New Statistical Methods for Drawing Inference Based on High Dimensional Regression Models

by

Zhe Fei

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2019

Doctoral Committee:

Professor Yi Li, Chair
Professor Moulinath Banerjee
Associate Professor Jian Kang
Professor Ji Zhu

Zhe Fei

feiz@umich.edu

ORCID iD: 0000-0001-9568-2857

To My Parents, Mei and Scarlett

ACKNOWLEDGEMENTS

My journey thus far dates back 7 years ago when I came to the Department of Biostatistics at the University of Michigan to pursue my degrees. It has been a wonderful part of my life with tremendous support coming from so many people, including my advisor, dissertation committee members, colleagues, friends, and family.

First, my deepest appreciation goes to my advisor Dr. Yi Li for his mentorship, not only in research, but also in life. He has always been pointing me to the right direction for achieving excellence. He has taught me how to become an independent researcher, through critical thinking, focus on details, value in innovation and collaboration, among many other aspects. He has been the perfect figure that I aim to become when I start my own career. His professionalism has deeply impacted me going forward, which is necessary toward any success in my own career.

My sincere appreciation also goes to Drs. Moulinath Banerjee, Jian Kang, and Ji Zhu for serving as my dissertation committee members. They have provided invaluable advice and support throughout my graduate study. Drs. Banerjee and Zhu have been my mentors since I started my dissertation, and have offered their expertise for all of my dissertation chapters. Dr. Kang has been routinely giving me critical and objective suggestions on a weekly basis. They have helped substantially improve the presentation of this dissertation. I am also grateful toward their kind and useful advice for my career path.

I would also like to thank my colleagues and friends for their help toward this dissertation. Here is everyone listed in alphabetical order: Kevin He, Zihuai He, Yan-

ming Li, Emily Morris, Lu Tang, Lu Xia, Fan Wu, and Yuan Yang. I am grateful for their generosity, especially, to Kevin He, who has been genuinely helpful in my research with insightful discussions. Additionally, I would like to thank my department, the faculty and staff, fellow students and friends for a supportive and encouraging environment for study, research and personal growth.

Last, the ultimate motivation and support come from my family, including my parents, Huiping Fei and Jushui Sun, my wife, Mei Mei, and my loved one, Scarlett. I would like to dedicate this thesis to them for their persevering and unconditional love.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

LIST OF APPENDICES

Appendix

ABSTRACT

Quantifying the uncertainty of estimated parameters in high dimensional sparse models gives critical insights and valuable information in analyzing various types of big data. Yet it possesses some unique difficulties and has been drawing numerous research attention over the past years. The goal of high dimensional inference is to provide accurate point estimators of the unknown parameters with tractable limiting distributions, which leads to confidence intervals, significance testing, and other uncertainty measures. In this dissertation, we propose a novel estimation procedure, along with a nonparametric variance estimator, which is adaptive to a wide range of regression models and outcome types to draw reliable inferences for the model parameters. Comparisons are made with several existing methods, and advantages of our procedure are shown both in simulation studies and real data applications. Our method is successfully applied to multiple genomic data sets with continuous, binary, and survival outcomes.

CHAPTER I

Introduction

Quantifying the uncertainty of estimated parameters in high dimensional sparse models gives critical insights and valuable information in analyzing various types of big data. Yet it possesses some unique difficulties and has been drawing numerous research attention over the past years. The goal of high dimensional inference is to provide accurate point estimators of the unknown parameters with tractable limiting distributions, which leads to confidence intervals, significance testing, and other uncertainty measures. In this dissertation, we propose a novel estimation procedure, along with a nonparametric model-free variance estimator, which is adaptive to a wide range of regression models and outcome types to draw reliable inferences for the model parameters.

In Chapter II, we started with high dimensional linear models and derived a smoothed estimator of the whole coefficient vector whose components were asymptotically unbiased and normal. Our procedure was based on multi-sample splitting and selection assisted partial regression so that the estimator enjoyed both low dimensional least square properties and variance reduction from the effect of bagging. Our numerical studies provided finite sample performances of the proposed procedure including consistency and coverage probabilities, as well as the advantages when comparing with the de-biased type of estimators. In an application to multiple myeloma

patients data, we identified 2 significant gene probes out of 789 potential predictors in association with the disease severity.

In Chapter III, we extended our method to generalized linear models while improving the procedure in several aspects. We generalized the re-sampling scheme to simple data splitting and studied the effect of different split proportions. We derived the consistent variance estimation and corrected for its bias due to finite number of re-samples. We explored the different selection methods to be applied in the procedure and showed the consistency in estimation and inference regardless of using LASSO, SCAD, MCP, or others. By fitting a high dimensional logistic regression with the proposed procedure, we found 9 significant gene-environment interactions among $13,663$ covariates that differentiate lung cancer patients versus controls.

In Chapter IV, we used the censored quantile regression framework to model survival outcomes with unknown censoring. Censored quantile regressions, an alternative to Cox proportional hazards model, were powerful in detecting the covariate effects at extreme tails, and thus provided complete information of the outcome distribution. In the context of high dimensional censored quantile regressions, our work pioneered in simultaneous estimation and inference for all model parameters, which was of significant importance in survival analysis with big data. We solved the theoretical challenges when extending the method to high dimensional censored quantile regressions. We successfully applied our method to analyze the survival of lung cancer patients with large number of potential predictors, and detected more significant SNP effects than other models.

CHAPTER II

Drawing Inferences for High Dimensional Linear Models: A Selection-assisted Partial Regression and Smoothing Approach

2.1 Introduction

Consider the classical linear model:

$$\boldsymbol{Y} = \boldsymbol{X}\beta^0 + \boldsymbol{\varepsilon} \tag{2.1}$$

where $\boldsymbol{Y} = (y_1, y_2, ..., y_n)^{\mathrm{T}}$ is the n-vector of the response variable; $\boldsymbol{X} = (X_1, X_2, ..., X_p)$ is the $n \times p$ design matrix that consists of p covariate vectors X_j's; $\boldsymbol{X}$ can also be written as $\boldsymbol{X} = (\mathbf{x}_1^{\mathrm{T}}, \mathbf{x}_2^{\mathrm{T}}, .., \mathbf{x}_n^{\mathrm{T}})^{\mathrm{T}}$, where $\mathbf{x}_i = (x_{i1}, ..., x_{ip})$ represents the p-vector of covariates for the i^{th} individual; $\beta^0 = (\beta_1^0, ..., \beta_p^0)^{\mathrm{T}}$ is the true parameter vector of interest; $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)^{\mathrm{T}}$ is the random noise vector and $\mathbf{E}(\boldsymbol{\varepsilon}) = \mathbf{0}_n$.

In the traditional low-dimensional setting when $n > p$, it is well known that least squares estimator $\hat{\beta}_{LS} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$ converges to a normal distribution centered at β^0, which provides exact estimation and inferences through explicitly computable p-values and confidence intervals. On the other hand, when $n < p$, the least squares estimation would fail because the sample covariance matrix $\widehat{\boldsymbol{\Sigma}} =$

$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}/n$ is singular. However the $n < p$ problem has become increasingly relevant over the past two decades with the common availability of high-throughput data. The goal is often to find a parsimonious model to explain the response in the presence of massive covariates. A number of selection and estimation methods including LASSO (*Tibshirani*, 1996), Adaptive LASSO (*Zou*, 2006), SCAD (*Fan and Li*, 2001), ISIS (*Fan and Lv*, 2008), among others, are available.

More recently, interest in the statistical community has shifted to making reliable inferences in high-dimensional models. Researchers have been trying to tackle the problem from different angles. One direction is to make inferences based on the *selected* model, i.e. the one that is chosen by a given variable selection procedure. *Wasserman and Roeder* (2009) proposes a multi-stage procedure that is based on data splitting to separate selection and inference; *Berk et al.* (2013) provides conservative confidence intervals for the selected variables by defining a set of candidate models; *Lee and Taylor* (2014); *Lee et al.* (2016) develops the conditional symptomatic of the coefficient estimates, given the selected model. The second direction is to estimate and make inferences of the low-dimensional parameters in the high dimensional models. *Belloni et al.* (2013, 2014) propose a double selection procedure instead of a single selection step to estimate and construct confidence regions for a regression parameter of primary interest. Some other works propose estimators and inferences based on penalized estimation. A typical example is the bias correction method based on LASSO (*Zhang and Zhang*, 2014; *Van de Geer et al.*, 2014; *Javanmard and Montanari*, 2014), which provides point estimation and confidence intervals for the model parameters. There is also work by *Ning and Liu* (2017) that proposes hypothesis tests and confidence regions based on the de-correlated score function and test statistic.

These approaches have their merits and demerits. While *Wasserman and Roeder* (2009); *Lee and Taylor* (2014); *Lee et al.* (2016) aim at exact inference for post-selection estimates, it is confined to the selected model from the *"first step."* Thus,

flaws in the initial model-selection step, cannot be rectified in subsequent steps. The limitation of requiring perfect model selection is improved in *Belloni et al.* (2014), meanwhile, *Wasserman and Roeder* (2009); *Meinshausen et al.* (2009) recommend *not* performing selection and estimation on the same data set. On the other hand, the performance of the original de-biased LASSO estimator relies heavily on the accuracy of estimating the precision matrix, i.e. $\mathbf{\Sigma}^{-1}$, which plays an unduly crucial role in the estimation and inference subsequently. In *Javanmard and Montanari* (2014), they relaxed the required accuracy of estimating $\mathbf{\Sigma}^{-1}$ (the matrix M in their paper), instead they set M as to minimize the error term and the variance of the target Gaussian limit.

In this chapter we propose a novel approach to consistently estimate β^0, provide p-values for all covariates, and compute confidence intervals for any fixed subset of parameters in high-dimensional linear models. The approach, coined *Selection-assisted Partial Regression and Smoothing* (SPARES), possesses asymptotic unbiasedness and asymptotic normality. Our idea takes advantage of the multisample-splitting method in *Meinshausen et al.* (2009), which defines a p-value for each predictor from each sample-splitting and then aggregates these p-values to declare a single p-value per feature. One possible criticism of this approach is that the p-values and the aggregation have a certain arbitrary angle to them: for example, features not selected in each sample-split subsample are all assigned a p-value 1. In contrast, our SPARES estimator utilizes partial regression to estimate β^0 in each sample-split followed by a natural smoothing step. In each data split, our procedure provides an estimate of $\beta^0_j,\ j = 1, 2, .., p$ regardless of whether it was chosen by the selection procedure. Such idea of attaching variable j to the selected variables is also used in *Belloni et al.* (2014). Then we average over the variation of the selection and sample-split to obtain a smoothed estimator. For these reasons, SPARES is *not* a post model-selection method. Furthermore, our approach avoids the need to estimate the high-dimensional

precision matrix.

Our approach stands out from the majority of related works (*Wasserman and Roeder*, 2009; *Zhang and Zhang*, 2014; *Van de Geer et al.*, 2014; *Javanmard and Montanari*, 2014; *Belloni et al.*, 2014; *Ning and Liu*, 2017) in that it is neither restricted to a fixed realization of the selected model nor limited to a certain selection procedure. The smoothing accomplished through multisample-splitting ensures that the $\widehat{\beta}_j$'s are asymptotically normal with negligible bias while the standard errors can be readily estimated via a nonparametric delta method (*Efron*, 2014). Consequently, inferences can be made for each and every $\beta_j^0, j = 1, 2, .., p$ without having to confront the curse of dimensionality. As shown in the data applications, our method is advantageous in giving uncertainty measures (such as p-values) to all high dimensional coefficients at once.

The rest of this chapter is organized as follows. Section 2.2 describes the SPARES estimator and Section 2.3 develops its theoretical properties. Section 2.4 shows how to draw inferences through SPARES, including confidence intervals and significance tests. Section 2.5 discusses the extension to a subvector of β^0 with a fixed dimension. In Section 2.6 we conduct simulations to examine the performance of SPARES and present comparisons to de-biased LASSO methods. Section 2.7 comprises two real data applications and Section 2.8 summarizes the merit of this work and pinpoints future research.

2.2 Proposed Method

Let $[p] = \{1, 2, .., p\}$ denote the set of integers for any positive p. For a vector V of length p, denote the entry corresponding to subscript $j \in [p]$ by V_j or $(V)_j$; for a square matrix $\mathbf{\Sigma} = \mathbf{\Sigma}_{p\times p}$, denote the entry corresponding to subscripts $j, k \in [p]$ by $\mathbf{\Sigma}_{jk}$ or $(\mathbf{\Sigma})_{jk}$ for clarity if necessary; for a subset $S \subset [p]$, denote the sub-design matrix $X_S = (X_j)_{j\in S}$ and the sub-covariance matrix $\mathbf{\Sigma}_S = (\mathbf{\Sigma}_{jk})_{j,k\in S}$. The projection

matrix of X_S is denoted as $H_S = X_S(X_S^{\mathrm{T}}X_S)^{-1}X_S^{\mathrm{T}}$. The active set of β^0 is $S_{0,n} = \{j \in [p] : \beta_j^0 \neq 0\}$.

One-time SPARE: We first introduce the estimation of β^0 through Selection-assisted Partial Regression (SPARE) on a single data-split. Given data $D_n = (\boldsymbol{X}, \boldsymbol{Y})$ as in model (2.1) and a generic selection procedure $\mathcal{S}_\lambda$ with parameter λ, we first split D_n into two halves D_1 and D_2, with $|D_1| = \lfloor n/2 \rfloor$, $|D_2| = \lceil n/2 \rceil$, the floor and ceiling of it. Denote the subset of variables selected by $\mathcal{S}_\lambda$ on D_2 as $S = \mathcal{S}_\lambda(D_2)$. Next on $D_1 = (\boldsymbol{X}^1, \boldsymbol{Y}^1)$, the partial regression estimator for $\beta_j^0,\ j \in [p]$ is

$$\widetilde{\beta}_j = \left\{ (\boldsymbol{X}^1_{S\cup j}{}^{\mathrm{T}} \boldsymbol{X}^1_{S\cup j})^{-1} \boldsymbol{X}^1_{S\cup j}{}^{\mathrm{T}} \boldsymbol{Y}^1 \right\}_j , \tag{2.2}$$

which is the coefficient estimate corresponding to $\boldsymbol{X}_j^1$ from the least squares regression of $\boldsymbol{Y}^1$ on $\boldsymbol{X}^1_{S\cup j}$. Moreover, (2.2) can be written as $\widetilde{\beta}_j = \left\{ \boldsymbol{X}_j^{1\mathrm{T}} (I_{n/2} - H^1_{S\setminus j}) \boldsymbol{X}_j^1 \right\}^{-1} \boldsymbol{X}_j^{1\mathrm{T}} (I_{n/2} - H^1_{S\setminus j}) \boldsymbol{Y}^1$ in the partial regression formulation.

Let $S_C = [p] \setminus S$, we can write the one-time SPARE estimator compactly as

$$\widetilde{\beta}(D_1, S) = \begin{pmatrix} \widetilde{\beta}_S \\ \widetilde{\beta}_{S_C} \end{pmatrix} = \begin{pmatrix} (\boldsymbol{X}_S^{1\mathrm{T}} \boldsymbol{X}_S^1)^{-1} \boldsymbol{X}_S^{1\mathrm{T}} \boldsymbol{Y}^1 \\ \left[\mathrm{diag}\{ \boldsymbol{X}^1_{S_C}{}^{\mathrm{T}} (I_{n/2} - H_S^1) \boldsymbol{X}^1_{S_C} \} \right]^{-1} \boldsymbol{X}^1_{S_C}{}^{\mathrm{T}} (I_{n/2} - H_S^1) \boldsymbol{Y}^1 \end{pmatrix} . \tag{2.3}$$

The rationale for SPARE to work is that given a subset of important predictors $S \subset [p]$ that is close to the active set $S_{0,n}$, the partial regression estimator (2.2) would be a fine estimator that is close to the truth β_j^0, for all $j \in [p]$. In fact, as long as $S \supset S_{0,n}$, (2.2) would be an unbiased estimator for β_j^0, regardless of $j \in S$ or not. However, given the large number of predictors, the one-time SPARE estimator is highly variable, and heavily depends on the selected S and the specific split of data.

SPARES: To overcome this difficulty, we introduce its smoothed version, the SPARES estimator, which is derived from multisample-splitting and repeated applications of SPARE. For a large enough B and each $b = 1, 2, .., B$, we first draw a

sample of size $n/2$, with replacement, from the full data and denote it as D_1^b. When n is odd, we interpret $n/2$ as $\lfloor n/2 \rfloor$. Let $I_1 = \{i_1, i_2, ..., i_{n/2}\}, 1 \leq i_k \leq n$ be the collection of indices of the observations in D_1^b. Next, we collect the observations that are not drawn in D_1^b as D_2^b with index set $I_2 = [n] \setminus I_1$. Thus $I_1 \cup I_2 = [n]$ and $I_1 \cap I_2 = \emptyset$. Now the application of SPARE by (2.3) is $\widehat{\beta}^b = \widetilde{\beta}(D_1^b, S^b)$, where $S^b = \mathcal{S}_\lambda(D_2^b)$; the final step is to average over all $\widehat{\beta}^b$'s,

$$\widehat{\beta}_{\text{SPARES}} = \frac{1}{B} \sum_{b=1}^{B} \widehat{\beta}^b. \tag{2.4}$$

In terms of the computational cost, each of the one-time SPARE has the same time complexity as one run of LASSO ($O(np^2)$), and the cost of the SPARES procedure is B times that. But with the help of parallel computing, we could largely reduce the computation time by any desired factor K depending on the computing tool. Thus the time complexity of SPARES is $O(Bnp^2/K)$, a multiple of one-time LASSO proportional to the number of re-samples. Empirically the total time cost of the SPARES procedure is linear in $p \log n$.

In the rest of the chapter, we will always use $\widetilde{\beta}$ for the one-time SPARE estimator and $\widehat{\beta}$ for the SPARES estimator. Both the one-time SPARE and the SPARES possess the asymptotic unbiasedness and normality, but SPARES is much more stable due to the smoothing effect from multisample-splitting, which we will explore in depth throughout the rest of this chapter.

2.3 Theoretical Results

2.3.1 One-time SPARE

We first establish the asymptotic property of the one-time SPARE estimator under the following assumptions.

(A1) Randomness of Data: In model (2.1), $\varepsilon_i \perp \mathbf{x}_i$; ε_i's are i.i.d. random errors with mean zero, finite variance σ^2 and finite third absolute moment $\mathbf{E}|\varepsilon_i|^3 \leq \rho_0$; $\boldsymbol{X} = (\mathbf{x}_1^{\mathrm{T}}, ..., \mathbf{x}_n^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{x}_i$'s are i.i.d. mean zero sub-Gaussian random vectors in $\mathbf{R}^p$ with covariance matrix $\boldsymbol{\Sigma}_{p\times p}$, whose eigenvalues are bounded,

$$0 < c_{\min} \leq \lambda_{\min}(\boldsymbol{\Sigma}) \leq \lambda_{\max}(\boldsymbol{\Sigma}) \leq c_{\max} < \infty.$$

$\mathbf{x}_i$'s also have finite component-wise third absolute moments $\forall j,\ \mathbf{E}|x_{ij}|^3 \leq \rho_1$.

(A2) Order of Model Parameters: There exist constants $0 < c_1 \leq 1, c_\beta > 0$ such that $s_0 = |S_{0,n}| = O(n^{c_1})$, $\max_j |\beta_j^0| \leq c_\beta$.

(A3) Sure Screening Property: There exists a sequence $\{\lambda_n\}_{n\geq 1}$ and constants $0 < \eta < 1$, $c_2 > 2c_1$ such that $|\widehat{S}_{n,\lambda_n}|/n \leq \eta$, and

$$P(\widehat{S}_{n,\lambda_n} \supset S_{0,n}) \geq 1 - o(n^{-c_2-1}) \quad \text{as} \quad n \to \infty.$$

Here $\widehat{S}_{n,\lambda_n}$ denotes the selected set of variables with sample size n and tuning parameter λ_n.

Remark II.1. The sure screening property is met in *Fan and Lv* (2008); *Fan and Song* (2010), and is guaranteed with the right order of tuning parameter λ using LASSO (*Bach*, 2008). More specifically, by *Fan and Lv* (2008); *Fan and Song* (2010), in addition to assumptions (A1) and (A2), the following conditions are required for the sure screening property to hold:

- $\mathrm{Var}(\boldsymbol{Y}) = O(1)$, and for some $\kappa \geq 0$ and $c_0, c_3 > 0$, $\min_{j\in S_0} |\beta_j^0| \geq c_0/n^\kappa$ and

$$\min_{\beta_j \neq 0} \left|\mathrm{cov}\left(\beta_j^{-1}\boldsymbol{Y}, \boldsymbol{X}_j\right)\right| \geq c_3;$$

- $\log p = O(n^\xi)$ for some $0 < \xi < 1 - 2\kappa$.

When $\kappa \geq 1/3$, the sparsity requirement implied by *Fan and Lv* (2008), $s_0 = o(n^\theta)$ for some $0 < \theta < 1-2\kappa$, is stronger than that in *Javanmard and Montanari* (2018), which is $s_0 = o(n/(\log p)^2)$. When $\kappa < 1/3$, the comparison between the two conditions are inconclusive. See conditions 1-4 in *Fan and Lv* (2008) for more details.

In (A1), only a moment condition is required on the error terms and a sub-Gaussian distribution for the covariates. For comparisons, while the asymptotic normality of the whole $p-$dimensional de-biased estimator is not guaranteed for non-Gaussian errors, a central limit theorem argument can be used to obtain approximate Gaussianity of components of fixed dimension (*Bühlmann et al.*, 2014). Thus the inference for any fixed low-dimensional parameter is still valid for these types of methods under sub-Gaussian errors with finite moment conditions. In (A2), there is no direct assumption on the order of p, however, it is implied through (A3), a condition made directly on the selection method. One reason for such an assumption, instead of more basic ones like the order of p or the covariance structure of the predictors, is that selection only plays an assistive role in our method; the estimation part is in fact low-dimensional and therefore does not directly require typical high-dimensional conditions.

Theorem II.2. *Given model (2.1) and assumptions (A1)-(A3), consider the one-time SPARE estimator* $\widetilde{\beta} = (\widetilde{\beta}_1, \widetilde{\beta}_2, .., \widetilde{\beta}_p)^{\mathrm{T}}$ *as defined in (2.3). Denote* $m = \lfloor n/2 \rfloor$, $\widetilde{\sigma}_j^2 = \sigma^2 \left({\boldsymbol{X}^1_{S\cup j}}^T X^1_{S\cup j}/m \right)^{-1}_{jj}$. *Then* $\forall j \in [p]$, *as* $m \to \infty$,

$$\sqrt{m}(\widetilde{\beta}_j - \beta_j^0)/\widetilde{\sigma}_j \to N(0,1). \tag{2.5}$$

Remark II.3. Note that we could always let the quantity of interest in (6) to be zero whenever $S_0 \not\subset S$, whose probability goes to zero by (A3). Thus we only need to show the convergence when the event $S_0 \subset \widehat{S}$ holds.

The proof is presented in Appendix A.

2.3.2 SPARES

Given the high volume of predictors in the model (2.1), the one-time estimator is expected to be noisy and unstable, especially for all the $j \notin S_{0,n}$ that are the majority of the $p-$vector β^0. In contrast, the SPARES estimator is more stable as it smooths over both estimation and selection. As the SPARES introduces extra dependency between the selections S^b's and the partial regression estimates, the following condition, which is stronger than "sure screening", is required for the desired theoretical property.

(B3). Selection Consistency: There exists a sequence $\{\lambda_n\}_{n\geq 1}$ and constants $0 < \eta < 1$, $c_2 > 2c_1$ such that $|\widehat{S}_{n,\lambda_n}|/n \leq \eta$, and

$$\mathbf{P}(\widehat{S}_{n,\lambda_n} = S_{0,n}) \geq 1 - o(n^{-c_2-1}) \quad \text{as} \quad n \to \infty.$$

The selection consistency is often met under certain sparsity conditions depending on the selection method (*Zhao and Yu*, 2006; *Zhang*, 2010). Take LASSO for example, the selection consistency property is guaranteed under $s_0 = O(n^{c_1})$ and $s_0 \log p = o(n^{c_3})$ for some $0 < c_1 < c_3 < 1$, along with irrepresentable condition and others.

Theorem II.4. *Given model (2.1) and assumptions (A1,A2,B3), consider the SPARES estimator* $\widehat{\beta}_{\text{SPARES}} = (\widehat{\beta}_1, ..., \widehat{\beta}_p)^T$ *as defined in (2.4). For each* j*, there exist random variables* Z_j^0, Δ_j*, such that as* $n, B \to \infty$*,*

$$\sqrt{n}(\widehat{\beta}_j - \beta_j^0) = Z_j^0 + \Delta_j, \quad Z_j^0/\sigma_j \to N(0,1), \quad \Delta_j = o_p(1),$$

where $\sigma_j^2 = \sigma^2 \left(\mathbf{\Sigma}^{-1}_{S_{0,n}\cup j}\right)_{jj}$ *is bounded.*

The proof is presented in Appendix A along with some useful lemmas. The difficulties in deriving the theoretical properties of the SPARES estimator arise primarily

from the randomness of S^b's, the selected subsets of variables from subsamples of the original data. It is unclear whether a standard bootstrap theorem can be applied to such random sets since the uniform control that one obtains under Donsker-type conditions in empirical process theory is absent. Consequently, assumptions weaker than selection consistency are not effective in controlling the randomness of the S^b's. Meanwhile our simulations suggest the validity of SPARES when only (A3) holds instead of (B3). Under assumption (B3), the asymptotic variance of ours converges to the best variance of an unbiased estimator of β_j^0 under the reduced model

$$\boldsymbol{Y} = \boldsymbol{X}_{S_0 \cup j} \beta^0_{S_0 \cup j} + \varepsilon.$$

Such bound is smaller than the semi-parametric information bound that involves all p covariates (*Belloni et al.*, 2014; *Van de Geer et al.*, 2014). Nevertheless the sets of conditions for the mentioned works and ours are quite different that they might not be directly comparable.

2.4 Inference by SPARES

2.4.1 Estimator of Standard Errors

As shown in Theorem (II.4), $\widehat{\beta}_j$ converges to a normal distribution whose variance depends on the unknown active set $S_{0,n}$. We propose an implementable approach to estimating the standard error of $\widehat{\beta}_j$ using Theorem 1 of *Efron* (2014), see also *Wager et al.* (2014) and Theorem 9 of *Wager and Athey* (2018). We denote the estimator as $\widehat{\text{se}}_j^B$. For the b^{th} bootstrap data, D_1^b, we re-write the index set as $I_1^b = (i_{b1}, i_{b2}, ..., i_{n/2})$. For $i = 1, 2, ..., n$ define $I_{bi} = \#\{i_{bk} = i\}$, the number of times that the i^{th} observation appears in the b^{th} re-sample. The vector $I^b = (I_{b1}, I_{b2}, ..., I_{bn})$ then follows a multinomial distribution with $n/2$ draws on n outcomes each having probability $1/n$, whose mean vector and covariance matrix are

$$I^b \sim \left(\frac{1}{2}\mathbf{1}_n, \frac{1}{2}\mathbf{I}_n - \frac{1}{2n}\mathbf{1}_n^{\mathrm{T}}\mathbf{1}_n\right)$$

where $\mathbf{1}_n$ the (column) vector of n 1's and $\mathbf{I}_n$ the $n \times n$ identity matrix. The nonparametric delta method estimator of the standard error is then given by:

$$\widehat{\mathrm{se}}_j^B = \left(\sum_{i=1}^{n} \widehat{\mathrm{cov}}_{ij}^2\right)^{1/2}, \tag{2.6}$$

where

$$\widehat{\mathrm{cov}}_{ij} = \sum_{b=1}^{B}(I_{bi} - \overline{I_{\cdot i}})(\widehat{\beta}_j^b - \widehat{\beta}_j)/B$$

is the bootstrap covariance between I_{bi} and $\widehat{\beta}_j^b$, and $\overline{I_{\cdot i}} = \sum_{b=1}^{B} I_{bi}/B$.

