, A. Keech, P.M. Kearney, et al. Efficacy and safety of cholesterol-lowering treatment: prospective meta-analysis of data from 90056 participants in 14 randomised trials of statins. *The Lancet*, 366(9493):1267–1278, 2005.

A. Belloni, D. Chen, V. Chernozhukov, and C. Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80 (6):2369–2429, 2012.

A. Belloni, V. Chernozhukov, and K. Kato. Uniform post-selection inference for least absolute deviation regression and other z-estimation problems. *Biometrika*, 102(1): 77–94, 2014.

Alexandre Belloni, Victor Chernozhukov, and Kengo Kato. Valid post-selection inference in high-dimensional approximately sparse quantile regression models. 2013. preprint, `http://arxiv.org/abs/1312.7186`.

C. Berzuini, H. Guo, S. Burgess, and L. Bernardinelli. Bayesian mendelian randomization. 2017. preprint,`https://arxiv.org/pdf/1608.02990`.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

J. Bowden, G. Davey Smith, and S. Burgess. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology*, 44(2):512–525, 2015.

J. Bowden, G. Davey Smith, P. C. Haycock, et al. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology*, 40(4):304–314, 2016.

Jelena Bradic, Jianqing Fan, and Jiancheng Jiang. Regularization for cox's proportional hazards model with np-dimensionality. *Annals of statistics*, 39(6):3092, 2011.

P. Bühlmann and S. van De Geer. *Statistics for high-dimensional data: methods, theory and applications.* Springer Science & Business Media, 2011.

F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.

T. T. Cai and Z. Guo. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of Statistics*, 45(2):615–646, 2017.

E. Candes and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351, 2007.

B. P. Carlin and A. E. Gelfand. Approaches for empirical Bayes confidence intervals. *Journal of the American Statistical Association*, 85(409):105–114, 1990.

B. P. Carlin and A. E. Gelfand. A sample reuse method for accurate parametric empirical Bayes confidence intervals. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 189–200, 1991.

G. Casella. Empirical Bayes Gibbs sampling. *Biostatistics*, 2(4):485–500, 2001.

A. Chatterjee and S. N. Lahiri. Asymptotic properties of the residual bootstrap for lasso estimators. *Proceedings of the American Mathematical Society*, 138(12):4497–4509, 2010.

A. Chatterjee and S. N. Lahiri. Bootstrapping lasso estimators. *Journal of the American Statistical Association*, 106(494):608–625, 2011.

A. Chatterjee and S. N. Lahiri. Rates of convergence of the adaptive lasso estimators to the oracle distribution and higher order refinements by the bootstrap. *Annals of Statistics*, 41(3):1232–1259, 2013.

V. Chernozhukov, D. Chetverikov, and K. Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819, 2013.

V. Chernozhukov, D. Chetverikov, M. Demirer, et al. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 2017.

G. Davey Smith and S. Ebrahim. Mendelian randomization: genetic variants as instruments for strengthening causal inference in observational studies. *Bio-Social Surveys: Current Insight and Future Promise*, pages 1–428, 2008.

G. Davey Smith and G. Hemani. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human Molecular Genetics*, 23(R1):R89–R98, 2014.

H. Deng and C.-H. Zhang. Beyond gaussian approximation: Bootstrap for maxima of sums of independent random vectors. 2017. preprint, `http://arxiv.org/abs/1705.09528`.

R. Dezeure, P. Bühlmann, and C.-H. Zhang. High-dimensional simultaneous inference with the bootstrap. 2016. preprint, `http://arxiv.org/abs/1606.03940`.

S. G. Donald and W. K. Newey. Choosing the number of instruments. *Econometrica*, 69(5):1161–1191, 2001.

J. Dupuis, C. Langenberg, I. Prokopenko, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nature Genetics*, 42 (2):105–116, 2010.

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

Jianqing Fan and Runze Li. Variable selection for cox's proportional hazards model and frailty model. *Annals of Statistics*, pages 74–99, 2002.

E. X. Fang, Y. Ning, and H. Liu. Testing and confidence intervals for high dimensional proportional hazards models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016.

Ethan X. Fang, Yang Ning, and Han Liu. Testing and confidence intervals for high dimensional proportional hazards model. *Journal of the Royal Statistical Society, Series B*, pages 1415 – 1437, 2017.

A. Feller and A. Gelman. Hierarchical models for causal effects. *Emerging Trends in the Social and Behavioral Sciences: An Interdisciplinary, Searchable, and Linkable Resource*, 2015.

E. I. George and R. E. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

E. I. George and R. E. McCulloch. Approaches for Bayesian variable selection. *Statistica Sinica*, pages 339–373, 1997.

C. L. Haase, A. Tybjaerg-Hansen, B. G. Nordestgaard, et al. Plasma HDL cholesterol and risk of myocardial infarction: a Mendelian randomisation study. *The Lancet*, 380 (9841):572–580, 2012.

G. Hemani, J. Zheng, K. H. Wade, et al. MR-base: a platform for systematic causal inference across the phenome using billions of genetic associations. *bioRxiv*, 2016. doi: 078972.

J. Huang and C.-H. Zhang. Estimation and selection via absolute penalized convex minimization and its multistage adaptive applications. *Journal of Machine Learning Research*, 13(Jun):1839–1864, 2012.

Jian Huang, Tingni Sun, Zhiliang Ying, Yi Yu, and Cun-Hui Zhang. Oracle inequalities for the lasso in the cox model. *Annals of statistics*, 41(3):1142, 2013.

H. Ishwaran and J. S. Rao. Spike and slab variable selection: frequentist and bayesian strategies. *The Annals of Statistics*, 33(2):730–773, 2005.

J. Jankova and S. van de Geer. Confidence intervals for high-dimensional inverse covariance estimation. *Electronic Journal of Statistics*, 9(1):1205–1229, 2015.

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15:2869–2909, 2014a.

A. Javanmard and A. Montanari. Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory*, 60(10):6522–6554, 2014b.

A. Javanmard and A. Montanari. De-biasing the Lasso: Optimal Sample Size for Gaussian Designs. pages 1–32, 2015. preprint, http://arxiv.org/abs/1508.02757.

R. Kahn. Follow-up report on the diagnosis of diabetes mellitus: the expert committee on the diagnosis and classifications of diabetes mellitus. *Diabetes Care*, 26(11):3160, 2003.

H. Kang, A. Zhang, T. T. Cai, et al. Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *Journal of the American Statistical Association*, 111(513):132–144, 2016.

K. Knight and W. Fu. Asymptotics for lasso-type estimators. *The Annals of Statistics*, 28(5):1356–1378, 2000.

Shengchun Kong and Bin Nan. Non-asymptotic oracle inequalities for the high-dimensional cox regression via lasso. *Statistica Sinica*, 24(1):25, 2014.

D. A. Lawlor, R. M. Harbord, J. A. C. Sterne, et al. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in Medecine*, 27(8):1133–1163, 2008.

R. A. Levine and G. Casella. Implementations of the Monte Carlo EM algorithm. *Journal of Computational and Graphical Statistics*, 10(3):422–439, 2001.

Wei Lin and Jinchi Lv. High-dimensional sparse additive hazards regression. *Journal of the American Statistical Association*, 108(501):247–264, 2013.

E. Mammen. Bootstrap and wild bootstrap for high dimensional linear models. *The Aannals of Statistics*, pages 255–285, 1993.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, pages 1436–1462, 2006.

N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, pages 246–270, 2009.

X.-L. Meng and S. Schilling. Fitting full-information item factor models and an empirical investigation of bridge sampling. *Journal of the American Statistical Association*, 91(435):1254–1267, 1996.

R. Mitra and C.-H. Zhang. The benefit of group sparsity in group inference with de-biased scaled group lasso. *Electronic Journal of Statistics*, 10(2):1829–1873, 2016.

N. N. Narisetty and X. He. Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, 42(2):789–817, 2014.

S. Reid, R. Tibshirani, and J. Friedman. A study of error variance estimation in lasso regression. *Statistica Sinica*, 26:35–67, 2016.

Z. Ren, T. Sun, C.-H. Zhang, et al. Asymptotic normality and optimalities in estimation of large gaussian graphical models. *The Annals of Statistics*, 43(3):991–1026, 2015.

V. Ročková and E. I. George. Emvs: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association*, 109(506):828–846, 2014.

V. Ročková and E. I. George. The Spike-and-Slab Lasso. *Journal of the American Statistical Association*, 2016.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688 – 701, 1974.

M. S. Sabatine, R. P. Giugliano, A. C. Keech, et al. Evolocumab and clinical outcomes in patients with cardiovascular disease. *New England Journal of Medicine*, 2017.

N. Sattar, D. Preiss, H. M. Murray, et al. Statins and risk of incident diabetes: a collaborative meta-analysis of randomised statin trials. *The Lancet*, 375(9716):735–742, 2010.

A. F. Schmidt and F. Dudbridge. Mendelian randomization with egger pleiotropy correction and weakly informative bayesian priors. *International journal of epidemiology*, 2017.

M. Spindler. Lasso for instrumental variable selection: A replication study. *Journal of Applied Econometrics*, 31(2):450–454, 2016.

J. Splawa-Neyman. On the application of probability theory to agricultural experiments. *Statistical Science*, 5(4):465–472, 1990.

M. Stephens and D. J. Balding. Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, 10(10):681–690, 2009.

T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

J. R. Thompson, C. Minelli, J. Bowden, et al. Mendelian randomization incorporating uncertainty about pleiotropy. *Statistics in medicine*, 36(29):4627–4645, 2017.

R. Tibshirani. Regression selection and shrinkage via the lasso. *Journal of the Royal Statistical Society B*, 58(1):267–288, 1996.

Robert Tibshirani. The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395, 1997.

S. van de Geer, P. Bühlmann, Y. Ritov, et al. On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42(3):1166–1202, 2014.

S. A. van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645, 2008.

H. van Kippersluis and C. A Rietveld. Pleiotropy-robust Mendelian randomization. *International Journal of Epidemiology*, 2017.

T. J. VanderWeele, E. J. Tchetgen Tchetgen, M. Cornelis, et al. Methodological challenges in Mendelian randomization. *Epidemiology*, 25(3):427–435, 2014.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. pages 210–268, 2010. preprint, `http://arxiv.org/abs/1011.3027`.

M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.

C. J. Willer, E. M. Schmidt, S. Sengupta, G. M Peloso, et al. Discovery and refinement of loci associated with lipid levels. *Nature Genetics*, 45(11):1274–1283, 2013.

F. Ye and C.-H. Zhang. Rate minimaxity of the lasso and dantzig selector for the lq loss in lr balls. *Journal of Machine Learning Research*, 11:3519–3540, 2010.

C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010.

C.-H. Zhang. Statistical inference for high-dimensional data. *In Very High Dimensional Semiparametric Models, Mathematisches Forschungsinstitut Oberwolfach*, (48), 2011.

C.-H. Zhang and J. Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):1567–1594, 2008.

C.-H. Zhang and S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 76(1):217–242, 2014.

X. Zhang and G. Cheng. Simultaneous inference for high-dimensional linear models. *Journal of the American Statistical Association*, 0(0):1–12, 2017.

P. Zhao and B. Yu. On model selection consistency of Lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

Appendix A

Appendices

A.1 Proofs in Chapter 2

A.1.1 Supportive lemmas

Lemma A.1.1. *Under Conditions 2.3.1 - 2.3.5, if* $v(t) \in \mathbb{R}^p$ *satisfies* $\max_{1\le i\le n}\sup_{t\in[0,\tau]}|X_i(t)v(t)-\mu(t,\beta_0)v(t)| = O_P(K)$, *then (i)*

$$\max_{1\le j,k\le p}\sup_{t\in[0,\tau]}\left|\overline{Var(X)}(t,\beta_0)-F(t,\beta_0)\right|_{j,k} = O_P(K_1^2K_2\sqrt{\frac{\log p}{n}}).$$

(ii)

$$\max_{1\le j\le p}\sup_{t\in[0,\tau]}\left|[\overline{Var(X)}(t,\beta_0)-F(t,\beta_0)]v(t)\right|_j = O_P(K_1K_2K\sqrt{\frac{\log p}{n}}).$$

Proof of Lemma A.1.1. First define

$$G_n(t,\beta_0)=\frac{1}{n}\sum_{i=1}^{n}\{X_i(t)-\mu(t,\beta_0)\}^{\otimes 2}\gamma_{ni}(t,\beta_0),\ t\in[0,\tau]. \tag{A.1}$$

By (2.14) and (A.1), we have

$$\begin{aligned}&\left[\overline{\mathrm{Var}(X)}(t,\beta_0)-F(t,\beta_0)\right]\\ &=\underbrace{\left[\overline{\mathrm{Var}(X)}(t,\beta_0)-G_n(t,\beta_0)\right]}_{T_1}+\underbrace{\left[G_n(t,\beta_0)-\left(\frac{s^{(2)}(t,\beta_0)}{s^{(0)}(t,\beta_0)}-\mu^{\otimes 2}(t,\beta_0)\right)\right]}_{T_2}\end{aligned}$$

For T_1, note that

$$\overline{\mathrm{Var}(X)}(t,\beta_0)=G_n(t,\beta_0)-(\bar{X}_n(t,\beta_0)-\mu(t,\beta_0))^{\otimes 2}.$$

It is easy to see that, $\forall 1\le j,k\le p$,

$$|T_1|_{j,k}\le\sup_{t\in[0,\tau]}\left\{|\bar{X}_n(t,\beta_0)-\mu(t,\beta_0)|_j|\bar{X}_n(t,\beta_0)-\mu(t,\beta_0)|_k\right\}.$$

By SLLN, for $\forall t \in [0, \tau]$,

$$S^{(0)}(t, \beta_0) = s^{(0)}(t, \beta_0) + o_P(1). \tag{A.2}$$

By Condition 2.3.2 and 2.3.3,

$$\begin{aligned}
&\max_{1\le i\le n} \max_{1\le j\le p} \sup_{t\in[0,\tau]} \left| \frac{(X_{i,j}(t) - \mu_j(t, \beta_0))Y_i(t)e^{X_i(t)\beta_0}}{S^{(0)}(t, \beta_0)} \right| \\
&= \max_{1\le i\le n} \max_{1\le j\le p} \sup_{t\in[0,\tau]} \left| \frac{(X_{i,j}(t) - \mu_j(t, \beta_0))Y_i(t)e^{X_i(t)\beta_0}}{s^{(0)}(t, \beta_0)} \right| + o_P(1) = O_P(K_1K_2).
\end{aligned}$$

We have proved that, with large probability $\bar{X}_n(t, \beta_0) - \mu(t, \beta_0)$ is an average of bounded independent random variables. By Hoeffding's inequality,

$$\max_{1\le j\le p} \sup_{t\in[0,\tau]} |\bar{X}_n(t, \beta_0) - \mu(t, \beta_0)|_j = O_P\left(K_1K_2\sqrt{\frac{\log p}{n}}\right)$$

Thus,

$$\max_{1\le j,k\le p} \sup_{t\in[0,\tau]} |T_1|_{j,k} = O_P\left(K_1K_2K\sqrt{\frac{\log p}{n}}\right).$$

For T_2, by (A.2), Conditions 2.3.2 and 2.3.3, we have

$$\begin{aligned}
&\max_{1\le i\le n} \max_{1\le j,k\le p} \sup_{t\in[0,\tau]} \left| \frac{[X_i(t) - \mu(t, \beta_0)]_j[X_i(t) - \mu(t, \beta_0)]_kY_i(t)e^{X_i(t)\beta_0}}{S^{(0)}(t, \beta_0)} \right| \\
&= \max_{1\le i\le n} \max_{1\le j,k\le p} \sup_{t\in[0,\tau]} \left| \frac{[X_i(t) - \mu(t, \beta_0)]_j[X_i(t) - \mu(t, \beta_0)]_kY_i(t)e^{X_i(t)\beta_0}}{s^{(0)}(t, \beta_0)} \right| + o_P(1) \\
&= O_P(K_1K_2K).
\end{aligned}$$

And

$$\begin{aligned}
&\mathbb{E}\left[[X_{i,j}(t) - \mu(t, \beta_0)_j][X_i(t) - \mu(t, \beta_0)]_kY_i(t)e^{X_i(t)\beta_0}\right] \\
&= s^{(2)}_{j,k}(t, \beta_0) - \mu_j(t, \beta_0)\mu_k(t, \beta_0)s^{(0)}(t, \beta_0).
\end{aligned}$$

Thus, by Hoeffding's inequality, $\forall 1 \le j \le p$,

$$\max_{1\le j,k\le p} \sup_{t\in[0,\tau]} |T_2|_{j,k} = O_P\left(K_1K_2K\sqrt{\frac{\log p}{n}}\right).$$

Second statement in Lemma A.1.1 can be proved similarly. □

To facilitate the proof of the subsequent lemmas, we define a function $\langle \cdot, \cdot \rangle(\theta)$ for some $\theta(t) = (\theta_1(t), \ldots, \theta_n(t))^T$, $t \in [0, \tau]$ and show that it is a semi-inner product with some operational properties.

Let $f(t) = (f_1(t), \ldots, f_n(t))$ and $g(t) = (g_1(t), \ldots, g_n(t))$ for $t \in [0, \tau]$. Let $\Delta f_{i,l}(t) = f_i(t) - f_l(t)$, $1 \le i, l \le n$ and $\Delta g_{i,l}(t)$ be similarly defined. Define

$$\langle f, g \rangle(t, \theta) = \frac{\sum\limits_{1 \le i,l \le n} Y_i(t) Y_l(t) e^{\theta_i(t) + \theta_l(t)} \Delta f_{i,l}(t) \Delta g_{i,l}(t)}{\sum\limits_{1 \le i,l \le n} 2 Y_i(t) Y_l(t) e^{\theta_i(t) + \theta_l(t)}} \tag{A.3}$$

$$\langle f, g \rangle(\theta) = \frac{1}{n} \int_0^\tau \langle f, g \rangle(t, \theta) d\bar{N}(t). \tag{A.4}$$

Let $\|f\|^2(t, \theta) = \langle f, f \rangle(t, \theta)$ and $\|f\|^2(\theta) = \langle f, f \rangle(\theta)$.