As emphasized in *Efron* (2014), the merit of smoothing the SPARE estimator is to convert a "jumpy" selection-based estimator $\widehat{\beta}^b$ into a smooth version of $\widehat{\beta}$. It is pointed out in *Wager et al.* (2014) that the nonparametric delta method standard error estimator tends to be biased upwards when the number of bootstraps is small. They proposed an alternative bias-corrected version of (2.6):

$$\widehat{\mathrm{se}}_U^B = \left\{(\widehat{\mathrm{se}}^B)^2 - \frac{n}{2B^2}\sum_{b=1}^{B}(\widehat{\beta}^b - \widehat{\beta})^2\right\}^{1/2} \tag{2.7}$$

Note that (2.7) converges to (2.6) as $B \to \infty$. The original version (2.6) would require $B = O(n^{1.5})$ to reduce Monte Carlo noise down to the level of sampling noise, while (2.7) only requires $B = O(n)$. Moreover, our experience shows that the unbiased version does converge to the empirical standard error faster than the original one.

2.4.2 Confidence Intervals and P-values

Following previous discussion, the asymptotic $1-\alpha$ confidence interval for each β_j^0 is given by

$$\left(\widehat{\beta}_j - \Phi^{-1}(1-\alpha/2)\widehat{\mathrm{se}}_j^B \ , \ \widehat{\beta}_j + \Phi^{-1}(1-\alpha/2)\widehat{\mathrm{se}}_j^B\right),$$

where Φ^{-1} is the inverse CDF of the standard normal distribution. The p-value of testing $H_0 : \beta_j = 0$ is

$$p_j = 2 \times \left\{1 - \Phi\left(|\widehat{\beta}_j|/\widehat{\mathrm{se}}_j^B\right)\right\}. \tag{2.8}$$

2.5 Extension of SPARES to a Subvector $\beta^{(1)}$ with a Fixed Dimension

It is natural to extend our procedure to a subvector $\beta^{(1)}$ of β^0 with a fixed dimension $p_1 \geq 2$. Without loss of generality, assume that $\beta^{(1)} = \beta^0_{S^{(1)}} = (\beta_1^0, \beta_2^0, .., \beta_{p_1}^0)^{\mathrm{T}}$ with $|S^{(1)}| = p_1$. Accordingly, we modify the SPARE estimator in (2.2) to be

$$\widehat{\beta}^b_{S^{(1)}} = \left\{ ({\boldsymbol{X}^b_{S^b \cup S^{(1)}}}^{\mathrm{T}} \boldsymbol{X}^b_{S^b \cup S^{(1)}})^{-1} {\boldsymbol{X}^b_{S^b \cup S^{(1)}}}^{\mathrm{T}} \boldsymbol{Y}^b \right\}_{S^{(1)}},$$

which gives a corresponding SPARES estimator for $\beta^{(1)}$:

$$\widehat{\beta}^{(1)} = \frac{1}{B}\sum_{b=1}^{B} \widehat{\beta}^b_{S^{(1)}}. \tag{2.9}$$

The corresponding extension of Theorem II.4 is stated below.

Theorem II.5. *Consider model (2.1) under assumptions (A1,A2,B3), and a fixed finite subset $S^{(1)} \subset \{1, 2, .., p\}$ with $|S^{(1)}| = p_1$. Let $\widehat{\beta}^{(1)}$ be the SPARES estimator for $\beta^{(1)} = \beta^0_{S^{(1)}}$ as defined in (2.9). There exist random vectors $Z^{(1)}, \Delta^{(1)}$, such that as*

$n, B \to \infty$,

$$\sqrt{n}(\widehat{\beta}^{(1)} - \beta^{(1)}) = Z^{(1)} + \Delta^{(1)}, \quad {\boldsymbol{\Sigma}^{(1)}}^{-1/2} Z^{(1)} \to N(0, \mathbf{I}_{p_1}), \quad \Delta^{(1)} = o_p(\mathbf{1}_{p_1}),$$

and $\boldsymbol{\Sigma}^{(1)} = \sigma^2 \left(\boldsymbol{\Sigma}^{-1}_{S_{0,n} \cup S^{(1)}} \right)_{S^{(1)}}$ *is positive definite.*

Remark II.6. There is also a direct extension of the one-dimensional nonparametric delta method for estimating the variance-covariance matrix of $\widehat{\beta}^{(1)}$, $\widehat{\boldsymbol{\Sigma}}^{(1)} = \widehat{\mathrm{COV}}_{(1)}^{\mathrm{T}} \widehat{\mathrm{COV}}_{(1)}$, where

$$\widehat{\mathrm{COV}}_{(1)} = \left(\widehat{\mathrm{cov}}_1^{(1)}, \widehat{\mathrm{cov}}_2^{(1)}, .., \widehat{\mathrm{cov}}_n^{(1)} \right)^{\mathrm{T}}$$
$$\widehat{\mathrm{cov}}_i^{(1)} = \sum_{b=1}^{B} (I_{bi} - I_{\cdot i})(\widehat{\beta}^b_{S^{(1)}} - \widehat{\beta}^{(1)})/B.$$

The extension to a subvector $\beta^{(1)}$ with a fixed dimension allows us to derive confidence regions for a subset of variables of interest and test for contrasts of certain predictors.

2.6 Simulation Studies

We designed all simulation scenarios based on the linear model (2.1) with $\boldsymbol{X} = (X_1, ...X_p) = (\mathbf{x}_1^{\mathrm{T}}, ..., \mathbf{x}_n^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)^{\mathrm{T}}$, assuming $\mathbf{x}_i$'s i.i.d. $\sim N(\mathbf{0}_p, \boldsymbol{\Sigma}_{p\times p})$ and ε_i's i.i.d. $\sim N(0, 1)$. A total of 200 simulated datasets were generated for each simulation configuration.

We first illustrate the advantage of using SPARES over one-time SPARE. We set sample size $n = 200$, number of predictors $p = 300$, and $s_0 = 3$ nonzero signals with $\boldsymbol{\Sigma}_{p\times p}$ being the identity matrix. As shown in Table (A.1), over 200 replications, the biases of both approaches are negligible on average, but the standard errors of SPARES are much smaller than those of one-time SPARE, which results in higher power and more accurate inferences. Thus we recommend SPARES in practice.

In subsection 6.1, we explore the performance of SPARES under various settings, including different correlation structures of $\boldsymbol{X}$, strong and weak signals strength, and stress tests with ultrahigh dimensionality. In subsection 6.2, we compare SPARES with two de-biased LASSO estimators, LASSO-Pro from *Van de Geer et al.* (2014) and SSLASSO from *Javanmard and Montanari* (2014).

2.6.1 Performance of SPARES under Various Settings

We will go over three examples, all of which assume the linear model (2.1) as truth, but with different parameters.

Example 1. Let sample size $n = 150$, number of predictors $p = 300$, number of nonzero signals $s_0 = 5$, and a fixed realization of β^0 where $S_{0,n} = \{66, 97, 145, 166, 173\}$ was a fixed realization of s_0 draws without replacement from $[p]$ and $\beta^0_{S_{0,n}} = (1, 0.6, -1, -0.6, 1)$. We examined three commonly used correlation structures: identity; first-order autoregressive (AR(1)) with $\rho = 0.5$; compound symmetry (CS) with $\rho = 0.5$. LASSO was used as the selection procedure $\mathcal{S}_\lambda$, while λ was chosen by cross-validation. As summarized in Table (2.1), for both nonzero signals and noise variables, the bias of SPARES estimator was well controlled while the SE estimates were very close to the empirical ones. Consequently, the coverage probabilities of the 95% confidence intervals were at the nominal level. In addition, the variable selection frequency based on p-values of SPARES was higher for true signals and much lower for noise variables compared to selection by LASSO. Notice that for identity and AR(1) correlation structures, the selection frequencies of the true signals were uniformly close to 1, suggesting "sure screening" condition was met and thus the better coverage probabilities. Therefore the simulation result validates our claim that SPARES works under "sure screening" assumption.

Example 2. Let $n = 150,\ p = 500$, and

- Example 2.1: $s_0 = 15$, $\boldsymbol{\Sigma}_{p\times p} = \text{diag}(\boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_{10})$, where each $\boldsymbol{\Sigma}_k$ was 50×50

with an AR(1) correlation structure, $(\mathbf{\Sigma}_k)_{ij} = (0.1k - 0.1)^{|i-j|},\ k = 1, 2, .., 10$. The active set $S_{0,n}$ was a fixed realization of s_0 draws without replacement from $[p]$, and $\beta^0_{S_{0,n}}$ was a fixed realization of s_0 i.i.d. Uniform $U[0, 2]$ variables;

- Example 2.2: $s_0 = 20$, $\mathbf{\Sigma}_{p\times p} = \text{diag}(\mathbf{\Sigma}_1, ..., \mathbf{\Sigma}_{10})$, where each $\mathbf{\Sigma}_k : (\mathbf{\Sigma}_k)_{ij} = (0.3)^{|i-j|}$. The non-zero signals are assigned effect sizes $\beta^0_{50k-45} = 0.2k,\ \beta^0_{50k-15} = -0.2k$ for $k = 1, 2, ..., 10$.

We applied SPARES with LASSO (10-fold cross validation to choose λ) as the model selection procedure, and reported the simulation averages of $\widehat{\beta}_{\text{SPARES}}$, along with confidence intervals, mean biases, coverage probabilities, and type I errors for testing $H_0 : \beta^0_j = 0$. The results are summarized in Figures (A.1) and (A.2). For the true signals $j \in S_{0,n}$, the proposed method worked well regardless of the correlation, with negligible biases and close-to-nominal coverage probabilities. On the other hand, the biases for the estimates of noise variables were enlarged when they were highly correlated with non-zero signals. The estimated coverage probabilities and type I errors deviated more from the nominal level consequently. The type I error became negligible when the effect size was over 1. Coupled with an observation that the bias was larger for the noise variables that were correlated with moderate non-zero signals, our takeaway was that the magnitude of bias was a combination of selection errors as well as correlations with true signals.

Example 3 serves as a "stress test" to illustrate how SPARES handle large datasets with a number of "weak signals". We let $n = 500$, $p = 1000,\ 5000$ and 10000, and $s_0 = 205$. Within the 205 non-zero signals, 5 are of sizes $0.2, 0.4, 0.6, 0.8, 1$, and the rest 200 are fixed random realizations from the uniform distribution $U[(-0.2, -0.1) \cup (0.1, 0.2)]$. The multivariate normal distribution with mean zero and the AR(1) correlation structure with $\rho = 0.5$ is applied to generate $\boldsymbol{X}$'s. As summarized in Table (2.2), the SPARES estimator remains nearly unbiased for both strong and weak signals. The coverage probabilities of strong signals are close to the nominal level 0.95,

while those for weak and zero signals are above 0.9 on average. This demonstrates that SPARES is rather reliable and robust even for large datasets with a number of weak signals.

2.6.2 Comparisons with De-biased LASSO Estimators

We compared SPARES with different versions of de-biased LASSO estimators in Example 4, where the active set $S_{0,n} \subset \{1, 2, .., p\}$ was a fixed random realization with size $|S_{0,n}| = 5$, and $\beta^0_{S_{0,n}}$ was a fixed realization of 5 i.i.d. random variables from uniform $U[0.5, 2]$. The size of the active set is reduced to 5 for clearer comparison and display of the result. Three correlation structures are considered for completeness:

- Example 4.1: Identity $\mathbf{\Sigma}_{p\times p} = \mathbf{I}_{p\times p}$;
- Example 4.2: AR(1) $\mathbf{\Sigma}_{p\times p} : (\mathbf{\Sigma})_{jk} = (0.8)^{|j-k|}$;
- Example 4.3: Compound symmetry $\mathbf{\Sigma}_{p\times p} : (\mathbf{\Sigma})_{jk} = 0.5$.

The estimated biases and coverage probabilities were shown in Table (2.3) and Figure (A.3), where LASSO-Pro was proposed in *Van de Geer et al.* (2014) and SSLASSO was from *Javanmard and Montanari* (2014).

Across the board, SPARES gave less biased point estimates for the true signals, and provided reliable confidence intervals around the nominal level for both true signals and noise variables. In contrast, both LASSO-Pro and SSLASSO had visible discrepancies between the true signals and noise variables. While LASSO-Pro had lower-than-nominal level coverages for the true signals, it performed even worse in Example 4.1, probably due to the fact that the node-wise LASSO was not ideal when estimating the precision matrix when $\mathbf{\Sigma}_{p\times p}$ was an identity matrix. As far as SSLASSO was concerned, the confidence intervals for the noise variables were too conservative, while the coverages for the true signals in Example 4.2 were considerably low.

In summary, the performance of SPARES aligned well with the theoretical expectations, especially for the active set $S_{0,n}$. We did observe, however, some false-positives when the noise variables were highly correlated with those in the active set. Nevertheless, compared with the de-biased LASSO methods, SPARES showed substantial improvement by providing less biased estimates with more accurate coverage probabilities close to the nominal level.

2.7 Data Examples

2.7.1 Riboflavin Production Data

We applied our method to analyze a dataset on riboflavin (vitamin B_2) production by bacillus subtilis, made public by *Bühlmann et al.* (2014) and analyzed by *Meinshausen et al.* (2009), *Bühlmann et al.* (2014), *Van de Geer et al.* (2014) and *Javanmard and Montanari* (2014). The data contained $n = 71$ samples and $p = 4088$ covariates, measuring the logarithm of the expression levels of 4088 genes. The response variable was the logarithm of the riboflavin production rate.

We related the response to the gene expressions using the linear model (2.1). We checked the collinearity among the genes, and their pairwise correlations are plotted on the left panel of Figure (A.4). We further normalized the genes so that their effect sizes are comparable. The LASSO was used as the variable selection method, and we let $B = 1000$ be the number of re-samples. Assisted by the LASSO selection, we derive the SPARES estimator $\widehat{\beta}$, the standard error estimates as in (2.6), and the p-values as in (2.8). With a standard Bonferroni correction to adjust FWER to the 5% significance level, we identified four genes that were significantly associated with the response, namely YCKE_at, XHLA_at, YXLD_at, and YDAR_at. If the FWER were set at 10%, one more gene, YCGN_at, would be included. The confidence intervals for the top 5 genes are displayed on the right panel of Figure (A.5), with the point

estimates shown in Table (2.4). By contrast, the results from other methods were less informative. For example, with a 5% FWER, the multisample-splitting method proposed in *Meinshausen et al.* (2009) identified YXLD_at, *Van de Geer et al.* (2014) claimed none, and *Javanmard and Montanari* (2014) only detected YXLD_at and YXLE_at, which are highly correlated themselves.

Our results had biological interpretations that are confirmed by the literature. It was reported that XHLA_at was involved in cell lysis upon induction of PbsX (*Kunst et al.*, 1997), increasing the capability to produce recombinant extracellular digestive enzymes that results in riboflavin production (7.04 in *Mander and Liu* (2010)). YCKE_at, formally named as bglC, was also responsible for the production of certain enzyme, Aryl-phospho-beta-D-glucosidase, and had extracellular protein secretory functions (*Schallmey et al.*, 2004). YXLD_at, together with YXLE_at, was important for negative regulation of sigma Y activity (*Tojo et al.*, 2003).

2.7.2 Multiple Myeloma Genomic Data

We analyzed a cancer genomic data with $n = 163$ multiple myeloma patients. Our interest lay in detecting the association between the β-2 microglobulin (B2M) and gene expressions. B2M is a small membrane protein produced by malignant myeloma cells, indicating the severity of disease. Identifying genes that are related to B2M is clinically important as it helps construct molecular prognostic tools for early diagnosis of disease.

We first used KEGG (*Carlson*, 2015) to identify gene pathways that are related to cancer development and progression, as well as some identified upstream genes that may regulate B2M. In total, there were $p = 789$ unique probes belonging to these pathways. We took the logarithm transformation for both the B2M test value and the gene expressions as our response and predictors for model (2.1). We applied SPARES with LASSO as the selection method, and $B = 500$ re-samples were drawn

for smoothing.

Our method offers additional biological insight compared to the other methods. As shown in Table (2.5) and Figure (A.6), it identified two significant probes at 5% FWER after the Bonferroni correction, namely 204171_at (RPS6KB1) and 202076_at (BIRC2). In contrast, the two de-biased LASSO estimators identified no significant probes. Both detected genes are highly associated with malignant tumor cells: RPS6KB1, member of the ribosomal protein S6 kinase (RPS6K) family, altercation/mutation has been related to numerous types of cancer including breast cancer, colon cancer, non-small-cell lung cancer, and prostate cancer (*Sinclair et al.*, 2003; *Van der Hage et al.*, 2004; *Slattery et al.*, 2011; *Zhang et al.*, 2013; *Cai et al.*, 2015); BIRC2, whose encoded protein is a member of inhibitors of apoptotic proteins (IAPs) that inhibits apoptosis by binding to tumor necrosis factor receptor-associated factors TRAF1 and TRAF2 (*Saleem et al.*, 2013), has been related to lung cancer and lymphoma (*Wang et al.*, 2010; *Rahal et al.*, 2014).

2.8 Conclusion

We have proposed a new framework of estimation and inference for the high-dimensional linear models (2.1), and shown the proposed SPARES estimator is asymptotically unbiased and normal, giving accurate and reliable component-wise inferences. The key improvement, compared to the existing works, lies in these aspects. SPARES converts the high-dimensional problem of estimating the p-vector β^0 to the low dimensional case by Selection-assisted Partial Regression. Thus we avoid the curse of dimensionality on estimation and inference. SPARES is applicable to general selection methods including LASSO, SCAD, screening, boosting, and etc., as long as they possess the desired selection consistency property, which is likely to be loosened to sure screening property in practice as suggested in the extensive simulation study. SPARES is not sensitive to the tuning parameter λ in $\mathcal{S}_\lambda$, since it is not directly used

for estimation, but only involved in the selection. Hence, our method has minimal requirements on extra model parameters and is almost robust toward selection of tuning parameters. This framework can be naturally extended to other non-linear regression models, such as generalized linear model and Cox model, through two general steps. First, we perform data-splitting on the original data, and then do selection on one half of the data followed by fitting low-dimensional model on the other half of the data using partial regression; Second, we repeat the first step many times and average over all estimates to form a smoothed estimate. We will report this work elsewhere.

Supporting Information

Additional supporting information can be found in Appendix A, including Proofs, Tables, and Figures referenced in Sections 2.2-2.7.

Software R implementation of SPARES is available on-line at `https://github.com/feizhe/SPARES`, along with the simulation examples.

Table 2.1: Performance of SPARES under simulation Example 1 with three correlation structures: Identity, AR(1) and Compound Symmetry (CS). The last column "-" represents the averages for all noise variables. **Freq $\mathcal{S}_\lambda$** is the selection frequency by LASSO; **Freq SPARES** is the selection frequency by p values of SPARES with 0.1 FDR control; **Empirical SE** is the empirical standard error.

	Index j	66	97	145	166	173	-
	β_j^0	1	0.6	-1	-0.6	1	0
Identity	Bias ($\times 10^{-3}$)	16	-1	-2	2	7	-1
	Average $\widehat{\text{se}}_j^B$	0.110	0.111	0.109	0.111	0.110	0.111
	Empirical SE	0.117	0.109	0.104	0.113	0.124	0.109
	Cov Prob (%)	91.5	94.0	95.0	96.0	91.5	94.8
	Freq $\mathcal{S}_\lambda$	1	0.956	1	0.965	1	0.059
	Freq SPARES	1	0.97	1	0.99	1	0.003
AR(1)	Bias ($\times 10^{-3}$)	-6	2	7	10	-1	0
	Average $\widehat{\text{se}}_j^B$	0.115	0.116	0.114	0.115	0.116	0.115
	Empirical SE	0.125	0.108	0.114	0.120	0.108	0.114
	Cov Prob (%)	93.5	96.0	95.0	92.5	96.5	94.5
	Freq $\mathcal{S}_\lambda$	0.998	0.938	1.000	0.929	1.000	0.046
	Freq SPARES	1	0.925	1	0.905	1	0.001
CS	Bias ($\times 10^{-3}$)	-12	-30	6	7	-14	-7
	Average $\widehat{\text{se}}_j^B$	0.151	0.149	0.152	0.150	0.150	0.154
	Empirical SE	0.165	0.161	0.168	0.162	0.163	0.154
	Cov Prob (%)	92.5	91.5	89.4	92.0	92.0	94.5
	Freq $\mathcal{S}_\lambda$	0.986	0.742	0.958	0.651	0.988	0.045
	Freq SPARES	1	0.775	1	0.795	1	0.005

Table 2.2: Performance of SPARES under simulation Example 3. Tables from top to bottom correspond to $p = 1000$, 5000 and 10000. Last two columns are averages over small and zero signals.

Index	36	272	376	568	915	Small	0's
β^0	0.200	0.400	0.600	0.800	1.000		0.000
				$p = 1000$			
Bias	0.013	-0.006	0.014	-0.002	-0.014	0.005	0.004
Avg SE	0.093	0.093	0.093	0.093	0.093	0.093	0.093
Emp SE	0.099	0.098	0.098	0.093	0.097	0.094	0.094
Cov Prob	0.960	0.920	0.930	0.930	0.940	0.907	0.908
Sel freq	0.045	0.418	0.930	1.000	1.000	0.021	0.002
				$p = 5000$			
Bias	-0.005	0.009	0.010	0.003	0.004	0.004	0.000
Avg SE	0.093	0.093	0.095	0.094	0.094	0.094	0.094
Emp SE	0.092	0.096	0.098	0.099	0.112	0.095	0.096
Cov Prob	0.960	0.930	0.960	0.910	0.920	0.905	0.935
Sel freq	0.022	0.390	0.906	0.999	1.000	0.015	0.001
				$p = 10000$			
Bias	-0.003	0.003	0.006	0.008	-0.025	0.005	0.000
Avg SE	0.094	0.094	0.094	0.095	0.094	0.095	0.095
Emp SE	0.094	0.096	0.101	0.103	0.093	0.096	0.097
Cov Prob	0.950	0.940	0.930	0.930	0.950	0.902	0.939
Sel freq	0.015	0.313	0.860	0.996	1.000	0.012	0.000

Table 2.3: Comparisons of SPARES with LASSO-Pro and SSLASSO under Example 4. The rows consist of 5 true signals and the average of zero signals. In each cell, top number is for SPARES; middle number is for LASSO-Pro; lower number is for SSLASSO.

		Example 4.1		Example 4.2		Example 4.3	
Index	β_j^0	Bias ($\times 10^{-3}$)	Cov Prob (%)	Bias ($\times 10^{-3}$)	Cov Prob (%)	Bias ($\times 10^{-3}$)	Cov Prob (%)
		-1.77	90.5	10.43	92.5	-0.35	96.5
		-81.78	70.5	-44.09	86	-38.43	92.5
78	1.07	-79.33	90.5	-101.95	84.5	-113.72	92.5
		-1.04	96.5	9.70	92	2.44	95
		-80.28	76	-44.54	87	-32.42	89
102	1.04	-77.72	93.5	-99.66	82	-105.60	92
		-1.62	94	15.58	93.5	-4.67	96.5
		-89.43	71.5	-47.57	88.5	-40.39	91.5
242	1.19	-88.69	87.5	-104.25	84	-115.51	92
		-0.14	94	2.98	96.5	2.01	95
		-75.87	81	-41.40	88	-30.61	91
359	1.43	-80.91	94	-98.14	85	-107.5	89
		-3.57	95.5	0.54	93	5.88	91.5
		-84.86	75	-60.80	88	-24.20	86.5
380	0.62	-85.73	89.5	-111.11	81.5	-99.26	90.5
		-0.46	95	0.65	94.82	3.26	95.16
		-0.40	97	3.16	96.46	5.24	96.34
-	0	-0.27	99.5	4.15	99.69	26.88	99.94

Table 2.4: Analysis of the riboflavin genomic data. $\widehat{\beta}$ is the SPARES estimator; p-values are adjusted by Bonferroni correction (multiplied by p). The top 10 and bottom 10 most/least significant genes are tabulated.

Gene	$\widehat{\beta}$	SE	Adjusted p-value
YCKE_at	0.37	0.06	< 0.001
XHLA_at	0.48	0.09	< 0.001
YXLD_at	-0.53	0.11	0.01
YDAR_at	-0.28	0.06	0.01
YCGN_at	-0.31	0.07	0.09
RPLJ_at	-0.26	0.06	0.10
YQIZ_at	-0.25	0.06	0.13
YCDH_at	-0.27	0.07	0.15
SPOIISA_at	0.25	0.06	0.35
YRPE_at	-0.25	0.07	0.63
...			
YXAL_at	-2×10^{-4}	0.09	1
XPT_at	-1.6×10^{-4}	0.07	1
YOZG_at	-2.9×10^{-4}	0.14	1
YOJB_at	1.7×10^{-4}	0.10	1
YBCL_at	-1.8×10^{-4}	0.11	1
YJAX_at	1.3×10^{-4}	0.09	1
YOSE_at	1.1×10^{-4}	0.11	1
YUNA_at	4.9×10^{-5}	0.07	1
YISO_at	1.7×10^{-5}	0.08	1

Table 2.5: Analysis of the Multiple Myeloma genomic data. The top 6 and bottom 6 most/least significant genes are tabulated.

Gene	$\widehat{\beta}$	SE	Adjusted p
204171_at (RPS6KB1)	-0.20	0.042	0.002
202076_at (BIRC2)	-0.17	0.041	0.037
220414_at	-0.20	0.05	0.14
220394_at	-0.18	0.05	0.59
206493_at	-0.19	0.06	0.63
209878_s_at	-0.17	0.05	0.69
...			
207924_x_at	5×10^{-4}	0.07	1
205289_at	-4.4×10^{-4}	0.06	1
203591_s_at	4.7×10^{-4}	0.07	1
224229_s_at	2.4×10^{-4}	0.06	1
217576_x_at	2.5×10^{-4}	0.07	1
201656_at	2.5×10^{-4}	0.08	1

CHAPTER III

Estimation and Inference for High Dimensional Generalized Linear Models: A Split and Smoothing Approach

3.1 Introduction

Lung cancer is the leading cause of cancer-related deaths in the United States, among both men and women (*US Department of Health and Human Services*, 2004; *Parkin et al.*, 2005). Understanding the molecular mechanisms of lung cancer is a focus of current basic and translational research. The Boston Lung Cancer Study Cohort (BLCSC) (*Christiani*, 2017) is a cancer epidemiology cohort of over 11,000 lung cancer cases enrolled at Massachusetts General Hospital and the Dana-Farber Cancer Institute from 1992 to present. In addition, controls are recruited at the hospital from healthy friends and nonblood-related family members (usually spouses) of the patients. This is the first and most comprehensive lung cancer survivor cohort with a long follow-up period, which has been growing with more patients recruited every year. For both groups, large scale data of various types, including gene expression, methylation, SNP, and CT imaging, have been measured and recorded. The rich data generated from the BLCSC cohort allow powerful translational research and exploration of potential predictors for lung cancer.

Using a target gene approach, this chapter analyzes high dimensional SNP data from 708 lung cancer patients and 751 controls, with more than $6,800$ SNPs from 15 cancer related genes, along with important demographic variables, such as age, gender, race, education level, and smoking status. Our goal is to model the binary lung cancer indicator as the outcome and to estimate and test the effects of the potential predictors that could explain the differences between the cases and controls. Since smoking plays a vital role in lung cancer, we are especially interested in the interaction terms between SNPs and the smoking status in addition to their main effects.

It has been challenging to construct confidence intervals, perform statistical tests and assign uncertainty measures in sparse high dimensional models (*Dezeure et al.*, 2015). The high dimensionality impedes accurate estimation of all the potential predictors, and evaluation of the uncertainty of the estimators. The high dimensionality considered in this chapter includes but is not limited to the usual case of "$p > n$," such as $n = 500$ samples with $p = 1000$ covariates. Even in a "$p < n$" setting such as $n = 1000$ samples with $p = 500$ covariates, direct applications of the GLM framework would lead to ambiguous and meaningless estimations and inferences. Alternatively, penalized regressions have been widely used to deal with high dimensionality (*Friedman et al.*, 2010; *Van de Geer*, 2008; *Candès and Tao*, 2007; *Lv and Fan*, 2009; *Huang et al.*, 2008; *Zou and Hastie*, 2005).The estimators from penalized regressions are shrunk and thus "irregular" as their asymptotics become difficult to track. There has been considerable success in drawing inferences based on penalized regressions, mostly for linear models (*Zhang and Zhang*, 2014; *Javanmard and Montanari*, 2014; *Bühlmann et al.*, 2014; *Dezeure et al.*, 2015). Meanwhile, there are limited works in the GLM settings. In sparse high dimensional GLMs, *Bühlmann et al.* (2014) offered the generalization of de-sparsified LASSO, while *Ning and Liu* (2017) proposed the decorrelated score tests for penalized M-estimators. In the presence of high dimen-

sional controls, *Belloni et al.* (2014, 2016) proposed a post double selection procedure for estimation and inference; *Lee et al.* (2016) characterized the distribution of a post-LASSO-selection estimator conditioned on the *selected variables*, but only for linear regressions. The performance of most of these methods depends heavily on multiple tuning parameters, and their optimal choices are often not apparent in practice. As we will see in the data analysis, the application of existing works is also limited to the scale of data due to computation burdens.

We propose a novel approach of simultaneous estimation and inference for high dimensional generalized linear models that aims to resolve the aforementioned limitations. We first introduce a one-time estimator by splitting the data into two halves, using one half to select a subset of important variables as the "candidates." On the other half, we fit a low dimensional GLM with the union of the parameter of interest (or a low dimensional subset of the coefficient vector) and the candidate set of variables (*Belloni et al.*, 2016). The one-time estimator for the parameter of interest is then from the fitted low dimensional GLM. While the one-time estimator is unbiased and asymptotically normal under mild conditions, it is highly variable, and heavily dependents on the specific one-time selection. Therefore, we further propose the smoothed estimator by repeating the previous procedure a large number of times and averaging the resulting estimators. The smoothed estimator is proved to possess the same desired theoretical properties with improved efficiency and practical performance. Our approach is shortened as SSGLM, where "SS" stands for "splitting and smoothing." Thus, our idea takes advantage of the multi sample-splitting method in *Meinshausen et al.* (2009), and the bagging idea (*Bühlmann and Yu*, 2002; *Friedman and Hall*, 2007; *Efron*, 2014), and is therefore fundamentally different from penalized regressions. In this way, we reduce the high dimensional inference problem into low dimensional estimations that are free of penalization/regularization. As variable/model selection only plays an assistive role, our procedure is not sensitive to the

tuning parameters, which is a major drawback of the existing methods (*Bühlmann et al.*, 2014; *Ning and Liu*, 2017). Furthermore, we derive the variance estimator using the non-parametric delta method adapted to the splitting and smoothing procedure (*Efron*, 2014; *Wager and Athey*, 2018), which is free of the parametric model (GLM in this case) and achieves variance reduction from the effect of bagging (*Bühlmann and Yu*, 2002). Our framework also facilitates hypothesis testing or drawing inferences on predetermined contrasts in the presence of high dimensional nuisance parameters.

The rest of the chapter is organized as follows. Section 3.2 describes the SS-GLM and Section 3.3 introduces the theoretical properties. Section 3.4 provides the inferential procedure and Section 3.5 extends it to accommodate any subvectors of parameters of interest. Section 3.6 provides simulations and comparisons with the existing methods. Section 3.7 reports the results of the analysis of the BLCSC SNP data. We conclude the chapter with a brief discussion.

3.2 Method

3.2.1 Notation

We denote the observations as $(Y_i, \mathbf{x}_i)$ for $i = 1, 2, .., n$, where $\mathbf{x}_i = (x_{i1}, x_{i2}, .., x_{ip})$ is the $1 \times p$ covariate vector and the outcome distribution belongs to a linear exponential family, which includes Normal, Bernoulli, Poisson, and other distributions,

$$f(Y_i|\theta_i) = \exp\left\{Y_i\theta_i - A(\theta_i) + c(Y_i)\right\}, \tag{3.1}$$

where θ_i is the parameter relating to the mean. In this chapter, we consider the canonical link with $\theta_i = \overline{\mathbf{x}}_i\beta$, where $\overline{\mathbf{x}}_i = (1, \mathbf{x}_i)$ and $\beta = (\beta_0, \beta_1, .., \beta_p)^T$ includes an intercept term. Specifically,

$$\mathbf{E}\left(Y_i\right) = \mu_i = A'(\theta_i) = g^{-1}\Big(\overline{\mathbf{x}}_i\beta\Big), \tag{3.2}$$

and $\mathbf{V}(Y_i) = A''(\theta_i) = \nu(\mu_i)$, where μ_i's are the mean of the responses Y_i's and g is the link function. The collection of all n observations is denoted as (Y, X), where $Y = (Y_1, .., Y_n)^T$ and $X = (\mathbf{x}_1^T, .., \mathbf{x}_n^T)^T$. In addition, we write $X = (X_1, .., X_p)$ with p column vectors and $\overline{X} = (\mathbf{1}, X)$ to include the $1 \times n$ column vector $\mathbf{1}$.

In the BLCSC SNP data, the outcome of interest is the binary lung cancer indicator, and the covariate vector $\mathbf{x}_i$ includes the demographic variables, the SNP variables, and the interactions between the SNPs and smoking. The parameterization of the assumed logistic regression is

$$g(\mu_i) = \text{logit}(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right), \text{ with } A(\theta_i) = \log\left(1 + e^{\theta_i}\right).$$

We write the full log-likelihood of model (3.1) and (3.2) as

$$\ell(\beta) = \ell(\beta; Y, X) = \frac{1}{n}\sum_{i=1}^{n}\left\{Y_i\theta_i - A(\theta_i)\right\} = \frac{1}{n}\sum_{i=1}^{n}\left\{Y_i(\overline{\mathbf{x}}_i\beta) - A(\overline{\mathbf{x}}_i\beta)\right\}.$$

The score and the observed information are

$$U(\beta) = \frac{1}{n}\overline{X}^T\left\{Y - A'(\overline{X}\beta)\right\}; \quad \widehat{I} = \widehat{I}(\beta) = \frac{1}{n}\overline{X}^T V \overline{X},$$

where $V = \text{diag}\{\nu(\mu_1), .., \nu(\mu_n)\}$, and whenever a univariate function such as $A(\cdot)$ is applied to a vector, it denotes the vector of values of the function applied to each entry of the argument.

As our method involves low-dimensional estimation based on subsets of the covariates, we introduce some notation with respect to an index set $S \subset [p] = \{1, 2, .., p\}$. We write the subvectors $\mathbf{x}_{iS} = (x_{ij})_{j\in S}$ and $\overline{\mathbf{x}}_{iS} = (1, \mathbf{x}_{iS})$, and submatrices $X_S = (X_j)_{j\in S}$ and $\overline{X}_S = (\mathbf{1}, X_S)$. Given a set $S \subset [p]$ and an index $j \in [p]$, we define $S_{+j} = \{j\} \cup S$ and $S_{-j} = S \setminus \{j\}$. In addition, we let $S_{+0} = S_{-0} = S$ when concerning the intercept. Furthermore, we write $\beta_S = (\beta_0, \beta_j)_{j\in S}$, which always includes the

intercept and thus is of length $1 + |S|$.

The working log-likelihood with respect to (Y, X_S) and β_S is

$$\ell_S(\beta_S) = \ell(\beta_S; Y, X_S) = \frac{1}{n} \sum_{i=1}^{n} \{Y_i(\overline{\mathbf{x}}_{iS}\beta_S) - A(\overline{\mathbf{x}}_{iS}\beta_S)\}.$$

Similarly, $U_S(\beta_S) = \frac{1}{n}\overline{X}_S{}^T(Y - A'(\overline{X}_S\beta_S))$; $\widehat{I}_S = \widehat{I}_S(\beta_S) = \frac{1}{n}\overline{X}_S{}^T V_S \overline{X}_S$, where $V_S = \text{diag}\{A''(\overline{\mathbf{x}}_{iS}\beta_S), .., A''(\overline{\mathbf{x}}_{iS}\beta_S)\}$. Now we write out the expected information with respect to β as $I = \mathbf{E}_\beta(\nabla^2\ell(\beta))$, and the *sub-information* I_S is the submatrix of I with rows and columns corresponding to S. Note I_S can also be written as I_{SS} so that the subscripts reflect both rows and columns. The truth of β is denoted as $\beta^* = (\beta_0^*, \beta_1^*, .., \beta_p^*)$, $I^* = \mathbf{E}_{\beta^*}(\nabla^2\ell(\beta^*))$, and I_S^* is the submatrix of I^* analog to I_S. Lastly, we define the partial information for $j \in [p]$ and given $S \subset [p]$ as

$$I_{j|S} = I_{jj} - I_{jS_{-j}} I_{S_{-j}S_{-j}}^{-1} I_{S_{-j}j}, \tag{3.3}$$

where $I_{jj}, I_{jS_{-j}}$, and $I_{S_{-j}S_{-j}}$ are the entry, subvector, and submatrix of I with respect to the respective subscripts.

3.2.2 Proposed SSGLM Estimator

Assume the data (Y, X) follows the generalized linear model (3.1) and (3.2) with the true parameter vector $\beta = \beta^*$. We first define the one-time SSGLM estimator $\widetilde{\beta} = (\widetilde{\beta}_j)_{j=0,1,..,p}$ based on a single data split (**Algorithm 3.1**). We split the data into two halves D_1 and D_2, with sample sizes $|D_1| = n_1$, $|D_2| = n_2$, $n_1 + n_2 = n$. For example, $n_1 = n_2 = n/2$. Next on D_2, we select a subset of important covariates $S \subset [p]$, $s = |S| < n_1 - 1$ via a selection scheme $\mathcal{S}_\lambda$, where λ is the regularization parameter. The selected set S is used as the candidate set of covariates for performing low dimensional estimation on D_1. On $D_1 = (Y^1, X^1)$, and for each $j \in [p]$, we fit a low dimensional GLM by regressing Y^1 on $X^1_{S_{+j}}$, where the set $S_{+j} = \{j\} \cup S$ is

as defined before. Denoting the Maximum Likelihood Estimator (MLE) of the fitted model as $\widetilde{\beta}^1$, we define the one-time estimator as $\widetilde{\beta}_j = \left(\widetilde{\beta}^1\right)_j$, which is the entry of $\widetilde{\beta}^1$ corresponding to covariate X_j. Meanwhile, we denote $\widetilde{\beta}_0$ as the intercept estimator from the MLE of $Y^1 \sim X^1_S$. With some abuse of notation that $S_{+0} = \{0\} \cup S = S$, the one-time SSGLM estimator is

$$\widetilde{\beta}^1 = \operatorname{argmin} \ell_{S_{+j}}(\beta_{S_{+j}}) = \operatorname{argmin} \ell(\beta_{S_{+j}}; Y^1, X^1_{S_{+j}}); \tag{3.4}$$

$$\widetilde{\beta}_j = \left(\widetilde{\beta}^1\right)_j; \quad \widetilde{\beta} = (\widetilde{\beta}_0, \widetilde{\beta}_1, .., \widetilde{\beta}_p). \tag{3.5}$$

If the outcome is linear (*Fei et al.*, 2018), (3.4) and (3.5) have an explicit form $\widetilde{\beta}_j = \left\{ (X^1_{S_{+j}}{}^{\mathrm{T}} X^1_{S_{+j}})^{-1} X^1_{S_{+j}}{}^{\mathrm{T}} Y^1 \right\}_j$.

The SSGLM estimator is defined by repeating the previous procedure B times and averaging over the B estimates from (3.4,3.5) (**Algorithm 3.2**). More specifically, for each $b = 1, 2, .., B$, we randomly split the data into two halves D^b_1 and D^b_2, with fixed sample sizes $|D^b_1| = n_1$ and $|D^b_2| = n_2$. In other words, the data splitting proportion $q = n_1/n,\ 0 < q < 1$ is a fixed constant. Denote the selected candidate set of variables by $\mathcal{S}_\lambda$ on D^b_2 as S^b, and the one-time estimator by (3.4-3.5) as $\widetilde{\beta}^b = (\widetilde{\beta}^b_0, \widetilde{\beta}^b_1, .., \widetilde{\beta}^b_p)$. The smoothed estimator is

$$\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1, .., \widehat{\beta}_p), \text{ where } \widehat{\beta}_j = \frac{1}{B} \sum_{b=1}^{B} \widetilde{\beta}^b_j. \tag{3.6}$$

3.3 Theoretical Results

3.3.1 One-time Estimator

We first establish the asymptotic property of the one-time estimator under the following assumptions.

Algorithm 3.1 One-time SSGLM Estimator

Require: A GLM regression model, a selection procedure $\mathcal{S}_\lambda$

Input: Data (Y, X), split proportion $q \in (0, 1)$

Output: Coefficient estimator $\widetilde{\beta}$

1: Split the data into two halves D_1 and D_2, with sample sizes $|D_1| = qn$, $|D_2| = (1-q)n$
2: Apply $\mathcal{S}_\lambda$ on D_2 to select a subset of important covariates $S \subset [p]$
3: **for** $j = 0, 1, .., p$ **do**
4: Define $S_{+j} = \{j\} \cup S$, and fit the GLM of Y^1 regressing on $X^1_{S_{+j}}$, where $D_1 = (Y^1, X^1)$
5: Denote the coefficient estimator in previous step as $\widetilde{\beta}^1$
6: Define $\widetilde{\beta}_j = \left(\widetilde{\beta}^1\right)_j$, which is the coefficient for covariate X_j ($\widetilde{\beta}_0$ represents the intercept)
7: **end for**
8: Define $\widetilde{\beta} = (\widetilde{\beta}_0, \widetilde{\beta}_1, .., \widetilde{\beta}_p)$

Algorithm 3.2 SSGLM Estimator

Require: A GLM regression model, a selection procedure $\mathcal{S}_\lambda$

Input: Data (Y, X), split proportion $q \in (0, 1)$, number of re-samples B

Output: Coefficient estimator $\widehat{\beta}$

1: **for** $b = 1, 2, .., B$ **do**
2: Run Algorithm 3.1 with random data split
3: Denote the output estimator as $\widetilde{\beta}^b = (\widetilde{\beta}^b_0, \widetilde{\beta}^b_1, .., \widetilde{\beta}^b_p)$
4: **end for**
5: Define $\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1, .., \widehat{\beta}_p)$, where $\widehat{\beta}_j = \frac{1}{B}\sum_{b=1}^{B} \widetilde{\beta}^b_j$

(A1). The eigenvalues of the expected information matrix at β^* are bounded:

$$0 < c_{\min} \leq \lambda_{\min}(I^*) \leq \lambda_{\max}(I^*) \leq c_{\max} < \infty.$$

In addition, for any $i \in [n],\ j \in [p],\ |x_{ij}| \leq \rho_0,\ \mathbf{E}\,|Y_i|^3 \leq \rho_1$.

(A2). Order of Model Parameters: There exist constants $0 < c_1 \leq 1, c_\beta > 0$ such that $s_0 = |S_0| = O(n^{c_1})$, $\max_j |\beta_j^*| \leq c_\beta$.

(A3). Sure Screening Property: There exist a sequence $\{\lambda_n\}_{n\geq 1}$ and constants $0 < \eta < 1$, $c_2 > 2c_1$ such that $|\hat{S}_{n,\lambda_n}|/n \leq \eta$, and

$$P(\hat{S}_{n,\lambda_n} \supset S_0) \geq 1 - o(n^{-c_2-1}) \quad \text{as} \quad n \to \infty.$$

Here $\hat{S}_{n,\lambda_n}$ denotes the selected set of variables with sample size n and tuning parameter λ_n.

Remark III.1. The sure screening property for GLMs has been established in *Fan et al.* (2009); *Fan and Song* (2010). In addition to (A1) and (A2), the following conditions are sufficient for the sure screening property by Theorem 4 in *Fan and Song* (2010):

- The second derivative of $A(\theta)$ is continuous and positive;
- There exists $c_0, \kappa > 0$, such that for $j \in S_0$, $|\mathrm{cov}\,(A'(\theta), X_j)| \geq c_0 n^{-\kappa}$;

It is worth pointing out that the aforementioned conditions imply that the order of p can be as large as $\log p = o(n^{1-2\kappa})$, while providing a stronger tail probability (exponentially small in n) than what is required in assumption (A3).

Assumption (A1) is a standard condition on the eigenvalues and norms of the observed data. Assumption (A2) specifies the order of the sparsity and the effect

sizes. While there is no direct assumption on the order of p, it is implied through Assumption (A3) as stated in Remark III.1.

Theorem III.2. *Given model (3.1,3.2) and assumptions (A1)-(A3), consider the one-time estimator* $\widetilde{\beta} = (\widetilde{\beta}_0, \widetilde{\beta}_1, .., \widetilde{\beta}_p)^{\mathrm{T}}$ *as defined in (3.4-3.5). For any* $j \in \{0\} \cup [p]$, *denote* $\widetilde{\sigma}_j^2 = \left\{ \widehat{I}_{S_{+j}}^{-1} \right\}_{jj}$, *as* $n_1, n \to \infty$,

$$\sqrt{n_1}(\widetilde{\beta}_j - \beta_j^*)/\widetilde{\sigma}_j \to N(0,1).$$

3.3.2 SSGLM Estimator

In practice, the one-time estimator is highly variable as p increases, making it difficult to separate signals from noise variables in the inferential step. In contrast, the smoothed estimator is much more consistent as it averages over both estimation and selection. However, it introduces dependency among the selected S^b's. The following condition, which is stronger than "sure screening," is required for the desired theoretical property.

(B3). Selection Consistency: There exists a sequence $\{\lambda_n\}_{n\geq 1}$ and constants $0 < \eta < 1$, $c_2 > 2c_1$ such that $|\hat{S}_{n,\lambda_n}|/n \leq \eta$, and

$$\mathbf{P}(\hat{S}_{n,\lambda_n} = S_0) \geq 1 - o(n^{-c_2-1}) \quad \text{as} \quad n \to \infty.$$

Theorem III.3. *Given model (3.1,3.2) and assumptions (A1,A2,B3), consider the smoothed estimator* $\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_p)^T$ *as defined in (3.6). For each* j, *as* $n, B \to \infty$,

$$\sqrt{n}(\widehat{\beta}_j - \beta_j^*)/\sqrt{I_{j|S_0}^*} \to N(0,1),$$

where $I_{j|S_0}^*$ *is defined as in (3.3) with the truth* β^*, *which is bounded away from both zero and infinity.*

The proofs of Theorems III.2 and III.3 are provided in Appendix B, along with some useful lemmas.

3.4 Inference by SSGLM

As shown in Theorem (III.3), $\widehat{\beta}_j$ converges to a normal distribution with the variance depending on the unknown active set S_0. We can accurately estimate the variance using the infinitesimal jackknife developed by *Efron* (2014); *Wager et al.* (2014); *Wager and Athey* (2018). For $i = 1, 2, ..., n$ and $b = 1, 2, .., B$, let $J_{bi} \in \{0, 1\}$ denote whether the i^{th} observation appears in the b^{th} sub-sample D_1^b, and $J_{\cdot i} = \left(\sum_{b=1}^{B} J_{bi}\right)/B$ is the average. Then the variance estimator for $\widehat{\beta}_j$ is

$$\widehat{V}_j = \frac{n-1}{n}\left(\frac{n}{n-n_1}\right)^2 \sum_{i=1}^{n} \widehat{\text{cov}}_{ij}^2, \tag{3.7}$$

where

$$\widehat{\text{cov}}_{ij} = \frac{1}{B}\sum_{b=1}^{B}\left(J_{bi} - J_{\cdot i}\right)\left(\widetilde{\beta}_j^b - \widehat{\beta}_j\right).$$

The term $n(n-1)/(n-n_1)^2$ is a finite-sample correction regarding the sub-sampling scheme (*Wager and Athey*, 2018). They have shown that the variance estimator is consistent as $B \to \infty$, in the sense that $\widehat{V}_j / \mathbf{V}\left(\widehat{\beta}_j\right) \xrightarrow{p} 1$.

Moreover, with a finite B, we propose a bias correction version of the variance estimator:

$$\widehat{V}_j^B = \widehat{V}_j - \frac{n}{B^2}\frac{n_1}{n-n_1}\sum_{b=1}^{B}(\widetilde{\beta}_j^b - \widehat{\beta}_j)^2. \tag{3.8}$$

The derivation is analogous to that in Section 4.1 of *Wager et al.* (2014). The advantage of using (3.8) instead of (3.7) is that (3.7) requires $B = O(n^{1.5})$ to reduce the Monte Carlo noise down to the sampling noise level, while (3.8) only requires $B = O(n)$ (*Wager et al.*, 2014).

Thus the asymptotic $100(1-\alpha)\%$ confidence interval for each β_j^* is given by

$$\left(\widehat{\beta}_j - \Phi^{-1}(1-\alpha/2)\sqrt{\widehat{V}_j^B}\ ,\ \widehat{\beta}_j + \Phi^{-1}(1-\alpha/2)\sqrt{\widehat{V}_j^B}\right),$$

where Φ is the CDF of the standard normal distribution. The p-value of testing $H_0 : \beta_j^* = 0$ is

$$2 \times \left\{1 - \Phi\left(|\widehat{\beta}_j| / \sqrt{\widehat{V}_j^B}\right)\right\}.$$

3.5 Extension to a Subvector of Coefficients with a Fixed Dimension

An extension of SSGLM to estimating a subvector $\beta^{(1)}$ of β^* with a fixed dimension allows us to derive confidence regions for a subset of covariates and to test for contrasts of interest. Denote $\beta^{(1)} = \beta^*_{S^{(1)}}$ with $|S^{(1)}| = p_1 \geq 2$, which is finite and does not increase with n and p. Accordingly, the SSGLM estimator for $\beta^{(1)}$ is presented in **Algorithm 3.3**, and the extension of Theorem III.3 is stated below.

Theorem III.4. *Given model (3.1,3.2) under assumptions (A1,A2,B3), and a fixed finite subset $S^{(1)} \subset \{1, 2, .., p\}$ with $|S^{(1)}| = p_1$. Let $\widehat{\beta}^{(1)}$ be the smoothed estimator for $\beta^{(1)} = \beta^*_{S^{(1)}}$ as defined in (3.3). Then as $n, B \to \infty$,*

$$\sqrt{n}\left\{I^*_{S^{(1)}|S_0}\right\}^{-1/2}\left(\widehat{\beta}^{(1)} - \beta^{(1)}\right) \to N(0, \mathrm{I}_{p_1}),$$

where $I^*_{S^{(1)}|S_0} = I^*_{S^{(1)}S^{(1)}} - I^*_{S^{(1)}S_{01}}\left\{I^*_{S_{01}S_{01}}\right\}^{-1} I^*_{S_{01}S^{(1)}}, \quad S_{01} = S_0 \setminus S^{(1)}.$

There is a direct extension of the one-dimensional nonparametric delta method

for estimating the variance-covariance matrix of $\widehat{\beta}^{(1)}$, $\widehat{\Sigma}^{(1)} = \widehat{\mathrm{COV}}_{(1)}^{\mathrm{T}} \widehat{\mathrm{COV}}_{(1)}$, where

$$\widehat{\mathrm{COV}}_{(1)} = \left(\widehat{\mathrm{cov}}_1^{(1)}, \widehat{\mathrm{cov}}_2^{(1)}, .., \widehat{\mathrm{cov}}_n^{(1)} \right)^{\mathrm{T}}, \text{ with}$$
$$\widehat{\mathrm{cov}}_i^{(1)} = \sum_{b=1}^{B} (J_{bi} - J_{\cdot i})(\widehat{\beta}^b_{S^{(1)}} - \widehat{\beta}^{(1)})/B.$$

Therefore, we are equipped to test the following hypothesis, where Q is a $q \times p_1$ matrix and r is a $q \times 1$ vector, $H_0 : Q\beta^{(1)} = r$. The Wald type test statistic is

$$T = \left(Q\widehat{\beta}^{(1)} - r \right)^T \left[Q\widehat{\Sigma}^{(1)} Q^T \right]^{-1} \left(Q\widehat{\beta}^{(1)} - r \right), \tag{3.9}$$

which follows the Chi-square distribution with degree of freedom q under the null. We would reject H_0 if T is larger than the critical value.

Algorithm 3.3 SSGLM for Subvector $\beta^{(1)}$

Require: A GLM regression model, a selection procedure $\mathcal{S}_\lambda$

Input: Data (Y, X), split proportion $q \in (0, 1)$, number of re-samples B, subvector $\beta^{(1)}$ with support $S^{(1)}$

Output: Coefficient estimator $\widehat{\beta}^{(1)}$

1: **for** $b = 1, 2, .., B$ **do** Split the data into two halves D_1 and D_2, with sample sizes $|D_1| = qn$, $|D_2| = (1 - q)n$
2: Apply $\mathcal{S}_\lambda$ on D_2 to select a subset of important covariates $S \subset [p]$
3: Fit the GLM of Y^1 regressing on $X^1_{S^{(1)} \cup S}$, where $D_1 = (Y^1, X^1)$
4: Denote the coefficient estimator in previous step as $\widetilde{\beta}^{(1)}$
5: Define $\widetilde{\beta}^b_{S^{(1)}} = \left(\widetilde{\beta}^{(1)} \right)_{S^{(1)}}$, which is the part estimating $\beta^{(1)}$
6: **end for**
7: Define $\widehat{\beta}^{(1)} = \left(\sum_{b=1}^{B} \widetilde{\beta}^b_{S^{(1)}} \right) /B$

3.6 Simulations

We have conducted numerical studies to investigate the performance of the proposed SSGLM procedure under various settings, and to compare with two existing methods, the de-biased LASSO for GLMs (*Van de Geer et al.*, 2014; *Dezeure et al.*,

2015) and the decorrelated score test (*Ning and Liu*, 2017). We investigated the role of $q = n_1/n$, the split proportion, in fitting SSGLM; we explored various selection methods used in SSGLM and their effects on the estimation and inference; we implemented SSGLM for both logistic and Poisson regressions; and we assessed its performance through calculating the power and type I errors. Some of the most challenging simulation settings (*Bühlmann et al.*, 2014) were examined, as both the indexes in the active set and the non-zero effect sizes were randomized, and different covariance structures were used.

Example 1 investigates the performance of SSGLM while tuning the split proportion in the procedure. We make data splitting with $n_1 = qn,\ q = 0.1, 0.2, ..., 0.9$. We set $n = 500,\ p = 1000,\ s_0 = 10$ with the identity covariance matrix. The indexes in the active set S_0 are randomly pick from $[p]$, and the non-zero effects $\beta_j^*, j \in S_0$ are generated from Unif$[(-1.5, -0.5) \cup (0.5, 1.5)]$. For each q, the objective function is defined by the mean squared errors (MSE), denote $\widehat{\beta}_j^{(k)}$ as the smoothed estimator for β_j from k-th simulation, $k = 1, 2, .., K$,

$$\text{MSE}_j = \frac{1}{K}\sum_{k=1}^{K}(\widehat{\beta}_j^{(k)} - \beta_j^*)^2, \quad \text{MSE}_{\text{avg}} = \frac{1}{p}\sum_{j=1}^{p}\text{MSE}_j.$$

From Figure (3.1),, the minimum MSE is achieved around $q = 0.5$, recommending half-sample in practice.

Example 2 implements a number of selection methods in SSGLM and their impacts on the estimation and inference. There are five procedures being compared: LASSO, SCAD, MCP, Elastic net, and Bayesian LASSO. Five-fold cross-validation is used for the parameter tuning in each selection procedure. We assume a Poisson model with $n = 300$, $p = 400$, and $s_0 = 5$. The results are summarized in Table (3.1). By comparing the bias, the coverage probability, and the mean squared error, we conclude that while the average selected set sizes might differ among the selection methods,

there is little impact on the smoothed estimators and the resulting inferences.