Lemma A.1.2. *The following inequalities hold for any given* $\theta(t) \in \mathbb{R}^n$, $t \in [0, \tau]$, $\langle f, g \rangle(t, \theta)$ *and* $\langle f, g \rangle(\theta)$ *defined in (A.3) and (A.4).*

$$|\langle f, g \rangle(t, \theta)| \le \|f\|(t, \theta) \|g\|(t, \theta) \text{ and } |\langle f, g \rangle(\theta)| \le \|f\|(\theta) \|g\|(\theta).$$

Proof of Lemma A.1.2. It is easy to prove $\langle f, g \rangle(\theta)$ satisfies:

$$\langle f, g \rangle(\theta) = \langle g, f \rangle(\theta)$$

$$\langle af, g \rangle(\theta) = a \langle f, g \rangle(\theta)$$

$$\langle f + z, g \rangle(\theta) = \langle f, g \rangle(\theta) + \langle z, g \rangle(\theta)$$

$$\langle f, f \rangle(\theta) \ge 0.$$

Thus, $\langle f, g \rangle(\theta)$ is a semi-product and Cauchy-Schwarz inequality follows. The proof for $\langle f, g \rangle(t, \theta)$ is exactly the same. □

Lemma A.1.3. *Assume Conditions 2.3.2-2.3.3. For some* $\theta'(t) \in \mathbb{R}^n$, $t \in [0, \tau]$, *let*

$$c_{i,l}(t) = \theta'_i(t) - \theta_{0,i}(t) + \theta'_l(t) - \theta_{0,l}(t) - 2[\bar{\theta}'_n(t, \beta_0) - \bar{\theta}_{0,n}(t, \beta_0)].$$

If $\max\limits_{1 \le i,l \le n} \sup\limits_{t \in [0,\tau]} |c_{i,l}(t)| = o_P(1)$ *and* $\max\limits_{1 \le i,l \le n} \sup\limits_{t \in [0,\tau]} |\Delta g_{i,l}(t)| = O_P(K)$, *then*

$$|\langle f, g \rangle(t, \theta') - \langle f, g \rangle(t, \theta_0)| \le O_P(K) \|f\|(t, \theta_0) \overline{Var(\theta_0 - \theta')}(t, \beta_0)^{1/2}. \tag{A.5}$$

$$|\langle f, g \rangle(\theta') - \langle f, g \rangle(\theta_0)| \le O_P(K) \|f\|(\theta_0) \overline{Var(\theta_0 - \theta')}(\beta_0)^{1/2}. \tag{A.6}$$

Proof of Lemma A.1.3. Without loss of generality, we only prove the more general result (A.6). To simplify the proof of this lemma, we first declare some notations.

$$\omega_i(t) = Y_i(t)e^{\theta_{0,i}(t)-\bar{\theta}_{0,n}(t,\beta_0)} \qquad \omega'_i(t) = Y_i(t)e^{\theta'_i(t)-\bar{\theta}'_n(t,\beta_0)}$$

$$\omega_{..}(t) = n^{-2}\sum_{1\le i,l\le n} 2\omega_i(t)\omega_l(t) \quad \omega'_{..}(t) = n^{-2}\sum_{1\le i,l\le n} 2\omega'_i(t)\omega'_l(t).$$

By definition (A.4),

$$\begin{aligned}\langle f,g\rangle(\theta') &= \frac{1}{n}\int_0^\tau \frac{n^{-2}\sum_{1\le i,l\le n}\Delta f_{i,l}(t)\Delta g_{i,l}(t)\omega'_i(t)\omega'_l(t)}{\omega'_{..}(t)}d\bar{N}(t),\\ &= \frac{1}{n}\int_0^\tau \frac{n^{-2}\sum_{1\le i,l\le n}\Delta f_{i,l}(t)\Delta g'_{i,l}(t)\omega_i(t)\omega_l(t)}{\omega_{..}(t)}d\bar{N}(t) = \langle f,g'\rangle(\theta_0),\end{aligned}$$

where

$$\Delta g'_{i,l}(t) = \Delta g_{i,l}(t)e^{c_{i,l}(t)}\frac{\omega_{..}(t)}{\omega'_{..}(t)}.$$

By Lemma A.1.2, we have

$$|\langle f,g'\rangle(\theta_0) - \langle f,g\rangle(\theta_0)| = |\langle f,g'-g\rangle(\theta_0)| \le \|f\|(\theta_0)\|g'-g\|(\theta_0). \tag{A.7}$$

Now we derive an upper bound for $\|g'-g\|(\theta_0)$. First note that

$$\begin{aligned}(\Delta g'_{i,l}(t) - \Delta g_{i,l}(t))^2 &= \Delta g^2_{i,l}(t)\left(\frac{e^{c_{i,l}(t)}\omega_{..}(t) - \omega_{..}(t) + \omega_{..}(t) - \omega'_{..}(t)}{\omega'_{..}(t)}\right)^2\\ &\le \Delta g^2_{i,l}(t)\frac{2(e^{c_{i,l}(t)}\omega_{..}(t) - \omega_{..}(t))^2 + 2(\omega_{..}(t) - \omega'_{..}(t))^2}{(\omega'_{..}(t))^2}.\end{aligned} \tag{A.8}$$

By the inequality that $(e^a - e^b)/(a-b) \le e^{|a|\vee|b|}$ for $\forall a,b$, we have for $\forall t\in[0,\tau]$,

$$|e^{c_{i,l}(t)} - 1| \le e^{|c_{i,l}(t)|}|c_{i,l}(t)|. \tag{A.9}$$

As a result, for $\forall t\in[0,\tau]$,

$$\left|\omega'_{..}(t) - \omega_{..}(t)\right| = \left|n^{-2}\sum_{1\le i,l\le n}2\omega_i(t)\omega_l(t)(e^{c_{i,l}(t)}-1)\right| \le n^{-2}\sum_{1\le i,l\le n}2\omega_i(t)\omega_l(t)e^{|c_{i,l}(t)|}|c_{i,l}(t)|.$$

From the above inequality, we can get following two conclusions. For $\forall t\in[0,\tau]$,

$$\left|\omega'_{..}(t) - \omega_{..}(t)\right|/\omega_{..}(t) = O_P\left(\max_{1\le i,l\le n}\sup_{t\in[0,\tau]} e^{|c_{i,l}(t)|}|c_{i,l}(t)|\right) = o_P(1). \tag{A.10}$$

$$\begin{aligned}\left|\omega'_{..}(t) - \omega_{..}(t)\right| &\le \sqrt{n^{-2}\sum_{1\le i,l\le n}2\omega_i(t)\omega_l(t)}\sqrt{n^{-2}\sum_{1\le i,l\le n}2\omega_i(t)\omega_l(t)e^{2|c_{i,l}(t)|}c^2_{i,l}(t)}\\ &= \sqrt{\omega_{..}(t)}\sqrt{n^{-2}\sum_{1\le i,l\le n}2\omega_i(t)\omega_l(t)e^{2|c_{i,l}(t)|}c^2_{i,l}(t)}.\end{aligned} \tag{A.11}$$

For $\forall t \in [0, \tau]$, we have

$$
\begin{aligned}
&(\Delta g'_{i,l}(t) - \Delta g_{i,l}(t))^2 \\
&\le \frac{2\Delta g_{i,l}^2(t)}{(\omega'_{\cdot\cdot}(t))^2}\left[c_{i,l}^2(t)e^{2|c_{i,l}(t)|}\omega_{\cdot\cdot}^2(t) + \omega_{\cdot\cdot}(t)n^{-2}\sum_{1\le i,l\le n} 2\omega_i(t)\omega_l(t)e^{2|c_{i,l}(t)|}c_{i,l}^2(t)\right] \\
&= 2\Delta g_{i,l}^2(t)c_{i,l}^2(t)(1+o_P(1)) + \frac{2(1+o_P(1))\Delta g_{i,l}^2(t)\left[n^{-2}\sum_{1\le i,l\le n} 2\omega_i(t)\omega_l(t)c_{i,l}^2(t)\right]}{\omega_{\cdot\cdot}(t)} \\
&\le 2\Delta g_{i,l}^2(t)(1+o_P(1))\left[c_{i,l}^2(t) + \frac{n^{-2}\sum_{1\le i,l\le n} 2\omega_i(t)\omega_l(t)c_{i,l}^2(t)}{\omega_{\cdot\cdot}(t)}\right],
\end{aligned}
\tag{A.12}
$$

where the first step is by plugging (A.9) and (A.11) into (A.8), the second step is by (A.10) and the last step is by simple calculation.

Since $\max_{1\le i,l,\le n}\sup_{t\in[0,\tau]}|\Delta g_{i,l}(t)| = O_P(K)$, (A.12) implies that

$$
\begin{aligned}
\|g' - g\|^2(\theta_0) &= \frac{1}{n}\int_0^\tau \frac{n^{-2}\sum_{1\le i,l\le n}\omega_i(t)\omega_l(t)(\Delta g'_{i,l}(t) - \Delta g_{i,l}(t))^2}{\omega_{\cdot\cdot}(t)}d\bar{N}(t) \\
&= \frac{O_P(2K^2)}{n}\left[\int_0^\tau \frac{n^{-2}\sum_{1\le i,l\le n}\omega_i(t)\omega_l(t)c_{i,l}^2(t)}{\omega_{\cdot\cdot}(t)}d\bar{N}(t)\right. \\
&\quad \left. + \int_0^\tau \frac{n^{-2}\sum_{1\le i,l\le n}\omega_i(t)\omega_l(t)c_{i,l}^2(t)}{\omega_{\cdot\cdot}(t)}d\bar{N}(t)\right] \\
&= \frac{O_P(4K^2)}{n}\int_0^\tau \frac{n^{-2}\sum_{1\le i,l\le n}\omega_i(t)\omega_l(t)c_{i,l}^2(t)}{\omega_{\cdot\cdot}(t)}d\bar{N}(t).
\end{aligned}
$$

Note that

$$
c_{i,l}^2(t) \le 2[\theta_{0,i}(t) - \bar{\theta}_{0,n}(t,\beta_0) - \theta'_i(t) + \bar{\theta}'_n(t,\beta_0)]^2 + 2[\theta_{0,l}(t) - \bar{\theta}_{0,n}(t,\beta_0) - \theta'_l(t) + \bar{\theta}'_n(t,\beta_0)]^2.
$$

Thus, we have

$$
\begin{aligned}
&\frac{1}{n}\int_0^\tau \frac{n^{-2}\sum_{i,l}\omega_i(t)\omega_l(t)c_{i,l}^2(t)}{\omega_{\cdot\cdot}(t)}d\bar{N}(t) \\
&\le \frac{2}{n}\int_0^\tau \frac{\sum_{1\le i\le n}Y_i(t)e^{X_i(t)\beta_0}[\theta_{0,i}(t) - \bar{\theta}_{0,n}(t,\beta_0) - \theta'_i(t) + \bar{\theta}'_n(t,\beta_0)]^2}{\sum_{1\le i\le n}Y_i(t)e^{X_i(t)\beta_0}}d\bar{N}(t) \\
&= \frac{2}{n}\int_0^\tau \overline{\mathrm{Var}(\theta_0 - \theta')}(t,\beta_0)d\bar{N}(t) \\
&= 2\overline{\mathrm{Var}(\theta_0 - \theta')}(\beta_0).
\end{aligned}
$$

Thus from (A.7), we proved that

$$|\langle f,g\rangle(\theta') - \langle f,g\rangle(\theta_0)| \le O_P(K)\|f\|(\theta_0)\overline{\mathrm{Var}(\theta_0-\theta')}(\beta_0)^{1/2}.$$

which gives the inequality in (A.6).

$\square$

Lemma A.1.4. *Suppose* $z'(T_{(l)}),\ l=1,\ldots,q$ *is a solution to (2.8) and* $z(t)$ *is computed via (2.9). For* $\lambda' \asymp \sqrt{\frac{\log p}{n}}$ *we have*

$$\overline{Var(z)}(T_{(l)},\widehat{\beta}^{(init)}) \le \frac{\max_{j\le p} F_{j,j}(T_{(l)},\beta_0)+o_P(1)}{\|a_0\|_\infty - \lambda'}. \tag{A.13}$$

Proof. By (2.9),

$$a_{0,j} - \overline{\mathrm{Cov}(z',X_j)}(T_{(l)},\widehat{\beta}^{(init)}) \le \lambda'.$$

For $c\in\mathbb{R}$,

$$\begin{aligned}\overline{\mathrm{Var}(z')}(t,\widehat{\beta}^{(init)}) &\ge \overline{\mathrm{Var}(z')}(t,\widehat{\beta}^{(init)}) + c(a_{0,j}-\lambda') - \overline{\mathrm{Cov}(z',X_j)}(t,\widehat{\beta}^{(init)})\\ &\ge \min_{z'(t)}\{\overline{\mathrm{Var}(z')}(t,\widehat{\beta}^{(init)}) + c(a_{0,j}-\lambda') - \overline{\mathrm{Cov}(z',X_j)}(t,\widehat{\beta}^{(init)})\}\\ &= c(a_{0,j}-\lambda') - \frac{c^2}{4}\overline{\mathrm{Var}(X_j)}(t,\widehat{\beta}^{(init)}).\end{aligned}$$

Optimizing this bound over c, we arrive at

$$\overline{\mathrm{Var}(z')}(t,\widehat{\beta}^{(init)}) \ge \frac{a_{0,j}-\lambda'}{\overline{\mathrm{Var}(X_j)}(t,\widehat{\beta}^{(init)})}.$$

with $c = 2(a_{0,j}-\lambda')/\overline{\mathrm{Var}(X_j)}(t,\widehat{\beta}^{(init)})$.

By Lemma A.1.3,

$$\overline{\mathrm{Var}(X_j)}(t,\widehat{\beta}^{(init)}) = \overline{\mathrm{Var}(X_j)}(t,\beta_0) + \overline{\mathrm{Var}(X_j)}(t,\beta_0)^{1/2}\overline{\mathrm{Var}(Xh)}(t,\beta_0)^{1/2}O_P(K_1). \tag{A.14}$$

Note that

$$\begin{aligned}\left|\overline{\mathrm{Var}(Xh)}(t,\beta_0)\right| &\le |h^T F(t,\beta_0)h| + |h^T\overline{\mathrm{Var}(X)}(t,\beta_0) - F(t,\beta_0))h|\\ &\le \|h\|_2^2\Lambda_{\max}(F(t,\beta_0)) + \|h\|_1 \max_{1\le j\le p}|\overline{\mathrm{Var}(X)}(t,\beta_0)h - F(t,\beta_0)h|_j\\ &= O_P(s\lambda^2 + s^2\lambda^3),\end{aligned}$$

where the last step is by Lemma A.1.1. Thus, using (A.14) and Lemma A.1.1 (ii), we have

$$\begin{aligned}\max_{j\le p}\overline{\mathrm{Var}(X_j)}(t,\widehat{\beta}^{(init)}) &\le \max_{j\le p}\sup_{t\in[0,\tau]}\overline{\mathrm{Var}(X_j)}(t,\beta_0)+O_P(s^{1/2}\lambda)\\ &=\max_{j\le p}\sup_{t\in[0,\tau]}F_{j,j}(t,\beta_0)+o_P(1).\end{aligned}$$

□

A.1.2 Proof of Lemmas in Section 2.3.2

Proof of Lemma 2.3.4. Theorem 3.2 in Huang et al. (2013) and the relationship among compatibility and invertibility factors and restricted eigenvalue discussed at the top of page 9 of Huang et al. (2013) impies that for $\lambda \asymp \sqrt{\frac{\log p}{n'}}$,

$$\|h\|_1 = O_P(s\lambda) \text{ and } \|h\|_2 = O_P(s\lambda^2),$$

under Conditions 2.3.1-2.3.2.

□

Proof of Lemma 2.3.8. Define $z_0(T_{(l)}) = X(T_{(l)})u_0(T_{(l)},\beta_0)$ for $u_0(T_{(l)},\beta_0)$ defined in Condition 2.3.7, $l=1,\dots,q$. By Condition 2.3.6, $F^{-1}(T_{(i)},\beta_0)$ is well-defined. Consider

$$z_i^*(T_{(l)}) = \frac{S^{(0)}(T_{(l)},\widehat{\beta}^{(init)})}{S^{(0)}(T_{(l)},\beta_0)}e^{-X_i(T_{(l)})h}[z_{0,l}(T_{(l)})-\bar{z}_{0,n}(T_{(l)},\beta_0)]. \tag{A.15}$$

We show that $z^*(T_{(l)})\in\mathbb{R}^n$ defined in (A.15) is a feasible solution to (2.8) for $l = 1,\dots,q$. First note that

$$\bar{z}_n^*(T_{(l)},\widehat{\beta}^{(init)}) = \sum_{i=1}^n \gamma_{ni}(T_{(l)},\widehat{\beta}^{(init)})z_i^*(T_{(l)}) = \sum_{i\in R(0)}\gamma_{ni}(T_{(l)},\beta_0)[z_{0,i}(T_{(l)})-\bar{z}_{0,n}(T_{(l)},\beta_0)] = 0.$$

One can see that

$$\begin{aligned}\overline{\mathrm{Cov}(z^*,X)}(T_{(l)},\widehat{\beta}^{(init)})-a_0 =&\sum_{i=1}^n\gamma_{ni}(T_{(l)},\widehat{\beta}^{(init)})[z_i^*(T_{(l)})-\bar{z}_n^*(T_{(l)},\widehat{\beta}^{(init)})]X_i(T_{(l)})-a_0\\ =&\sum_{i=1}^n\gamma_{ni}(T_{(l)},\widehat{\beta}^{(init)})z_i^*(T_{(l)})X_i(T_{(l)})-a_0\\ =&\sum_{i=1}^n\gamma_{ni}(T_{(l)},\beta_0)[z_{0,i}(T_{(l)})-\bar{z}_{0,n}(T_{(l)},\beta_0)]X_i(T_{(l)})-a_0\\ =&u_0(T_{(l)},\beta_0)\left[\overline{\mathrm{Var}(X)}(T_{(l)},\beta_0)-F(T_{(l)},\beta_0)\right],\end{aligned}$$

where $u_0(t, \beta_0) = F^{-1}(t, \beta_0)a_0$. Condition 2.3.7 and Lemma A.1.1 (i) together imply that

$$\sup_{t\in[0,\tau]} \left\| u_0(t, \beta_0)^T \left[\overline{\mathrm{Var}(X)}(t, \beta_0) - F(t, \beta_0) \right] \right\|_\infty = O_P\left(K_1 K_2 K_3 \sqrt{\frac{\log p}{n}} \right).$$

Moreover, it is easy to see that

$$\max_{1\le i\le n} \max_{1\le l\le q} \left| z_i^*(T_{(l)}) \right| \le \max_{1\le i\le n} \max_{1\le l\le q} \left| z_{0,i}(T_{(l)}) - \bar{z}_{0,n}(T_{(l)}, \beta_0) \right| + O_P(s\lambda) = O_P(K_3).$$

□

Proof of Lemma 2.3.5. Note that

$$\sqrt{n}\xi(0, \beta_0) = \sum_{i=1}^n \int_0^\tau \xi_i(t, \beta_0) dN_i(t),$$

where $\xi_i(t, \beta_0) = n^{-1/2}[z_i(t) - \bar{z}_n(t, \beta_0)]$.

First note that $z_i(t)$ $t \in [T_{(l)}, T_{(l+1)})$ is a function of $(X(T_{(l)}), Y(T_{(l)}), \widehat{\beta}^{(init)})$. Since $(X(t), Y(t))$ are predictable processes and $\widehat{\beta}^{(init)}$ is $\mathcal{F}_0$ measurable, $_i(t)$ is a $\mathcal{F}_t$ predictable process for $t \in [0, \tau]$. Moreover, since $z(t) = z'(t)\overline{\mathrm{Var}(z)}(t, \widehat{\beta}^{(init)})$, Lemma A.1.4 implies that

$$\begin{aligned} \max_{i\in R(0)} \sup_{t\in[0,\tau]} |\xi_i(t, \beta_0)| &\le \max_{i\in R(0)} \sup_{t\in[0,\tau]} n^{-1/2} |z_i'(t) - \bar{z}_n'(t, \beta_0)| \overline{\mathrm{Var}(z)}(t, \widehat{\beta}^{(init)}) \\ &\le \sup_{t\in[0,\tau]} 2K_4 n^{-1/2} \left[\frac{\max_{j\le p} F_{j,j}(t, \beta_0)}{\|a_0\|_\infty - \lambda} + o_P(1) \right], \end{aligned} \tag{A.16}$$

where $\sup_{t\in[0,\tau]} \max_{j\le p} F_{j,j}(t, \beta_0) \le c_2^*$ by Condition 2.3.4.