Example 3 assumes the following Poisson model for count data, for $i = 1, 2, .., n$:

$$\log\Big(\mathbf{E}\left(Y_i|\mathbf{x}_i\right)\Big) = \beta_0 + \mathbf{x}_i\beta.$$

We set $n = 400, p = 500,\ s_0 = 6$, with non-zero coefficients between 0.5 and 1, and three different correlation structures: Identity; AR(1) with $\Sigma_{ij} = \rho^{|i-j|}, \rho = 0.5$; Compound Symmetry with $\Sigma_{ij} = \rho^{I(i\neq j)}, \rho = 0.5$. The results are summarized in Table (3.2), as SSGLM provides nearly unbiased component-wise estimation and accurate standard errors, which leads to coverage probabilities that are close to the nominal level. Meanwhile, the non-zero signals are selected with a very high probability in the procedure, suggesting our assumptions (A3) or (B3) are well met in this case.

Example 4 assumes the following logistic regression for binary outcomes, with $n = 400, p = 500$, and $s_0 = 4$.

$$\text{logit}\Big(\mathbf{P}(Y_i = 1|\mathbf{x}_i)\Big) = \beta_0 + \mathbf{x}_i\beta. \tag{3.10}$$

We show the performance of SSGLM when estimating and drawing inferences for the subvector $\beta^{(1)} = \beta_{S_0}$, as a whole. The results are summarized in Tables (3.3,3.4). Our method gives nearly unbiased estimates under different correlation structures and reliable testing power of the low-dimensional contrasts.

Example 5 compares our method with the de-biased LASSO estimator (*Van de Geer et al.*, 2014) and the decorrelated score test (*Ning and Liu*, 2017) through the testing power and the type I error. We again assume the logistic model (3.10) with $n = 200,\ p = 300,\ s_0 = 3, \beta^*_{S_0} = (2, -2, 2)$ with AR(1) correlation structures. From Table (3.5), our method gives the highest testing power while maintaining the type I error around the nominal 0.05 level, while the de-biased LASSO estimators outperform the decorrelated score tests to some extent.

In summary, we have provided numerical evidence that SSGLM performs well when using half sample split, and is robust to various selection methods. We have illustrated its performance under both Poisson and Logistic regressions, and for either single β_j's or a subvector $\beta^{(1)}$. More importantly, the comparison with existing methods shows the clear advantage of our method in terms of the power and the type I error.

3.7 Data Example

A number of studies have aimed to understand the molecular causes of the lung cancer heterogeneity. Identifying the genes and pathways involved, determining how they relate to the biologic behavior of lung cancer and their utility as diagnostic and therapeutic targets are important basic and translational research issues (*Larsen and Minna*, 2011). Recent studies have revealed extensive genetic diversity both between and within tumors. This heterogeneity affects key cancer pathways, driving phenotypic variation, and poses a significant challenge to personalized cancer medicine (*Burrell et al.*, 2013; *Fisher et al.*, 2013).

A subset of the Boston Lung Cancer Study Cohort (*Christiani* (2017)) contains of $n = 1,459$ individuals, among which 708 are lung cancer patients and 751 are controls. The cleaned data consists of $6,829$ SNPs, along with important demographic variables including age, gender, race, education level, and smoking status (Table 3.6). Since smoking plays a vital role in lung cancer, we are particularly interested in the interactions between the SNPs and smoking status, in addition to the main effects.

We assumed a high-dimensional logistic model with the binary lung cancer indicator as the outcome; the demographic variables, the SNPs and the interactions between all SNPs and the smoking status give a total of $p = 13,663$ covariates. The SSGLM is fitted with $B = 1,000$ re-samples. The partial result is shown in Table (3.7), as we list the top coefficients sorted by their p-values. The SNP names starting

with "AX," and the prefix "SAX" indicates the covariate is the interaction between the SNP "AX.xxx" and the smoking status. We identified 9 significant coefficients after Bonferroni correction, all of which are interaction terms (Table (3.7)). This result provides strong evidence of the gene-environmental interactions in addition to the main SNP effects among the lung cancer patients that has rarely been reported before. These nine SNPs come from three genes, TUBB, ERBB2, and TYMS. The presence of TUBB mutations has been associated with both poor treatment response to paclitaxel-containing chemotherapy and shortened overall survival in patients with advanced non-small-cell lung cancer (NSCLC) (*Monzó et al.*, 1999; *Kelley et al.*, 2001). *Rosell et al.* (2001) has proposed using the presence of TUBB mutations as a basis for selecting initial chemotherapy for patients with advanced NSCLC. In contrast, intragenic ERBB2 kinase mutations occur more often in the adenocarcinoma subtype of lung cancer (*Stephens et al.*, 2004; *Beer et al.*, 2002). Finally, advanced NSCLC patients with low/negative thymidylate synthase (TYMS) have better response to Pemetrexed–Based Chemotherapy and longer progression free survival (*Wang et al.*, 2013).

For comparisons, we applied the de-sparsified estimator for GLM (*Bühlmann et al.*, 2014). Direct applications of the "lasso.proj" function in the "hdi" R package (*Dezeure et al.*, 2015) were not feasible given the size of the data. Instead, we used a shorter sequence of the candidate λ values and 5-fold instead of 10-fold cross validation for the node-wise LASSO in the procedure, which still cost about one day of CPU time. After correcting for multiple testing, there were two significant coefficients, both of which were interaction terms corresponding to SNPs AX.35719413_C and AX.83477746_A. Both SNPs were from the TUBB gene and the first SNP was also identified by our method.

To validate our findings, we applied the prediction accuracy measures for nonlinear models proposed in *Li and Wang* (2018). We calculated the R^2, the proportion of

variation in Y explained, for the models we chose to compare. The five models chosen and their respective R^2's were: **Model 1.** $(R^2 = 0.0938)$ the baseline model including only the demographic variables; **Model 2.** $(R^2 = 0.1168)$ the baseline model plus the significant interactions after Bonferroni correction as the top ones from Table (3.7); **Model 3.** $(R^2 = 0.1181)$ the baseline model plus the iterations in Model 2 and their corresponding main effects; **Model 4.** $(R^2 = 0.1018)$ the baseline model plus the significant interactions from the de-sparsified LASSO method; **Model 5.** $(R^2 = 0.1076)$ Model 4 plus the corresponding main effects. In summary, Model 2 based on our method would explain 25% more variation in Y (from 0.0938 to 0.1168), while Model 4 based on the de-sparsified LASSO method only explains 8.5% more variation (from 0.0938 to 0.1018). We also plotted the ROC curves of models 1, 2, and 4 (Figure 3.2) and their AUCs were $0.645, 0.69, 0.668$, respectively.

Our method also provides estimation and uncertainty measures for any pre-specified subvectors of parameters. Past literature has identified several SNPs as potential risk factors for lung cancer. We studied a controversial SNP, rs3117582 from the TUBB gene on chromosome 6. This SNP was identified in association with lung cancer risk in a case/control study by *Wang et al.* (2008), while on the other hand, *Wang et al.* (2009) found no evidence of association between the SNP and risk of lung cancer *among never-smokers*. Our goal was to test this SNP and its interaction with smoking together with all the other covariates under the high dimensional logistic model. Without loss of generality, we denoted the coefficients corresponding to rs3117582 and its interaction term as $\beta^{(1)} = (\beta_1, \beta_2)$. To test the overall effect of rs3117582, the null hypothesis was $H_0 : \beta_1 = \beta_2 = 0$. From the proposed method, we got

$$(\widehat{\beta}_1, \widehat{\beta}_2) = (-0.067, 0.005),\ \widehat{\mathrm{COV}}\,(\beta_1, \beta_2) = \begin{pmatrix} 0.44, & -0.43 \\ -0.43, & 0.50 \end{pmatrix}.$$

While the main effect of the SNP rs3117582 was small, the interaction with smoking

was even more negligible. The test statistic of the overall effect was $T = 0.062 \sim \chi^2(2)$ by (3.9), and the p-value is 0.97. We conclude that rs3117582 is not significantly associated with lung cancer regardless of smoking status in this dataset.

3.8 Conclusion

We have proposed a novel procedure for estimation and inference in high dimensional generalized linear models. We have shown the SSGLM estimator is asymptotically unbiased and normal, which leads to reliable inferences for any low dimensional parameters. Our method utilizes the partial regression idea, which estimates the parameter of interest together with a subset of important covariates, to avoid the common disadvantages caused by penalized regression approaches. Furthermore, our estimator is based on multi-sample splitting and smoothing so that it is robust to the selection variability and enjoys the variance reduction from the bagging effect. Unlike the existing methods (*Belloni et al.* (2014); *Van de Geer et al.* (2014); *Javanmard and Montanari* (2014); *Ning and Liu* (2017)) that require certain conditions on more than one tuning parameters, our method is not sensitive to the λ used for the selection. Hence, our method has minimal requirements on extra model parameters. For the same reason, we have shown that our method is adaptive to a wide range of model selection procedures that gives robust estimation and inferential results. The variance of the proposed estimator is derived from the non-parametric delta method applied to the re-samples, which is free of the regression model and is consistent both theoretically and in simulations. The assumptions on the selection may limit our approach to sparse models and certain data structures. Weakening such conditions has great potential to broaden the applications and is our future work.

Figure 3.1: Average MSEs of all covariates at split proportions q's from 0.1 to 0.9.

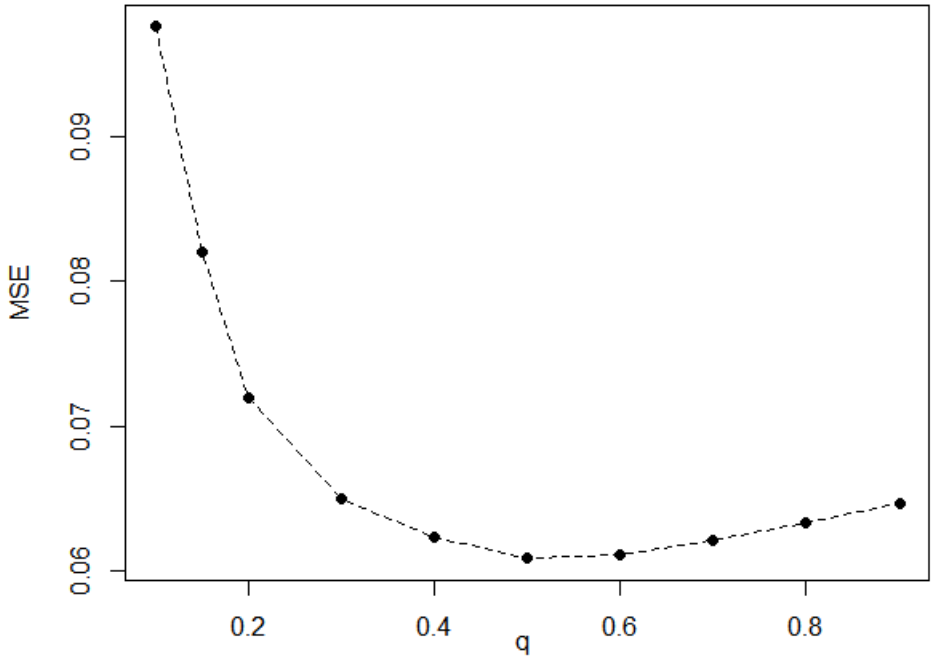


Figure 3.2: ROC curves of the three selected models.

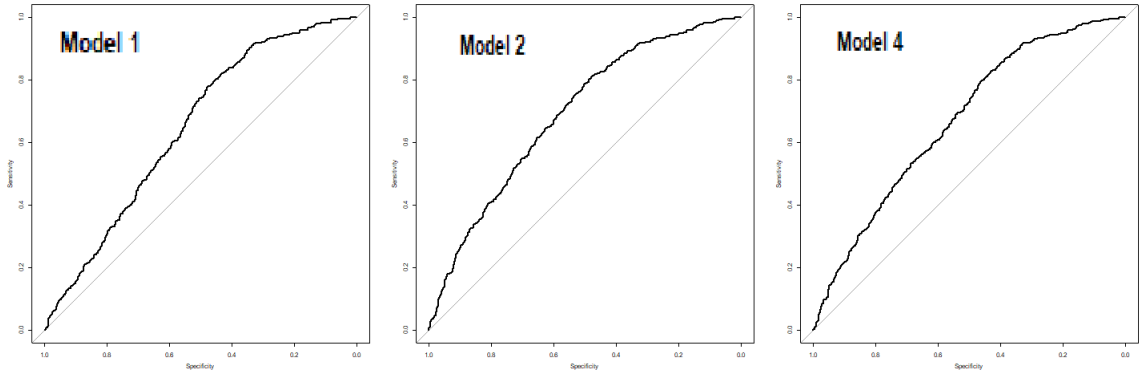


Table 3.1: Comparisons of different selection procedures to implement our proposed method. First column is the indexes of the non-zero signals. Last row for the selection frequency is the average number of covariates being selected by each procedure. Last row for the coverage probability is the average coverage probability of all covariates.

Bias	β^*	LASSO	SCAD	MCP	EN	Bayesian
12	0.4	0.003	0.003	0.003	0.003	0.001
71	0.6	0.007	0.008	0.008	0.008	-0.010
351	0.8	-0.001	0.001	0	0	0.001
377	1.0	-0.005	-0.005	-0.006	-0.005	0.001
386	1.2	0.002	0.001	0.001	0.001	0.004
Selection frequency		LASSO	SCAD	MCP	EN	Bayesian
12		0.59	0.55	0.49	0.60	0.60
71		0.93	0.92	0.90	0.95	0.94
351		0.99	0.99	0.99	1.00	1.00
377		1.00	1.00	1.00	1.00	1.00
386		1.00	1.00	1.00	1.00	1.00
Average #		23.12	13.15	10.89	10.31	7.98
Coverage Prob		LASSO	SCAD	MCP	EN	Bayesian
12		0.90	0.90	0.91	0.91	0.95
71		0.94	0.94	0.95	0.94	0.94
351		0.95	0.95	0.95	0.94	0.95
377		0.94	0.93	0.93	0.94	0.92
386		0.94	0.95	0.95	0.95	0.94
Average		0.93	0.94	0.94	0.94	0.94
MSE		LASSO	SCAD	MCP	EN	Bayesian
12		0.111	0.110	0.110	0.109	0.106
71		0.104	0.103	0.102	0.102	0.101
351		0.103	0.103	0.103	0.103	0.100
377		0.101	0.100	0.100	0.100	0.109
386		0.097	0.096	0.096	0.096	0.102
Average		0.105	0.104	0.103	0.103	0.102

Table 3.2: SSGLM under Poisson regression and three correlation structures. The last column summarizes the average of all noise variables.

	Index	Int	74	109	347	358	379	438	-
	β^*	1.000	0.810	0.595	0.545	0.560	0.665	0.985	0
Identity	Bias	-0.010	0	0	0.001	0.005	0.005	0.006	0
	Avg SE	0.050	0.035	0.034	0.035	0.035	0.034	0.035	0.034
	Emp SE	0.064	0.036	0.038	0.031	0.033	0.038	0.036	0.036
	Cov prob	0.870	0.920	0.900	0.960	0.990	0.910	0.950	0.936
	Sel freq	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.015
AR(1)	Bias	0.006	0.003	-0.002	-0.001	-0.001	-0.005	0.003	0
	Avg SE	0.052	0.035	0.035	0.035	0.035	0.035	0.035	0.035
	Emp SE	0.056	0.031	0.041	0.035	0.037	0.037	0.037	0.036
	Cov prob	0.930	0.970	0.890	0.960	0.950	0.930	0.960	0.937
	Sel freq	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.015
CS	Bias	-0.003	-0.005	0.004	-0.002	0.005	-0.004	-0.001	0.001
	Avg SE	0.033	0.043	0.043	0.042	0.043	0.043	0.044	0.042
	Emp SE	0.038	0.046	0.044	0.052	0.040	0.047	0.043	0.044
	Cov prob	0.960	0.900	0.930	0.900	0.970	0.910	0.950	0.934
	Sel freq	1.000	1.000	0.999	0.997	0.998	0.999	1.000	0.016

Table 3.3: SSGLM under Logistic regression, with estimation and inference for the subvector $\beta^{(1)} = \beta_{S_0}$. The oracle estimator is from the low dimensional GLM knowing the true set S_0. The empirical covariance matrix is with respect to the simulation replications.

Index	218	242	269	517	Index	218	242	269	517
Truth	-2	-1	1	2	Truth	-2	-1	1	2
					Identity				
$\widehat{\beta}^{(1)}$	-2.048	-1.043	0.999	2.096	Oracle	-1.995	-1.026	0.973	2.043
$\widehat{\Sigma}^{(1)}$	0.146	0.010	-0.009	-0.020	Empirical	0.155	0.006	-0.009	-0.027
	0.010	0.134	-0.004	-0.011		0.006	0.129	-0.011	-0.015
	-0.009	-0.004	0.134	0.009		-0.009	-0.011	0.152	0.010
	-0.020	-0.011	0.009	0.143		-0.027	-0.015	0.010	0.134
					AR(1)				
$\widehat{\beta}^{(1)}$	-2.073	-1.014	1.002	2.110	Oracle	-2.024	-0.991	0.977	2.062
$\widehat{\Sigma}^{(1)}$	0.145	0.012	-0.011	-0.023	Empirical	0.141	0.012	-0.016	-0.028
	0.012	0.137	-0.006	-0.011		0.012	0.112	-0.006	0
	-0.011	-0.006	0.135	0.010		-0.016	-0.006	0.129	0.009
	-0.023	-0.011	0.010	0.147		-0.028	0	0.009	0.136
					CS				
$\widehat{\beta}^{(1)}$	-2.095	-1.033	1.070	2.102	Oracle	-2.037	-1.024	1.027	2.028
$\widehat{\Sigma}^{(1)}$	0.223	-0.026	-0.048	-0.063	Empirical	0.192	-0.030	-0.044	-0.045
	-0.026	0.208	-0.043	-0.047		-0.030	0.187	-0.037	-0.044
	-0.048	-0.043	0.207	-0.028		-0.044	-0.037	0.165	-0.011
	-0.063	-0.047	-0.028	0.224		-0.045	-0.044	-0.011	0.179

Table 3.4: SSGLM under Logistic regression, with rejection rates of testing the contrasts.

H_0	Truth	Identity	AR(1)	CS
$\beta^*_{218} + \beta^*_{517} = 0$	0	0.05	0.04	0.03
$\beta^*_{242} + \beta^*_{269} = 0$	0	0.06	0.04	0.025
$\beta^*_{218} + \beta^*_{269} = 0$	-1	0.56	0.57	0.42
$\beta^*_{242} + \beta^*_{517} = 0$	1	0.55	0.58	0.48
$\beta^*_{242} = 0$	-1	0.83	0.80	0.61
$\beta^*_{269} = 0$	1	0.74	0.81	0.70
$\beta^*_{218} = 0$	-2	1	1	1
$\beta^*_{517} = 0$	2	1	1	1

Table 3.5: Comparisons of SSGLM, Lasso-pro, and Decorrelated score in power and Type I error. AR(1) correlation structure with different ρ's are examined.

	Power			Type I error
Index	10	20	30	0's
$\rho = 0.25$ Proposed	0.920	0.930	0.950	0.049
Lasso-pro	0.900	0.930	0.900	0.042
Dscore	0.790	0.880	0.890	0.177
$\rho = 0.4$ Proposed	0.940	0.960	0.965	0.049
Lasso-pro	0.920	0.910	0.920	0.043
Dscore	0.770	0.905	0.840	0.175
$\rho = 0.6$ Proposed	0.940	0.950	0.880	0.054
Lasso-pro	0.850	0.750	0.850	0.045
Dscore	0.711	0.881	0.647	0.268
$\rho = 0.75$ Proposed	0.863	0.847	0.923	0.060
Lasso-pro	0.690	0.670	0.650	0.053
Dscore	0.438	0.843	0.530	0.400

Table 3.6: Demographic characteristics of the BLCSC SNP data.

Case	0	1
Race		
White	726	668
Black	5	22
Other	20	18
Education		
<High school	64	97
High school	211	181
>High school	476	430
Age		
Mean(sd)	59.7(10.6)	60(10.8)
Gender		
Female	460	437
Male	291	271
Pack years		
Mean(sd)	18.8(25.1)	46.1(38.4)
Smoking		
Ever	498	643
Never	253	65

Table 3.7: SSGLM fitted to the BLCSC SNP data. SNP variables start with "AX"; interaction terms start with "SAX"; "Smoke" is the binary smoking status indicator. Rows are sorted by p-values.

Variable	$\widehat{\beta}$	SE	T	P-value	Adjusted P	Sel freq
SAX.88887606_T	0.33	0.02	17.47	$< 10^{-3}$	< 0.01	0.08
SAX.11279606_T	0.53	0.06	8.23	$< 10^{-3}$	< 0.01	0.00
SAX.88887607_T	0.29	0.04	6.97	$< 10^{-3}$	< 0.01	0.01
SAX.15352688_C	0.56	0.08	6.90	$< 10^{-3}$	< 0.01	0.01
SAX.88900908_T	0.54	0.09	5.95	$< 10^{-3}$	< 0.01	0.02
SAX.88900909_T	0.51	0.09	5.69	$< 10^{-3}$	< 0.01	0.02
SAX.32543135_C	0.78	0.14	5.49	$< 10^{-3}$	< 0.01	0.25
SAX.11422900_A	0.32	0.06	5.24	$< 10^{-3}$	< 0.01	0.09
SAX.35719413_C	0.47	0.10	4.63	$< 10^{-3}$	0.049	0.00
SAX.88894133_C	0.43	0.10	4.53	$< 10^{-3}$	0.08	0.00
SAX.11321564_T	0.47	0.11	4.44	$< 10^{-3}$	0.12	0.00
...						
AX.88900908_T	0.40	0.11	3.84	$< 10^{-3}$	1.00	0.00
Smoke	0.89	0.23	3.82	$< 10^{-3}$	1.00	-
...						

CHAPTER IV

Simultaneous Estimation and Inference for High Dimensional Censored Quantile Regression Via Fused Multi-sample Splitting

4.1 Introduction

Lung cancer is the most common cancer-related cause of death worldwide. Understanding the molecular mechanisms on lung cancer survival is a focus of current translational research. The Boston Lung Cancer Study Cohort (BLCSC) (*Christiani*, 2017) is a cancer epidemiology cohort of over 11,000 lung cancer cases enrolled at Massachusetts General Hospital and the Dana-Farber Cancer Institute from 1992 to present. This is the first and most comprehensive lung cancer survivor cohort with a long follow-up period, which has been growing with more patients recruited every year. On a subset of the cohort, the SNP information have been measured and recorded.

Using a target gene approach, we identified a high dimensional SNP data set of 674 lung cancer patients, with measurements of over $2,000$ SNPs from 14 well-known cancer related genes, along with important demographic variables, such as age, gender, race, education level, and smoking status. Our goal is to model the survival times with censoring as the outcome and to estimate and test the effects of

potential predictors on lung cancer survival.

Quantile regression (*Koenker and Bassett Jr*, 1978) has emerged as an efficient way of linking the whole outcome distribution to the covariates, and a useful alternative regression strategy for survival analysis. Quantile regression is especially powerful in detecting the covariate effect at extreme tails, and thus provides more complete information of the outcome distribution. Censored quantile regression (CQR) for randomly censored survival data has been well studied in the finite p case, where p is the number of covariates (*Portnoy* (2003); *Peng and Huang* (2008) among others). High dimensional censored quantile regression (HDCQR), on the other hand, is still an area with growing research interests. The complexity of simultaneous estimation and inference based on HDCQR arises from censoring, extreme quantiles, and other high dimensional inference challenges. *Wang et al.* (2012) deals with variable selection for quantile regressions with ultra-high dimension. *He et al.* (2013) provides variable screening for HDCQR that can handle censoring. *Zheng et al.* (2018) proposes to model the HDCQR with sequential estimation and penalization based on a stochastic integral based estimating equation. They consider two types of penalties, a Lasso type L_1 penalty for sparse estimation and an Adaptive Lasso type penalty to reduce the bias. Their method aims to select a sparse subset of covariates based on the penalized coefficient estimators, but the inferences remain unsolved. Although there have been considerable success in high dimensional inferences for linear and non-linear models (*Zhang and Zhang* (2014); *Bühlmann et al.* (2014); *Javanmard and Montanari* (2014); *Belloni et al.* (2014); *Ning and Liu* (2017); *Fei et al.* (2018) among others), the counterpart to properly handle survival outcomes has been lacking. *Belloni et al.* (2018) provides valid post-selection inference in high dimensional quantile regression models for fixed quantiles, but could not handle censoring and survival outcomes. *Shows et al.* (2010) provides sparse estimation and inference for censored median regression, but with fixed number of predictors.

The goal of this paper is to provide simultaneous estimation and inference based on high dimensional sparse censored quantile regression models, thus to properly analyze the survival outcomes with censoring and when the number of covariates is much larger than the sample size. This is, to our knowledge, the first work that achieves these goals within the given context. Our proposed method uses multi-sample splitting and smoothing techniques to convert the challenging high dimensional estimation problem to a series of low dimensional estimations (*Fei et al.*, 2018). Splitting the original data into two equal halves, we first apply some variable selection procedure to choose a subset of important covariates on one half of data. Next on the other half of data, we fit low dimensional CQRs only using the union of selected subset and each covariate of interest as regressors. If the selected subset is an superset of the sparse active set, then the resulting coefficient estimator of the covariate of interest is unbiased, whether the true effect is non-zero or not. The estimator based on a single split is highly variable, due to the variation in selection, the random data split, and the reduced sample size comparing to the large number of parameters to be estimated. Thus we perform multi-sample splitting and average the resulting estimators. The split and aggregation procedure, named as Fused-HDCQR, gives unbiased estimation of the whole coefficient vector over an interval of quantile values. Each coefficient estimator is shown to converge weakly to a mean zero Gaussian process in the quantile interval. We further derive a model-free variance estimator based on the functional delta method and the multi-sampling splitting properties (*Efron*, 2014; *Wager and Athey*, 2018). The variance estimator is asymptotically consistent, and possesses satisfying empirical performance.

As our fused estimator takes into account the variation in the model selection, it is not post model selection inference (*Belloni et al.*, 2018). Our procedure aims to recover the sparse model that associates the survival outcome with the predictors. By combining low-dimensional selection-assisted estimation and multiple re-samples

and aggregation, our procedure offers both unbiased point estimators and accurate uncertainty measures.

Section 4.2 introduces our method, and Section 4.3 details the asymptotic properties of the proposed estimator. Section 4.4 derives the non-parametric variance estimation procedure and the inferential procedure; Sections 4.5 conducts simulation studies, and Section 4.6 applies the proposed method to analyze the BLCSC SNP data.

4.2 Method

Let T denote a survival outcome and C a right censoring time. We assume C is independent of T given $\widetilde{\mathbf{Z}}$, a $(p-1)\times 1$ covariate vector ($p>1$). Let $X=\min\{T,C\}, \Delta = I(T\le C)$, and $\mathbf{Z}=(1,\widetilde{\mathbf{Z}}^{\mathrm{T}})^{\mathrm{T}}$. The observed data is n i.i.d. copies of $(X,\Delta,\mathbf{Z})$, denoted as $\{(X_i,\Delta_i,\mathbf{Z}_i), i=1,2,..,n\}$.

Define the τ-th conditional quantile of $Y=\log T$ given $\mathbf{Z}$ as $Q_Y(\tau|\mathbf{Z})=\inf\{t: \mathrm{P}(Y\le t|\mathbf{Z})\ge\tau\}$, which is often modeled by a linear quantile regression as:

$$Q_Y(\tau|\mathbf{Z})=\mathbf{Z}^{\mathrm{T}}\beta^*(\tau),\ \tau\in(0,\tau_U], \tag{4.1}$$

where $\beta^*(\tau)$ is a p-dimensional vector of unknown coefficients at each τ, and $0<\tau_U<1$. $\beta^*(\tau)$ is believed to be sparse, that is $q=\left|\cup_{\tau\in(0,\tau_U]}S_\tau\right|=o(n)$, where $S_\tau=\{j\in[p]:\beta_j^*(\tau)\neq 0\}$. Detailed assumptions on the sparsity will be explored later. The goal of this paper is to accurately estimate $\beta_j^*(\tau)$ for $\tau\in(0,\tau_U]$ and for each $j=1,2,..,p$, and their standard errors.