To utilize the martingale CLT, we first check

$$\begin{aligned} \langle \sqrt{n}\xi(0, \beta_0), \sqrt{n}\xi(0, \beta_0) \rangle(t) &= \sum_{i=1}^n \int_0^\tau \xi_i^2(t, \beta_0) d\Lambda_i(t) \\ &= \int_0^\tau \overline{\mathrm{Var}(z)}(t, \beta_0) S^{(0)}(t, \beta_0) d\Lambda_0(t) \to F_z([0, \tau], \beta_0), \end{aligned}$$

in probability by Condition 2.3.4. It remains to check the Lindeberg condition

$$\sum_{i=1}^n \int_0^\tau \xi_i^2(t, \beta_0) \mathbb{I}_{\{|\xi_i(t,\beta_0)|>\epsilon\}} d\Lambda_i(t) \to 0 \text{ in probability fo all } \epsilon > 0. \tag{A.17}$$

The left hand side of (A.17) is bounded by

$$
\begin{aligned}
&\mathbb{1}_{\{\max_{1\le i\le n}\sup_{t\in[0,\tau]}|\xi_i(t,\beta_0)|>\epsilon\}}\sum_{1\le i\le n}\int_0^\tau \xi_i^2(t,\beta_0)d\Lambda_i(t)\\
&=\mathbb{1}_{\{\max_{1\le i\le n}\sup_{t\in[0,\tau]}|z_i(t)-\bar{z}_n(t,\beta_0)|>\sqrt{n}\epsilon\}}\sum_{1\le i\le n}\int_0^\tau \xi_i^2(t,\beta_0)d\Lambda_i(t)\\
&\to 0 \text{ as } n\to\infty,
\end{aligned}
$$

due to (A.16). □

Proof of Lemma 2.3.6. First by Lemma A.1.3 and Lemma 2.3.4,

$$\overline{\mathrm{Cov}(z)}(\widehat{\beta}^{(init)})=\overline{\mathrm{Cov}(z)}(\beta_0)+O_P(K_4 s^{1/2}\lambda).$$

Since $\overline{\mathrm{Cov}(z)}(t,\beta_0)$ is a bounded process, by Lemma 3.3(i) in Huang et al. (2013),

$$\overline{\mathrm{Cov}(z)}(\beta_0)-F_z([0,\tau];\beta_0)=O_P(n^{-1/2}).$$

(2.18) is proved by combining above two equations. □

Proof of Lemma 2.3.7. For any $t\in\{T_{(1)},\dots,T_{(q)}\}$, by the proof of Lemma 2.3.8,

$$\overline{\mathrm{Var}(z')}(t,\widehat{\beta}^{(init)})\le\overline{\mathrm{Var}(z^*)}(t,\widehat{\beta}^{(init)})=a_0^T F^{-1}(t,\beta_0)a_0+o_P(1),$$

where the last step is due to Lemma A.1.3. And the upper bound is proved. On the other hand, since $z'(t)$ satisfies the constraints in (2.8), with large probability it holds that

$$a_{0,j}-\lambda'\le\overline{\mathrm{Cov}(z',X_j)}(t,\widehat{\beta}^{(init)})\le a_{0,j}+\lambda'.$$

Thus, we can obtain that

$$
\begin{aligned}
\overline{\mathrm{Cov}(z',Xu_0)}(t,\widehat{\beta}^{(init)})&=\sum_{j=1}^p\overline{\mathrm{Cov}(z',X_j)}(t,\widehat{\beta}^{(init)})u_{0,j}(t)\\
&\ge\sum_{u_{0,j}(t)\ge0}(a_{0,j}-\lambda')u_{0,j}(t)+\sum_{u_{0,j}(t)<0}(a_{0,j}+\lambda')u_{0,j}(t)\\
&=a_0^T u_0(t)-\lambda\|u_0(t)\|_1.
\end{aligned}
$$

By Cauchy-Schwarz inequality, we arrive at

$$\overline{\mathrm{Var}(z')}^{1/2}(t,\widehat{\beta}^{(init)})\overline{\mathrm{Var}(Xu_0)}^{1/2}(t,\widehat{\beta}^{(init)})\ge a_0^T F^{-1}(t,\beta_0)a_0-\lambda\|u_0(t)\|_1.$$

By Lemma A.1.1 and Lemma A.1.3, we have $\overline{\mathrm{Var}(Xu_0)}(t, \widehat{\beta}^{(init)}) = a_0^T F(t, \beta_0) a_0 + o_P(1)$ and

$$.\overline{\mathrm{Var}(z')}^{1/2}(t, \widehat{\beta}^{(init)}) \geq (a_0^T F^{-1}(t, \beta_0) a_0)^{1/2} - o_P(1).$$

□

A.1.3 Proof of theorems in Section 2.3.1

Proof of Theorem 2.3.1. To see the expansion in (2.11), we reparamaterize $D(\theta_0)$ as $\xi(h, \beta_0)$:

$$\xi(h, \beta_0) := \frac{1}{n} \sum_{i=1}^{n} \int_0^{\tau} \left[z_i(t) - \frac{\sum_{i=1}^{n} z_i(t) Y_i(t) e^{X_i(t)\beta_0 + X_i(t)h - z_i(t) a_0^T h}}{\sum_{i=1}^{n} Y_i(t) e^{X_i(t)\beta_0 + X_i(t)h - z_i(t) a_0^T h}} \right] dN_i(t). \quad \text{(A.18)}$$

Then we have

$$\begin{aligned}
\sqrt{n} D\big(\theta_0 + \delta n^{-1/2}\big) &= \sqrt{n} D(\theta_0) + \delta \int_0^1 \dot{D}(\theta_0 + x\delta n^{-1/2}) dx \\
&= \sqrt{n} \xi(h, \beta_0) + \delta \int_0^1 \dot{D}(\theta_0 + x\delta n^{-1/2}) dx \\
&= \sqrt{n} \xi(0, \beta_0) + \sqrt{n} \int_0^1 h^T \dot{\xi}(xh, \beta_0) dx + \delta \int_0^1 \dot{D}(\theta_0 + x\delta n^{-1/2}) dx \\
&= \sqrt{n} \xi(0, \beta_0) + \delta F_z([0, \tau]; \beta_0) + \underbrace{\sqrt{n} h^T \dot{\xi}(0, \widehat{\beta}^{(init)})}_{Rem_1} \\
&\quad + \underbrace{\sqrt{n} \left[\int_0^1 h^T \dot{\xi}(xh, \beta_0) dx - h^T \dot{\xi}(0, \widehat{\beta}^{(init)}) \right]}_{Rem_2} \\
&\quad + \underbrace{\delta \int_0^1 \dot{D}(\theta_0 + x\delta n^{-1/2}) dx - \delta F_z([0, \tau]; \beta_0)}_{Rem_3}. \qquad \text{(A.19)}
\end{aligned}$$

Taking derivative of (A.18) with respect to h, we can obtain that

$$\begin{aligned}
h^T \dot{\xi}(0, \widehat{\beta}^{(init)}) &= \frac{1}{n} \int_0^{\tau} \left[\overline{\mathrm{Cov}(z, X)}(t, \widehat{\beta}^{(init)}) h - \overline{\mathrm{Cov}(z, z a_0^T)}(t, \widehat{\beta}^{(init)}) h \right] d\bar{N}(t) \\
&= \frac{1}{n} \int_0^{\tau} \left[\overline{\mathrm{Cov}(z, X - z a_0^T)}(t, \widehat{\beta}^{(init)}) h \right] d\bar{N}(t) \\
&= \frac{1}{n} \int_0^{\tau} \overline{\mathrm{Var}(z)}(t, \widehat{\beta}^{(init)}) [\overline{\mathrm{Cov}(z', X)}(t, \widehat{\beta}^{(init)}) - a_0^T] h d\bar{N}(t) \\
&= \frac{1}{n} \sum_{l=1}^{q} \overline{\mathrm{Var}(z)}(T_{(l)}, \widehat{\beta}^{(init)}) [\overline{\mathrm{Cov}(z', X)}(T_{(l)}, \widehat{\beta}^{(init)}) - a_0^T] h
\end{aligned}$$

It follows from (2.8) and Lemma A.1.4 that

$$\begin{aligned}
&|Rem_1| \\
&\le \max_{1\le l\le q} \sqrt{n} \left\| \overline{\mathrm{Cov}(Z', X)}(T_{(l)}, \widehat{\beta}^{(init)}) - a_0^T \right\|_\infty \|h\|_1 \left| \frac{1}{n} \sum_{l=1}^q \overline{\mathrm{Var}(Z)}(T_{(l)}, \widehat{\beta}^{(init)}) \right| \\
&= O_P(\sqrt{n} s\lambda\lambda') \left| \frac{1}{n} \int_0^\tau \overline{\mathrm{Var}(Z)}(t, \widehat{\beta}^{(init)}) d\bar{N}(t) \right| \\
&= o_P(1),
\end{aligned}$$

where the last step is due to (2.18) and Lemma 2.3.6.

For Rem_2, note that

$$h^T \dot{\xi}(xh, \beta_0) = \langle z, Xh\rangle(\theta_x) \text{ and } h^T \dot{\xi}(0, \widehat{\beta}^{(init)}) = \langle z, Xh\rangle(\widehat{\theta}^{(init)}),$$

where $\theta_x(t) = X(t)\beta_0 + xX(t)h - xZ(t)a_0^T h$.

We apply Lemma A.1.3 with $\theta'(t) = \theta_x(t)$, $f(t) = X(t)h$ and $g(t) = z(t)$. Then

$$\begin{aligned}
&c_{i,l}(t) \\
&= x(X_i(t)h - z_i(t)a_0^T h) + x(X_l(t)h - z_l(t)a_0^T h) - 2x(\bar{X}_n(t, \beta_0)h - \bar{z}_n(t, \beta_0)a_0^T h).
\end{aligned}$$

Since

$$|a_0^T h| \le \|a_0\|_2 \|h\|_2 = O_P(s^{1/2}\lambda), \tag{A.20}$$

together with Conditions 2.3.2-2.3.3 and (2.8),

$$\sup_{t\in[0,\tau]} \max_{1\le i,l\le n} |c_{i,l}(t)| = O_P(K_1 s\lambda) + \sup_{t\in[0,\tau]} \max_{1\le j\le p} F_{j,j}(t, \beta_0) O_P(K_4 s^{1/2}\lambda) = o_P(1).$$

Thus, (A.5) implies that

$$|\langle z, Xh\rangle(\theta_x) - \langle z, Xh\rangle(\theta_0)| = O_P(K_4)\|Xh\|(\theta_0)\overline{\mathrm{Var}(Xh - za_0^T h)}(\beta_0)^{1/2}.$$

On the right hand side,

$$\begin{aligned}
&\|Xh\|^2(\theta_0) = \langle h, \ddot{\ell}([0,\tau]; \beta_0)h\rangle \\
&\le h^T F([0,\tau]; \beta_0)h + \left| h^T(\ddot{\ell}([0,\tau]; \beta_0) - F([0,\tau]; \beta_0))h \right| \\
&\le \|h\|_2^2 \Lambda_{\max}(F([0,\tau]; \beta_0)) + \|h\|_1 \left\| (\ddot{\ell}([0,\tau]; \beta_0) - F([0,\tau]; \beta_0))h \right\|_\infty \\
&= O_P(c_1^* s\lambda^2 + K_1 s^2 \lambda^3),
\end{aligned}$$

where the last step is due to Condition 2.3.2 and 2.3.4 and Lemma 2.3.4 and A.1.1. Moreover,

$$\begin{aligned}\overline{\mathrm{Var}(Xh - za_0^T h)}(\beta_0) &\leq 2\overline{\mathrm{Var}(Xh)}(\beta_0) + 2(a_0^T h)^2\overline{\mathrm{Var}(z)}(\beta_0) \\ &= 2\langle h, \ddot{\ell}([0,\tau];\beta_0)h\rangle + 2(a_0^T h)^2\overline{\mathrm{Var}(z)}(\beta_0) \\ &= O_P(s\lambda^2) + O_P(c_1^* s\lambda^2),\end{aligned}$$

where the last step is due to (A.20) and Lemma A.1.3. Similarly,

$$\begin{aligned}\left|\langle z, Xh\rangle(\widehat{\theta}^{(init)}) - \langle z, Xh\rangle(\theta_0)\right| &= O_P(K_4)\|Xh\|(\theta_0)\|\overline{\mathrm{Var}(Xh)}(\beta_0)^{1/2} \\ &= O_P(K_4)h^T\ddot{\ell}([0,\tau],\beta_0)h = O_P(c_1^* s\lambda^2).\end{aligned}$$

Thus,

$$|Rem_2| \leq \left|h^T\dot{\xi}(xh,\beta_0) - h^T\dot{\xi}(0,\widehat{\beta}^{(init)})\right|\sqrt{n}\int_0^1 dx = O_P(\sqrt{n}s\lambda^2) = o_P(1). \tag{A.21}$$

For Rem_3, we can see that

$$\dot{D}(\theta_0 + x\delta) = \|z\|^2(\tilde{\theta}),$$

where $\tilde{\theta}(t) = X(t)\widehat{\beta}^{(init)} + z(t)(x\delta n^{-1/2} - a_0^T h)$. Note that

$$\tilde{c}_{i,l}(t) = X_i(t)h + z_i(t)(x\delta n^{-1/2} - a_0^T h) + X_l(t)h + z_l(t)(x\delta n^{-1/2} - a_0^T h) = o_P(s\lambda).$$

Applying Lemma A.1.3 again, we have

$$|\|z\|^2(\tilde{\theta}) - \|z\|^2(\theta_0)| \leq \|z\|(\theta_0)\overline{\mathrm{Var}(Xh + z(x\delta n^{-1/2} - a_0^T h))}(\theta_0)^{1/2}O_P(K_4).$$

By similar arguments as for Rem_2 and Condition 2.3.4, we have

$$\delta\|z\|^2(\tilde{\theta}) = \delta\|z\|^2(\theta_0) + O_P(\delta s^{1/2}\lambda) = \delta F_z([0,\tau];\beta_0) + o_P(1).$$

Putting our arguments together, we have

$$\sqrt{n}D\big(\theta_0 + \delta n^{-1/2}\big) = \sqrt{n}\xi(0,\beta_0) + \delta F_z([0,\tau];\beta_0) + o_P(1).$$

□

Proof of Corollary 2.3.2. We first show the MLE $\widehat{\theta}^{(LDP)}$ defined via (2.5) satisfies $\widehat{\theta}^{(LDP)} = \theta_0 + O_P(n^{-1/2})$. By the convexity of the log-likelihood, it is sufficient to show that there exist $B > 0$, such that $\forall B' > B$

$$\inf_{\epsilon\in\{\pm 1\}} \mathbb{P}\left\{(B' - B)n^{-1/2}\epsilon D(\theta_0 + Bn^{-1/2}\epsilon) > 0\right\} \geq 1 - \epsilon, \text{ for large enough } n.$$

By the results in Theorem 2.3.1(i),

$$D(\theta_0 + Bn^{-1/2}\epsilon) = \xi(0, \beta_0) + Bn^{-1/2}\epsilon F_z([0, \tau]; \beta_0) + o_P(n^{-1/2}).$$

As a result,

$$\begin{aligned}&\inf_{\epsilon\in\{\pm 1\}} \mathbb{P}\left\{(B' - B)n^{-1/2}\epsilon D(\theta_0 + Bn^{-1/2}\epsilon) > 0\right\}\\ &\geq \mathbb{P}\left\{\sqrt{n}\xi(0, \beta_0) > -BF_z([0, \tau]; \beta_0)\right\} - o_P(1).\end{aligned}$$

Since $\xi(0, \beta_0)$ is a locally bounded martingale, it is easy to show $\xi(0, \beta_0) = O_P(n^{-1/2})$ and the right hand side of the above inequality equals $1 - o_P(1)$.

Since $\widehat{\theta}^{(LDP)} - \theta_0 = O_P(n^{-1/2})$, by the results in Theorem 2.3.1(i), the equation $D(\theta) = 0$ yields that

$$0 = D(\widehat{\theta}^{(LDP)}) = \xi(0, \beta_0) + F_z([0, \tau]; \beta_0)(\widehat{\theta}^{(LDP)} - \theta_0) + o_P(n^{-1/2}).$$

Then we have

$$\sqrt{n}(\widehat{\theta}^{(LDP)} - \theta_0) = -F_z^{-1}([0, \tau]; \beta_0)\xi(0, \beta_0) + o_P(n^{-1/2}).$$

With Lemma 2.3.5 and Theorem 2.3.1 (ii), we apply Slutsky's lemma to obtain that

$$\sqrt{n\hat{F}_z(\widehat{\beta}^{(init)})}(\widehat{\theta}^{(LDP)} - \theta_0) \xrightarrow{D} N(0, 1).$$

For $\widehat{\theta}^{(OS)}$ defined in (2.10), we have

$$\widehat{\theta}^{(OS)} - \theta_0 = \langle a_0, h\rangle - D(\widehat{\theta}^{(init)})/\hat{F}_z(\widehat{\beta}^{(init)}).$$

We expand $D(\widehat{\theta}^{(init)})$ as in Theorem 2.3.1 (i), which gives

$$\begin{aligned}\widehat{\theta}^{(OS)} - \theta_0 &= \langle a_0, h\rangle - \xi(0, \beta_0)/\hat{F}_z(\widehat{\beta}^{(init)}) - \langle a_0, h\rangle + o_P(n^{-1/2})\\ &= -\xi(0, \beta_0)/\hat{F}_z(\widehat{\beta}^{(init)}) + o_P(n^{-1/2}).\end{aligned}$$

Again by Slutsky's Lemma, we arrive at

$$\sqrt{n\hat{F}_z(\widehat{\beta}^{(init)})}(\widehat{\theta}^{(OS)} - \theta_0) \xrightarrow{D} N(0,1).$$

□

A.2 Proofs for theorems and lemmas in Section 3

A.2.1 Some technical lemmas

To facilitate the proofs of lemmas in the paper, we first prove two technical lemmas. To simplify our notations, let $\hat{u}_S = \hat{\beta}_S - \beta_S$, $\hat{u}^*_S = \hat{\beta}^*_S - \hat{\beta}_S$, $W^n_S = X^T_S\epsilon/n$, $W^*_S = X^T_S\hat{\epsilon}^*/n$ and $S_{\backslash j} = S\backslash\{j\}$.

Lemma A.2.1 (Symmetrization). *Assume that Conditions 3.3.1 - 3.3.4 hold true,*

$$\lambda > \frac{16\sigma}{1-\kappa}\sqrt{\frac{2\log p}{n}} \text{ and } n \geq \frac{32\sigma^2}{\lambda^2(1-\kappa)^2}.$$

Then we have

$$\mathbb{P}\left(\max_{j\in S^c}\left|\frac{x_j^T P_S^{\perp}\epsilon}{n\lambda}\right| > \frac{1-\kappa}{2}\right) \leq 4\exp(-c_1\log p) + \frac{c_2}{n}, \tag{A.22}$$

$$\mathbb{P}\left(\max_{j\in S}\left|(\Sigma^n_{S,S})^{-1}_{j,\cdot}W^n_S\right| > 8\sigma\sqrt{\frac{2\log p}{C_{\min}n}}\right) \leq 4\exp(-c_1\log p) + \frac{c_2}{n}, \tag{A.23}$$

for some $c_1, c_2 > 0$.

Proof. In this proof, we apply standard symmetrization techniques.