Let $N(t)=I(\log X\le t,\Delta=1)$, $\Lambda_T(t|\mathbf{Z})=-\log(1-\mathrm{P}(\log T\le t|\mathbf{Z}))$, and $H(u)=-\log(1-u)$. $M(t)=N(t)-\Lambda_T(t\wedge\log X|\mathbf{Z})$ is a martingale process, and hence $\mathbf{E}\left(M(t)|\mathbf{Z}\right)=0$. Let $N_i(t)$ and $M_i(t)$ be sample analogs of $N(t)$ and $M(t)$, $i=1,2,..,n$.

Unlike the locally concerned quantile regression, in which only some fixed quantile values are of interest, we would like to adopt the globally concerned CQR framework (*Zheng et al.*, 2015) in this paper. By specifying a quantile interval $[\tau_L, \tau_U]$, the globally concerned framework provides more comprehensive estimation and inference to the assumed model (4.1). In detail, we define a grid of quantile values that covers the interval, $\Gamma_m = \{\tau_0, \tau_1, .., \tau_m\}$, where $\tau_0 = \nu > 0$ and $\tau_m = \tau_U$. By allowing m to increase with n, Γ_m becomes a fine grid that well approximates $\beta^*(\tau)$ on the interval $[\tau_L, \tau_U]$. In addition, we restrict $\tau_0 = \nu$, instead of $\tau_0 = 0$, to circumvent the singularity problem with censored quantile regression at $\tau = 0$, as detailed in assumption (A1). In practice, ν should be chosen such that only a small proportion of the X_i's below the fitted ν-th quantiles are censored. On the other hand, $\tau_U < 1$ is away from 1 to avoid the identifiability issue at upper quantiles due to censoring. The theoretical constraints and the practical selection of τ_U is discussed in *Peng and Huang* (2008). Equally spaced grid Γ_m with small grid size is assumed, as our procedure relies on the sequential selection and estimation on the grid points to derive the desired theoretical properties. Therefore a functional form of $\beta^*(\tau)$ over the interval $[\tau_L, \tau_U]$ can be obtained by extending our estimator to a right-continuous piecewise-constant function that only jumps at the grid points.

4.2.1 Preliminaries

Low dimensional CQR It has been a well-studied problem for censored quantile regressions with finite p, which can be dated back to *Powell* (1986). There are two popular approaches, *Portnoy* (2003) and *Peng and Huang* (2008), where the former developed a recursively re-weighted estimation procedure with follow-up works in *Neocleous et al.* (2006); *Portnoy and Lin* (2010). *Peng and Huang* (2008) introduced estimating equations for $\beta^*(\tau)$ in model (4.1) for fixed p and conditionally independent censoring. Their procedure uses empirical process and stochastic inte-

gral techniques to derive the asymptotic properties including uniform consistency and weak convergence. Their estimating equation takes the form

$$n^{1/2}\mathbf{U}_n(\beta,\tau)=0, \tag{4.2}$$

where

$$\mathbf{U}_n(\beta,\tau)=n^{-1}\sum_{i=1}^{n}\mathbf{Z}_i\left(N_i\left(\theta_i(\tau)\right)-\int_0^{\tau}I[\log X_i\geq\theta_i(u)]dH(u)\right);$$

$$\theta_i(\tau)=\mathbf{Z}_i^{\mathrm{T}}\beta(\tau).$$

Let $\mathbf{u}(\beta,\tau)=\mathbf{E}\left\{\mathbf{U}_n(\beta,\tau)\right\}$, the martingale property gives $\mathbf{u}(\beta^*,\tau)=0,\ \forall\tau\in\Gamma_m$. Furthermore, (4.2) is solved sequentially for $\beta(\tau_k),\tau_k\in\Gamma_m$ through the following monotone estimating equation:

$$n^{-1/2}\sum_{i=1}^{n}\mathbf{Z}_i\left(N_i\left(\theta_i(\tau_k)\right)-\sum_{r=0}^{k-1}\int_{\tau_r}^{\tau_{r+1}}I[\log X_i\geq\check{\theta}_i(\tau_r)]dH(u)\right)=0,$$

where $\check{\beta}(\tau_k),\check{\theta}_i(\tau_k)$ denote the estimators from *Peng and Huang* (2008). Due to the monotonicity of $\theta_i(\tau)$ in τ, $\check{\beta}(\tau)$ can be solved efficiently via L_1-minimization and is shown to be uniformly consistent and converges weakly to a mean zero Gaussian process for $\tau\in[\tau_L,\tau_U]$.

Variable selection in high dimensional censored quantile regression There have been several works for variable selection in high dimensional quantile regressions that have seen relative success. *Zheng et al.* (2013) proposed an adaptive penalized quantile regression estimator that could select the true sparse model with probability converging to 1. *Fan et al.* (2014) studied the penalized quantile regression with a weighted L_1-penalty in the ultra-high dimensional setting. However, neither work dealt with censoring and thus was not applicable to survival outcomes. On the other

hand, *He et al.* (2013) proposed quantile-adaptive variable screening for high dimensional data with censoring, which was designed for a few fixed quantiles and achieved sure screening property.

Recently, *Zheng et al.* (2018) proposed estimation and variable selection for high dimensional CQR using penalization and sequential estimation. They considered two types of penalties, Lasso type L_1 penalty, and Adaptive Lasso type penalty to reduce the bias by the L_1 penalty. Their first estimator (L-HDCQR) incorporated the L_1 penalty into the stochastic integral based estimating equation, which stemmed from *Peng and Huang* (2008) for low dimensional CQR. The L-HDCQR estimator had a uniform convergence rate of $\sqrt{q\log(p\wedge n)/n}$, and resulted in "sure screening" variable selection with high probability. Furthermore, the second estimator based on Adaptive Lasso penalties (AL-HDCQR) had the estimation bias reduced to the order $\sqrt{q\log(n)/n}$, and achieved "selection consistency" with proper order of tuning parameters.

4.2.2 Proposed Fused-HDCQR

We use *Peng and Huang* (2008)'s procedure to fit low dimensional CQRs, which yields simultaneous estimation of the fine grid Γ_m while only requiring conditionally independent censoring. The proposed fused estimator is derived from multi-sample splitting, which separates selection and estimation, and achieves variance reduction through the effect of bagging. More importantly, its asymptotic variance can be consistently estimated in a non-parametric way, as will be shown in Section 4.4.

Given data $D = \{(X_i, \Delta_i, \mathbf{Z}_i), i = 1, 2, .., n\}$, a grid of τ values, $\Gamma_m = \{\tau_0, \tau_1, .., \tau_m\}$ with $\tau_0 = \nu, \tau_m = \tau_U$, and a variable selection procedure for HDCQR with extra parameter λ denoted by $\mathcal{S}_\lambda$:

1. Choose a fixed tuning parameter λ_n (not necessarily optimal): on the full data D, apply $\mathcal{S}_\lambda$ with K-fold cross-validation, and let $\lambda_n = \lambda_{\min}$, which gives the

minimum cross validation error.

2. Let B be a large positive number, and for each $b = 1, 2, .., B$, repeat the following steps;

   (i) Randomly split the data into two equal halves, D_1 and D_2

   (ii) On D_2, apply $\mathcal{S}_\lambda$ with the chosen λ_n on the τ-grid Γ_m, to select a subset of important covariates, denoted as $\widehat{S}^b_{\lambda_n}$, or $\widehat{S}^b$ for short.

   (iii) On D_1, for each $j = 1, 2, .., p$, fit the following low dimensional CQR with respect to the subset of covariates $\widehat{S}^b_{+j} = \{j\} \cup \widehat{S}^b$, and denote the estimator as $\widetilde{\beta}_{\widehat{S}^b_{+j}}(\tau)$.

$$Q_Y(\tau|\mathbf{Z}_{\widehat{S}^b_{+j}}) = \mathbf{Z}^{\mathrm{T}}_{\widehat{S}^b_{+j}} \beta_{\widehat{S}^b_{+j}}(\tau).$$

   (iv) Define the b-th estimator of $\beta^*_j(\tau)$ as the entry in $\widetilde{\beta}_{\widehat{S}^b_{+j}}(\tau)$ that is the coefficient for variable Z_j, $\widetilde{\beta}^b_j(\tau) = \left(\widetilde{\beta}_{\widehat{S}^b_{+j}}(\tau)\right)_j$.

3. Smoothing: the final estimator of $\beta^*_j(\tau), j = 1, 2, .., p$ is

$$\widehat{\beta}_j(\tau_k) = \frac{1}{B}\sum_{b=1}^{B} \widetilde{\beta}^b_j(\tau_k),\ \tau_k \in \Gamma_m; \quad \widehat{\beta}_j(\tau) = \widehat{\beta}_j(\tau_k),\ \tau_{k-1} \le \tau < \tau_k,\ k = 1, 2, .., m. \tag{4.3}$$

Remark IV.1. Since the intercept term is always included in CQR models, its corresponding index $1 \in \widehat{S}$ for all selections, and $\widehat{S}_{+1} = \widehat{S}$. Thus the intercept estimator $\widetilde{\beta}_1(\tau)$ is defined as the component in $\widetilde{\beta}_{\widehat{S}}(\tau)$ for any $\widehat{S}$.

Remark IV.2. Several procedures can be used as $\mathcal{S}_\lambda$: the screening method in *He et al.* (2013) for fast computation; the L-HDCQR for detecting any non-zero effects in the interval $[\tau_L, \tau_U]$; the AL-HDCQR for most accurate selections, among others. Take L-HDCQR for example, the selected sets are defined as $\{j : \max_k |\hat{\gamma}_j(\tau_k)| > a_0, \tau_k \in$

$\Gamma_m\}$, where $\hat{\gamma}_j(\tau_k)$'s are the L-HDCQR estimates, and $a_0 > 0$ is a predetermined threshold.

In Section (4.3), we show the asymptotic properties of the Fused-HDCQR estimator. Next, in Section (4.4), we present the non-parametric variance estimator for Fused-HDCQR and on how to make inferences accordingly.

4.3 Theoretical Properties

4.3.1 Notation and regularity conditions

For any vector $\boldsymbol{\delta} \in \mathbf{R}^p$ and a subset $S \in [p]$, S^C is the complementary set, and define $\|\boldsymbol{\delta}\|_{r,S} = \|\boldsymbol{\delta}_S\|_r$, the l_r-norm of the sub-vector $\boldsymbol{\delta}_S$. We assume the following regularity conditions:

(A1) There exists a quantile ν and some constant c such that

$$n^{-1}\sum_{i=1}^{n} I\left(\log C_i \leq \mathbf{Z}_i^{\mathrm{T}}\beta^*(\nu)\right)(1-\Delta_i) \leq cn^{-1/2}$$

holds for sufficiently large n.

(A2) (*Bounded covariates*) $\|\mathbf{Z}\|_\infty \leq C_0$, for some constant C_0.

(A3) (*Bounded densities*) Let $F_T(t|\mathbf{Z}) = \mathrm{P}\left(\log T \leq t|\mathbf{Z}\right)$, $\Lambda_T(t|\mathbf{Z}) = -\log\left(1 - F_T(t|\mathbf{Z})\right)$, $F(t|\mathbf{Z}) = \mathrm{P}\left(\log X \leq t|\mathbf{Z}\right)$, and $G(t|\mathbf{Z}) = \mathrm{P}\left(\log X \leq t, \Delta = 1|\mathbf{Z}\right)$. Also, define $f(t|\mathbf{Z}) = dF(t|\mathbf{Z})/dt$, and $g(t|\mathbf{Z}) = dG(t|\mathbf{Z})/dt$.

(a) There exist constants $\underline{f}$, $\overline{f}$, $\underline{g}$ and $\overline{g}$ such that

$$\underline{f} \leq \inf_{\mathbf{z},\tau\in[\tau_L,\tau_U]} f(\mathbf{z}^{\mathrm{T}}\beta^*(\tau)|\mathbf{z}) \leq \sup_{\mathbf{z},\tau\in[\tau_L,\tau_U]} f(\mathbf{z}^{\mathrm{T}}\beta^*(\tau)|\mathbf{z}) \leq \overline{f},$$

$$\underline{g} \leq \inf_{\mathbf{z},\tau\in[\tau_L,\tau_U]} g(\mathbf{z}^{\mathrm{T}}\beta^*(\tau)|\mathbf{z}) \leq \sup_{\mathbf{z},\tau\in[\tau_L,\tau_U]} g(\mathbf{z}^{\mathrm{T}}\beta^*(\tau)|\mathbf{z}) \leq \overline{g}.$$

(b) There exist constant $\kappa > 0$ and A such that $\forall |t| \leq \kappa$,

$$\sup_{\mathbf{z}, \tau \in [\tau_L, \tau_U]} \left| f(\mathbf{z}^{\mathrm{T}} \beta^*(\tau) + t | \mathbf{z}) - f(\mathbf{z}^{\mathrm{T}} \beta^*(\tau) | \mathbf{z}) \right| \leq A|t|,$$

$$\sup_{\mathbf{z}, \tau \in [\tau_L, \tau_U]} \left| g(\mathbf{z}^{\mathrm{T}} \beta^*(\tau) + t | \mathbf{z}) - g(\mathbf{z}^{\mathrm{T}} \beta^*(\tau) | \mathbf{z}) \right| \leq A|t|.$$

(A4) (*Sparsity and dimensionality*) Let $S_\tau = \{j \in [p] : \beta_j^*(\tau) \neq 0\}$, $S^* = \bigcup_{\tau \in [\tau_L, \tau_U]} S_\tau = \left\{ j : \sup_{\tau \in [\tau_L, \tau_U]} |\beta_j^*(\tau)| > 0 \right\}$. We assume $n/p = o(1), \log p = o(n^{1/2})$, and $q = |S^*| = o(n)$.

(A5) There exists a sequence $\{\lambda_n\}_{n \to \infty}$ and constants $0 \leq c_1 < 1/2, 0 < K_1 \leq 1, K_2 > 0$ such that the selections $\widehat{S}^b$'s by $\mathcal{S}_{\lambda_n}$ with sample size n satisfy $|\widehat{S}^b| \leq K_1 n^{c_1}$, and

$$\mathrm{P}\left(\widehat{S}^b = S^* \right) \geq 1 - K_2 (p \vee n)^{-1}.$$

(A6) Let $\tilde{\mu}(\tau) = \mathbf{E}\, I\left(\log X > \mathbf{Z}^{\mathrm{T}} \beta^*(\tau) \right)$, then there exists a positive constant L, such that $|\beta_j^*(\tau_1) - \beta_j^*(\tau_2)| \leq L|\tau_1 - \tau_2|$ and $|\tilde{\mu}(\tau_1) - \tilde{\mu}(\tau_2)| \leq L|\tau_1 - \tau_2|$, for all $\tau_1, \tau_2 \in (\nu, \tau_U]$ and $1 \leq j \leq p$.

(A7) (*Restricted eigenvalue condition*) Let A_τ denote the restricted set

$$\{\boldsymbol{\delta} \in \mathbf{R}^p : \|\boldsymbol{\delta}\|_{1, S_\tau^c} \leq ((c_0 + 1)/(c_0 - 1)) \|\boldsymbol{\delta}\|_{1, S_\tau}, \|\boldsymbol{\delta}\|_{0, S_\tau^c} \leq n\},$$

and A_S denote

$$\{\boldsymbol{\delta} \in \mathbf{R}^p : \|\boldsymbol{\delta}\|_{1, S^c} \leq ((c_0 + 1)/(c_0 - 1)) \|\boldsymbol{\delta}\|_{1, S}, \|\boldsymbol{\delta}\|_{0, S^c} \leq n\},$$

for some constant $c_0 > 1$. We can see that $A_\tau \subset A_S$, for all $\tau \in (0, \tau_U]$. Then $0 < \lambda_{\min} \leq \inf_{\boldsymbol{\delta} \in A_S, \boldsymbol{\delta} \neq 0} \boldsymbol{\delta}^{\mathrm{T}} \mathbf{E}\left[\mathbf{Z}_i \mathbf{Z}_i^{\mathrm{T}} \right] \boldsymbol{\delta} / \|\boldsymbol{\delta}\|^2$.

(A8) Let $\epsilon_n = \tau_k - \tau_{k-1}, \tau_k \in \Gamma_m$, $k = 1, 2, .., m$. The grid size satisfies $\sqrt{n}\epsilon_n = o(1)$.

Assumption (A1) requires that the number of censored observations below the ν-th quantile does not exceed $cn^{1/2}$. Since $\mathbf{Z}^{\mathrm{T}}\beta^*(0) = -\infty$ under model (4.1), which corresponds to the 0-th quantile of the survival time, (A1) is satisfied if the lower bound of the censoring time C's support is greater than 0, which is common and reasonable in real world applications. As recommended in *Zheng et al.* (2018), ν is chosen such that only a small proportion of the observed survival times below the fitted ν-th quantile are censored. (A2) assumes the covariates are uniformly bounded. (A3) is a condition on the data distribution that assures the positiveness of $f(t|\mathbf{Z})$ between $\mathbf{Z}^{\mathrm{T}}\beta^*(\tau_L)$ and $\mathbf{Z}^{\mathrm{T}}\beta^*(\tau_U)$, which is essential for the identifiability of $\beta^*(\tau)$ for $\tau < \tau_U$. (A4) restricts the sparsity of $\beta^*(\tau)$, as well as the order of data dimensions. The sparsity is also implied through (A5) that $q \leq K_1 n^{c_1}$. (A5) characterizes the selection properties by $\mathcal{S}_\lambda$, in which the limiting probability going to 1 is an asymptotic property that does not take into account the possibly high variation of a single selection with finite sample. Thus even with the so-called asymptotic "selection consistency" property, it is still crucial to take into account the variation in selection, which yields more efficient estimators. See the simulation in *Fei et al.* (2018) that illustrates the difference in efficiency. Meanwhile, for example the selection consistency is guaranteed by the AL-HDCQR procedure in *Zheng et al.* (2018) with beta-min condition, restriction on sparsity, and some other conditions. (A6) characterizes the smoothness of $\beta^*(\tau)$. (A7) is a typical assumption in high dimensional statistics literature (*Belloni and Chernozhukov*, 2011; *Bickel et al.*, 2009; *Fan et al.*, 2014). (A8) details the fineness of the grid Γ_m, which is needed for the weak convergence of Fused-HDCQR estimator and is also required in Theorem 2 of *Peng and Huang* (2008).

4.3.2 Fused-HDCQR Estimator

Theorem IV.3. *Consider the Fused-HDCQR estimator (4.3). Under assumptions (A1)-(A8), for any $j \in [p]$,*

$$\sqrt{n}\left(\widehat{\beta}_j(\tau) - \beta_j^*(\tau)\right)$$

converges weakly to a mean zero Gaussian process for $\tau \in [\tau_L, \tau_U]$.

The proof and a couple of lemmas are presented in Appendix (C).

4.4 Inferences Based on Fused-HDCQR

The results from previous section indicate that deriving the analytical form of the variance of Fused-HDCQR estimator is difficult, as the covariance function of the limiting Gaussian process involves unknown active set S^* and conditional density functions $f(t|\mathbf{Z})$ and $g(t|\mathbf{Z})$. Alternatively, we propose a model-free variance estimator based on functional delta method and multi-sampling splitting properties (*Efron*, 2014; *Fei et al.*, 2018).

First, we let $J_{bi} \in \{0, 1\}$ be the indicator of whether the i^{th} observation appears in the b^{th} sub-sample D_1^b, and $J_{\cdot i} = \left(\sum_{b=1}^{B} J_{bi}\right)/B$ is the average. Next, we define the re-sampling covariances between J_{bi} and $\widetilde{\beta}_j^b(\tau_k)$ at $\tau_k \in \Gamma_m$ for $b = 1, 2, .., B$ as

$$\widehat{\mathrm{cov}}_{ij}(\tau_k) = \frac{1}{B}\sum_{b=1}^{B}\left(J_{bi} - J_{\cdot i}\right)\left(\widetilde{\beta}_j^b(\tau_k) - \widehat{\beta}_j(\tau_k)\right);$$

$$\widehat{\mathrm{Cov}}_j(\tau_k) = \left(\widehat{\mathrm{cov}}_{1j}(\tau_k),\, \widehat{\mathrm{cov}}_{2j}(\tau_k),\, ..,\, \widehat{\mathrm{cov}}_{nj}(\tau_k)\right)^{\mathrm{T}}.$$

Let $n_1 = |D_1^b|$, the asymptotic covariance estimator between $\widehat{\beta}_j(\tau_k)$ and $\widehat{\beta}_j(\tau_\ell)$ is

$$\widehat{\mathrm{Cov}}_j(\tau_k, \tau_\ell) = \frac{n-1}{n}\left(\frac{n}{n-n_1}\right)^2 \sum_{i=1}^{n} \widehat{\mathrm{cov}}_{ij}(\tau_k)\widehat{\mathrm{cov}}_{ij}(\tau_\ell) = \frac{n(n-1)}{(n-n_1)^2}\widehat{\mathrm{Cov}}_j^{\mathrm{T}}(\tau_k)\widehat{\mathrm{Cov}}_j(\tau_\ell),$$

where the multiplier $n(n-1)/(n-n_1)^2$ is a finite-sample correction for the sub-

sampling D_1^b's. Thus the asymptotic variance estimator for $\widehat{\beta}_j(\tau_k)$ is

$$\widehat{V}_j(\tau_k) = \frac{n(n-1)}{(n-n_1)^2} \sum_{i=1}^{n} \widehat{\text{cov}}_{ij}^2(\tau_k). \tag{4.4}$$

It is shown in *Wager and Athey* (2018) that the variance estimator is consistent as $n, B \to \infty$, in the sense that $\widehat{V}_j(\tau_k)/\text{Var}\left(\widehat{\beta}_j(\tau_k)\right) \xrightarrow{p} 1$. Furthermore, we propose a finite B bias correction to the above variance (4.4) as below,

$$\widehat{V}_j^B(\tau_k) = \widehat{V}_j(\tau_k) - \frac{n}{B^2}\frac{n_1}{n-n_1} \sum_{b=1}^{B} \left(\widetilde{\beta}_j^b(\tau_k) - \widehat{\beta}_j(\tau_k)\right)^2, \quad \tau_k \in \Gamma_m \tag{4.5}$$

where the correction term is a multiplier of the re-sampling variance of $\widetilde{\beta}_j^b(\tau_k)$'s. While the two variance estimators (4.4) and (4.5) are equivalent as $B \to \infty$, and both are asymptotically unbiased to the truth, the former requires $B = O(n^{1.5})$ to reduce the Monte Carlo noise down to the sampling noise, and the latter only requires $B = O(n)$ (*Wager et al.*, 2014). By assumption (A6) and the Lipschitz continuity of the covariance function, we can extend the variance estimator from the grid points $\tau_k \in \Gamma_m$ to the interval $\tau \in [\tau_L, \tau_U]$ by defining $\widehat{V}_j^B(\tau) = \widehat{V}_j^B(\tau_{k-1})$, for $\tau_{k-1} \leq \tau < \tau_k$.

To make inferences of the estimated parameters in model (4.1), by Theorem IV.3, $\widehat{\beta}_j(\tau)$ converges weakly to some Gaussian process on $[\tau_L, \tau_U]$. Thus we define the asymptotic $100(1-\alpha)\%$ confidence interval for $\beta_j^*(\tau)$ at any $\tau \in [\tau_L, \tau_U]$ as

$$\left(\widehat{\beta}_j(\tau) - \Phi^{-1}(1-\alpha/2)\sqrt{\widehat{V}_j^B(\tau)}\ ,\ \widehat{\beta}_j(\tau) + \Phi^{-1}(1-\alpha/2)\sqrt{\widehat{V}_j^B(\tau)}\right),$$

where $\widehat{V}_j^B(\tau)$ is the variance estimator in (4.5), and Φ is the CDF of the standard normal distribution. The p-values of testing $H_0 : \beta_j^*(\tau) = 0$ for each $\tau \in \Gamma_m$ are

$$2 \times \left\{1 - \Phi\left(|\widehat{\beta}_j(\tau)|/\sqrt{\widehat{V}_j^B(\tau)}\right)\right\}.$$

4.5 Simulations

The simulation studies aim to assess the finite sample performance of the proposed Fused-HDCQR method. We first consider examples with true $\beta^*(\tau)$ invariant in τ, then explore the examples with some $\beta_j^*(\tau)$'s changing with τ.

Example 1. With sample size $n = 300$ and the number of covariates $p = 500$, the event times are generated following

$$\log T_i = \widetilde{\mathbf{Z}}_i^{\mathrm{T}}\mathbf{b} + \varepsilon_i, \quad i = 1, 2, .., n,$$

where the coefficient vector $\mathbf{b}$ are sparse with $\mathbf{b}_{20} = 0.25, \mathbf{b}_{40} = 0.5, \mathbf{b}_{60} = 0.75, \mathbf{b}_{80} = 1, \mathbf{b}_{100} = 1.25$, $\mathbf{b}_j = 0$ for all other j's, and $\varepsilon_i \sim N(0, 1)$. Therefore, the truth is $\beta^*(\tau) = (Q_\varepsilon(\tau), \mathbf{b}^{\mathrm{T}})^{\mathrm{T}}$ for all $\tau \in (0, 1)$, where $Q_\varepsilon(\tau)$, τ-th quantile of the distribution of ε, is the intercept. The censoring time is generated independently as $\log C_i = N(0, 16) + N(-5, 1) + N(8, 0.25)$, which gives a censoring rate around 33%. Two covariate distributions are examined, i) $\tilde{Z}_{j,i}$'s are i.i.d. $Unif(-1, 1)$ for $j = 1, 2, .., p$; ii) $\widetilde{\mathbf{Z}}_i$'s follow multivariate normal distribution $N_p(0, \mathbf{\Sigma})$, with $\mathbf{\Sigma} = (\sigma_{k\ell})_{p\times p}$, $\sigma_{k\ell} = 0.5^{|k-\ell|}$ the AR(1) correlation structure.

Example 2. With $n = 200, p = 300$, the event times follow

$$\log T_i = \widetilde{\mathbf{Z}}_i^{\mathrm{T}}\mathbf{b} + 1.5\tilde{Z}_{3,i}\varepsilon_i, \tag{4.6}$$

where $\mathbf{b}_{20} = 1, \mathbf{b}_{40} = 1.5, \mathbf{b}_{60} = 1, \mathbf{b}_{80} = 1.5$ and $\mathbf{b}_j = 0$ for all other j's, and $\varepsilon_i \sim N(0, 1)$. We first generate $\acute{\mathbf{Z}}_i \sim N_p(0, \mathbf{\Sigma})$ as in Example 1 case ii), and then let $\widetilde{\mathbf{Z}}_i = \acute{\mathbf{Z}}_i$, except for the third covariate $\tilde{Z}_{3,i} = |\acute{Z}_{3,i}| + 0.5$. Therefore $\beta_1^*(\tau) = 0, \beta_4^*(\tau) = 1.5Q_\varepsilon(\tau)$, and $\beta_j^*(\tau) = \mathbf{b}_{j+1}$, for all other j's. The censoring time is generated following the same distribution as in Example 1.

Example 3. With $n = 300, p = 400$, the event times follow

$$\log T_i = \widetilde{\mathbf{Z}}_i^{\mathrm{T}}\mathbf{b} + \phi_1(\xi_i)\tilde{Z}_{1,i} + \phi_4(\xi_i)\tilde{Z}_{4,i},$$

where $\mathbf{b}_8 = 2, \mathbf{b}_{15} = 1.5, \mathbf{b}_{25} = 1.5$ and $\mathbf{b}_j = 0$ for all other j's, $\xi_i \sim N(0,1)$, and ϕ_1, ϕ_4 are monotone functions shown in Figure (4.1). We first generate $\acute{\mathbf{Z}}_i \sim N_p(0, \mathbf{\Sigma})$ as in Example 1 case ii), and then let $\widetilde{\mathbf{Z}}_i = \acute{\mathbf{Z}}_i$, except $\tilde{Z}_{1,i} = |\acute{Z}_{1,i}|+0.5$ and $\tilde{Z}_{4,i} = |\acute{Z}_{4,i}|+0.5$. Therefore $\beta_1^*(\tau) = 0, \beta_2^*(\tau) = \phi_1(\tau), \beta_5^*(\tau) = \phi_4(\tau)$, and $\beta_j^*(\tau) = \mathbf{b}_{j+1}$, for all other j's. The censoring time is generated following the same distribution as in Example 1.