Let

$$Q_{1,j} = \frac{x_j^T P_S^{\perp}\epsilon}{n\lambda} = \frac{1}{n}\sum_{i=1}^n \frac{(x_j^T P_S^{\perp})_i\epsilon_i}{\lambda},$$

$(\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_n)$ be an independent copy of $(\epsilon_1, \ldots, \epsilon_n)$. $\tilde{Q}_{1,j} = x_j^T P_S^{\perp}\tilde{\epsilon}/(n\lambda)$ and $\omega_1, \ldots, \omega_n$ be a Rademacher sequence. Note that

$$\max_{j\in S^c}\mathbb{E}\left[\frac{(x_j^T P_S^{\perp})_i\epsilon_i}{\lambda}\right] = 0,$$

$$\max_{j\in S^c}\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\left(\frac{(x_j^T P_S^{\perp})_i\epsilon_i}{\lambda}\right)^2\right] = \frac{1}{n\lambda^2}\left\|P_S^{\perp}x_j\right\|_2^2\sigma^2 \leq \frac{\sigma^2}{\lambda^2} \equiv C_1.$$

We apply symmetrization inequalities (Problem 14.5 in Bühlmann and van De Geer (2011)), which gives for $\forall\ 0 < \eta < 1$,

$$\mathbb{P}\left(\max_{j\in S^c}|Q_{1,j}| > t\right) \leq \frac{\mathbb{P}\left(\max_{j\in S^c}|Q_{1,j} - \tilde{Q}_{1,j}| > (1-\eta)t\right)}{1 - C_1/n\eta^2t^2}$$
$$\leq \frac{2\mathbb{P}\left(\max_{j\in S^c}\left|\frac{1}{n}\sum_{i=1}^n \frac{(x_j^T P_S^{\perp})_i\epsilon_i\omega_i}{\lambda}\right| > (1-\eta)t/2\right)}{1 - C_1/n\eta^2t^2}.$$

For $n \geq 2C_1/(\eta^2t^2)$, we have

$$\mathbb{P}\left(\max_{j\in S^c}|Q_{1,j}| > t\right) \leq 4\mathbb{P}\left(\max_{j\in S^c}\left|\frac{1}{n}\sum_{i=1}^n \frac{(x_j^T P_S^{\perp})_i\epsilon_i\omega_i}{\lambda}\right| > (1-\eta)t/2\right). \tag{A.24}$$

Conditioning on ϵ, by McDiarmid's inequality, we have

$$\mathbb{P}\left(\max_{j\in S^c}\left|\frac{1}{n}\sum_{i=1}^n \frac{(x_j^T P_S^{\perp})_i\epsilon_i\omega_i}{\lambda}\right| > (1-\eta)t/2\Big|\epsilon\right) \leq \sum_{j\in S^c}\exp\left\{-\frac{2((1-\eta)t/2)^2}{\sum_{i=1}^n\left[2(x_j^T P_S^{\perp})_i\epsilon_i/(n\lambda)\right]^2}\right\}. \tag{A.25}$$

Moreover, by Chebyshev's inequality, $\forall j \in S^c$

$$\max_{j\in S^c}\mathbb{P}\left(\sum_{i=1}^n\left[2(x_j^T P_S^{\perp})_i\epsilon_i/(n\lambda)\right]^2 - \frac{4\sigma^2}{n\lambda^2} > t\right) \leq \max_{j\in S^c}\frac{16\|P_S^{\perp}x_j\|_4^4\mathbb{E}[\epsilon_i^4]}{(n\lambda)^4t^2}$$
$$\leq \frac{16(K_0 + K_0\kappa)^2M_0}{n^3\lambda^4t^2},$$

where the last step is due to

$$\max_{j\in S^c}\|P_S^{\perp}x_j\|_4^4 \leq \max_{j\in S^c, i\leq n}(x_j^T P_S^{\perp})_i^2\|P_S^{\perp}x_j\|_2^2$$
$$\leq n\max_{j\in S^c, i\leq n}\left|x_{i,j} - x_{i,S}(X_S^TX_S)^{-1}X_S^Tx_j\right|^2$$
$$\leq n\left[\max_{i,j}|x_{i,j}| + \max_{i,j}|x_{i,j}|\|\Sigma_{S^c,S}^n(\Sigma_{S,S}^n)^{-1}\|_\infty\right]^2$$
$$\leq n(K_0 + K_0\kappa)^2.$$

Thus, for any constant C_2,

$$\mathbb{P}\left(\sum_{i=1}^n\left[2(x_j^T P_S^{\perp})_i\epsilon_i/(n\lambda)\right]^2 > \frac{4\sigma^2}{n\lambda^2} + \frac{C_2(K_0 + K_0\kappa)\sqrt{M_0}}{n\lambda^2}\right) \leq \frac{16}{nC_2^2} \to 0.$$

And hence by (A.24) and (A.25), for $n \geq 8\sigma^2/(\lambda^2\eta^2(1-\kappa)^2)$,

$$\mathbb{P}\left(\max_{j\in S^c}|Q_{1,j}| > \frac{1-\kappa}{2}\right) \leq 4(p-s)\exp\left\{-\frac{2[(1-\eta)(1-\kappa)/4]^2n\lambda^2}{4\sigma^2 + C_2(K_0 + K_0\kappa)\sqrt{M_0}}\right\} + \frac{16}{nC_2^2}.$$

Take $C_2 = \frac{4\sigma^2}{(K_0 + K_0\kappa)\sqrt{M_0}}$ and $\eta = 1/2$. Then for $\lambda > \frac{16\sigma}{1-\kappa}\sqrt{\frac{2\log p}{n}}$ and $n \geq \frac{32\sigma^2}{\lambda^2(1-\kappa)^2}$, we have

$$\mathbb{P}\left(\max_{j \in S^c} |Q_{1,j}| > \frac{1-\kappa}{2}\right) \leq 4\exp\left(-c_1 \log p\right) + \frac{c_2}{n},$$

for some $c_1, c_2 > 0$.

(ii) Now consider $Q_{2,j} = \sum_{i=1}^n (\Sigma_{S,S}^n)_{j,S}^{-1} X_{i,S}^T \epsilon_i / n$. Previous arguments still applies with

$$\begin{aligned}
\frac{1}{n}\max_{j \in S} \mathbb{E}\left[\sum_{i=1}^n (X_{i,S}(\Sigma_{S,S}^n)_{S,j}^{-1})^2 \epsilon_i^2\right] &\leq \frac{1}{n}\max_{j \in S} \left\| X_S (\Sigma_{S,S}^n)_{\cdot,j}^{-1} \right\|_2^2 \mathbb{E}\left[\epsilon_i^2\right] \\
&= \max_{j \in S} e_j^T (\Sigma_{S,S}^n)^{-1} e_j \sigma^2 \leq \frac{\sigma^2}{C_{\min}}, \\
\max_{j \in S} \mathrm{Var}\left[\sum_{i=1}^n \left(2(\Sigma_{S,S}^n)_{j,S}^{-1} X_{i,S}^T \epsilon_i / n\right)^2\right] &\leq \max_{j \in S} \frac{M_0}{n^4} \sum_{i=1}^n \left(2(\Sigma_{S,S}^n)_{j,S}^{-1} X_{i,S}^T\right)^4 \\
&\leq \max_{j \in S} \frac{16 M_0}{n^4} \sum_{i=1}^n \left(\|(\Sigma_{S,S}^n)_{j,S}^{-1}\|_1 \|X_{i,S}^T\|_\infty\right)^4 \\
&\leq \frac{16 M_0 K_1^4 K_0^4}{n^3}.
\end{aligned}$$

Thus,

$$\mathbb{P}\left(\max_{j \in S} |Q_{2,j}| > 8\sigma\sqrt{\frac{2\log p}{C_{\min} n}}\right) \leq 4\exp(-c_3 \log p) + \frac{c_4}{n}.$$

□

Lemma A.2.2 (Selection consistency of the bootstrapped Lasso)**.** *Formally, the bootstrapped Lasso estimator* $\hat{\beta}^*$ *is defined via*

$$\hat{\beta}^* = \arg\min_{b \in \mathbb{R}^p} \frac{1}{2n}\|y^* - Xb\|_2^2 + \lambda\|b\|_1. \tag{A.26}$$

Define a restricted Lasso problem with observations (X_S, y^*)*:*

$$\check{\beta}_S^* = \arg\min_{b \in \mathbb{R}^s} \frac{1}{2n}\|y^* - X_S b_S\|_2^2 + \lambda\|b_S\|_1 \text{ and } \check{\beta}_{S^c}^* = 0. \tag{A.27}$$

Define T_j^* *as*

$$T_j^* = x_j^T\left(X_S(X_S^T X_S)^{-1} sgn(\check{\beta}_S^*) + \frac{P_S^\perp \hat{\epsilon}^*}{n\lambda}\right). \tag{A.28}$$

If $\hat{S} \subseteq S$*,* $\max_{j \in S^c} |T_j^*| < 1$*, and* $\Sigma_{S,S}^n$ *is invertible, then* $\check{\beta}^*$ *in (A.27) is the unique solution to the bootstrapped Lasso (A.26) and* $\hat{S}^* \subseteq S$*.*

Proof. In the event that $\{\hat{S} \subseteq S\}$, $\hat{\beta}_{S^c} = 0$. By the KKT condition of $\check{\beta}^*_S$ in (A.27),

$$\Sigma^n_{S,S}(\check{\beta}^*_S - \hat{\beta}_S) - W^*_S + \lambda sgn(\check{\beta}^*_S) = 0.$$

If $|T^*_j| < 1$ for $\forall\ j \in S^c$, then there exists $sgn(\check{\beta}^*_{S^c})$ such that

$$\Sigma^n_{S^c,S}(\check{\beta}^*_S - \hat{\beta}_S) - W^*_{S^c} + \lambda sgn(\check{\beta}^*_{S^c}) = 0.$$

And hence there exists $sgn(\check{\beta}^*)$ such that $\check{\beta}^*$ in (A.27) is a solution to

$$\Sigma^n(\check{\beta}^* - \hat{\beta}) - W^* + \lambda sgn(\check{\beta}^*) = 0,$$

which is the KKT condition of the bootstrapped Lasso (A.26). By Lemma 1 in Wainwright (2009), $\check{\beta}^*$ is an optimal solution to the bootstrapped Lasso problem (A.26). Moreover, $\check{\beta}^*$ is the unique solution, since $\Sigma^n_{S,S}$ is invertible and $|T^*_j| < 1$ for $\forall\ j \in S^c$. This implies that $\hat{S}^* \subseteq S$. □

Lemma A.2.3 (Bounds on the ℓ_2-norm and ℓ_∞-norm of $(\Sigma^n_{S,S})^{-1}$)**.** *Under Conditions 3.4.1 - 3.4.4, we have the following results.*

(i) For $c_1 > 4$, $c_2 > 0$ and $n > c_1 s$, with probability at least $1 - 2\exp(-c_2 n) \to 1$,

$$\Lambda_{\max}((\Sigma^n_{S,S})^{-1}) \le \frac{4}{C_{\min}}. \tag{A.29}$$

(ii) Let $c_n = (\sqrt{s} \vee \sqrt{\log p})/\sqrt{n}$ and $C_n = 4\sqrt{s}c_n/(1-2c_n)^2 = O((s \vee \sqrt{s\log p})/\sqrt{n})$. With probability at least $1 - 2\exp(-\log p/2)$,

$$\|(\Sigma^n_{S,S})^{-1}\|_\infty \le K_1 (1 + C_n). \tag{A.30}$$

Proof of Lemma A.2.3. Let $\tilde{X} = X\Sigma^{-1/2}$ and $\tilde{\Sigma}^n = \tilde{X}^T\tilde{X}/n$. Then $\tilde{\Sigma}^n = I_{p\times p}$.

By Corollary 5.35 of Vershynin (2010), with probability at least $1 - 2\exp(-x^2/2)$,

$$\left(1 - \sqrt{\frac{s}{n}} - \frac{x}{\sqrt{n}}\right)^2 \le \Lambda_{\min}(\tilde{\Sigma}^n_{S,S}) \le \Lambda_{\max}(\tilde{\Sigma}^n_{S,S}) \le \left(1 + \sqrt{\frac{s}{n}} + \frac{x}{\sqrt{n}}\right)^2. \tag{A.31}$$

For $x = \sqrt{n}/2 - \sqrt{s}$ and $n \gg 4s$, with probability at least $1 - 2\exp(-(\sqrt{n}/2 - \sqrt{s})^2/2) \to 1$,

$$\Lambda_{\min}(\tilde{\Sigma}^n_{S,S}) \ge 1/4.$$

And hence,

$$\begin{aligned}
&\Lambda_{\max}((\tilde{\Sigma}^n_{S,S})^{-1}) = \Lambda^{-1}_{\min}(\tilde{\Sigma}^n_{S,S}) \leq 4 \\
&\Lambda_{\max}((\Sigma^n_{S,S})^{-1}) = \|(\Sigma^n_{S,S})^{-1}\|_2 \leq \|\Sigma^{-1/2}_{S,S}\|_2 \|(\tilde{\Sigma}^n_{S,S})^{-1}\|_2 \|\Sigma^{-1/2}_{S,S}\|_2 \leq \frac{4}{C_{\min}}.
\end{aligned}$$

Moreover,

$$\begin{aligned}
\|(\tilde{\Sigma}^n_{S,S})^{-1}\|_\infty &\leq 1 + \|(\tilde{\Sigma}^n_{S,S})^{-1} - I\|_\infty \\
&\leq 1 + \sqrt{s}\|(\tilde{\Sigma}^n_{S,S})^{-1} - I\|_2 \\
&\leq 1 + \sqrt{s}\Lambda_{\max}((\tilde{\Sigma}^n_{S,S})^{-1}) - I) \\
&\leq 1 + \sqrt{s}(\Lambda^{-1}_{\min}(\tilde{\Sigma}^n_{S,S}) - 1).
\end{aligned}$$

Taking $x = \sqrt{s} \vee \sqrt{\log p}$ in (A.31), we have with probability $1 - \exp(-\log p/2)$,

$$\begin{aligned}
\Lambda^{-1}_{\min}(\tilde{\Sigma}^n_{S,S}) - 1 &\leq \left(1 - 2\frac{\sqrt{s} \vee \sqrt{\log p}}{\sqrt{n}}\right)^{-2} - 1 \\
&= \frac{4(\sqrt{s} \vee \sqrt{\log p})/\sqrt{n} - 4(\sqrt{s} \vee \sqrt{\log p})^2/n}{\left(1 - 2\frac{\sqrt{s}\vee\sqrt{\log p}}{\sqrt{n}}\right)^2} \\
&\leq C_n/\sqrt{s}.
\end{aligned}$$

Putting these arguments together, we have

$$\begin{aligned}
\left\|(\Sigma^n_{S,S})^{-1}\right\|_\infty &\leq \left\|\Sigma^{-1/2}_{S,S}(\tilde{\Sigma}^n_{S,S})^{-1}\Sigma^{-1/2}_{S,S}\right\|_\infty \\
&\leq \|\Sigma^{-1/2}_{S,S}\|^2_\infty \left\|(\tilde{\Sigma}^n_{S,S})^{-1}\right\|_\infty \\
&= K_1(1 + C_n).
\end{aligned}$$

□

Lemma A.2.4 (Consistency of variance estimator)**.** *Assume that* $n \gg s\log p$ *and* $\lambda \asymp \sqrt{\log p/n}$. *If either (i) Conditions 3.3.1 - 3.3.5 hold true, or (ii) Conditions 3.4.1 - 3.4.5 hold true, then we have*

$$\hat{\sigma}^2 = \sigma^2 + o_P(1), \tag{A.32}$$

for $\hat{\sigma}^2$ *defined in (3.7).*

Proof of Lemma A.2.4. (i) For deterministic designs with ϵ satisfying Condition 3.3.4, we have

$$\|X_S^T\epsilon/n\|_\infty = O_P(\lambda),$$

By (A.23) in Lemma A.2.1, In the event that (3.17) holds, we have

$$\begin{aligned}\|\hat{u}_S\|_2^2 &\le 2\|(\Sigma^n_{S,S})^{-1}W_S^n\|_2^2 + 2\lambda^2\|(\Sigma^n_{S,S})^{-1}sgn(\hat{\beta}_S)\|_2^2\\ &\le 2s\|(\Sigma^n_{S,S})^{-1}W_S^n\|_\infty^2 + \frac{2\lambda^2}{C^2_{\min}}\|sgn(\hat{\beta}_S)\|_2^2\\ &= O_P\left(\frac{s\log p}{n}\right).\end{aligned}$$

Therefore, $\|\hat{u}_S\|_1 \le \sqrt{s}\|\hat{u}_S\|_2 = O_P(s\sqrt{\log p/n})$ and $|\epsilon^T X_S\hat{u}_S| \le n\|W_S^n\|_\infty\|\hat{u}_S\|_1 = O_P(s\log p)$, by a similar proof as for (A.23). Moreover, by the KKT condition of the Lasso (3.1),

$$\hat{u}_S^T\Sigma^n_{S,S}\hat{u}_S = \hat{u}_S^T\left(W_S^n - \lambda sgn(\hat{\beta}_S)\right) \le \|\hat{u}_S\|_1\|W_S^n - \lambda sgn(\hat{\beta}_S)\|_\infty = O_P\left(\frac{s\log p}{n}\right). \tag{A.33}$$

Note that $|\hat{S}| \le |S| \ll n$. Hence,

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n-|\hat{S}|}\|y - X\hat{\beta}\|_2^2\\ &= \frac{1}{n-|\hat{S}|}\left(\|\epsilon\|_2^2 + \|X\hat{u}\|_2^2 - 2\epsilon^T X\hat{u}\right) \qquad (A.34)\\ &= \sigma^2 + O_P\left(\frac{1}{n} + \frac{s\log p}{n}\right) + o_P(1)\\ &= \sigma^2 + o_P(1).\end{aligned}$$

(ii) For the Gaussian designs with Gaussian errors (Condition 3.4.4), we have

$$\mathbb{P}\left(\|X_S^T\epsilon\|_\infty > x|X_S\right) \le s\exp\left(-\frac{x^2}{2n\sigma^2}\right).$$

In the event that (3.29) holds,

$$\begin{aligned}\|\hat{u}_S\|_2^2 &\le 2\|(\Sigma^n_{S,S})^{-1}W_S^n\|_2^2 + 2\lambda^2\|(\Sigma^n_{S,S})^{-1}sgn(\hat{\beta}_S)\|_2^2\\ &= O_P\left(\frac{8}{C_{\min}}\left(s\|W_S^n\|_\infty^2 + s\lambda^2\right)\right)\\ &= O_P\left(\frac{s\log p}{nC_{\min}}\right).\end{aligned}$$

Hence, $\|\hat{u}_S\|_1 = O_P(s\lambda)$. By (A.33), (A.34) and (A.29) in Lemma A.2.3,

$$\hat{\sigma}^2 = \sigma^2 + O_P\left(\frac{1}{n} + \frac{s\log p}{n}\right) = \sigma^2 + o_P(1).$$

□

A.2.2 Proof of lemmas and theorems in Section 3

Now we are ready to prove the lemmas in the paper.

Proof of Lemma 3.3.1. Firstly, we use Lemma 1 - Lemma 3 in Wainwright (2009) to prove that (3.17) holds with large probablity. Consider a restricted Lasso problem

$$\check{\beta}_S = \arg\min_{b\in\mathbb{R}^s} \frac{1}{2n}\|y - X_S b_S\|_2^2 + \lambda\|b_S\|_1 \text{ and } \check{\beta}_{S^c} = 0. \tag{A.35}$$

Define T_j as

$$T_j = x_j^T\left(X_S(X_S^TX_S)^{-1}sgn(\check{\beta}_S) + \frac{P_S^{\perp}\epsilon}{n\lambda}\right). \tag{A.36}$$

By Lemma 1 of Wainwright (2009), if $\Sigma^n_{S,S}$ is invertible and $|T_j| < 1$ for $\forall j \in S^c$, then the $\check{\beta}$ is the unique solution to the Lasso with $\hat{S} \subseteq S$. Note that

$$\begin{aligned}\max_{j\in S^c}|T_j| &\le \|X_{S^c}^TX_S(X_S^TX_S)^{-1}sgn(\check{\beta}_S)\|_\infty + \max_{j\in S^c}\left|\frac{x_j^TP_S^{\perp}\epsilon}{n\lambda}\right|\\ &\le \left\|X_{S^c}^TX_S(X_S^TX_S)^{-1}\right\|_\infty + \underbrace{\max_{j\in S^c}\left|\frac{x_j^TP_S^{\perp}\epsilon}{n\lambda}\right|}_{Q_1}.\end{aligned}$$

We use standard symmetrization techniques to prove that $Q_1 \le (1-\kappa)/2$ with large probability (see Lemma A.2.1 for detailed results). By Condition 3.3.2 and (A.22) in Lemma A.2.1, there exists some $c_1, c_2 > 0$ such that

$$\mathbb{P}\left(\max_{j\in S^c}|T_j| > \frac{\kappa+1}{2}\right) \le \mathbb{P}\left(Q_1 > \frac{1-\kappa}{2}\right) \le 4\exp\left(-c_1\log p\right) + \frac{c_2}{n},$$

for λ in (3.16). Together with Condition 3.3.1, we have $\hat{S} \subseteq S$ with probability greater than $4\exp\left(-c_1\log p\right) + c_2/n$. By the KKT condition of the Lasso (3.1), $\hat{S} \subseteq S$ implies that

$$\hat{u}_S = (\Sigma^n_{S,S})^{-1}W^n_S - \lambda(\Sigma^n_{S,S})^{-1}sgn(\hat{\beta}_S). \tag{A.37}$$

Then we have

$$\begin{aligned}\|\hat{u}_S\|_\infty &= \|(\Sigma^n_{S,S})^{-1}W^n_S - \lambda(\Sigma^n_{S,S})^{-1}sgn(\hat{\beta}_S)\|_\infty \\ &\le \underbrace{\|(\Sigma^n_{S,S})^{-1}W^n_S\|_\infty}_{Q_2} + \lambda\|(\Sigma^n_{S,S})^{-1}\|_\infty .\end{aligned}$$

By (A.23) in Lemma A.2.1 and Condition 3.3.3, there exists some $c_3, c_4 > 0$ such that

$$\mathbb{P}\left(\|\hat{u}_S\|_\infty > K_1\lambda + 8\sigma\sqrt{\frac{2\log p}{C_{\min} n}}\right) \le 4\exp(-c_3\log p) + \frac{c_4}{n}.$$

□

Proof of Lemma 3.3.2. Consider the bootstrapped Lasso problem (A.26). By Condition 3.3.1 and Lemma A.2.2, for T^*_j defined in (A.28),

$$\begin{aligned}\mathbb{P}(\hat{S}^* \subseteq S) &\ge \mathbb{P}\left(\max_{j\in S^c}|T^*_j| < 1, \hat{S} \subseteq S\right) \\ &= \mathbb{P}\left(\max_{j\in S^c}|T^*_j| < 1\right) - \mathbb{P}(\hat{S} \not\subseteq S)\end{aligned} \tag{A.38}$$

Note that

$$\max_{j\in S^c}|T^*_j| \le \max_{j\in S^c}\left\|x_j^T X_S(X_S^T X_S)^{-1}sgn(\check{\beta}^*_S)\right\|_1 + \max_{j\in S^c}\underbrace{\left|\frac{x_j^T P_S^\perp \hat{\epsilon}^*}{n\lambda}\right|}_{Q^*_{1,j}}.$$

In view of (3.5), $Q^*_{1,j}$ is a Gaussian variable with mean zero and variance no larger than $\hat{\sigma}^2/(n\lambda^2)$, $\forall j \in S^c$, conditioning on $\hat{\sigma}^2$. Thus,

$$\begin{aligned}\mathbb{P}\left(\max_{j\in S^c}|Q^*_{1,j}| \ge \frac{1-\kappa}{2}\right) &\le \mathbb{P}\left(\max_{j\in S^c}|Q^*_{1,j}| \ge \frac{1-\kappa}{2}|\hat{\sigma}^2 \le 2\sigma^2\right) + \mathbb{P}(\hat{\sigma}^2 \ge 2\sigma^2) \\ &\le 2(p-s)\exp\left\{-\frac{n\lambda^2(1-\kappa)^2}{16\sigma^2}\right\} + o(1),\end{aligned}$$

where the last step is due to the consistency of $\hat{\sigma}^2$ in (3.7) (see Lemma A.2.4). Condition 3.3.2 implies that $\max_{j\in S^c}\|x_j^T X_S(X_S^T X_S)^{-1}sgn(\check{\beta}^*_S)\|_1 \le \kappa$. For $\lambda > \frac{4\sigma}{1-\kappa}\sqrt{\frac{2\log p}{n}}$ and some $c_1 > 0$, we have

$$\mathbb{P}\left(|T^*_j| \le \frac{1+\kappa}{2} < 1\right) \ge 1 - 2\exp(-c_1\log p) - o(1).$$

By Lemma 3.3.1 and (A.38), $\mathbb{P}(\hat{S} \subseteq S) \to 1$ and hence

$$P(\hat{S}^* \subseteq S) \ge 1 - 2\exp(-c_1\log p) - o(1).$$

By the KKT condition of the bootstrapped Lasso (A.26), in the event that $\{\hat{S}^* \subseteq S\}$,

$$\hat{u}^*_S = (\Sigma^n_{S,S})^{-1} W^*_S - \lambda (\Sigma^n_{S,S})^{-1} sgn(\hat{\beta}^*_S). \tag{A.39}$$

Therefore,

$$\begin{aligned} \|\hat{u}^*_S\|_\infty &\leq \|(\Sigma^n_{S,S})^{-1} W^*_S\|_\infty + \|\lambda (\Sigma^n_{S,S})^{-1} sgn(\hat{\beta}^*_S)\|_\infty \\ &\leq \|(\Sigma^n_{S,S})^{-1} W^*_S\|_\infty + \lambda K_1. \end{aligned} \tag{A.40}$$

Again using the Gaussian property of $\hat{\epsilon}^*$, there exists some $c_2 > 0$ such that

$$\begin{aligned} &\mathbb{P}\left(\|(\Sigma^n_{S,S})^{-1} W^*_S\|_\infty \geq \frac{2\sigma}{\sqrt{C_{\min}}} \sqrt{\frac{\log p}{n}} \right) \\ &\leq \mathbb{P}\left(\|(\Sigma^n_{S,S})^{-1} W^*_S\|_\infty \geq \frac{2\sigma}{\sqrt{C_{\min}}} \sqrt{\frac{\log p}{n}} \Big| \hat{\sigma}^2 \leq 2\sigma^2, \hat{S}^* \subseteq S \right) + \mathbb{P}(\hat{\sigma}^2 > 2\sigma^2) + \mathbb{P}(\hat{S}^* \nsubseteq S) \\ &\leq 2\exp(-c_2 \log p) + o(1). \end{aligned}$$

Together with (A.40), the proof is completed.

□

Proof of Lemma 3.3.3. In the event that (3.17) holds true, (A.37) holds true and hence we can rewrite *Remainder* in (3.13) as

$$Remainder = \lambda \left(e_j^T - \frac{z_j^T X}{z_j^T x_j} \right)_S (\Sigma^n_{S,S})^{-1} [sgn(\hat{\beta}_S) - sgn(\beta_S)]. \tag{A.41}$$

Let $\tilde{S}$ be the set of small coefficients, such that $\tilde{S} = \{j : 0 < |\beta_j| < g_1(\lambda) + g_1'(\lambda)\}$ for $g_1(\lambda)$ and $g_1'(\lambda)$ in (3.17) and (3.19) respectively. $\|\hat{u}_S\|_\infty \leq g_1(\lambda)$ further implies that for $\forall j \in S \backslash \tilde{S}$,

$$\begin{aligned} &\hat{\beta}_j = \beta_j + \hat{u}_j > \beta_j - \max_j |\hat{u}_j| > \beta_j - g_1(\lambda) > g_1'(\lambda), \text{ for } \beta_j > g_1(\lambda) + g_1'(\lambda). \\ &\hat{\beta}_j = \beta_j + \hat{u}_j < \beta_j + \max_j |\hat{u}_j| < \beta_j + g_1(\lambda) < -g_1'(\lambda), \text{ for } \beta_j < -[g_1(\lambda) + g_1'(\lambda)]. \end{aligned}$$

Therefore, if (3.17) holds true, then

$$sgn(\beta_j) = sgn(\hat{\beta}_j) \text{ and } |\hat{\beta}_j| > g'(\lambda) \text{ for } j \in S \backslash \tilde{S}. \tag{A.42}$$

The sign inconsistency of the Lasso estimator $\hat{\beta}$ only occurs on $\tilde{S}$ and hence

$$\left\| sgn(\hat{\beta}_S) - sgn(\beta_S) \right\|_1 \leq 2\tilde{s}. \tag{A.43}$$

By (A.41), we have

$$\begin{aligned}|Remainder| &\le \lambda \left\| \left(e_j^T - \frac{z_j^T X_S}{z_j^T x_j} \right)_S (\Sigma^n_{S,S})^{-1} \right\|_\infty \left\| sgn(\hat\beta_S) - sgn(\beta_S) \right\|_1 \\ &\le \lambda \left\| \frac{z_j^T X_{S_{\setminus j}}}{z_j^T x_j} \right\|_\infty \|(\Sigma^n_{S,S})^{-1}\|_\infty \left\| sgn(\hat\beta_S) - sgn(\beta_S) \right\|_1 \\ &\le \frac{2K_1 \lambda \lambda_j \tilde s}{z_j^T x_j / n},\end{aligned}$$

where the last step is due to Condition 3.3.3, (A.58) and (A.43). The proof for (3.22) is completed by the fact that (3.17) holds with probability going to 1.

For the bootstrap version, define an oracle Lasso estimator computed with the bootstrap samples. Formally,

$$\hat\beta_S^{(*,o)} = \hat\beta_S + \left(\Sigma^n_{S,S}\right)^{-1} \left[W_S^* - \lambda sgn(\beta_S)\right] \quad \text{and} \quad \hat\beta_{S^c}^{(*,o)} = 0. \tag{A.44}$$

If $\hat S \subseteq S$ and $\hat S^* \subseteq S$, we can plug in $\hat\beta_S^{(*,o)}$ and obtain that

$$\begin{aligned}\hat\beta_j^{(*,DB)} - \hat\beta_j =& \frac{z_j^T \hat\epsilon^*}{z_j^T x_j} + \left(e_j^T - \frac{z_j^T X}{z_j^T x_j} \right)_S (\hat\beta_S^{(*,o)} - \hat\beta_S) + \left(e_j^T - \frac{z_j^T X}{z_j^T x_j} \right)_S \left(\hat\beta_S^* - \hat\beta_S^{(*,o)} \right) \\ =& \underbrace{\frac{z_j^T \hat\epsilon^*}{z_j^T x_j} - \left(e_j^T - \frac{z_j^T X}{z_j^T x_j} \right)_S (\Sigma^n_{S,S})^{-1} W_S^*}_{Noise^*} + Bias \\ & \underbrace{+ \lambda \left(e_j^T - \frac{z_j^T X}{z_j^T x_j} \right)_S (\Sigma^n_{S,S})^{-1} \left(sgn(\hat\beta_S^*) - sgn(\beta_S) \right)}_{Remainder^*}.\end{aligned}$$

In view of (3.19) and (A.42) , we have $sgn(\hat\beta_j^*) = sgn(\hat\beta_j) = sgn(\beta_j)$ for $j \in S \backslash \tilde S$. Hence,

$$\|sgn(\hat\beta_S^*) - sgn(\beta_S)\|_1 \le 2\tilde s. \tag{A.45}$$

Together with (A.58) and Condition 3.3.3, it holds that

$$|Remainder^*| \le \frac{2K_1 \tilde s \lambda \lambda_j}{z_j^T x_j / n} = o_P(1),$$

in the event that $\{\hat S \subseteq S,\ \hat S^* \subseteq S\}$, which holds with large probability by Lemma 3.3.1 and 3.3.2. $\square$

Proof of Lemma 3.4.1. (i) Let

$$t_j^o = x_j - X_S \Sigma_{S,S}^{-1} \Sigma_{S,j}, \text{ for } j \in S^c. \tag{A.46}$$

For $\check{\beta}_S$ and T_j defined in (A.35) and (A.36) respectively, we can rewrite T_j as

$$T_j = \underbrace{(t_j^o)^T \left(X_S (X_S^T X_S)^{-1} sgn(\check{\beta}_S) + \frac{P_S^{\perp} \epsilon}{n\lambda} \right)}_{E_{1,j}} + \Sigma_{j,S} \Sigma_{S,S}^{-1}.$$

Conditioning on X_S and ϵ, t_j^o is a Gaussian random variable with mean 0 and variance at most $\Sigma_{j,j}$. Thus,

$$\begin{aligned} \mathrm{Var}(E_{1,j} | X_S, \epsilon) &\le \Sigma_{j,j} \left\| (X_S (X_S^T X_S)^{-1} sgn(\check{\beta}_S) + \frac{P_S^{\perp} \epsilon}{n\lambda} \right\|_2^2 \\ &\le \Sigma_{j,j} \left[(sgn(\check{\beta}_S))^T (X_S^T X_S)^{-1} sgn(\check{\beta}_S) + \|\epsilon\|_2^2 / (n\lambda)^2 \right]. \end{aligned} \tag{A.47}$$

Define an event

$$\begin{aligned} \mathcal{B}_1 = &\Big\{ \|\epsilon\|_2^2 / n \le 2\sigma^2, \ \Lambda_{\max}((\Sigma_{S,S}^n)^{-1}) \le \frac{4}{C_{\min}}, \|(\Sigma_{S,S}^n)^{-1}\|_\infty \le (1 + C_n) K_1 \text{ for} \\ &C_n \text{ in Lemma A.2.3 (ii)} \Big\}. \end{aligned}$$

$\mathcal{B}_1$ implies that

$$sgn(\check{\beta}_S)^T (X_S^T X_S)^{-1} sgn(\check{\beta}_S) \le \|sgn(\check{\beta}_S)\|_2^2 \|(X_S^T X_S)^{-1}\|_2 \le \frac{s}{n} \Lambda_{\max}((\Sigma_{S,S}^n)^{-1}) \le \frac{4s}{nC_{\min}}.$$

Thus, by (A.47) and Condition 3.4.4, in $\mathcal{B}_1$,

$$\max_{j \in S^c} \mathrm{Var}(E_{1,j}) \le C^* \left(\frac{4s}{nC_{\min}} + \frac{2\sigma^2}{n\lambda^2} \right). \tag{A.48}$$

Thus, by Lemma A.2.3 and Condition 3.4.5,

$$\begin{aligned} \mathbb{P}\left(\max_{j \in S^c} |E_{1,j}| > x \right) &\le \mathbb{P}\left(\max_{j \in S^c} |E_{1,j}| > x, \mathcal{B}_1 \right) + \mathbb{P}(\mathcal{B}_1^c) \\ &\le 2(p - s) \exp\left\{ - \frac{x^2}{2C^* (\frac{4s}{nC_{\min}} + \frac{2\sigma^2}{n\lambda^2})} \right\} + \frac{c_1}{n} \\ &\quad + 2\exp(-c_2 \log p) + 2\exp(-c_3 n), \end{aligned}$$

for some constant $c_1, c_2, c_3 > 0$. Let $x = (1 - \kappa)/2$ and solve

$$\frac{x^2}{2C^* (\frac{4s}{nC_{\min}} + \frac{2\sigma^2}{n\lambda^2})} \ge 2 \log p.$$

For

$$s\lambda^2 \leq \frac{\sigma^2 C_{\min}}{2} \quad \text{and } \lambda \geq \frac{8\sigma}{1-\kappa}\sqrt{\frac{C^* \log p}{n}},$$

there exists some $c_1 > 0$, such that

$$\mathbb{P}\left(\max_{j\in S^c} |T_j| > \frac{1+\kappa}{2}\right) \leq \frac{c_1}{n} + 2\exp(-c_4 \log p) + 2\exp(-c_3 n),$$

for some constant $c_1, c_3, c_4 > 0$.

(ii) The second task is to bound $\|\hat{u}_S\|_\infty$. In the event that $\{\hat{S} \subseteq S\}$,

$$\begin{aligned}\|\hat{u}_S\|_\infty &\leq \|(\Sigma^n_{S,S})^{-1} W^n_S\|_\infty + \lambda\|(\Sigma^n_{S,S})^{-1} sgn(\hat{\beta}_S)\|_\infty \\ &\leq \underbrace{\|(\Sigma^n_{S,S})^{-1} W^n_S\|_\infty}_{E_2} + \underbrace{\lambda\|(\Sigma^n_{S,S})^{-1}\|_\infty}_{E_3}.\end{aligned}$$

For E_2, conditioning on X, $(\Sigma^n_{S,S})^{-1} W^n_S$ is a Gaussian random vector with mean 0 and variance $(\sigma^2/n)(\Sigma^n_{S,S})^{-1}$. And hence,

$$\mathbb{P}\left(E_2 > x\right) \leq \mathbb{P}\left(E_2 > x, \mathcal{B}_1\right) + \mathbb{P}\left(\mathcal{B}_1^c\right) \leq 2s\exp\left(-\frac{nC_{\min}x^2}{8\sigma^2}\right) + \mathbb{P}\left(\mathcal{B}_1^c\right).$$

Lemma A.2.3 implies that

$$\mathbb{P}\left(E_3 > \lambda(1+C_n)K_1\right) \leq 2\exp(-c_5 \log p),$$

for some $c_5 > 0$. Using part (i) of the proof, we can obtain that for some $c_6, c_7, c_8 > 0$,

$$\begin{aligned}&\mathbb{P}\left(\|\hat{u}_S\|_\infty \leq 4\sigma\sqrt{\frac{\log p}{C_{\min} n}} + \lambda(1+C_n)K_1\right) \\ &\geq 1 - \frac{c_6}{n} - 2\exp(-c_7 n) - 2\exp(-c_8 \log p).\end{aligned}$$

□

Proof of Lemma 3.4.2. (i) Define an event

$$\mathcal{B}_2 = \left\{\max_{j\in S^c} |T_j| < 1 \text{ for } T_j \text{ defined in (A.36) and } \Lambda_{\max}((\Sigma^n_{S,S})^{-1}) \leq \frac{4}{C_{\min}}\right\}.$$

Since $\mathcal{B}_2$ implies $\{\hat{S} \subseteq S\}$, for T^*_j defined in (A.28), Lemma A.2.2 implies that

$$\mathbb{P}(\hat{S}^* \subseteq S) \geq \mathbb{P}\left(\max_{j\in S^c} |T^*_j| < 1, \mathcal{B}_2\right).$$

For t_j^o defined in (A.46), we have

$$\max_{j\in S^c}|T_j^*| \le \max_{j\in S^c}\left|x_j^T\left(X_S(X_S^TX_S)^{-1}sgn(\check{\beta}_S^*)+\frac{P_S^{\perp}\hat{\epsilon}^*}{n\lambda}\right)\right|$$
$$\le \max_{j\in S^c}\underbrace{\left|(t_j^o)^T\left(X_S(X_S^TX_S)^{-1}sgn(\check{\beta}_S^*)+\frac{P_S^{\perp}\hat{\epsilon}^*}{n\lambda}\right)\right|}_{E_{1,j}^*}+\left\|\Sigma_{S^c,S}\Sigma_{S,S}^{-1}\right\|_\infty. \tag{A.49}$$

Recall that under the bootstrap resampling plan (3.5) and (3.6), $y_i^* \sim N(x_{i,\hat{S}}\hat{\beta}_{\hat{S}}, \hat{\sigma}^2)$ conditioning on $(X, \hat{\beta}, \hat{S}, \hat{\sigma}^2)$.

For $E_{1,j}^*$, we first show that $X_S(X_S^TX_S)^{-1}sgn(\check{\beta}_S^*)$ is independent of t_j^o in (A.46) $\forall j \in S^c$, in the event of $\mathcal{B}_2$. Note that by Lemma 1 in Wainwright (2009), $\mathcal{B}_2$ implies that $\check{\beta}$ in (A.35) is the unique solution to the Lasso (3.1). As a result, $\hat{\beta}$ is a function of (X_S, ϵ) and $\hat{S} \subseteq S$. $\hat{S} \subseteq S$ further implies that $\check{\beta}_S^*$ in (A.27) is a function of $(X_S, \hat{\beta}, \hat{\epsilon}^*)$. Therefore, the following arguments hold true:

$$\mathcal{B}_2 \subseteq \{\hat{\beta} \text{ is a function of } (X_S,\epsilon),\ \check{\beta}_S^* \text{ in (A.27) is a function of } (X_S,\hat{\beta},\hat{\epsilon}^*)\}$$
$$\subseteq \{\hat{\beta} \text{ is a function of } (X_S,\epsilon),\ \hat{\sigma}^2 \text{ in (3.7) is a function of } (X_S,\epsilon),$$
$$\check{\beta}_S^* \text{ in (A.27) is a function of } (X_S,\epsilon,\hat{\sigma},\xi)\}$$
$$\subseteq \{\hat{\beta} \text{ is a function of } (X_S,\epsilon),\ \hat{\sigma}^2 \text{ in (3.7) is a function of } (X_S,\epsilon),$$
$$\check{\beta}_S^* \text{ in (A.27) is a function of } (X_S,\epsilon,\xi)\}$$
$$\subseteq \{X_S(X_S^TX_S)^{-1}sgn(\check{\beta}_S^*) \text{ is a function of } (X_S,\epsilon,\xi)\}. \tag{A.50}$$

Moreover, $\mathcal{B}_2 \cap \{\hat{\sigma}^2 \le 2\sigma^2, \|\xi\|_2^2 \le 2n\}$ implies that

$$\left\|X_S(X_S^TX_S)^{-1}sgn(\check{\beta}_S^*)+\frac{P_S^{\perp}\hat{\epsilon}^*}{n\lambda}\right\|_2^2 \le \frac{4s}{C_{\min}n}+\frac{\hat{\sigma}^2\|\xi\|_2^2}{n^2\lambda^2} \le \frac{4s}{C_{\min}n}+\frac{4\sigma^2}{n\lambda^2}. \tag{A.51}$$

Thus,

$$\mathbb{P}\left(\max_{j\in S^c}E_{1,j}^* \ge \frac{1-\kappa}{2},\ \mathcal{B}_2, \hat{\sigma}^2 \le 2\sigma^2, \|\xi\|_2^2 \le 2n\right)$$
$$\le \mathbb{P}\left(\max_{j\in S^c}E_{1,j}^* \ge \frac{1-\kappa}{2},\ \text{(A.50) and (A.51) hold true}\right)$$
$$\le 2(p-s)\exp\left\{-\frac{(1-\kappa)^2}{8C^*(\frac{s}{nC_{\min}}+\frac{\sigma^2}{n\lambda^2})}\right\}$$
$$\le 2\exp(-c_1\log p),$$

for $n \gg s \log p$ and $\lambda \geq \frac{4\sigma}{1-\kappa}\sqrt{\frac{\log p}{n}}$.