Table (4.1) summarizes the result from Example 1, where the coefficient estimates are unbiased and the standard errors agree with the empirical standard deviations. It leads to proper coverage probabilities and good testing power when the effect size is large enough. Moreover, the estimation and inference from our proposed method are reliable for small signals even when their selection frequencies are low.

Under the setting of Example 2, we first compare the performance of two proposed variance estimations, (4.4) and (4.5). As shown in Figure (4.2), both estimators converge to the empirical truth as B increases, while the bias corrected version (4.5) converges much faster than the original version (4.4) and becomes stable for B less than 200. Table (4.2) summarizes the result by Fused-HDCQR comparing to the oracle estimation, which is based on low dimensional CQR knowing the true active set, and the standard errors are estimated via bootstrap (*Peng and Huang*, 2008). While the performance on invariant coefficients remains accurate, the estimation and inference for $\beta_4(\tau)$ is reliable as well in terms of little bias and proper coverage probabilities at different τ values.

Table (4.5) summarizes the estimation and inference by Fused-HDCQR comparing to the oracle results in Example 3. Even with two varying coefficients $\beta_2^*(\tau) = \phi_1(\tau), \beta_5^*(\tau) = \phi_4(\tau)$, our proposed procedure still provides accurate point estimations

and reliable inferences and testing powers at different τ values. The Fused estimator and confidence intervals for $\beta_2^*(\cdot)$ and $\beta_5^*(\cdot)$ are shown in Figure (4.1).

4.6 Data Application

Finding significant genetic variants that are associated with patients' survival has been one of the main themes in modern translational cancer studies. The BLCSC provides a rich data set measuring over $40,000$ SNP variations among $n = 674$ lung cancer patients. We are specifically interested in modeling the effects of those SNP that belong to certain high risk cancer-related genes, and have extracted $2,002$ SNPs located on 14 previously documented genes. In addition, we take into account the demographic variables, including age, gender, race, education level and smoking status (Table (4.8)). The longest survival time is over 23 years (8584 days) while the shortest event time is only 13 days. The censoring rate is 0.23, which is assumed to be independent of T.

We chose $[\tau_L, \tau_U] = [0.2, 0.7]$ so that only 2.4% of observations were censored below the τ_L-th quantile and there were sufficient data to estimate the survival distribution at τ_U-th quantile, with $\epsilon_n = 1/80$ to form the τ-grid

$$\Gamma_m = \{\tau_1 = 0.2, \tau_2 = 0.2125, .., \tau_m = 0.7\}$$

of length $m = 41$. We used L-HDCQR for variable selection and $B = 700$ as the number of re-samples, which was sufficiently large compared to the sample size. First step was to fit the initial HDCQR at $\tau = \tau_L$ with 5-fold cross validation to choose a fixed λ_n. Next, we ran Fused-HDCQR with the chosen λ_n on Γ_m.

For the sake of conciseness, we summarized and reported the result at 6 equally spaced τ values $\{0.2, 0.3, .., 0.7\}$ instead of the whole grid Γ_m. As we were especially interested in the risk factors for the high risk patients group (small τ), in Table

(4.9), we ranked the SNPs based on their p-values at $\tau = 0.2$. After adjusting for multiple testing using Bonferroni correction, there were still 87 significant SNP effects at level 0.05 and $\tau = 0.2$. Thus we only reported the top 10 significant SNPs, as well as the bottom 3 least significant ones to illustrate that our approach was comprehensive in estimating and drawing inferences for all potential predictors. From Table (4.9), the effect of smoking was around -0.4 across different quantiles, with decreased significance level for larger quantiles. This suggested that while the effect of smoking was constant, there were less data to support its significance. On the other hand, both the point estimation and the significance level of some top SNPs varied for different τ values. For examples, for τ from 0.2 to 0.7, the effect of SNP AX.83072863_A dropped from 2.58 to 0.28 and the standard errors increased from 0.23 to 0.39; the effect of SNP AX.83265037_A dropped from 2.33 to -0.11 with standard errors from 0.24 to 0.31. This suggested strong evidence of heterogeneous SNP effects for different risk groups. Therefore, if other quantiles were of interest, investigators could rank the SNPs based on their respective p-values. Furthermore, we mapped the 87 significant SNPs at $\tau = 0.2$ to their corresponding genes and ranked them by their respective numbers of significant SNPs in the parenthesis over total number of SNPs for the gene. They were TP53 (18/321), ALK (9/163), BRCA1 (9/114), ERCC1 (9/167), RRM1 (8/174), ROS1 (7/294), ERBB2 (6/167), EGFR (5/261), BRAF (4/49), RET (4/38), and 4 others with numbers of SNPs less than 4. While there were overwhelming evidence that these genes are associated with lung cancer (*Toyooka et al.*, 2003; *Takeuchi et al.*, 2012; *Rosell et al.*, 2011; *Lord et al.*, 2002; *Zheng et al.*, 2007; *Sasaki et al.*, 2006; *Brose et al.*, 2002), our analysis provided more detailed information as to which SNPs and locations of the genes were associated with the lung cancer survival, as well as the effect sizes and their significance levels.

4.7 Conclusion

Motivated by analyzing survival data with censoring and a large number of potential predictors, we have proposed a framework based on high dimensional censored quantile regressions for simultaneous estimation and inference of the model parameters. Using censored quantile regressions for survival analysis is advantageous in several aspects comparing to Cox proportional hazards model, as it models the extreme quantiles of the outcome distribution and is more powerful in detecting significant heterogeneous effects at different quantiles. The Fused-HDCQR procedure consists of two components, point estimator and variance estimator of the model coefficients. The fused coefficient estimator is the average of a large number of estimators that are derived from multiple random sample splits. The model-free variance estimator is derived using functional delta method and the multi-sample splitting properties, and is corrected for the bias caused by the finite number of re-samples B. By defining a fine grid of quantile values on an interval of interest, we are able to provide comprehensive understanding of the conditional quantiles, as well as precise inferences for each predictor and quantile. Our procedure is straightforward to implement, and computationally efficient, especially when implemented with parallel computing for the multiple re-samples. We conclude that our procedure could be extended to other types of censoring, for example left truncation, as well as modeling time-dependent covariates or time-varying effects.

Table 4.1: Summary of fitting Fused-HDCQR in Example 1 with $n = 300, p = 500$, invariant effects, and two covariance structures.

	Oracle			Fused Est			SE			Emp SD			Cov Prob			Power			Freq
Index	$\tau = 0.3$	0.5	0.7	0.3	0.5	0.7	0.3	0.5	0.7	0.3	0.5	0.7	0.3	0.5	0.7	0.3	0.5	0.7	-
i) $Unif(-1, 1)$																			
Int	-0.49	0.02	0.55	-0.50	0.03	0.57	0.08	0.07	0.08	0.08	0.07	0.08	0.93	0.97	0.93	1.00	0.03	1.00	-
20	0.27	0.24	0.24	0.26	0.24	0.24	0.14	0.13	0.14	0.14	0.14	0.16	0.94	0.93	0.88	0.49	0.51	0.46	0.12
40	0.48	0.48	0.47	0.48	0.48	0.47	0.13	0.13	0.14	0.14	0.14	0.15	0.94	0.93	0.90	0.93	0.93	0.93	0.45
60	0.73	0.74	0.72	0.74	0.74	0.74	0.13	0.13	0.14	0.13	0.13	0.13	0.96	0.96	0.93	1.00	1.00	1.00	0.88
80	0.98	0.98	0.98	0.99	0.99	0.99	0.14	0.13	0.14	0.13	0.15	0.17	0.96	0.93	0.85	1.00	1.00	1.00	0.99
100	1.25	1.26	1.29	1.25	1.26	1.28	0.14	0.13	0.14	0.12	0.12	0.13	0.97	0.94	0.93	1.00	1.00	1.00	1.00
0's	-	-	-	0.00	0.00	-0.00	0.14	0.13	0.14	0.13	0.13	0.14	0.94	0.94	0.94	0.06	0.06	0.06	0.01
ii) AR(1), $N_p(0, \mathbf{\Sigma})$																			
Int	-0.50	0.03	0.56	-0.51	0.03	0.58	0.09	0.08	0.09	0.08	0.08	0.10	0.92	0.95	0.86	1.00	0.05	1.00	-
20	0.25	0.26	0.26	0.26	0.26	0.26	0.08	0.08	0.09	0.08	0.08	0.09	0.89	0.89	0.93	0.90	0.83	0.79	0.23
40	0.51	0.52	0.51	0.52	0.51	0.52	0.09	0.08	0.09	0.09	0.09	0.09	0.92	0.94	0.92	1.00	1.00	1.00	0.85
60	0.75	0.76	0.76	0.75	0.76	0.75	0.08	0.08	0.09	0.09	0.08	0.08	0.90	0.92	0.95	1.00	1.00	1.00	1.00
80	1.00	0.99	1.00	1.00	0.99	1.00	0.08	0.08	0.09	0.07	0.08	0.08	0.99	0.95	0.96	1.00	1.00	1.00	1.00
100	1.25	1.25	1.26	1.25	1.25	1.25	0.08	0.09	0.09	0.08	0.08	0.09	0.96	0.99	0.93	1.00	1.00	1.00	1.00
0's	-	-	-	0.00	0.00	0.00	0.08	0.08	0.09	0.08	0.08	0.09	0.94	0.94	0.94	0.06	0.06	0.06	0.01

Figure 4.1: Two heterogeneous effects and their estimation and confidence intervals. $\beta_2^*(\cdot)$ (Left), and $\beta_5^*(\cdot)$ (Right).

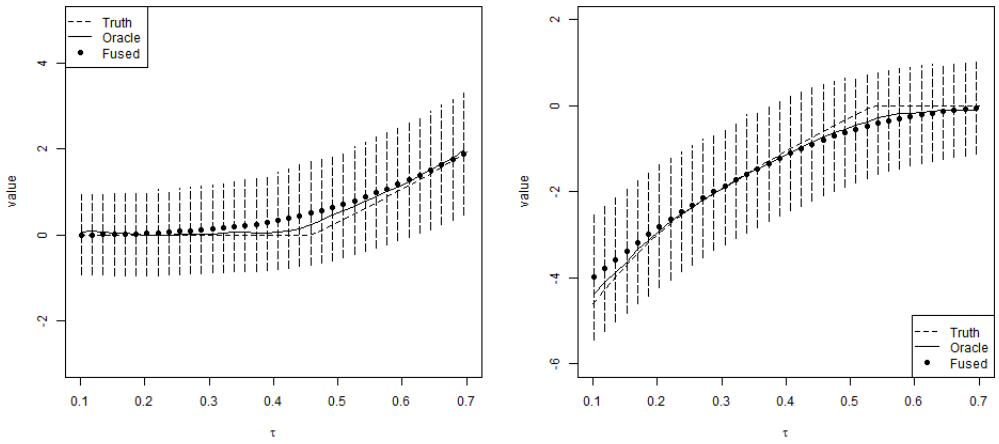


Figure 4.2: Two SE estimators with varying B, versus the empirical standard deviation in Example 2.

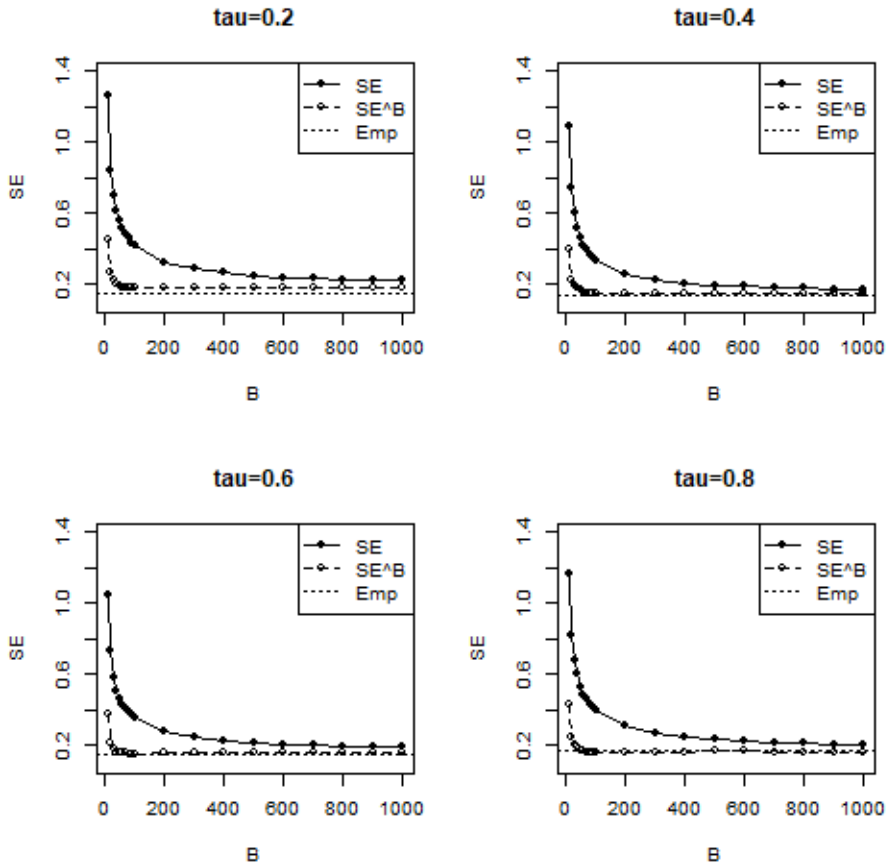


Table 4.2: Oracle estimation for Example 2, with $n = 200, p = 300$.

	Truth				Est				Boot SE				Emp sd			
Index	$\tau = 0.2$	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
Int	0	0	0	0	-0.02	-0.02	0.00	0.02	0.47	0.42	0.42	0.50	0.46	0.38	0.39	0.45
3	-1.26	-0.38	0.38	1.26	-1.18	-0.30	0.43	1.31	0.45	0.40	0.41	0.48	0.47	0.38	0.39	0.45
20	1	1	1	1	1.00	0.99	0.99	0.99	0.19	0.17	0.17	0.20	0.16	0.15	0.16	0.19
40	1.5	1.5	1.5	1.5	1.51	1.49	1.49	1.51	0.20	0.17	0.17	0.21	0.17	0.15	0.15	0.18
60	1	1	1	1	1.00	1.01	1.00	1.00	0.19	0.17	0.17	0.20	0.17	0.15	0.15	0.18
80	1.5	1.5	1.5	1.5	1.48	1.50	1.51	1.49	0.19	0.17	0.17	0.21	0.17	0.15	0.16	0.20

Table 4.3: Estimation and inference for $\beta_4(\tau)$.

	Truth	Est	SE	Emp sd	Cov Prob	Power
0.2	-1.26	-1.11	0.38	0.37	0.93	0.82
0.4	-0.38	-0.37	0.34	0.37	0.92	0.23
0.6	0.38	0.31	0.35	0.40	0.92	0.20
0.8	1.26	1.07	0.40	0.43	0.89	0.69

Table 4.4: All other covariates, the truth is displayed in the table for oracle estimation.

	Est				SE				Emp sd				
Index	$\tau = 0.2$	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8	Sel freq
Int	-0.14	0.02	0.13	0.33	0.40	0.35	0.36	0.42	0.39	0.38	0.40	0.44	0
20	1.00	1.01	1.01	1.00	0.16	0.14	0.15	0.18	0.16	0.15	0.14	0.15	0.94
40	1.49	1.49	1.49	1.50	0.16	0.14	0.15	0.18	0.16	0.15	0.16	0.20	1.00
60	1.02	1.01	1.01	1.01	0.16	0.15	0.15	0.17	0.16	0.14	0.14	0.16	0.95
80	1.48	1.49	1.50	1.51	0.15	0.14	0.15	0.17	0.16	0.15	0.15	0.18	1.00
0's	0.00	0.00	0.00	0.00	0.16	0.15	0.15	0.17	0.15	0.14	0.15	0.17	0.02

	Cov prob				Power			
τ	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
Int	0.91	0.92	0.93	0.86	0.09	0.08	0.07	0.14
20	0.94	0.91	0.96	0.96	1.00	1.00	1.00	1.00
40	0.92	0.92	0.95	0.94	1.00	1.00	1.00	1.00
60	0.96	0.96	0.95	0.95	1.00	1.00	1.00	1.00
80	0.92	0.93	0.94	0.95	1.00	1.00	1.00	1.00
0's	0.93	0.94	0.93	0.93	0.07	0.06	0.07	0.07

Table 4.5: Oracle estimation for Example 3, with $n = 300, p = 400$.

	Truth				Est				Boot SE				Emp sd			
τ	0.24	0.39	0.54	0.69	0.24	0.39	0.54	0.69	0.24	0.39	0.54	0.69	0.24	0.39	0.54	0.69
Int	0	0	0	0	-0.08	0.05	0.02	0.16	1.09	1.10	1.16	1.11	1.01	0.93	0.99	1.04
1	0	0	0.60	1.88	0.01	0.04	0.73	1.83	0.56	0.62	0.77	0.83	0.49	0.50	0.71	0.77
4	-2.57	-1.14	-0.01	0	-2.43	-1.10	-0.19	0.01	0.79	0.75	0.70	0.59	0.84	0.73	0.63	0.49
8	2	2	2	2	1.99	1.98	1.98	2.02	0.33	0.34	0.37	0.35	0.30	0.33	0.35	0.33
15	1.5	1.5	1.5	1.5	1.51	1.49	1.49	1.51	0.33	0.34	0.37	0.35	0.30	0.30	0.36	0.30
25	1.5	1.5	1.5	1.5	1.51	1.52	1.51	1.52	0.33	0.34	0.36	0.35	0.30	0.31	0.35	0.32

Table 4.6: Estimation by Fused-HDCQR.

	Est				SE				Emp sd				
τ	0.24	0.39	0.54	0.69	0.24	0.39	0.54	0.69	0.24	0.39	0.54	0.69	Sel freq
Int	-0.12	0.13	0.27	0.56	0.88	0.91	0.89	0.85	0.92	0.96	1.00	1.09	0.00
1	0.11	0.29	0.89	1.86	0.50	0.56	0.64	0.72	0.55	0.60	0.71	0.83	0.90
4	-2.42	-1.26	-0.50	-0.16	0.71	0.67	0.58	0.51	0.68	0.66	0.60	0.51	1.00
8	2.00	2.00	2.01	2.02	0.27	0.28	0.30	0.30	0.27	0.29	0.30	0.32	1.00
15	1.54	1.53	1.51	1.50	0.29	0.29	0.30	0.30	0.29	0.29	0.29	0.29	0.90
25	1.47	1.46	1.45	1.46	0.29	0.29	0.30	0.30	0.29	0.31	0.33	0.33	0.88
0's	0.00	-0.00	-0.00	-0.00	0.28	0.29	0.29	0.30	0.28	0.28	0.29	0.29	0.02

Table 4.7: Inference by Fused-HDCQR.

	Cov prob				Power			
τ	0.24	0.39	0.54	0.69	0.24	0.39	0.54	0.69
Int	0.89	0.92	0.89	0.77	0.11	0.08	0.11	0.23
1	0.94	0.91	0.87	0.90	0.06	0.09	0.27	0.72
4	0.92	0.92	0.87	0.91	0.94	0.46	0.14	0.09
8	0.95	0.94	0.94	0.91	1.00	1.00	1.00	1.00
15	0.96	0.95	0.92	0.90	1.00	1.00	1.00	1.00
25	0.93	0.93	0.92	0.92	1.00	1.00	1.00	0.99
0's	0.93	0.93	0.93	0.93	0.07	0.07	0.07	0.07

Table 4.8: Demographic table of the lung cancer SNP data.

Variable	Mean	SD	Count (%)
Age	60	10.8	-
Female (vs. male)	-	-	259 (38.4)
White (vs. non-white)	-	-	635 (94.2)
Education level			
< High school	-	-	93 (13.8)
High school	-	-	171 (25.4)
> High school	-	-	410 (60.8)
Smoking			
Never	-	-	64 (9.5)
Past	-	-	354 (52.5)
Current	-	-	256 (38)

Table 4.9: Analysis of BLCSC lung cancer patients data with Fused-HDCQR. The SNPs are sorted by their p-values at $\tau = 0.2$, corresponding to the high risk effects.

	Est	SE	P	Est	SE	P	Est	SE	P
$\tau =$		0.2			0.3			0.4	
Intercept	5.51	0.40	5.3E-44	5.86	0.42	3.7E-44	6.28	0.49	3.3E-37
Age	5.3E-3	5.9E-3	0.38	5.7E-3	6.1E-3	0.35	5.5E-3	7.1E-3	0.43
Education	0.02	0.09	0.79	0.05	0.11	0.62	0.10	0.12	0.4
Female	-0.15	0.14	0.29	-0.24	0.11	3.2E-2	-0.29	0.14	4.1E-2
Smoke	-0.42	0.09	4.0E-6	-0.43	0.13	9.7E-4	-0.46	0.15	2.3E-3
AX.83102092_A	3.24	0.25	2.1E-39	2.81	0.26	5.2E-28	2.39	0.27	3.4E-18
AX.83072863_A	2.58	0.23	5.2E-29	2.18	0.22	9.4E-24	1.78	0.23	4.7E-15
AX.13920550_G	2.99	0.27	5.5E-28	2.54	0.30	3.8E-17	2.15	0.34	1.8E-10
AX.13917089_A	2.97	0.28	4.0E-26	2.53	0.30	1.1E-16	2.12	0.35	9.3E-10
AX.83283864_C	2.30	0.22	1.3E-24	1.86	0.24	5.5E-15	1.45	0.23	1.1E-10
AX.83265037_A	2.33	0.24	1.2E-21	1.88	0.27	7.1E-12	1.43	0.31	4.5E-6
AX.82976133_A	2.10	0.22	2.4E-21	1.69	0.25	5.7E-12	1.30	0.26	7.3E-7
AX.83028383_C	2.74	0.29	5.1E-21	2.30	0.34	1.0E-11	1.85	0.34	5.0E-8
AX.36265829_T	2.57	0.27	6.3E-21	2.13	0.27	1.1E-15	1.72	0.31	2.2E-8
AX.83331966_A	2.73	0.29	1.5E-20	2.27	0.31	3.0E-13	1.87	0.35	9.4E-8
...									
AX.38313601_G	-1.8E-4	0.39	1.00	0.02	0.36	0.96	0.00	0.35	1.00
AX.83233196_A	3.1E-4	0.87	1.00	-0.40	0.88	0.65	-0.38	1.10	0.73
AX.15536107_A	-3.0E-6	0.08	1.00	-0.07	0.09	0.45	-0.09	0.12	0.47

	Est	SE	P	Est	SE	P	Est	SE	P
$\tau =$		0.5			0.6			0.7	
Intercept	7.14	0.67	2.9E-26	8.24	0.62	4.0E-40	9.16	0.48	9.9E-82
Age	4.1E-4	0.01	0.97	-9.4E-2	0.01	0.33	-0.02	0.007	9.7E-3
Education	0.14	0.12	0.27	0.12	0.11	0.3	0.11	0.09	0.19
Female	-0.40	0.17	2.0E-2	-0.36	0.17	3.2E-2	-0.32	0.14	1.7E-2
Smoke	-0.46	0.15	2.0E-3	-0.42	0.16	9.7E-3	-0.39	0.13	1.9E-3
AX.83102092_A	1.89	0.32	2.2E-9	1.38	0.30	2.8E-6	0.97	0.24	6.1E-5
AX.83072863_A	1.28	0.32	5.4E-5	0.77	0.34	2.3E-2	0.28	0.39	0.47
AX.13920550_G	1.67	0.37	6.1E-6	1.25	0.39	1.5E-3	0.84	0.34	1.3E-2
AX.13917089_A	1.63	0.37	1.1E-5	1.19	0.35	8.0E-4	0.79	0.30	8.8E-3
AX.83283864_C	1.04	0.25	3.2E-5	0.74	0.26	5.3E-3	0.37	0.25	0.14
AX.83265037_A	0.90	0.37	1.6E-2	0.36	0.37	0.32	-0.11	0.31	0.72
AX.82976133_A	0.85	0.39	2.8E-2	0.41	0.56	0.46	-0.05	0.45	0.9
AX.83028383_C	1.34	0.39	6.0E-4	0.81	0.37	2.8E-2	0.37	0.30	0.23
AX.36265829_T	1.31	0.35	2.3E-4	0.81	0.38	3.0E-2	0.38	0.32	0.24
AX.83331966_A	1.38	0.44	1.7E-3	0.90	0.44	3.9E-2	0.41	0.34	0.24
...									
AX.38313601_G	-0.02	0.41	0.97	-0.09	0.52	0.86	-0.12	0.51	0.82
AX.83233196_A	-0.73	1.13	0.52	-0.71	1.20	0.56	-0.96	1.32	0.47
AX.15536107_A	-0.11	0.14	0.42	-0.07	0.14	0.64	0.01	0.10	0.89

APPENDICES

APPENDIX A

Chapter II Supplementary Material

Lemmas and Proofs

Proof of Theorem II.2. Our estimator for β_j^0 by the one-time SPARE is

$$\tilde{\beta}_j = \left\{ (X^1_{S\cup j}{}^{\mathrm{T}} X^1_{S\cup j})^{-1} X^1_{S\cup j}{}^{\mathrm{T}} Y^1 \right\}_j .$$

Here $D_1 = (X^1, Y^1)$ with sample size $\lfloor n/2 \rfloor$, for notational simplicity, we denote $m = \lfloor n/2 \rfloor$ within this proof.

By (A3), with probability at least $1 - o(m^{-c_2-1})$, the selection $S \supset S_{0,n}$. Since the two halves of data D_1 and D_2 are mutually exclusive, $(X^1, Y^1) \perp S$. Thus given $S \supset S_{0,n}$ and X^1, the OLS estimator $\tilde{\beta}^1 = (X^1_{S\cup j}{}^T X^1_{S\cup j})^{-1} X^1_{S\cup j}{}^T Y^1$ is unbiased,

$$\begin{aligned}
&\mathbf{E}\left(\tilde{\beta}^1 \middle| S, X^1\right) \\
=&\mathbf{E}\left((X^1_{S\cup j}{}^T X^1_{S\cup j})^{-1} X^1_{S\cup j}{}^T X^1 \beta^0 \middle| S, X^1\right) + \mathbf{E}\left((X^1_{S\cup j}{}^T X^1_{S\cup j})^{-1} X^1_{S\cup j}{}^T X^1 \boldsymbol{\varepsilon}^1 \middle| S, X^1\right) \\
=&\mathbf{E}\left((X^1_{S\cup j}{}^T X^1_{S\cup j})^{-1} X^1_{S\cup j}{}^T X^1_{S\cup j} \beta^0_{S\cup j} \middle| S, X^1\right) + \mathbf{E}\left(\boldsymbol{\varepsilon}^1 \middle| S, X^1\right) \\
=&\beta^0_{S\cup j}.
\end{aligned}$$

In addition, $\mathrm{Var}\left(\tilde{\beta}^1 \middle| S, X^1\right) = \sigma^2 \Sigma^{-1}_{S\cup j}/m$, which is bounded by assumption (A1). Thus,

$$\sqrt{m}(\tilde{\beta}^1 - \beta^0_{S\cup j}) \Big| S, X^1 \xrightarrow{d} N(0, \sigma^2 \Sigma^{-1}_{S\cup j}).$$

Furthermore,

$$\sqrt{m}(\tilde{\beta}_j - \beta^0_j) \Big| S, X^1 \xrightarrow{d} N(0, \tilde{\sigma}^2_j),$$

where $\tilde{\sigma}^2_j = \sigma^2 \left(\Sigma^{-1}_{S\cup j}\right)_{jj}$.