We conclude that for $n \gg s \log p$ and $\lambda > \frac{4\sigma}{1-\kappa}\sqrt{\frac{\log p}{n}}$,

$$\begin{aligned}
&\mathbb{P}\left(\max_{j\in S^c} |T_j^*| \leq \frac{1+\kappa}{2}, \mathcal{B}_2\right)\\
&\geq \mathbb{P}\left(\max_{j\in S^c} |E_{1,j}^*| < \frac{1-\kappa}{2}, \mathcal{B}_2\right)\\
&= \mathbb{P}(\mathcal{B}_2) - \mathbb{P}\left(\max_{j\in S^c} |E_{1,j}^*| \geq \frac{1-\kappa}{2}, \mathcal{B}_2\right)\\
&\geq \mathbb{P}(\mathcal{B}_2) - \mathbb{P}\left(\max_{j\in S^c} |E_{1,j}^*| \geq \frac{1-\kappa}{2}, \mathcal{B}_2, \hat{\sigma}^2 \leq 2\sigma^2, \|\xi\|_2^2 \leq 2n\right)\\
&\quad - \mathbb{P}\left(\hat{\sigma}^2 > 2\sigma^2\right) - \mathbb{P}\left(\|\xi\|_2^2 > 2n\right)\\
&= 1 - o(1).
\end{aligned}$$

(ii) Let

$$\begin{aligned}
\mathcal{B}_3 = \Big\{&\hat{S}^* \subseteq S,\ \Lambda_{\max}((\Sigma_{S,S}^n)^{-1}) \leq \frac{4}{C_{\min}},\ \|(\Sigma_{S,S}^n)^{-1}\|_\infty \leq K_1(1+C_n) \text{ for}\\
&C_n \text{ in Lemma A.2.3 (ii)},\ \hat{\sigma}^2 \leq 2\sigma^2\Big\}. \qquad \text{(A.52)}
\end{aligned}$$

In $\mathcal{B}_3$, (A.39) holds true and we have

$$\begin{aligned}
\|\hat{u}_S^*\|_\infty &\leq \|(\Sigma_{S,S}^n)^{-1} W_S^*\|_\infty + \lambda \|(\Sigma_{S,S}^n)^{-1} sgn(\hat{\beta}_S^*)\|_\infty\\
&\leq \max_{j\in S} \underbrace{\left|\hat{\sigma}(\Sigma_{S,S}^n)^{-1}_{j,S} X_S^T \xi/n\right|}_{E_{2,j}^*} + \lambda \|(\Sigma_{S,S}^n)^{-1}\|_\infty.
\end{aligned}$$

In $\mathcal{B}_3$, for $\forall j \in S$,

$$\sum_{i=1}^n ((\Sigma_{S,S}^n)^{-1}_{j,\cdot} X_{i,S}^T/n)^2 = \|(\Sigma_{S,S}^n)^{-1}_{j,\cdot} X_S^T/n\|_2^2 \leq \frac{1}{n}\Lambda_{\max}((\Sigma_{S,S}^n)^{-1}) \leq \frac{4}{nC_{\min}}.$$

By the Gaussian property of ξ, in $\mathcal{B}_3$,

$$\mathbb{P}\left(\max_{j\in S} E_{2,j}^* > x\right) \leq 2s \exp(-\frac{nC_{\min}x^2}{16\sigma^2}).$$

$\mathcal{B}_3$ is a large probability event due to part (i) of the proof, Lemma A.2.3 and Lemma A.2.4. Putting these pieces together, we have

$$\mathbb{P}\left(\|\hat{u}_S^*\|_\infty > 4\sigma\sqrt{\frac{\log p}{C_{\min} n}} + K_1(1+C_n)\lambda\right) \leq 2\exp(-c_2 \log p) + o(1) \to 0,$$

for some $c_2 > 0$ and λ satisfying (3.30).

□

Proof of Theorem 3.3.4. To simplify the notations, let $S_{\backslash j} = S\backslash\{j\}$. For the terms in (3.13) and (3.20), let

$$\begin{aligned} &b_j = \mathit{Bias} \text{ in } (3.13),\ \ \mathit{Rem}_j = \mathit{Remainder} \text{ in } (3.13),\ \ \eta_j = \mathit{Noise} \text{ in } (3.13), \\ &\mathit{Rem}_j^* = \mathit{Remainder}^* \text{ in } (3.20),\ \ \eta_j^* = \mathit{Noise}^* \text{ in } (3.20). \end{aligned} \tag{A.53}$$

Define a version of pivots in (3.24) which is standardized and bias-removed:

$$R_j^o = \frac{z_j^T x_j}{\sigma\|z_j\|_2}(\hat{\beta}_j^{(DB)} - \beta_j - b_j) \ \text{ and } \ R_j^{(*,o)} = \frac{z_j^T x_j}{\sigma\|z_j\|_2}(\hat{\beta}_j^{(*,DB)} - \hat{\beta}_j - b_j). \tag{A.54}$$

We first find the limiting distribution for R_j^o and $R_j^{(*,o)}$.

Let ζ_j be a normalized version of η_j in (A.53):

$$\zeta_j = \frac{z_j^T x_j}{\sigma\|z_j\|_2}\eta_j = \sum_{i=1}^n \zeta_{i,j}, \tag{A.55}$$

where $\zeta_{i,j} = \frac{1}{\sigma\|z_j\|_2}\left(z_{i,j}\epsilon_i - (z_j^T x_j e_j^T - z_j^T X_S)(\Sigma_{S,S}^n)^{-1} X_{i,S}^T \epsilon_i / n\right)$.

Statement (3.22) in Lemma 3.3.3 implies that

$$R_j^o = O_P\left(\frac{nK_1\tilde{s}\lambda\lambda_j}{\sigma\|z_j\|_2}\right) + \frac{z_j^T x_j}{\sigma\|z_j\|_2}\eta_j = O_P\left(\frac{\sqrt{n}K_1\tilde{s}\lambda\lambda_j}{\sigma K_2}\right) + \zeta_j = o_P(1) + \zeta_j, \tag{A.56}$$

where the second step is due to Condition 3.3.5 and the last step is by our sample size condition in $\mathcal{A}_1$ (3.26). Note that ζ_j is a random variable with mean zero and variance s_n^2, where

$$\begin{aligned} s_n^2 &= Var(\zeta_j) \\ &= \frac{1}{\|z_j\|_2^2}\|z_j - (z_j^T x_j e_j^T - z_j^T X)_S(\Sigma_{S,S}^n)^{-1} X_S^T / n\|_2^2 \\ &= 1 + \frac{1}{n\|z_j\|_2^2}(z_j^T x_j e_j^T - z_j^T X)_S(\Sigma_{S,S}^n)^{-1}(z_j^T x_j e_j - X^T z_j)_S \\ &\quad - \frac{2}{n\|z_j\|_2^2} z_j^T X_S(\Sigma_{S,S}^n)^{-1}(z_j^T x_j e_j - X^T z_j)_S \\ &= 1 + \underbrace{\frac{3}{n\|z_j\|_2^2}(z_j^T x_j e_j^T - z_j^T X)_S(\Sigma_{S,S}^n)^{-1}(z_j^T x_j e_j - X^T z_j)_S}_{H_1} \\ &\quad - \underbrace{\frac{2}{n\|z_j\|_2^2} z_j^T x_j e_j^T(\Sigma_{S,S}^n)^{-1}(z_j^T x_j e_j - X^T z_j)_S}_{H_2}. \end{aligned} \tag{A.57}$$

Note that the KKT condition of (3.4) is

$$\left\|\frac{1}{n} z_j^T X_{-j}\right\|_\infty \le \lambda_j. \tag{A.58}$$

Therefore, H_1 in (A.57) can be bounded by

$$\begin{aligned} |H_1| &\le \frac{3\left\|(z_j^T x_j e_j^T - z_j^T X)_S/n\right\|_2^2 \left\|(\Sigma_{S,S}^n)^{-1}\right\|_2}{\|z_j\|_2^2/n} \\ &\le \frac{3s\|z_j^T X_{S_{\setminus j}}/n\|_\infty^2}{K_2 C_{\min}} \le \frac{3s\lambda_j^2}{K_2 C_{\min}}, \end{aligned} \tag{A.59}$$

where the second last step is by Conditions 3.3.1 and 3.3.5 and the last step is by (A.58). Similarly, H_2 in (A.57) can be bounded by

$$|H_2| \le \frac{2}{\|z_j\|_2^2}\left\|z_j^T X_{S_{\setminus j}}/n\right\|_2 \left\|(\Sigma_{S,S}^n)^{-1}\right\|_2 |x_j^T z_j| \le \frac{\sqrt{s}\lambda_j}{\sqrt{K_2}C_{\min}}. \tag{A.60}$$

Thus, for $n \gg s\log p$, we have

$$s_n^2 = 1 + o(1). \tag{A.61}$$

Now we check the Lyapunov condition, say

$$\lim_{n\to\infty} \frac{1}{s_n^4}\sum_{i=1}^n \mathbb{E}[|\zeta_{i,j}|^4] = 0.$$

Using Condition 3.3.4, we can obtain that

$$\begin{aligned} \sum_{i=1}^n \mathbb{E}[\zeta_{i,j}|^4] &\le \frac{\mathbb{E}[|\epsilon|^4]}{\sigma^4\|z_j\|_2^4}\sum_{i=1}^n |z_{i,j} - (z_j^T x_j e_j^T - z_j^T X)_S(\Sigma_{S,S}^n)^{-1}X_{i,S}^T/n|^4 \\ &\le \frac{M_0}{\sigma^4\|z_j\|_2^4}2^3\left(\sum_{i=1}^n |z_{i,j}|^4 + \sum_{i=1}^n |(z_j^T x_j e_j^T - z_j^T X)_S(\Sigma_{S,S}^n)^{-1}X_{i,S}^T/n|^4\right). \end{aligned} \tag{A.62}$$

For ease of notation, let $c_i = (z_j^T x_j e_j^T - z_j^T X)_S(\Sigma_{S,S}^n)^{-1}X_{i,S}^T/n,\ i = 1,\ldots,n$. Then we have

$$\begin{aligned} &\|c\|_2^2 = \frac{\|z_j\|_2^2}{3}H_1 \le \frac{ns\lambda_j^2}{C_{\min}}, \\ &\max_{i\le n}|c_i| \le s\left\|z_j^T X_{S_{\setminus j}}/n\right\|_\infty \left\|(\Sigma_{S,S}^n)^{-1}\right\|_\infty \max_{i,j}|x_{i,j}| \le K_1 K_0 s\lambda_j. \end{aligned}$$

As a consequence,

$$\sum_{i=1}^n |c_i|^4 \le \max_{i\le n}|c_i|^2\sum_{i=1}^n |c_i|^2 \le (K_1K_0s\lambda_j)^2\frac{ns\lambda_j^2}{C_{\min}} = \frac{nK_1^2K_0^2s^3\lambda_j^4}{C_{\min}}.$$

In view of (A.62), it holds that

$$\lim_{n\to\infty}\frac{1}{s_n^4}\sum_{i=1}^n \mathbb{E}[\zeta_{i,j}|^4] \le \lim_{n\to\infty}\frac{M_0 2^3(nK_1^2K_0^2s^3\lambda_j^4/C_{\min}+\|z_j\|_4^4)}{\sigma^4\|z_j\|_2^4(1-o(1))^2}$$
$$\le \lim_{n\to\infty}\frac{M_0 2^3K_1^2K_0^2s^3\lambda_j^4}{\sigma^4K_2^2C_{\min}n}=0,$$

as long as $s^3\lambda^4 \ll n$. For $s \ll n/\log p$ and $\lambda_j \asymp \sqrt{\frac{\log p}{n}}$, it is easy to check that $s^3\lambda^4 = O(n/\log p) \ll n$.

We have proved that

$$\zeta_j/s_n \xrightarrow{D} Z, \text{ for } Z \sim N(0,1).$$

Together with (A.56) and (A.61), we have

$$\sup_{c\in\mathbb{R}}\left|\mathbb{P}(R_j^o \le c)-\Phi(c)\right| = o_P(1). \tag{A.63}$$

For the bootstrap version, consider $R_j^{(*,o)}$ defined in (A.54). By (3.23) in Lemma 3.3.3,

$$R_j^{(*,o)} = O_P\left(\frac{nK_1\tilde{s}\lambda\lambda_j}{\sigma\|z_j\|_2}\right)+\frac{z_j^Tx_j}{\sigma\|z_j\|_2}\eta_j+o_P(1)$$
$$= O_P\left(\frac{\sqrt{n}K_1\tilde{s}\lambda\lambda_j}{\sigma K_2}\right)+\zeta_j^*+o_P(1)=o_P(1)+\zeta_j^*,$$

where

$$\zeta_j^* = \frac{z_j^Tx_j}{\sigma\|z_j\|_2}\eta_j^* \tag{A.64}$$

is a Gaussian variable with mean zero and variance $1+o_P(1)$ by Lemma A.2.4. This implies that

$$\sup_{c\in\mathbb{R}}\left|\mathbb{P}(R_j^{(*,o)} \le c)-\Phi(c)\right| = o_P(1). \tag{A.65}$$

Let $F_*(c)$ be the cumulative distribution function of ζ_j^*, i.e. $F_*(c)=\mathbb{P}(\zeta_j^*\le c)$. For $\forall v_1, v_2 > 0$ and $\forall \alpha \in (0,1)$,

$$\mathbb{P}\left(q_\alpha(R_j^{(*,o)})-z_\alpha > v_1\right) \le \mathbb{P}\left\{F_*\left(q_\alpha\left(R_j^{(*,o)}\right)\right) > F_*(z_\alpha+v_1)\right\}$$
$$\le \mathbb{P}\left\{\alpha > F_*(z_\alpha+v_1)\right\}$$
$$\le \mathbb{P}\left\{\alpha+v_2 > \Phi(z_\alpha+v_1)\right\}+$$
$$\mathbb{P}\left\{F_*(z_\alpha+v_1) \le \Phi(z_\alpha+v_1)-v_2\right\}$$
$$= \mathbb{P}\left\{\alpha+v_2 > \Phi(z_\alpha+v_1)\right\}+o(1),$$

where the first step is due to the monotonicity of F_*, the second step is by the definition of quantile function, and the last step is due to (A.65). By first taking $v_2 \to 0$, we have proved that for $\forall \alpha \in (0,1)$ and $\forall v_1 > 0$,

$$\mathbb{P}\left\{q_\alpha\left(R_j^{(*,o)}\right) - z_\alpha > v_1\right\} = o(1).$$

A matching lower bound can be proved by a completely analogous argument. Thus,

$$\sup_{\alpha\in(0,1)} \left|q_\alpha\left(R_j^{(*,o)}\right) - z_\alpha\right| = o_P(1). \tag{A.66}$$

To complete our proof, note that

$$R_j = \sigma R_j^o + \frac{z_j^T x_j}{\|z_j\|_2} b_j + o_P(1) \text{ and } R_j^* = \sigma R_j^{(*,o)} + \frac{z_j^T x_j}{\|z_j\|_2} b_j + o_P(1). \tag{A.67}$$

Together with (A.63) and (A.66), it holds that

$$\begin{aligned}
\sup_{\alpha\in(0,1)} \left|\mathbb{P}\left\{R_j \le q_\alpha(R_j^*)\right\} - \alpha\right| &\le \sup_{\alpha\in(0,1)} \left|\mathbb{P}\left\{R_j^o \le q_\alpha\left(R_j^{(*,o)}\right)\right\} - \alpha\right| \\
&\le \sup_{\alpha\in(0,1)} \left|\mathbb{P}\{R_j^o \le z_\alpha\} - \alpha\right| + o_P(1) \\
&= o_P(1).
\end{aligned}$$

Next, we prove the asymptotic normality of $R_j^{(DDB)}$ in (3.25). Note that $\hat{\beta}_j^{(DDB)}$ in (3.10) corrects $\hat{\beta}_j^{(*,DB)}$ with an estimated bias

$$\hat{b}_j = median\left(\hat{\beta}_j^{(*,DB)} - \hat{\beta}_j\right). \tag{A.68}$$

Due to (3.20), we can easily obtain that

$$\begin{aligned}
\frac{z_j^T x_j}{\sigma\|z_j\|_2}\hat{b}_j &= \frac{z_j^T x_j}{\sigma\|z_j\|_2} b_j + median\left(R_j^{(*,o)}\right) \\
&= \frac{z_j^T x_j}{\sigma\|z_j\|_2} b_j + z_{0.5} + o_P(1) \\
&= \frac{z_j^T x_j}{\sigma\|z_j\|_2} b_j + o_P(1),
\end{aligned}$$

where the second step is due to (A.66). By definition of $\hat{\beta}_j^{(DDB)}$(3.10) and R_j defined in (3.24),

$$\frac{z_j^T x_j}{\sigma\|z_j\|_2}(\hat{\beta}_j^{(DDB)} - \beta_j) = \frac{1}{\sigma} R_j - \frac{z_j^T x_j}{\sigma\|z_j\|_2}\hat{b}_j = R_j^o + o_P(1) = Z + o_P(1), \tag{A.69}$$

where the last step is due to (A.63) for $Z \sim N(0,1)$. For $R_j^{(DDB)}$ defined in (3.25), (A.69) and Lemma A.2.4 implies that

$$\mathbb{P}\left(R_j^{(DDB)} \le c\right) = \mathbb{P}(Z \le c) + o_P(1) = \Phi(c) + o_P(1).$$

□

Proof of Theorem 3.4.3. Under Gaussian designs, we still consider error decompositions as in (3.13) and (3.20). We use simplified notations described in (A.53).

In the event that (3.29) holds, we can obtain that

$$\begin{aligned} |Rem_j| &\le \lambda \left\| \left(e_j^T - \frac{z_j^T X_S}{z_j^T x_j} \right)_S (\Sigma^n_{S,S})^{-1} [sgn(\hat{\beta}_S) - sgn(\beta_S)] \right\|_1 \\ &\le \lambda \left\| \frac{z_j^T X_{S \setminus j}}{z_j^T x_j} \right\|_\infty \left\| (\Sigma^n_{S,S})^{-1} \right\|_\infty \left\| sgn(\hat{\beta}_S) - sgn(\beta_S) \right\|_1 \\ &= O_P\left(\frac{K_1 (1 + C_n) n \tilde{s} \lambda \lambda_j}{z_j^T x_j} \right), \end{aligned}$$

where the last step is by (A.58), Lemma A.2.3, Lemma 3.4.1 and the definition of $\tilde{s}$ in (3.32).