Next we show the uniform convergence of $\sqrt{m}(\tilde{\beta}_j - \beta^0_j)/\tilde{\sigma}_j$ with respect to j, S and X^1. From the partial regression formulation of $\tilde{\beta}_j$, if $S \supset S_{0,n}$,

$$\tilde{\beta}_j - \beta^0_j = \frac{X_j^{1\mathrm{T}}(I_m - H^1_{S\setminus j})\boldsymbol{\varepsilon}^1}{X_j^{1\mathrm{T}}(I_m - H^1_{S\setminus j})X^1_j} = \frac{m}{X_j^{1\mathrm{T}}(I_m - H^1_{S\setminus j})X^1_j} \frac{X_j^{1\mathrm{T}}(I_m - H^1_{S\setminus j})\boldsymbol{\varepsilon}^1}{m}. \quad \text{(A.1)}$$

By Lemma (A.1),

$$\frac{m}{X_j^{1\mathrm{T}}(I_m - H^1_{S\setminus j})X^1_j} = \left(\widehat{\Sigma}^{-1}_{S\cup j}\right)_{jj} \to \left(\Sigma^{-1}_{S\cup j}\right)_{jj},$$

and $\forall j, S$, $\left|\frac{m}{X_j^{1\mathrm{T}}(I_m - H^1_{S\setminus j})X^1_j}\right| \leq 2/c_{\min}$. Moreover, the second term of the right hand side in (A.1) is the mean of i.i.d. $\tilde{x}^1_{ij}\varepsilon^1_i$'s, where $(\tilde{x}^1_{ij})_{i=1,..,m} = X^1_j(I_m - H^1_{S\setminus j})$. Since $\mathbf{E}|\boldsymbol{\varepsilon}_i|^3 \leq \rho_0$ and $X^1_j(I_m - H^1_{S\setminus j})$ is the projection vector of X^1_j,

$$\mathbf{E}|X^1_j(I_m - H^1_{S\setminus j})|^3_\infty \leq \mathbf{E}|X^1_j|^3_\infty \leq \rho_1.$$

By the Berry-Esseen Theorem, $\forall j$, X and $S \supset S_{0,n}$,

$$|F_n(x) - \Phi(x)| \leq \left(\frac{2}{c_{\min}}\right)^3 \frac{C\rho_0\rho_1}{\tilde{\sigma}^3_j\sqrt{m}} \leq \frac{8c^{3/2}_{\max}C\rho_0\rho_1}{c^3_{\min}\sigma^3\sqrt{m}},$$

where $F_n(x)$ is the CDF of $\sqrt{m}(\tilde{\beta}_j - \beta^0_j)/\tilde{\sigma}_j$ and $\Phi(x)$ is the CDF of standard normal.

Thus as $m \to \infty$, with probability at least $1 - o(m^{-c_2-1})$,

$$\sqrt{m}(\tilde{\beta}_j - \beta_j^0)/\tilde{\sigma}_j \to N(0,1).$$

□

Proof of Theorem II.4. We first introduce the *oracle* SPARE estimators of β_j^0's, i.e. the ones we would compute if we knew the true active set $S_{0,n}$,

$$\hat{\beta}_j^0 = \left\{ (X_{S_{0,n}\cup j}{}^T X_{S_{0,n}\cup j})^{-1} X_{S_{0,n}\cup j}{}^T Y \right\}_j$$
$$\hat{\beta}_{j,S_{0,n}}^b = \left\{ (X_{S_{0,n}\cup j}^b{}^T X_{S_{0,n}\cup j}^b)^{-1} X_{S_{0,n}\cup j}^b{}^T Y^b \right\}_j,$$

which are estimations on the original data (X, Y) and the bootstrap half data D_1^b, respectively. Since $\hat{\beta}_j^0$ is the least square corresponding to X_j when regressing Y on $X_{S_{0,n}\cup j}$, we have for each j

$$W_j^0 = \sqrt{n}(\hat{\beta}_j^0 - \beta_j^0)/\sigma_j \xrightarrow{d} N(0,1) \quad \text{as} \quad n \to \infty, \tag{A.2}$$

where $\sigma_j^2 = \sigma^2 \left(\Sigma_{S_{0,n}\cup j}^{-1} \right)_{jj}$ that corresponds to subscript j. By Cauchy's interlacing theorem (Proposition A.3), $\sigma^2/c_{\max} \le \sigma_j^2 \le \sigma^2/c_{\min}$, and thus it is bounded away from zero and infinity.

Now we consider the behavior of the selections S^b's from D_2^b's. For each $b = 1, 2, ..., B$, the subsample D_2^b consists of $m_b \ge n/2$ distinct observations from the original data that are not drawn in the bootstrap half dataset D_1^b. In other words, D_2^b can be regarded as a sample of m_b i.i.d. observations from the population distribution. In addition, since m_b is independent of the observations, with a conditional argument

on m_b, the following holds for each b by (B3),

$$\begin{aligned}
&\mathbf{P}(S^b = S_{0,n}) \\
=&\int \mathbf{P}(S^b = S_{0,n}|m_b = m)\mathrm{d}\mathbf{P}(m) \\
\geq&\int \Big\{1 - o(m^{-c_2-1})\Big\}\mathrm{d}\mathbf{P}(m) \\
\geq&1 - o\{(n/2)^{-c_2-1}\} \\
=&1 - o(n^{-c_2-1}).
\end{aligned}$$

Next, we decompose $\hat{\beta}_j$ into two parts:

$$\begin{aligned}
\hat{\beta}_j =&\frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_j^b \\
=&\frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_{j,S_{0,n}}^b + \frac{1}{B}\sum_{b:S^b\neq S_{0,n}}\Big(\hat{\beta}_j^b - \hat{\beta}_{j,S_{0,n}}^b\Big),
\end{aligned} \tag{A.3}$$

and equivalently

$$\begin{aligned}
&\sqrt{n}(\hat{\beta}_j - \beta_j^0) \\
=&\sqrt{n}\Big(\frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_{j,S_{0,n}}^b - \beta_j^0\Big) + \frac{\sqrt{n}}{B}\sum_{b:S^b\neq S_{0,n}}\Big(\hat{\beta}_j^b - \hat{\beta}_{j,S_{0,n}}^b\Big) \\
\doteq&Z_j^0 + \Delta_j.
\end{aligned} \tag{A.4}$$

To show $\Delta_j = o_p(1)$, we write

$$\Delta_j = \frac{1}{B}\sum_{b=1}^{B}\mathbf{1}(S^b \neq S_{0,n})\sqrt{n}\Big(\hat{\beta}_j^b - \hat{\beta}_{j,S_{0,n}}^b\Big);$$

$$\Delta_j = \frac{1}{B}\sum_{b=1}^{B}\delta_b; \quad \delta_b \doteq \mathbf{1}(S^b \neq S_{0,n})\sqrt{n}\Big(\hat{\beta}_j^b - \hat{\beta}_{j,S_{0,n}}^b\Big).$$

By Corollary (A.2),

$$\begin{aligned}\mathbf{E}\delta_b =&\mathbf{P}(S^b \neq S_{0,n})\mathbf{E}\sqrt{n}\Big(\hat{\beta}_j^b - \hat{\beta}_{j,S_{0,n}}^b\Big)\\ =&o\Big(n^{-c_2-1}2C_\beta n^{c_1+\frac{1}{2}}\Big)\\ =&o\Big(n^{-c_2+c_1-\frac{1}{2}}\Big)\\ \to&0 \quad \text{as} \quad n\to\infty.\end{aligned}$$

Similarly,

$$\begin{aligned}\mathbf{Var}\delta_b =&\mathbf{P}(S^b \neq S_{0,n})\mathbf{E}n\Big(\hat{\beta}_j^b - \hat{\beta}_{j,S_{0,n}}^b\Big)^2\\ =&o\Big(n^{-c_2-1}4C_\beta^2 n^{2c_1+1}\Big)\\ =&o(n^{-c_2+2c_1})\\ \to&0 \quad \text{as} \quad n\to\infty.\end{aligned}$$

Thus $\delta_b = o_p(1)$ for all $b \in [B]$. Furthermore, since $\mathbf{E}\Delta_j = \mathbf{E}\delta_b$ and $\mathbf{Var}\Delta_j \leq \mathbf{Var}\delta_b$, we have $\Delta_j = o_p(1)$.

Next, we show the convergence of Z_j^0. Notice that

$$Z_j^0/\sigma_j = W_j^0 + \sqrt{n}\Big(\frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_{j,S_{0,n}}^b - \hat{\beta}_j^0\Big)/\sigma_j \doteq W_j^0 + T_n^B/\sigma_j.$$

By (A.2), we are only left to show $T_n^B = o_p(1)$. Define $t_{n,b} = \sqrt{n}\big(\hat{\beta}_{j,S_{0,n}}^b - \hat{\beta}_j^0\big)$, then $T_n^B = \sqrt{n}(\frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_{j,S_{0,n}}^b - \hat{\beta}_j^0) = \frac{1}{B}\sum_{b=1}^{B} t_{n,b}$. Recall that $\hat{\beta}_{j,S_{0,n}}^b$ is the bootstrap statistic of $\hat{\beta}_j^0$, so its conditional mean is $\hat{\beta}_j^0$ and conditional variance is

$$\hat{\sigma}^2\Big\{(X_{S_{0,n}\cup j}^T X_{S_{0,n}\cup j})^{-1}\Big\}_{jj} = \hat{\sigma}^2\Big(\widehat{\Sigma}_{S_{0,n}\cup j}^{-1}\Big)_{jj}/n \doteq \hat{\sigma}_j^2/n,$$

where $\hat{\sigma}^2 = \|(\mathrm{I}_n - H_{S_{0,n}})Y\|_2^2/n$ (*Freedman et al.* (1981)). Thus, conditional on the

data, $\{t_{n,b}\}_{b=1,2,..,B}$ are i.i.d. with

$$\mathbf{E}\big(t_{n,b}|(X^{(n)},Y^{(n)})\big)=0,\quad \mathbf{Var}\big(t_{n,b}|(X^{(n)},Y^{(n)})\big)=\hat{\sigma}_j^2=\hat{\sigma}^2\Big(\widehat{\Sigma}^{-1}_{S_{0,n}\cup j}\Big)_{jj}.$$

We now argue that with probability going to 1, $\hat{\sigma}_j^2$'s, $j=1,2,..,p$, are bounded. First, $\mathbf{P}(\hat{\sigma}^2<2\sigma^2)\to 1$ as $n\to\infty$. Then,

$$\Big(\widehat{\Sigma}^{-1}_{S_{0,n}\cup j}\Big)_{jj}\leq \lambda_{\max}(\widehat{\Sigma}^{-1}_{S_{0,n}\cup j})=1/\lambda_{\min}(\widehat{\Sigma}_{S_{0,n}\cup j}), \tag{A.5}$$

whenever $\lambda_{\min}(\widehat{\Sigma}_{S_{0,n}\cup j})>0$. Assumption (B3) implies $|S_{0,n}|/n\leq\eta$. By Lemma (A.4) from *Vershynin* (2010) and Lemma (A.7), letting $\epsilon=c_{\min}/2$ and $t^2=c_{\min}^2\eta/C$ for some constant C only depending on the sub-Gaussian norm $\|\mathbf{x}_i\|_{\psi_2}$, we have that with probability at least $1-2\exp(-c_{\min}^2\eta n^{\gamma_0}/C)$

$$\lambda_{\min}(\widehat{\Sigma}_{S_{0,n}\cup j})\geq\lambda_{\min}(\Sigma_{S_{0,n}\cup j})-c_{\min}/2\geq\lambda_{\min}(\Sigma)-c_{\min}/2\geq c_{\min}/2, \tag{A.6}$$

where the second inequality follows the interlacing property of the eigenvalues. Combining (A.5) and (A.6), $\Big(\widehat{\Sigma}^{-1}_{S_{0,n}\cup j}\Big)_{jj}\leq 2/c_{\min}$ with probability going to 1 exponentially fast in n, and consequently $\hat{\sigma}_j^2<4\sigma^2/c_{\min}$. Now define

$$\Omega_n=\{(X^{(n)},Y^{(n)})=(\mathbf{x}_i,y_i)_{i=1,2,..,n}:\hat{\sigma}_j^2<4\sigma^2/c_{\min},\forall j=1,2,...,p\}.$$

Since $p=O(n^{\gamma_1})$ for some $\gamma_1>1$, $\mathbf{P}\{(X^{(n)},Y^{(n)})\in\Omega_n\}\to 1$ as $n\to\infty$. Thus $\forall(X^{(n)},Y^{(n)})\in\Omega_n$, $\mathbf{Var}\big\{t_{n,b}|(X^{(n)},Y^{(n)})\big\}\leq 4\sigma^2/c_{\min}$. Furthermore,

$$\mathbf{Var}\big\{T_n^B|(X^{(n)},Y^{(n)})\big\}=\frac{1}{B^2}\sum_{b=1}^{B}\mathbf{Var}\big\{t_{n,b}|(X^{(n)},Y^{(n)})\big\}\leq\frac{4\sigma^2}{Bc_{\min}}$$

Thus, $\forall \delta, \zeta > 0$, $\exists N_0, B_0 > 0$ such that $\forall n > N_0, B > B_0$,

$$
\begin{aligned}
&\mathbf{P}(|T_n^B| \geq \delta) \\
\leq &\int_{\Omega_n} \mathbf{P}\left\{|T_n^B| \geq \delta \middle| (X^{(n)}, Y^{(n)})\right\} \mathrm{d}\mathbf{P}(X^{(n)}, Y^{(n)}) + \mathbf{P}\left\{(X^{(n)}, Y^{(n)}) \notin \Omega_n\right\} \\
\leq &\int_{\Omega_n} \frac{\mathbf{Var}\left\{T_n^B | (X^{(n)}, Y^{(n)})\right\}}{\delta^2} \mathrm{d}\mathbf{P}(X^{(n)}, Y^{(n)}) + \mathbf{P}\left\{(X^{(n)}, Y^{(n)}) \notin \Omega_n\right\} \\
\leq &\frac{4\sigma^2}{B_0 \delta^2 c_{\min}} \int_{\Omega_n} \mathrm{d}\mathbf{P}(X^{(n)}, Y^{(n)}) + \mathbf{P}\left\{(X^{(n)}, Y^{(n)}) \notin \Omega_n\right\} \\
\leq &\zeta/2 + \zeta/2 \\
\leq &\zeta.
\end{aligned}
$$

Finally, combining this with (A.2), we have

$$
Z_j^0/\sigma_j = W_j^0 + T_n^B/\sigma_j \xrightarrow{d} N(0,1) \quad \text{as} \quad B, n \to \infty.
$$

□

Proof of Theorem II.5. Follow the previous proof, we replace the arguments in j with those in $S^{(1)}$. The *oracle* estimators are

$$
\begin{aligned}
\hat{\beta}^0_{S^{(1)}} &= \left((X_{S_{0,n} \cup S^{(1)}}{}^T X_{S_{0,n} \cup S^{(1)}})^{-1} X_{S_{0,n} \cup S^{(1)}}{}^T Y \right)_{S^{(1)}} \\
\hat{\beta}^b_{S^{(1)}, S_{0,n}} &= \left((X^b_{S_{0,n} \cup S^{(1)}}{}^T X^b_{S_{0,n} \cup S^{(1)}})^{-1} X^b_{S_{0,n} \cup S^{(1)}}{}^T Y^b \right)_{S^{(1)}}.
\end{aligned}
$$

Notice that $|S^{(1)}| = p_1 = O(1)$, as $n \to \infty$, $|S_{0,n} \cup S^{(1)}| = O\big(|S_{0,n}|\big) = o(n)$, so that the above quantities are well-defined. Next

$$
W^{(1)} = \sqrt{n}\{\Sigma^{(1)}\}^{-1}(\hat{\beta}^0_{S^{(1)}} - \beta^0_{S^{(1)}}) \xrightarrow{d} N(0, \mathrm{I}_{p_1}) \quad \text{as} \quad n \to \infty,
$$

where $\Sigma^{(1)} = \sigma^2 \left(\Sigma^{-1}_{S_{0,n} \cup S^{(1)}} \right)_{S^{(1)}}$. Similar to (A.4), we decompose $\sqrt{n}(\hat{\beta}_{S^{(1)}} - \beta^0_{S^{(1)}})$

into three parts:

$$\sqrt{n}(\hat{\beta}_{S^{(1)}} - \beta^0_{S^{(1)}})$$
$$\doteq Z^{(1)} + \Delta_0^{(1)} + \Delta_1^{(1)}.$$

For the sake of space, we prefer not to write out these quantities, but it is straightforward analog that $\Delta_0^{(1)} = \Delta_1^{(1)} = o_p(\mathbf{1}_{p_1})$ and ${\Sigma^{(1)}}^{-1} Z^{(1)} - W^{(1)} = o_p(\mathbf{1}_{p_1})$ as well, which completes the proof. □

Technical details on useful definitions, lemmas and related proofs.

Lemma A.1. *Assume* $X = (X_1, ..., X_p) = (x_1^{\mathrm{T}}, ..., x_n^{\mathrm{T}})^{\mathrm{T}}$ *where* x_i*'s are i.i.d. copies of a sub-Gaussian random vector in* $\mathbf{R}^p$ *with covariance matrix* $\Sigma_{p\times p}$*, with*

$$0 < c_{\min} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq c_{\max} < \infty.$$

For any subset $S \subset \{1, 2, .., p\}$ *with* $|S| \leq \eta n$*,* $0 < \eta < 1$*, and* $\forall j \in S$*, with probability at least* $1 - 2\exp(-\frac{\varepsilon^2 \eta}{C_K} n)$*,*

$$\frac{c_{\min}}{2} \leq \frac{1}{n} X_j^{\mathrm{T}} (I_n - H_{S\setminus j}) X_j \leq c_{\max} + \frac{1 + c_{\min}}{2}$$

where $\varepsilon = \min(\frac{1}{2}, \frac{c_{\min}}{2})$ *and* C_K *is the constant depends only on the sub-Gaussian norm* $K = \|x_i\|_{\psi_2}$*.*

Corollary A.2. *Given model (1) and assumptions (A1,A2), consider the partial regression estimator on* (X, Y) *given subset* S*. If* $|S| \leq \eta n$*,* $0 < \eta < 1$*, then with probability at least* $1 - 2\exp(-\frac{\varepsilon^2 \eta}{C_K} n)$*,*

$$\hat{\beta}_j \leq C_\beta n^{c_1},$$

where C_β *depends on* $c_{\min}, c_{\max}, c_\beta$*.*

Proposition A.3 (Cauchy interlacing theorem)**.** *Let A be a symmetric $n \times n$ matrix. The $m \times m$ matrix B, where $m \leq n$, is called a compression of A if there exists an orthogonal projection P onto a subspace of dimension m such that $P^{\mathrm{T}}AP = B$. The Cauchy interlacing theorem states:*
if the eigenvalues of A are $\lambda_1 \leq ... \leq \lambda_n$, and those of B are $\nu_1 \leq ... \leq \nu_m$, then for all $j < m + 1$,

$$\lambda_j \leq \nu_j \leq \lambda_{n-m+j}$$

Proposition A.4 (Corollary 5.50 in *Vershynin* (2010))**.** *Consider a $n \times q$ matrix X whose rows $\mathbf{x}_i$'s are i.i.d. samples from a sub-Gaussian distribution in R^q with covariance matrix Σ, and let $\epsilon \in (0, 1), t \geq 1$. Denote the sample covariance matrix as $\widehat{\Sigma}_n = X^{\mathrm{T}}X/n$ Then with probability at least $1 - 2\exp(-t^2 q)$ one has*

$$\textit{If} \quad n \geq C(t/\epsilon)^2 q \quad \textit{then} \quad \|\widehat{\Sigma}_n - \Sigma\| \leq \epsilon.$$

Here $C = C_K$ depends only on the sub-Gaussian norm $K = \|\mathbf{x}_i\|_{\psi_2}$ of a random vector taken from this distribution.

Definition A.5. The sub-Gaussian norm of a random variable V is defined as

$$\|V\|_{\psi_2} = \sup_{k \geq 1} k^{-1/2}(E|V|^k)^{1/k}$$

then the sub-Gaussian norm of a random vector V in R^q is defined as

$$\|V\|_{\psi_2} = \sup_{x \in S^{q-1}} \|V^{\mathrm{T}}x\|_{\psi_2}$$

Remark A.6. Assume $V_0 = (v_1, v_2, ..., v_q)$ is a sub-Gaussian random vector in R^q, and $V_1 = (v_1, v_2, ..., v_r), r < q$ is the sub-vector of V_0. By taking $x = (x_1, .., x_r, 0, .., 0) \in S^{q-1}$, we have $\|V_1\|_{\psi_2} \leq \|V_0\|_{\psi_2}$.

Corollary A.7. *For two $n \times n$ positive definite matrices Σ_1 and Σ_2, if $\|\Sigma_1 - \Sigma_2\| \leq \epsilon$, then*

$$\lambda_{\min}(\Sigma_2) \geq \lambda_{\min}(\Sigma_1) - \epsilon$$

$$\lambda_{\max}(\Sigma_2) \leq \lambda_{\max}(\Sigma_1) + \epsilon.$$

Proof. On one hand, $\forall n-$vector X with $\|X\|_2 = 1$,

$$\begin{aligned}
\epsilon \geq & \|\Sigma_1 - \Sigma_2\| \\
\geq & \|(\Sigma_1 - \Sigma_2)X\|_2 \\
\geq & \|\Sigma_1 X\|_2 - \|\Sigma_2 X\|_2
\end{aligned}$$

then take X to be the eigenvector for $\lambda_{\min}(\Sigma_2)$, we have

$$\begin{aligned}
\lambda_{\min}(\Sigma_2) = & \|\Sigma_2 X\|_2 \\
\geq & \|\Sigma_1 X\|_2 - \epsilon \\
\geq & \lambda_{\min}(\Sigma_1) - \epsilon.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\lambda_{\max}(\Sigma_2) = & \|\Sigma_2\| \\
\leq & \|\Sigma_1\| + \|\Sigma_2 - \Sigma_1\| \\
\leq & \|\Sigma_1\| + \epsilon \\
= & \lambda_{\max}(\Sigma_1) + \epsilon
\end{aligned}$$

□

Proof of lemma (A.1). Note that

$$\frac{n}{X_j^T (I_n - H_{S\setminus j}) X_j}$$

is the $(j,j)^{\text{th}}$ entry of $\widehat{\Sigma}_S^{-1}$, where $\widehat{\Sigma}_S = (X_S^T X_S)/n$ is the sample covariance matrix

corresponds to subset S. Therefore

$$\frac{1}{\lambda_{\max}(\widehat{\Sigma}_S)} \leq \frac{n}{X_j^T(I_n - H_{S\setminus j})X_j} \leq \frac{1}{\lambda_{\min}(\widehat{\Sigma}_S)}.$$

Refer to Corollary 5.50 in *Vershynin* (2010) and choose $\varepsilon = \min(\frac{1}{2}, \frac{c_{\min}}{2})$. Then with probability at least $1 - 2\exp(-\frac{\varepsilon^2\eta}{C_K}n)$,

$$\|\widehat{\Sigma}_S - \Sigma_S\| \leq \varepsilon.$$

By Corollary (A.7) and Cauchy interlacing theorem,

$$\lambda_{\min}(\widehat{\Sigma}_S) \geq \lambda_{\min}(\Sigma_S) - \varepsilon \geq \lambda_{\min}(\Sigma) - \varepsilon \geq c_{\min}/2,$$

and

$$\lambda_{\max}(\widehat{\Sigma}_S) \leq \lambda_{\max}(\Sigma_S) + \varepsilon \leq \lambda_{\max}(\Sigma) + \varepsilon \leq c_{\max} + (1 + c_{\min})/2.$$

Thus, with high probability,

$$\frac{c_{\min}}{2} \leq \frac{1}{n}X_j^T(I_n - H_{S\setminus j})X_j \leq c_{\max} + \frac{1 + c_{\min}}{2}$$

□

Proof of Corollary (A.2). From Lemma (A.1), we can bound $\hat{\beta}_j$ as below:

$$\begin{aligned}
\hat{\beta}_j =& \frac{X_j^{\mathrm{T}}(I - H_{S\setminus j})Y}{X_j^{\mathrm{T}}(I - H_{S\setminus j})X_j} \\
=& \frac{n}{X_j^{\mathrm{T}}(I - H_{S\setminus j})X_j} \frac{X_j^{\mathrm{T}}(I - H_{S\setminus j})X_{S_{0,n}}\beta^0_{S_{0,n}}}{n} \\
\leq& \frac{2}{c_{\min}} \frac{c_\beta \sum_{k\in S_{0,n}} |X_j^{\mathrm{T}}(I - H_{S\setminus j})X_k|}{n} \\
\leq& \frac{2}{c_{\min}} c_\beta \big(c_{\max} + \frac{1 + c_{\min}}{2}\big) n^{c_1}.
\end{aligned}$$

Let $C_\beta = \frac{2c_\beta}{c_{\min}}\left(c_{\max} + \frac{1+c_{\min}}{2}\right)$, we complete the proof. □

Additional Simulation Results

Table A.1: Comparisons of SPARES and one-time SPARE based on 200 replications. Bias (SE) is displayed in each cell. LSE refers to least square estimation as if $S_{0,n}$ were known.

Index	β_j^0	SPARES	One-time SPARE	LSE
199	1.00	0.03(0.16)	-0.02(0.26)	0.03(0.16)
243	-1.00	-0.02(0.16)	0.03(0.26)	-0.02(0.16)
256	1.00	-0.002(0.16)	-0.007(0.26)	-0.002(0.16)
0's	0.00	0.000(0.16)	-0.001(0.26)	

Figure A.1: Performance of SPARES under simulation Example 2.1. X-axis is the variable index. **Topleft:** Average estimates and average CIs V.S. true signals. **Topright:** Bias of SPARES estimates for each j, red dots are non-zero signals, dashed lines indicate blocks of the predictors. **Bottomleft:** Coverage probability of β^0 for each j w.r.t. 0.95 norminal level. **Bottomright:** Empirical probability of not rejecting $H_0 : \beta_j^0 = 0$.

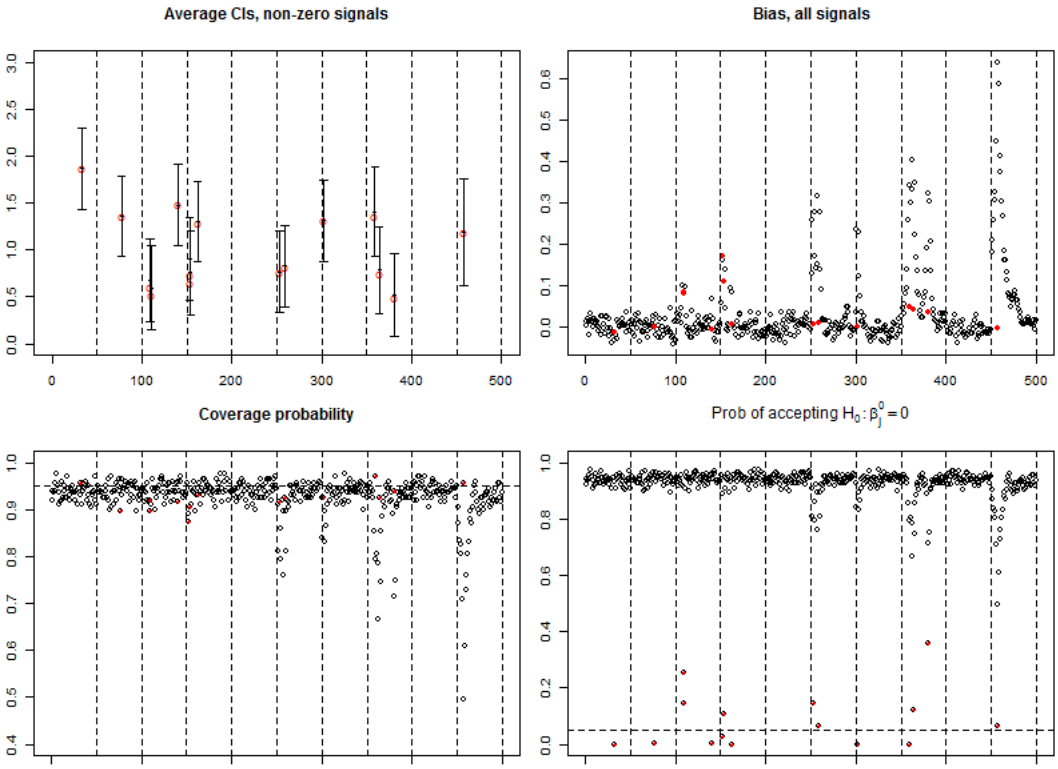


Figure A.2: Performance of SPARES under simulation Example 2.2.