By Lemma 5.3 of van de Geer et al. (2014), if $n \gg s_j \log p$,

$$\|z_j\|_2^2 / n = 1/(\Sigma^{-1})_{j,j} + o_P(1), \tag{A.70}$$

where $\max_{j \le p} (\Sigma^{-1})_{j,j} \le 1/C_*$ by Condition 3.4.4. Thus, for R_j^o defined in (A.54) and ζ_j in (A.55) we have

$$\begin{aligned} R_j^o &= O_P\left(\frac{n K_1 (1 + C_n) \tilde{s} \lambda \lambda_j}{\sigma \|z_j\|_2} \right) + \zeta_j + o_P(1) \\ &= O_P\left(\frac{\sqrt{n} K_1 (1 + C_n) \tilde{s} \lambda \lambda_j}{\sigma \sqrt{C_*}} \right) + \zeta_j + o_P(1) \\ &= o_P(1) + \zeta_j, \end{aligned} \tag{A.71}$$

where the last step is due to in $\mathcal{A}_2$ (3.33),

$$\begin{aligned} &\sqrt{n} \tilde{s} \lambda \lambda_j = \frac{\tilde{s} \log p}{n} = o(1), \;\; s \tilde{s} \lambda \lambda_j = \frac{s \tilde{s} \log p}{n} = o(1) \text{ and} \\ &\sqrt{s \log p} \tilde{s} \lambda \lambda_j = \sqrt{\frac{s \log p}{n} \frac{\tilde{s} \log p}{n}} = o(1). \end{aligned}$$

Next we show that for ζ_j in (A.71),

$$\sup_{c \in \mathbb{R}} |\mathbb{P}(\zeta_j \le c) - \Phi(c)| = o_P(1).$$

Conditioning on X, ζ_j is a Gaussian random variable with mean 0 and variance of the form (A.57). By (A.70) and Lemma A.2.3, we can similarly prove (A.59) and (A.60) by replacing K_2 with $C_*(1 - o_P(1))$ and replacing $C_{\min}$ with $C_{\min}/4 + o_P(1)$. Hence, $|s_n^2 - 1| = O_P(\sqrt{s}\lambda_j) = o_P(1)$. Then we have,

$$\mathbb{P}(\zeta_j \le c) = \mathbb{E}[\mathbb{P}(\zeta_j \le c|X)] = \mathbb{E}[\Phi(c/s_n)] = \Phi(c) + o_P(1).$$

For the bootstrap version $R_j^{(*,o)}$ in (A.54) and ζ_j^* in (A.64), Lemma 3.4.2 implies that

$$R_j^{(*,o)} = O_P\left(\frac{\sqrt{n}K_1(1+C_n)\tilde{s}\lambda\lambda_j}{\sigma\sqrt{C_*}}\right) + \zeta_j^* + o_P(1) = o_P(1) + \zeta_j^*. \tag{A.72}$$

Conditioning on X and ϵ, we can similarly prove ζ_j^* is a Gaussian random variable with mean 0 and variance $1 + o_P(1)$. Thus,

$$\sup_{c \in \mathbb{R}} \left|\mathbb{P}(R_j^{(*,o)} \le c) - \Phi(c)\right| = o_P(1).$$

Together with (A.67), (A.71), (A.72) and Lemma A.2.4,

$$\begin{aligned}
\sup_{\alpha \in (0,1)} \left|\mathbb{P}\left\{R_j \le q_\alpha(R_j^*)\right\} - \alpha\right| &\le \sup_{\alpha \in (0,1)} \left|\mathbb{P}\left\{R_j^o \le q_\alpha\left(R_j^{(*,o)}\right)\right\} - \alpha\right| \\
&\le \sup_{\alpha \in (0,1)} \left|\mathbb{P}\{R_j^o \le z_\alpha\} - \alpha\right| + o_P(1) \\
&= o_P(1).
\end{aligned}$$

The asymptotic normality of $R_j^{(DDB)}$ can be similarly proved as for the fixed design case and is omitted here. □

A.3 Proofs for Chapter 4

A.3.1 Proof of theorems and lemmas

Proof of Theorem 4.2.1. Let $H = (n\tau_n^2)^{-1} I_{J \times J}$, $\tilde{\Sigma}^n = \frac{1}{n}\begin{pmatrix} \hat{D}^T\hat{D} & \hat{D}^T Z \\ Z^T\hat{D} & Z^T Z \end{pmatrix}$ and $\tilde{\Sigma}_H^n = \frac{1}{n}\begin{pmatrix} \hat{D}^T\hat{D} & \hat{D}^T Z \\ Z^T\hat{D} & Z^T Z + nH \end{pmatrix}$. First we show the symmetric matrix $\tilde{\Sigma}_H^n$ is invertible for any

$\tau_n^2 > 0$.

$$\begin{aligned}\min_{u\in\mathbb{R}^{(J+1)},\|u\|_2=1} u^T\tilde{\Sigma}_H^n u &= \min\{\min_{u\in\mathbb{R}^{(J+1)},u_1=1} u^T\tilde{\Sigma}_H^n u, \min_{u\in\mathbb{R}^{(J+1)},u_1\neq 1} u^T\tilde{\Sigma}_H^n u\}\\ &= \min\{\|\hat{D}\|_2^2/n, \min_{u\in\mathbb{R}^{(J+1)},u_1\neq 1} u^T\tilde{\Sigma}_H^n u\}\\ &= \min\{\|\hat{D}\|_2^2/n, \min_{u\in\mathbb{R}^{(J+1)},u_1\neq 1} (u^T\tilde{\Sigma}^n u + u_{-1}^T H u_{-1})\}\\ &\geq \min\{\|\hat{D}\|_2^2/n, \min_{u\in\mathbb{R}^{(J+1)},u_1\neq 1} u_{-1}^T H u_{-1}\}\\ &> 0,\end{aligned}$$

where the first inequality is due to $\tilde{\Sigma}^n$ is a nonnegative definite matrix and the second inequality is by our assumption.

By Bayes rule, we can get

$$\begin{aligned}\begin{pmatrix}\hat{\beta}^{(SG)}\\ \hat{\alpha}^{(SG)}\end{pmatrix} &= (\tilde{\Sigma}_H^n)^{-1}\begin{pmatrix}\hat{D}^TY/n\\ \hat{Z}^TY/n + H\mu_\alpha\end{pmatrix}\\ &= (\tilde{\Sigma}_H^n)^{-1}\begin{pmatrix}\hat{D}^TY/n\\ Z^TY/n + H\alpha - H(\alpha-\mu_\alpha\mathbb{1}_J)\end{pmatrix}\\ &= \begin{pmatrix}\beta\\ \alpha\end{pmatrix} + (\tilde{\Sigma}_H^n)^{-1}\begin{pmatrix}\hat{D}^T\hat{\eta}/n\\ Z^T\hat{\eta}/n - H(\alpha-\mu_\alpha\mathbb{1}_J)\end{pmatrix}.\end{aligned}$$

Hence,

$$\begin{aligned}\hat{\beta}^{(SG)} &= \beta + \frac{1}{n}(\tilde{\Sigma}_H^n)^{-1}_{1,1}\hat{D}^T\hat{\eta} + \frac{1}{n}(\tilde{\Sigma}_H^n)^{-1}_{1,[J]}(Z^T\hat{\eta} + nH\,(\mu_\alpha\mathbb{1}_J - \alpha))\\ &= \beta + \underbrace{\frac{1}{n}(\tilde{\Sigma}_H^n)^{-1}_{1,1}\hat{D}^T\hat{\eta} + \frac{1}{n}(\tilde{\Sigma}_H^n)^{-1}_{1,[J]}Z^TP_{\hat{D}}\hat{\eta}}_{E_1} + \underbrace{\frac{1}{n}(\tilde{\Sigma}_H^n)^{-1}_{1,[J]}Z^TP_{\hat{D}}^{\perp}\hat{\eta}}_{E_2}\\ &- \underbrace{(\tilde{\Sigma}_H^n)^{-1}_{1,[J]}H(\alpha-\mu_\alpha\mathbb{1}_J)}_{E_3}. \qquad (A.73)\end{aligned}$$

By the matrix inverse formula and some simple algebra, we can get

$$(\Sigma_H^n)^{-1}_{[J],[J]} = \left(Z^TZ/n + H - \frac{Z^T\hat{D}\hat{D}^TZ}{\hat{D}^T\hat{D}}/n\right)^{-1} = \left(Z^TP_{\hat{D}}^{\perp}Z/n + H\right)^{-1} \qquad (A.74)$$

$$(\Sigma_H^n)^{-1}_{1,[J]} = -\frac{\hat{D}^TZ(\Sigma_H^n)^{-1}_{[J],[J]}}{\hat{D}^T\hat{D}} = -\frac{\hat{D}^TZ\left(Z^TP_{\hat{D}}^{\perp}Z/n + H\right)^{-1}}{\hat{D}^T\hat{D}}. \qquad (A.75)$$

Thus, for E_1 in (A.73), we have

$$\begin{aligned} E_1 &= \frac{1}{n}(\Sigma_H^n)_{1,1}^{-1}\hat{D}^T\hat{\eta} + \frac{(\Sigma_H^n)_{1,[J]}^{-1}Z^T\hat{D}\hat{D}^T\hat{\eta}}{n\hat{D}^T\hat{D}} \\ &= \left((\Sigma_H^n)_{1,1}^{-1} - (\Sigma_H^n)_{1,[J]}^{-1}(\Sigma_H^n)_{[J],[J]}(\Sigma_H^n)_{[J],1}^{-1}\right)\hat{D}^T\hat{\eta} \\ &= \frac{\hat{D}^T\hat{\eta}}{\hat{D}^T\hat{D}}, \end{aligned}$$

where the second equality can be seen from the first part of (A.75) and the last step is by the matrix inverse formula.

By (A.75), we have

$$\begin{aligned} |E_3| &= \frac{1}{n\tau_n^2\|\hat{D}\|_2^2}|\hat{D}^T Z(\tilde{\Sigma}_H^n)^{-1}(\alpha - \mu_\alpha \mathbb{1}_J)| \\ &\leq \frac{1}{n\tau_n^2\|\hat{D}\|_2^2}\|\hat{D}^T Z\|_2 \Lambda_{\max}((\tilde{\Sigma}_H^n)^{-1})\|\alpha - \mu_\alpha \mathbb{1}_J\|_2 \\ &\leq \frac{\|\hat{D}^T Z\|_2\|\alpha - \mu_\alpha \mathbb{1}_J\|_2}{\|\hat{D}\|_2^2} \\ &\leq \frac{\|\alpha - \mu_\alpha \mathbb{1}_J\|_2 \Lambda_{\max}^{1/2}(\Sigma_Z^n)}{\|\hat{D}\|_2/\sqrt{n}} \\ &\leq \frac{\|\alpha - \mu_\alpha \mathbb{1}_J\|_2 \Lambda_{\max}^{1/2}(\Sigma_Z^n)}{\|\hat{\gamma}\|_2 \Lambda_{\min}^{1/2}(\Sigma_Z^n)}, \end{aligned}$$

where the second inequality is due to

$$\Lambda_{\max}((\tilde{\Sigma}_H^n)^{-1}) \leq \Lambda_{\min}^{-1}((n\tau_n^2)^{-1}I_{J\times J}) = n\tau_n^2$$

and the last step is due to

$$\|\hat{D}\|_2/\sqrt{n} = \|Z\hat{\gamma}\|_2/\sqrt{n} \geq \|\hat{\gamma}\|_2 \Lambda_{\min}^{1/2}(\Sigma_Z^n).$$

Similarly,

$$|E_2| \leq \frac{\tau_n^2 \Lambda_{\max}^{1/2}(\Sigma_Z^n)\|Z^T P_{\hat{D}}^{\perp}\hat{\eta}\|_2}{\|\hat{D}\|_2/\sqrt{n}}.$$

□

Proof of Corollary 4.2.2. Note that

$$\|Z^T P_{\hat{D}}^{\perp}\hat{\eta}\|_2 \leq \sqrt{J}\max_j \|Z_j^T P_{\hat{D}}^{\perp}\hat{\eta}\|_2$$

and

$$\|Z_j^T P_{\hat{D}}^{\perp}\hat{\eta}\|_2 = \|Z_j^T P_{\hat{D}}^{\perp}(\epsilon + \beta P_Z^{\perp} v)\|_2 \le \|Z_j^T \epsilon\|_2 = O_P(\|Z_j\|_2 \sigma_\epsilon) = O_P(\sqrt{n}\Lambda_{\max}^{1/2}(\Sigma_Z^n)\sigma_\epsilon).$$

For the third term, it holds that

$$|\hat{D}^T\hat{\eta}| = |\hat{D}^T(\epsilon + \beta P_Z^{\perp} v)| = |\hat{D}^T \epsilon| = O_P(\|\hat{D}\|_2 \sigma_\epsilon).$$

$\square$

Proof of Theorem 4.2.3. Define a matrix $E^n \in \mathbb{R}^{J\times J}$, where $E^n_{j,\cdot} = \mathbb{1}_J^T/J$ for $j = 1, \ldots, J$. Let $H' = H(I_{J\times J} - E^n)$. First we show that $Z^T P_{\hat{D}}^{\perp} Z/n + H'$ is invertible.

$$\begin{aligned}
&\Lambda_{\min}(Z^T P_{\hat{D}}^{\perp} Z + nH') \\
&= \min_{\|u\|_2=1} u^T(Z^T P_{\hat{D}}^{\perp} Z + nH')u \\
&\ge \Lambda_{\min}(Z^T Z) + \min_{\|u\|_2=1}\{\tau_n^{-2} u^T(u - \bar{u}\mathbb{1}_J) - u^T Z^T P_{\hat{D}} Z u\} \\
&\ge \Lambda_{\min}(Z^T Z) + \min_{\|u\|_2=1}\{\tau_n^{-2}(u - \bar{u}\mathbb{1}_J)^T(u - \bar{u}\mathbb{1}_J) \\
&\quad - (u - \bar{u}\mathbb{1}_J)^T Z^T P_{\hat{D}} Z(u - \bar{u}\mathbb{1}_J) - \bar{u}^2 \mathbb{1}_J^T Z^T P_{\hat{D}} Z \mathbb{1}_J - 2\bar{u}\mathbb{1}_J^T Z^T P_{\hat{D}} Z(u - \bar{u}\mathbb{1}_J)\} \\
&\ge \Lambda_{\min}(Z^T Z) + \min_{\|u\|_2=1}\{[\tau_n^{-2} - \Lambda_{\max}(Z^T Z)](u - \bar{u}\mathbb{1}_J)^T(u - \bar{u}\mathbb{1}_J) \\
&- 2\bar{u}\mathbb{1}_J^T Z^T P_{\hat{D}} Z(u - \bar{u}\mathbb{1}_J) - \bar{u}^2\mathbb{1}_J^T Z^T P_{\hat{D}} Z\mathbb{1}_J\} \\
&\ge \Lambda_{\min}(Z^T Z) - \min_{\|u\|_2=1}\{2\bar{u}\mathbb{1}_J^T Z^T P_{\hat{D}} Z(u - \bar{u}\mathbb{1}_J) + \bar{u}^2\mathbb{1}_J^T Z^T P_{\hat{D}} Z\mathbb{1}_J\},
\end{aligned}$$

for $\tau_n^{-2} > n\Lambda_{\max}(\Sigma_Z^n)$. Moreover,

$$\|P_{\hat{D}} Z\mathbb{1}_J\|_2 = \frac{J\|\hat{D}\hat{D}^T\tilde{Z}\|_2}{\hat{D}^T\hat{D}} \le \|Z\mathbb{1}_J\|_2|\mathrm{Cor}(\hat{D}, \tilde{Z})|.$$

Since $\sqrt{nJ}\Lambda_{\max}^{1/2}(\Sigma_Z^n) \ge \|Z\mathbb{1}_J\|_2 \ge \sqrt{nJ}\Lambda_{\min}^{1/2}(\Sigma_Z^n)$ and $|\bar{u}| \le J^{-1/2}$, we can get

$$\begin{aligned}
\Lambda_{\min}(Z^T P_{\hat{D}}^{\perp} Z + nH') &\ge n\Lambda_{\min}(\Sigma_Z^n) - 2n|\mathrm{Cor}(\hat{D}, \tilde{Z})|\Lambda_{\max}(\Sigma_Z^n) - n\mathrm{Cor}^2(\hat{D}, \tilde{Z})\Lambda_{\max}(\Sigma_Z^n) \\
&\ge nr_n^*,
\end{aligned}$$

for $r_n^* = \Lambda_{\min}(\Sigma_Z^n) - 2|\mathrm{Cor}(\hat{D}, \tilde{Z})|\Lambda_{\max}(\Sigma_Z^n) - \mathrm{Cor}^2(\hat{D}, \tilde{Z})\Lambda_{\max}(\Sigma_Z^n)$.

To get $r_n^* > 0$, we need

$$0 < |\mathrm{Cor}(\hat{D}, \tilde{Z})| < \sqrt{\frac{\Lambda_{\min}(\Sigma_Z^n)}{\Lambda_{\max}(\Sigma_Z^n)} + 1} - 1.$$

Computing the MLE, we can get $\hat{\mu}_\alpha^{(SG*)} = \hat{\alpha}^{(SG*)}\mathbb{1}_J/J$.

$$\begin{aligned}
\begin{pmatrix} \hat{\beta}^{(SG*)} \\ \hat{\alpha}^{(SG*)} \end{pmatrix} &= (\tilde{\Sigma}_H^n - HE^n)^{-1} \begin{pmatrix} \hat{D}^T Y/n \\ \hat{Z}^T Y/n \end{pmatrix} \\
&= (\tilde{\Sigma}_H^n - HE^n)^{-1} \begin{pmatrix} \hat{D}^T Y/n \\ Z^T Y/n + H'\alpha - H'\alpha \end{pmatrix} \\
&= \begin{pmatrix} \beta \\ \alpha \end{pmatrix} + (\tilde{\Sigma}_H^n - HE_n)^{-1} \begin{pmatrix} \hat{D}^T \hat{\eta}/n \\ Z^T \hat{\eta}/n - H'\alpha \end{pmatrix},
\end{aligned}$$

where $H'\alpha = \tau_n^{-2}(\alpha - \bar{\alpha}\mathbb{1}_J)$.

By similar proof of Theorem 4.2.1, we can get desired results. □

Proof of Theorem 4.2.5. We first state and prove the following lemma. Let $B_\alpha = Q^n + H_\alpha$ and $c_n^* = \max\left\{\Lambda_{\max}((B_\alpha^{-1})_{\hat{S}_\alpha,\hat{S}_\alpha}), \Lambda_{\max}((B_\alpha^{-1})_{\hat{S}_\alpha^c,\hat{S}_\alpha^c})\right\}$.

Lemma A.3.1. *Suppose that the eigenvalues of Σ_Z^n are bounded from above. Let $\tau_{0n}^2 = n^{-1}$ and τ_{1n}^2 be a positive constant. If $\Lambda_{\min}(Q^n_{\hat{S}_\alpha,\hat{S}_\alpha}) > 0$ for some $\hat{S}_\alpha \neq [J]$, then*

$$c_n^* \leq \max\left\{\frac{1 + c_0^{-1}\Lambda_{\max}(\Sigma_Z^n)}{\Lambda_{\min}(Q^n_{\hat{S}_\alpha,\hat{S}_\alpha})}, c_0^{-1}\right\},$$

for some constant $c_1 > 0$.

Proof of Lemma A.3.1. Let $\pi_0(\hat{\alpha}_j^{(MG)}) = 1 - \pi_1(\hat{\alpha}_j^{(MG)})$. First note that by definition of $\hat{S}_\alpha^c$ and the condition on τ_{0n}^2, $\min_{j\in\hat{S}_\alpha^c}(H_\alpha)_{j,j} \geq \min_{j\in\hat{S}_\alpha^c}\frac{\pi_0(\hat{\alpha}_j^{(MG)})}{n\tau_{0n}^2} \geq c_0$ for some constant $c_0 > 0$.

$$\begin{aligned}
\Lambda_{\max}((B_\alpha^{-1})_{\hat{S}_\alpha,\hat{S}_\alpha}) &= \Lambda_{\max}\left(\left\{(B_\alpha)_{\hat{S}_\alpha,\hat{S}_\alpha} - (B_\alpha)_{\hat{S}_\alpha,\hat{S}_\alpha^c}(B_\alpha)^{-1}_{\hat{S}_\alpha^c,\hat{S}_\alpha^c}(B_\alpha)_{\hat{S}_\alpha^c,\hat{S}_\alpha}\right\}^{-1}\right) \\
&= \Lambda_{\min}^{-1}\left((B_\alpha)_{\hat{S}_\alpha,\hat{S}_\alpha} - (B_\alpha)_{\hat{S}_\alpha,\hat{S}_\alpha^c}(B_\alpha)^{-1}_{\hat{S}_\alpha^c,\hat{S}_\alpha^c}(B_\alpha)_{\hat{S}_\alpha^c,\hat{S}_\alpha}\right), \qquad \text{(A.76)}
\end{aligned}$$

where

$$\begin{aligned}
&(B_\alpha)_{\hat{S}_\alpha,\hat{S}_\alpha} - (B_\alpha)_{\hat{S}_\alpha,\hat{S}_\alpha^c}(B_\alpha)^{-1}_{\hat{S}_\alpha^c,\hat{S}_\alpha^c}(B_\alpha)_{\hat{S}_\alpha^c,\hat{S}_\alpha} \\
&= (H_\alpha)_{\hat{S}_\alpha,\hat{S}_\alpha} + Q^n_{\hat{S}_\alpha,\hat{S}_\alpha} - Q^n_{\hat{S}_\alpha,\hat{S}_\alpha^c}(Q^n_{\hat{S}_\alpha^c,\hat{S}_\alpha^c} + (H_\alpha)_{\hat{S}_\alpha^c,\hat{S}_\alpha^c})^{-1}Q^n_{\hat{S}_\alpha^c,\hat{S}_\alpha} \\
&\succeq (H_\alpha)_{\hat{S}_\alpha,\hat{S}_\alpha} + Q^n_{\hat{S}_\alpha,\hat{S}_\alpha} - Q^n_{\hat{S}_\alpha,\hat{S}_\alpha^c}(Q^n_{\hat{S}_\alpha^c,\hat{S}_\alpha^c} + c_0 I_{\hat{S}_\alpha^c,\hat{S}_\alpha^c})^{-1}Q^n_{\hat{S}_\alpha^c,\hat{S}_\alpha}.
\end{aligned}$$

Thus, (A.76) satisfies

$$(A.76) \leq \left\{ \min_{j \in \hat{S}_\alpha} (H_\alpha)_{j,j} + \Lambda_{\min}(Q^n_{\hat{S}_\alpha,\hat{S}_\alpha}) \left[1 - \frac{\Lambda_{\max}(Q^n_{\hat{S}^c_\alpha,\hat{S}^c_\alpha})}{c_0 + \Lambda_{\max}(Q^n_{\hat{S}^c_\alpha,\hat{S}^c_\alpha})} \right] \right\}^{-1}.$$

Thus,

$$\Lambda_{\max}((B_\alpha^{-1})_{\hat{S}_\alpha,\hat{S}_\alpha}) \leq \frac{1 + c_0^{-1}\Lambda_{\max}(Q^n_{\hat{S}^c_\alpha,\hat{S}^c_\alpha})}{\Lambda_{\min}(Q^n_{\hat{S}_\alpha,\hat{S}_\alpha})}.$$

Similarly, we can get

$$\Lambda_{\max}((B_\alpha^{-1})_{\hat{S}^c_\alpha,\hat{S}^c_\alpha}) \leq \left\{ c_0 + \Lambda_{\min}(Q^n_{\hat{S}^c_\alpha,\hat{S}^c_\alpha}) \frac{\min_{j\in\hat{S}_\alpha}(H_\alpha)_{j,j}}{\min_{j\in\hat{S}_\alpha}(H_\alpha)_{j,j} + \Lambda_{\max}(Q^n_{\hat{S}_\alpha,\hat{S}_\alpha})} \right\}^{-1} \leq c_0^{-1}.$$

Since $\Lambda_{\max}(Q^n_{\hat{S}_\alpha,\hat{S}_\alpha}) \leq \Lambda_{\max}(\Sigma^n_Z)$ which is a constant,

$$c_n^* \leq \max\left\{ \frac{1 + c_0^{-1}\Lambda_{\max}(\Sigma^n_Z)}{\Lambda_{\min}(Q^n_{\hat{S}_\alpha,\hat{S}_\alpha})}, c_0^{-1} \right\}.$$

□

Now we are ready to prove Theorem 4.2.5. Let H_α be a diagonal matrix with $(H_\alpha)_{j,j} = (1 - \pi_1(\hat{\alpha}_j^{(MG)})/(n\tau^2_{0n}) + \pi_1(\hat{\alpha}_j^{(MG)})/(n\tau^2_{1n})$ for $j = 1, \ldots J$ and $\tilde{\Sigma}^n_{H_\alpha} = \frac{1}{n}\begin{pmatrix} \hat{D}^T\hat{D} & \hat{D}^T Z \\ Z^T\hat{D} & Z^T Z + nH_\alpha^{-1} \end{pmatrix}$. Replacing H by H_α in the proof of Theorem 4.2.1, we can obtain that $\tilde{\Sigma}^n_{H_\alpha}$ is invertible for any $\tau^2_{0n}, \tau^2_{1n} > 0$. Then

$$\begin{aligned}\hat{\beta}^{(MG)} =& \beta + \underbrace{\frac{1}{n}(\tilde{\Sigma}^n_{H_\alpha})^{-1}_{1,1}\hat{D}^T\hat{\eta} + \frac{1}{n}(\tilde{\Sigma}^n_{H_\alpha})^{-1}_{1,[J]}Z^T P_{\hat{D}}\hat{\eta}}_{E'_1} + \underbrace{\frac{1}{n}(\tilde{\Sigma}^n_{H_\alpha})^{-1}_{1,[J]}Z^T P^{\perp}_{\hat{D}}\hat{\eta}}_{E'_2} \\ &+ \underbrace{(\tilde{\Sigma}^n_{H_\alpha})^{-1}_{1,[J]} H_\alpha \Big(\omega(\hat{\alpha}^{(MG)})\mu_\alpha - \alpha\Big)}_{E'_3},\end{aligned} \tag{A.77}$$

where $E'_1 = E_1$.

For E'_3, by matrix inverse formula as in (A.74) and (A.75),

$$\begin{aligned}|E'_3| \leq& \frac{1}{\|\hat{D}\|_2^2}\left|\hat{D}^T Z_{\hat{S}_\alpha}(B_\alpha^{-1})_{\hat{S}_\alpha,\hat{S}_\alpha}(H_\alpha)_{\hat{S}_\alpha,\hat{S}_\alpha}(\alpha - \mu_\alpha\omega(\hat{\alpha}^{(MG)}))_{\hat{S}_\alpha}\right| \\ &+ \frac{1}{\|\hat{D}\|_2^2}\left|\hat{D}^T Z_{\hat{S}_\alpha}(B_\alpha^{-1})_{\hat{S}_\alpha,\hat{S}^c_\alpha}(H_\alpha)_{\hat{S}^c_\alpha,\hat{S}^c_\alpha}(\alpha - \mu_\alpha\omega(\hat{\alpha}^{(MG)}))_{\hat{S}^c_\alpha}\right| \\ &+ \frac{1}{\|\hat{D}\|_2^2}\left|\hat{D}^T Z_{\hat{S}^c_\alpha}(B_\alpha^{-1})_{\hat{S}^c_\alpha,\hat{S}_\alpha}(H_\alpha)_{\hat{S}_\alpha,\hat{S}_\alpha}(\alpha - \mu_\alpha\omega(\hat{\alpha}^{(MG)}))_{\hat{S}_\alpha}\right| \\ &+ \frac{1}{\|\hat{D}\|_2^2}\left|\hat{D}^T Z_{\hat{S}^c_\alpha}(B_\alpha^{-1})_{\hat{S}^c_\alpha,\hat{S}^c_\alpha}(H_\alpha)_{\hat{S}^c_\alpha,\hat{S}^c_\alpha}(\alpha - \mu_\alpha\omega(\hat{\alpha}^{(MG)}))_{\hat{S}^c_\alpha}\right|.\end{aligned} \tag{A.78}$$

For the first term on the right hand side of (A.78), we have

$$
\begin{aligned}
&\frac{1}{\|\hat{D}\|_2^2}\left|\hat{D}^T Z_{\hat{S}_\alpha}(B_\alpha^{-1})_{\hat{S}_\alpha,\hat{S}_\alpha}(H_\alpha)_{\hat{S}_\alpha,\hat{S}_\alpha}(\alpha-\mu_\alpha\omega(\hat{\alpha}^{(MG)}))_{\hat{S}_\alpha}\right| \\
&\leq \frac{\|\hat{D}^T Z_{\hat{S}_\alpha}\|_2}{\|\hat{D}\|_2^2}\Lambda_{\max}((B_\alpha^{-1})_{\hat{S}_\alpha,\hat{S}_\alpha})\left\|(H_\alpha)_{\hat{S}_\alpha,\hat{S}_\alpha}(\alpha-\mu_\alpha\omega(\hat{\alpha}^{(MG)}))_{\hat{S}_\alpha}\right\|_2 \\
&= \frac{c_n^*\Lambda_{\max}^{1/2}(\Sigma_Z^n)}{\|\hat{D}\|_2/\sqrt{n}}\left\|(H_\alpha)_{\hat{S}_\alpha,\hat{S}_\alpha}(\alpha-\mu_\alpha\omega(\hat{\alpha}^{(MG)}))_{\hat{S}_\alpha}\right\|_2 \\
&\leq \frac{c_n^*\Lambda_{\max}^{1/2}(\Sigma_Z^n)}{\|\hat{\gamma}\|_2\Lambda_{\min}^{1/2}(\Sigma_Z^n)}\left\|(H_\alpha)_{\hat{S}_\alpha,\hat{S}_\alpha}(\alpha-\mu_\alpha\omega(\hat{\alpha}^{(MG)}))_{\hat{S}_\alpha}\right\|_2 \\
&\leq \frac{c_n^* c_{1n}\Lambda_{\max}^{1/2}(\Sigma_Z^n)}{\|\hat{\gamma}\|_2\Lambda_{\min}^{1/2}(\Sigma_Z^n)}\left\|(\alpha-\mu_\alpha\omega(\hat{\alpha}^{(MG)}))_{\hat{S}_\alpha}\right\|_2 \text{ for some } c_{1n}=o(1),
\end{aligned}
$$

where the last step is due to

$$
\max_{j\in\hat{S}_\alpha}(H_\alpha)_{j,j} \leq \max_{j\in\hat{S}_\alpha}\frac{\pi_0(\hat{\alpha}_j^{(MG)})}{n\tau_{0n}^2} = o(1).
$$

Also note that $\max_{j\in\hat{S}_\alpha^c}(H_\alpha)_{j,j}\leq 1$ and

$$
\omega(\hat{\alpha}_j) \leq \frac{1}{1+\min_{j\in\hat{S}_\alpha^c}\pi_0(\hat{\alpha}_j^{(MG)})n\tau_{1n}^2} = O(n^{-1}).
$$

For the second term on the right hand side of (A.78), we have

$$
\begin{aligned}
&\frac{1}{\|\hat{D}\|_2^2}\left|\hat{D}^T Z_{\hat{S}_\alpha}(B_\alpha^{-1})_{\hat{S}_\alpha,\hat{S}_\alpha^c}(H_\alpha)_{\hat{S}_\alpha^c,\hat{S}_\alpha^c}(\alpha_{\hat{S}_\alpha^c}-\omega(\hat{\alpha}^{(MG)})_{\hat{S}_\alpha^c}\mu_\alpha)\right| \\
&\leq \frac{\|\hat{D}^T Z_{\hat{S}_\alpha}\|_2}{\|\hat{D}\|_2^2}\Lambda_{\max}^{1/2}((B_\xi^{-1})_{\hat{S}_\alpha,\hat{S}_\alpha})\Lambda_{\max}^{1/2}((B_\xi^{-1})_{\hat{S}_\alpha^c,\hat{S}_\alpha^c})\left\|\alpha_{\hat{S}_\alpha^c}-c_{2n}\right\|_2 \\
&\leq \frac{c_n^*\Lambda_{\max}^{1/2}(\Sigma_Z^n)}{\|\hat{\gamma}\|_2\Lambda_{\min}^{1/2}(\Sigma_Z^n)}\left\|\alpha_{\hat{S}_\alpha^c}-c_{2n}\right\|_2,
\end{aligned}
$$

where $c_{2n}=O(n^{-1})$. Dealing the third and fourth term in (A.78) similarly, we have

$$
|E_3'| \leq \frac{2c_n^*\Lambda_{\max}^{1/2}(\Sigma_Z^n)}{\Lambda_{\min}^{1/2}(\Sigma_Z^n)}\left[\frac{c_{1n}\|(\alpha-\omega(\hat{\alpha}^{(MG)})\mu_\alpha)_{\hat{S}_\alpha}\|_2}{\|\hat{\gamma}\|_2}+\frac{\|\alpha_{\hat{S}_\alpha^c}-c_{2n}\|_2}{\|\hat{\gamma}\|_2}\right],
$$

for some $c_{1n}=o(1)$ and $c_{2n}=O(n^{-1})$. For E_2', we can similarly prove that

$$
\begin{aligned}
|E_2'| &\leq \frac{2c_n^*\|\hat{D}^T Z/n\|_2(\|Z_{\hat{S}_\alpha}^T P_{\hat{D}}^\perp\hat{\eta}\|_2+\|Z_{\hat{S}_\alpha^c}^T P_{\hat{D}}^\perp\hat{\eta}\|_2)}{\|\hat{D}\|_2^2} \\
&\leq \frac{4c_n^*\Lambda_{\max}^{1/2}(\Sigma_Z^n)\|Z^T P_{\hat{D}}^\perp\hat{\eta}\|_2}{\sqrt{n}\|\hat{D}\|_2}.
\end{aligned}
$$

□

A.3.2 Implementation details

In this section, we discuss the computation of the conditional posterior mean under prior (4.10)-(4.12). For the initial values, take $\hat{\beta}^{(0)} = 0$. For known μ_α and p_0, the Gibbs samples can be generated with respect to

$$\hat{\alpha}^{(t)} \sim N(\underline{\theta}_{\alpha,\xi}, \underline{\Sigma}_{\alpha,\xi}), \quad \hat{\beta}^{(t)} \sim N(\underline{\theta}_{\beta,\xi}, \underline{\sigma}^2_{\beta,\xi}), \quad \xi_j^{(t)} \sim Ber(\underline{p_j}),\ j = 1, \ldots, p, \tag{A.79}$$

$$\begin{aligned}(\sigma_\eta^{-2})^{(t)} \sim \text{Gamma}(\nu_1 + \frac{n}{2} + \frac{J}{2}, \nu_2 + \frac{1}{2}\|Y - \hat{D}\hat{\beta}^{(t-1)} - Z\hat{\alpha}^{(t)}\|_2^2 \\ + \sum_{j=1}^{J} \frac{(\hat{\alpha}_j^{(t)} - \mu_\alpha \xi_j^{(t)})^2}{2\tau_{0n}^2 + 2(\tau_{1n}^2 - \tau_{0n}^2)\xi_j^{(t)}}),\end{aligned} \tag{A.80}$$

where

$$\underline{\Sigma}_{\alpha,\xi} = \left(Z^T Z + (H_\xi^{(t)})^{-1}\right)^{-1} (\sigma_\eta^2)^{(t-1)},$$

$$\underline{\theta}_{\alpha,\xi} = \underline{\Sigma}_{\alpha,\xi}(Z^T(Y - \hat{D}\hat{\beta}^{(t-1)}) + (H_\alpha^{(t)})^{-1}\mu_\alpha),$$

for a diagonal matrix $H_\alpha^{(t)}$ with $(H_\alpha^{(t)})_{j,j} = (\tau_{0n}^2 + (\tau_{1n}^2 - \tau_{0n}^2)\xi_j^{(t)})$,

$$\underline{\theta}_{\beta,\xi} = \frac{\hat{D}^T(Y - Z\hat{\alpha}^{(t)})}{\hat{D}^T\hat{D}}, \quad \underline{\sigma}^2_{\beta,\xi} = \frac{(\sigma_\eta^2)^{(t-1)}}{\hat{D}^T\hat{D}}$$

and

$$\underline{p_j} = \frac{p_1\phi(\hat{\alpha}_j^{(t)}|\mu_\alpha, \tau_{1n}^2)(\sigma_\eta^2)^{(t-1)}}{p_1\phi(\hat{\alpha}_j^{(t)}|\mu_\alpha, \tau_{1n}^2)(\sigma_\eta^2)^{(t-1)}) + (1 - p_1)\phi(\hat{\alpha}_j^{(t)}|0, \tau_{0n}^2(\sigma_\eta^2)^{(t-1)})}$$

For unknown μ_α and p_0, we can apply the MCEM algorithm with some modifications.

Start with initial values $\hat{\mu}_\alpha^{(0)} = 0$ and $\hat{p}_1^{(0)} = 0.5$.

E-step: Generate $(\alpha_i^{(t)}, \beta_i^{(t)}, \xi_i^{(t)}, (\sigma_\eta^2)_i^{(t)})$, $i = 1, \ldots, m$, by running m rounds of Gibbs sampling as (A.79) and (A.80) replacing μ_α and p_1 with $\hat{\mu}_\alpha^{(t-1)}$ and $\hat{p}_1^{(t-1)}$ respectively.

M-step:

$$(\hat{\mu}_\alpha^{(t)}, \hat{p}_1^{(t)}) = \underset{(\mu_\alpha, p_1)}{\arg\max} \frac{1}{m}\sum_{i=1}^{m} \log p(\beta, \mu_\alpha, p_1|\mathcal{D}, \alpha_i^{(t)}, \beta_i^{(t)}, \xi_i^{(t)}, (\sigma_\eta^2)_i^{(t)}).$$

Specifically, the maximizers in the M-step take the form

$$\hat{\mu}_\alpha^{(t)} = \frac{\sum_{i=1}^m \sum_{j=1}^J \hat{\alpha}_{i,j}^{(t)}\hat{\xi}_{i,j}^{(t)}/(\sigma_\eta^2)_i^{(t)}}{\sum_{i=1}^m \sum_{j=1}^J \hat{\xi}_{i,j}^{(t)}/(\sigma_\eta^2)_i^{(t)}}$$

$$\hat{p}_1^{(t)} = \frac{1}{mJ}\sum_{i=1}^{m}\sum_{j=1}^{J} \hat{\xi}_{i,j}^{(t)}.$$

At the convergence of $(\hat{\mu}_{\alpha}^{(t)}, \hat{p}_1^{(t)})$, produce $\frac{1}{m}\sum_{i=1}^{m}\hat{\beta}_i^{(t)}$ as the final estimate of β.

A.3.3 Implementation with summary statistics

Suppose we observe $\hat{\Gamma}_j \in \mathbb{R}$ as the association estimate between the interested outcome Y and the j-th genetic variant Z_j, $\hat{\sigma}^2_{\Gamma,j} \in \mathbb{R}$ as the estimated variance of $\hat{\Gamma}_j$ and $\hat{\gamma}_j \in \mathbb{R}$ be the association estimate between the interested exposure D and Z_j. Let $\hat{\gamma}^{(2)}$ be a version of $\hat{\gamma}$ obtained from an independent sample $(Z^{(2)}, D^{(2)})$. In model (4.17) - (4.19), the sample moments used throughout the computation are replaced by the summary statistics as follows.

$$\begin{aligned} &Z^T Z = I_{J\times J}, \quad Z^T \hat{D} = \hat{\gamma}^{(2)}, \\ &Z^T Y = \hat{\Gamma}, \quad \hat{D}^T \hat{D} = (\hat{\gamma}^{(2)})^T \hat{\gamma}^{(2)}, \hat{D}^T Y = (\hat{\gamma}^{(2)})^T \hat{\Gamma}. \end{aligned}$$

We use $\hat{\sigma}^2_{\Gamma}$ as an estimate of σ^2_{Γ}. Thus, we are able to get an empirical Bayes estimator with summary statistics based on the algorithm in the previous section.