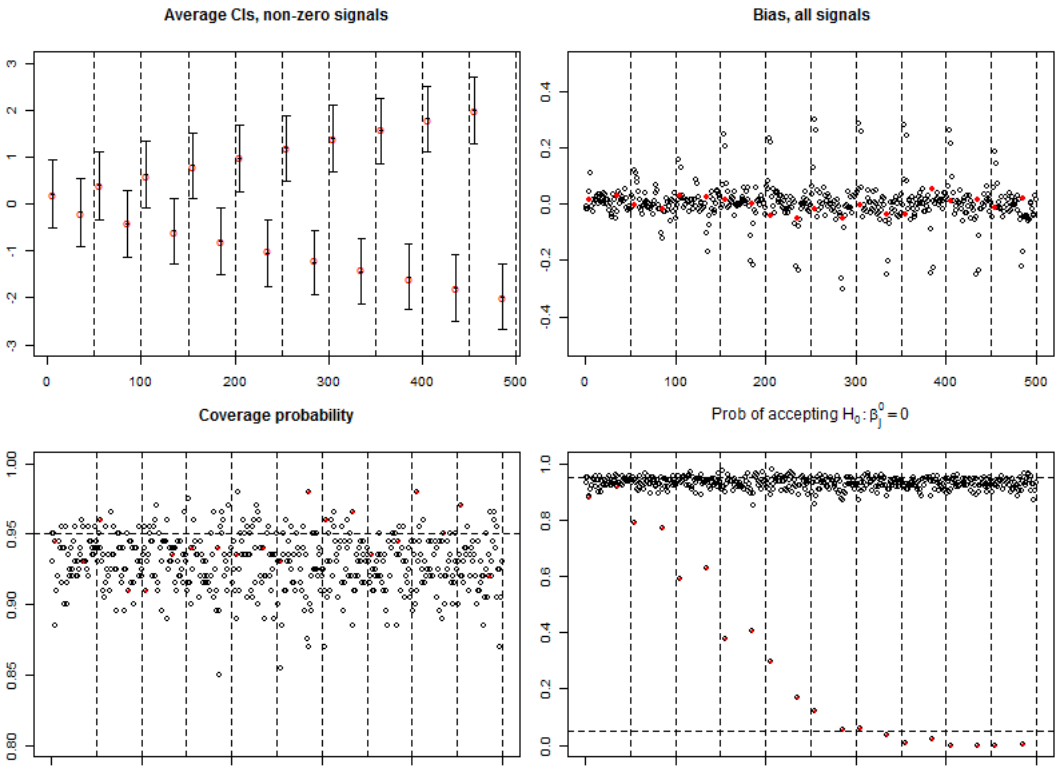


Figure A.3: Comparisons of SPARES with LASSO-Pro and SSLASSO under simulation Example 4. Left panels: Mean estimates from each method and the true signals. Right panels: Coverage probabilities for each $j \in S_{0,n}$ and 20 representatives of $j \notin S_{0,n}$.

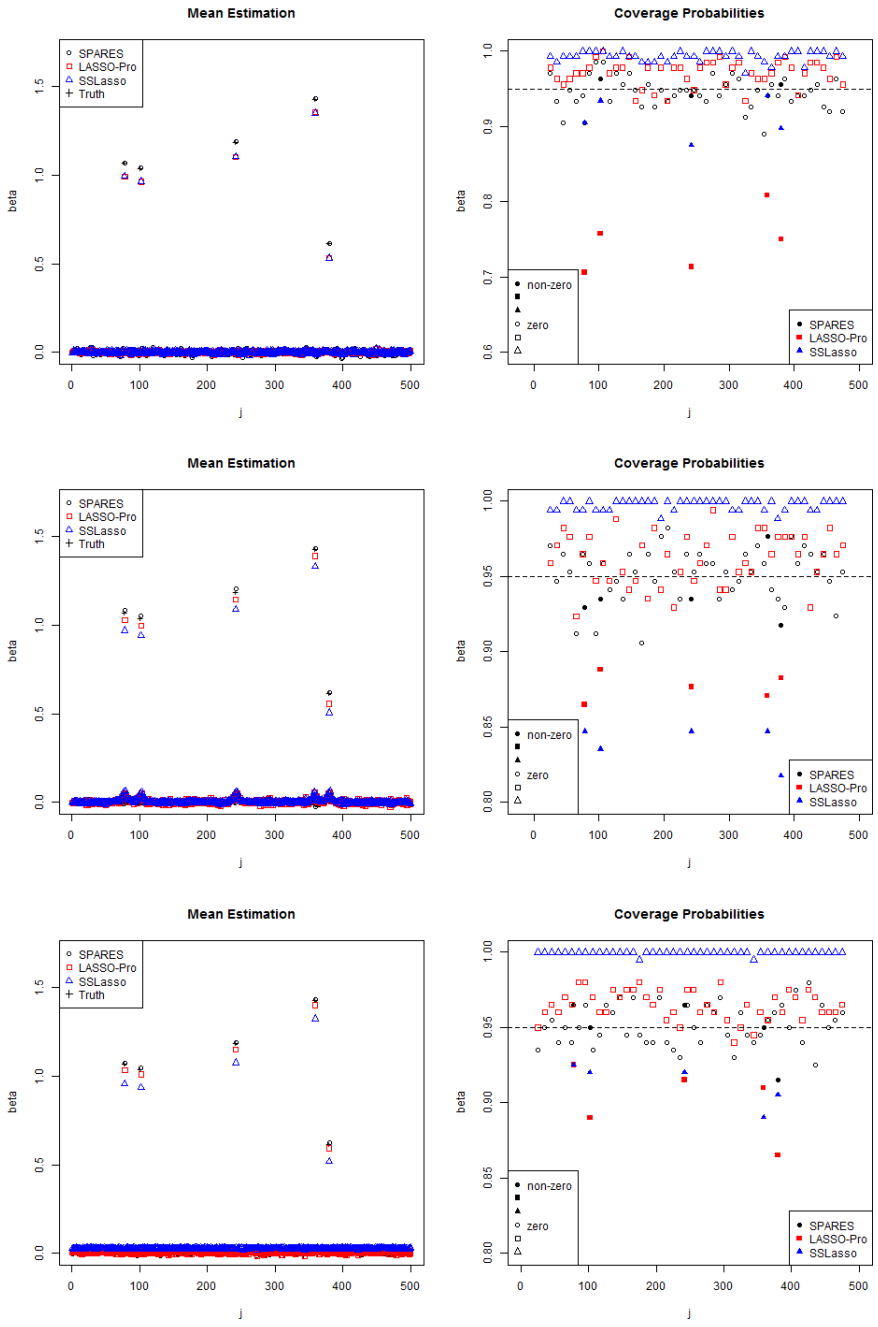


Figure A.4: Correlation among predictors: left panel - riboflavin data; right panel - multiple myeloma data.

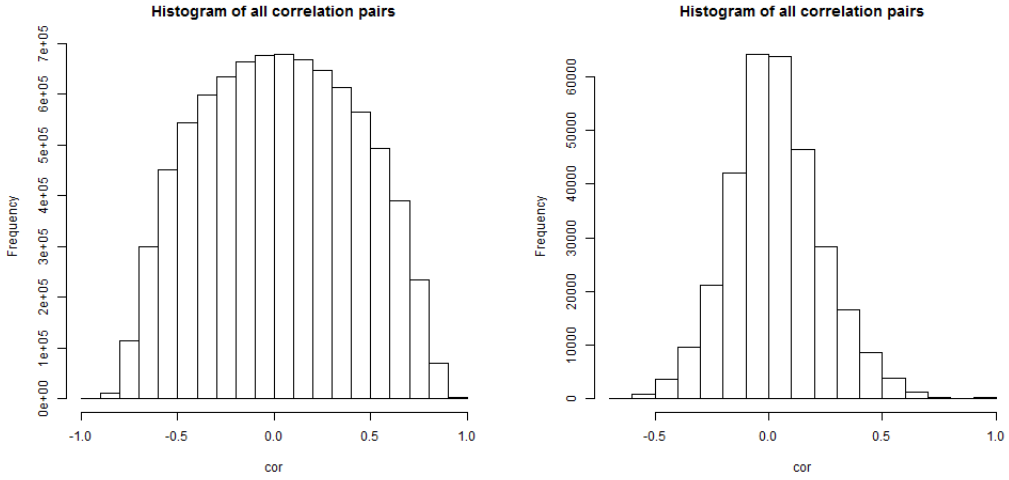


Figure A.5: Results of the riboflavin genomic data analysis. Left panel: selection frequency of each gene; Right panel: confidence intervals of the top five most significant genes.

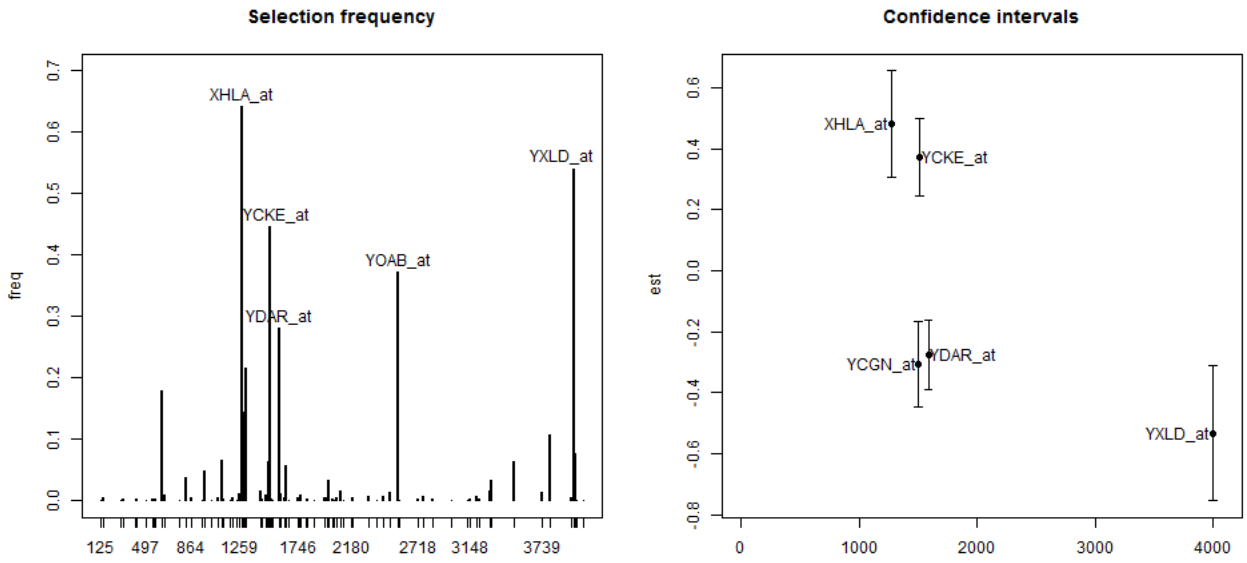
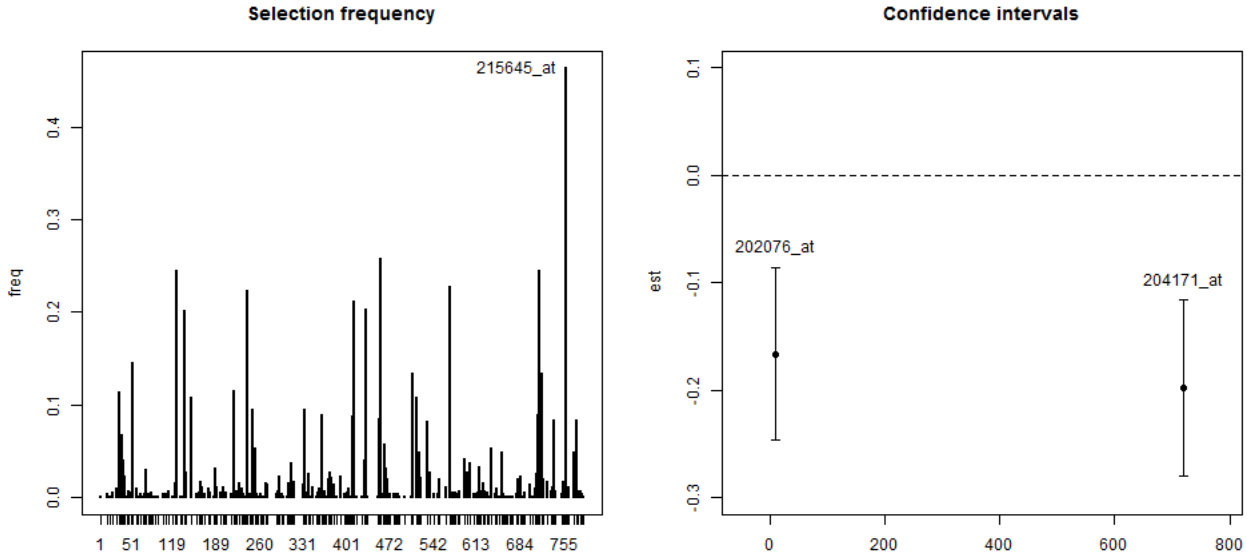


Figure A.6: Results of the Multiple Myeloma genomic data analysis. Left panel: selection frequency of each gene; Right panel: confidence intervals of the top two most significant genes.



APPENDIX B

Chapter III Supplementary Material

Proofs

Proof of Theorem III.2. From the data split, D_1 and D_2 are mutually exclusive, thus S, from D_2, is independent of $D_1 = (Y^1, X^1)$. Given X^1 and for $S \supset S_0$ under assumption (A3), the MLE $\widetilde{\beta}^1$ in (3.4) follows the classic low dimensional convergence. So does its scalar component $\widetilde{\beta}_j$. Thus the key is to show the asymptotic normality is uniform with respect to S, X^1 and j. We reiterate that

$$\widetilde{\beta}^1 = \operatorname{argmin} \ell_{S_{+j}}(\beta_{S_{+j}}) = \operatorname{argmin} \ell(\beta_{S_{+j}}; Y^1, X^1_{S_{+j}});$$
$$\widetilde{\beta}_j = \left(\widetilde{\beta}^1\right)_j; \quad \widetilde{\beta} = (\widetilde{\beta}_0, \widetilde{\beta}_1, .., \widetilde{\beta}_p).$$

When $S \supset S_0$, $\widetilde{\beta}^1$ satisfies $\sqrt{n_1}\widehat{I}^{1/2}_{S_{+j}}\left(\widetilde{\beta}^1 - \beta^*_{S_{+j}}\right) \mid S, X^1 \to N(0, \mathbf{I})$, where $\widehat{I}_{S_{+j}} = \overline{X}_{S_{+j}}{}^T V_{S_{+j}} \overline{X}_{S_{+j}}/n_1$, with $V_S = \operatorname{diag}\{A''(\overline{\mathbf{x}}_{iS}\beta_S), .., A''(\overline{\mathbf{x}}_{iS}\beta_S)\}$. Thus its component $\widetilde{\beta}_j$ follows $\sqrt{n_1}\left(\widetilde{\beta}_j - \beta^*_j\right)/\widetilde{\sigma}_j \mid S, X^1 \to N(0,1)$, where $\widetilde{\sigma}^2_j = \left\{\widehat{I}^{-1}_{S_{+j}}\right\}_{jj}$.

To derive the uniform convergence in j and S, we refer to Theorem 5 of *Niemiro* (1992), which gives a precise approximation of the convergence of M-estimators.

Treating $\widetilde{\beta}^1$ as the M-estimator, with our notations, there exists $0 \leq t < 1$,

$$\sqrt{n_1}\left(\widetilde{\beta}^1 - \beta^*_{S_{+j}}\right) = -\left\{I^*_{S_{+j}}\right\}^{-1}\sqrt{n_1}U_{S_{+j}}(\beta^*_{S_{+j}}) + O\left(n_1^{-(1+t)/4}(\log n_1)^{1/2}(\log\log n_1)^{(1+t)/4}\right). \tag{B.1}$$

Furthermore, in the GLM case, t can approach 1, meaning the remainder term can be of order close to $O(n_1^{-1/2})$. More importantly, the order of this remainder term only depends on the sample size n_1, but not on j or S.

Now we write

$$\sqrt{n_1}\left(\widetilde{\beta}_j - \beta^*_j\right)/\widetilde{\sigma}_j = \phi_{n_1} + \xi_{n_1},$$

where ϕ_{n_1} corresponds to the first term on the right hand side in (B.1) and ξ_{n_1} is the remainder term. When $S \supset S_0$, by assumption (A1) and Lemma (B.1), $\widetilde{\sigma}_j$ is bounded away from zero and infinity with probability going to 1 uniformly in j and S.

By the Berry-Esseen Theorem, under assumptions (A1) and (A2), $|F_{n_1}(x) - \Phi(x)| \leq \frac{C}{\sqrt{n_1}}$, where $F_{n_1}(x)$ is the CDF of ϕ_{n_1}, $\Phi(x)$ is the CDF of the stand normal, and C only depends on $c_{\min}, c_{\max}, \rho_0$, and ρ_1. Together with $\xi_{n_1} = O(n_1^{-t'})$ for some $t' < 1/2,\ t' \to 1/2$, we have

$$\sqrt{n_1}\left(\widetilde{\beta}_j - \beta^*_j\right)/\widetilde{\sigma}_j \to N(0,1).$$

□

Proof of Theorem III.3. We define the *oracle* estimators of β^*_j on the full data (Y, X) and the b-th subsample D_1^b respectively, where the candidate set is the true set S_0:

$$\beta^o_{S_{0+j}} = \operatorname{argmin} \ell_{S_{0+j}}(\beta_{S_{0+j}}) = \operatorname{argmin} \ell_{S_{0+j}}(\beta_{S_{0+j}}; Y, X_{S_{0+j}}),\ \beta^o_j = \{\beta^o_{S_{0+j}}\}_j;$$
$$\beta^b_{S_{0+j}} = \operatorname{argmin} \ell^b_{S_{0+j}}(\beta_{S_{0+j}}) = \operatorname{argmin} \ell_{S_{0+j}}(\beta_{S_{0+j}}; Y^{1(b)}, X^{1(b)}_{S_{0+j}}),\ \beta^b_j = \{\beta^b_{S_{0+j}}\}_j.$$

For each $j \in [p]$,

$$W_j^* = \sqrt{n}(\beta_j^o - \beta_j^*)/\sqrt{I_{j|S_0}^*} \xrightarrow{d} N(0,1) \quad \text{as} \quad n \to \infty, \tag{B.2}$$

where $I_{j|S_0}^*$ is defined as in (3.3). By Cauchy's interlacing theorem, $c_{\min} \le I_{j|S_0}^* \le c_{\max}$, and thus it is bounded away from zero and infinity.

With the oracle estimates β_j^b's, we have the following decomposition:

$$\begin{aligned} \sqrt{n}(\widehat{\beta}_j - \beta_j^*) &= \frac{1}{B}\sum_{b=1}^{B} \sqrt{n}(\widetilde{\beta}_j^b - \beta_j^*) \\ &= \frac{1}{B}\sum_{b=1}^{B} \sqrt{n}(\beta_j^b - \beta_j^*) + \frac{1}{B}\sum_{b:S^b \neq S_0} \sqrt{n}(\widetilde{\beta}_j^b - \beta_j^b) \doteq Z_j^* + \Delta_j. \end{aligned} \tag{B.3}$$

First we show $\Delta_j = o_p(1)$ by writing $\Delta_j = \frac{1}{B}\sum_{b=1}^{B} \delta_b$; $\quad \delta_b \doteq \mathbf{1}(S^b \neq S_0)\sqrt{n}\Big(\widetilde{\beta}_j^b - \beta_j^b\Big)$. By Corollary (B.2), $\mathbf{E}\,\delta_b = \mathbf{P}(S^b \neq S_0)\mathbf{E}\,\sqrt{n}\Big(\widetilde{\beta}_j^b - \beta_j^b\Big) = o\Big(n^{-c_2-1}2C_\beta n^{c_1+\frac{1}{2}}\Big) = o\Big(n^{-c_2+c_1-\frac{1}{2}}\Big) \to 0$ as $n \to \infty$. Similarly, $\mathbf{V}\delta_b = \mathbf{P}(S^b \neq S_0)\mathbf{E}\,n\Big(\widetilde{\beta}_j^b - \beta_j^b\Big)^2 = o\Big(n^{-c_2-1}4C_\beta^2 n^{2c_1+1}\Big) = o(n^{-c_2+2c_1}) \to 0$ as $n \to \infty$. Thus $\delta_b = o_p(1)$ for all $b \in [B]$. Furthermore, since $\mathbf{E}\,\Delta_j = \mathbf{E}\,\delta_b$ and $\mathbf{V}\Delta_j \le \mathbf{V}\delta_b$, we have $\Delta_j = o_p(1)$.

Next, we show the convergence of Z_j^*. Notice that

$$Z_j^*/\sqrt{I_{j|S_0}^*} = W_j^* + \sqrt{n}\Big(\frac{1}{B}\sum_{b=1}^{B} \beta_j^b - \beta_j^o\Big)/\sqrt{I_{j|S_0}^*} \doteq W_j^* + T_n^B/\sqrt{I_{j|S_0}^*}. \tag{B.4}$$

By (B.2) and that $I_{j|S_0}^*$ is bounded, we are only left to show $T_n^B = o_p(1)$. Define $t_{n,b} = \sqrt{n}\left(\beta_j^b - \beta_j^o\right)$, then $T_n^B = \Big(\sum_{b=1}^{B} t_{n,b}\Big)/B$. Since β_j^b is estimated on the random subsample D_1^b, its conditional mean is β_j^o and conditional variance is $\widehat{I}_{j|S_0}^{-1}/n_1$. Thus, conditional on the data $(Y,X) = (X^{(n)}, Y^{(n)})$, $\{t_{n,b}\}_{b=1,2,..,B}$ are i.i.d. with

$$\mathbf{E}\left(t_{n,b}|(X^{(n)}, Y^{(n)})\right) = 0, \quad \mathbf{V}\big(t_{n,b}|(X^{(n)}, Y^{(n)})\big) = \frac{n}{n_1}\widehat{I}_{j|S_0}^{-1} = \widehat{I}_{j|S_0}^{-1}/q.$$

By Lemma (B.1), with probability at least $1 - 2\exp(-\varepsilon^2\eta n/C_K)$ for some constant C_K and $\varepsilon = \min(\frac{1}{2}, \frac{c_{\min}}{2})$, $c_{\min}/2 \leq \widehat{I}_{j|S_0} \leq c_{\max} + (1 + c_{\min})/2$.

Therefore, $\widehat{I}^{-1}_{j|S_0} \leq 2/c_{\min}$ with probability going to 1 exponentially fast in n. Now define

$$\Omega_n = \{(X^{(n)}, Y^{(n)}) = (\mathbf{x}_i, y_i)_{i=1,2,..,n} : \widehat{I}^{-1}_{j|S_0} \leq 2/c_{\min}, \forall j \in [p]\}.$$

Since $p = O(n^{\gamma_1})$ for some $\gamma_1 > 1$, $\mathbf{P}\{(X^{(n)}, Y^{(n)}) \in \Omega_n\} \to 1$ as $n \to \infty$. Thus $\forall (X^{(n)}, Y^{(n)}) \in \Omega_n$, $\mathbf{V}\big\{t_{n,b}|(X^{(n)}, Y^{(n)})\big\} \lesssim 2/c_{\min}$. Furthermore,

$$\mathbf{V}\big\{T_n^B|(X^{(n)}, Y^{(n)})\big\} = \frac{1}{B^2}\sum_{b=1}^{B}\mathbf{V}\big\{t_{n,b}|(X^{(n)}, Y^{(n)})\big\} \leq \frac{2}{Bqc_{\min}}.$$

Thus, $\forall \delta, \zeta > 0$, $\exists N_0, B_0 > 0$ such that $\forall n > N_0, B > B_0$,

$$\begin{aligned}
&\mathbf{P}(|T_n^B| \geq \delta)\\
\leq &\int_{\Omega_n} \mathbf{P}\left\{|T_n^B| \geq \delta\big|(X^{(n)}, Y^{(n)})\right\} \mathrm{d}\mathbf{P}(X^{(n)}, Y^{(n)}) + \mathbf{P}\left\{(X^{(n)}, Y^{(n)}) \notin \Omega_n\right\}\\
\leq &\int_{\Omega_n} \frac{\mathbf{V}\left\{T_n^B|(X^{(n)}, Y^{(n)})\right\}}{\delta^2}\mathrm{d}\mathbf{P}(X^{(n)}, Y^{(n)}) + \mathbf{P}\left\{(X^{(n)}, Y^{(n)}) \notin \Omega_n\right\}\\
\leq &\frac{2}{B_0\delta^2 qc_{\min}}\int_{\Omega_n} \mathrm{d}\mathbf{P}(X^{(n)}, Y^{(n)}) + \mathbf{P}\left\{(X^{(n)}, Y^{(n)}) \notin \Omega_n\right\}\\
\leq &\zeta/2 + \zeta/2\\
\leq &\zeta.
\end{aligned}$$

Finally, combining this with (B.2), we have

$$Z_j^*/\sqrt{I^*_{j|S_0}} = W_j^* + T_n^B/\sqrt{I^*_{j|S_0}} \xrightarrow{d} N(0,1) \quad \text{as} \quad B, n \to \infty.$$

□

Proof of Theorem III.4. Follow the previous proof, we replace the arguments in j

with those in $S^{(1)}$. The *oracle* estimators are

$$\beta^o_{S^{(1)}\cup S_0} = \text{argmin}\, \ell_{S^{(1)}\cup S_0}(\beta_{S^{(1)}\cup S_0}; Y, X_{S^{(1)}\cup S_0}),\ \beta^o_{S^{(1)}} = \{\beta^o_{S^{(1)}\cup S_0}\}_{S^{(1)}};$$
$$\beta^b_{S^{(1)}\cup S_0} = \text{argmin}\, \ell_{S^{(1)}\cup S_0}(\beta_{S^{(1)}\cup S_0}; Y^{1(b)}, X^{1(b)}_{S^{(1)}\cup S_0}),\ \beta^b_{S^{(1)}} = \{\beta^b_{S^{(1)}\cup S_0}\}_{S^{(1)}}.$$

Notice that $|S^{(1)}| = p_1 = O(1)$, as $n \to \infty$, $|S_0 \cup S^{(1)}| = O\big(|S_0|\big) = o(n)$, so that the above quantities are well-defined. The oracle estimator satisfies

$$W^{(1)} = \sqrt{n}\left\{I^*_{S^{(1)}|S_0}\right\}^{-1/2}\left(\beta^o_{S^{(1)}} - \beta^*_{S^{(1)}}\right) \xrightarrow{d} N(0, \mathrm{I}_{p_1}) \quad \text{as} \quad n \to \infty,$$

Similar to (B.3,B.4), we decompose $\sqrt{n}(\widehat{\beta}_{S^{(1)}} - \beta^*_{S^{(1)}})$ into three parts:

$$\sqrt{n}(\widehat{\beta}_{S^{(1)}} - \beta^*_{S^{(1)}}) \doteq W^{(1)} + \Delta^{(1)}_0 + \Delta^{(1)}_1.$$

For the interest of space, we do not spell out these quantities and the derivations, but it is straightforward to show that $|\Delta^{(1)}_0|_1 = |\Delta^{(1)}_1|_1 = o_p(1)$, which completes the pro □

Lemmas

Lemma B.1. *Given model (3.1,3.2), and the corresponding information matrix at the truth β^*, $I^* = \mathbf{E}_{\beta^*}(\nabla^2\ell(\beta^*))$. Assume $0 < c_{\min} \le \lambda_{\min}(I^*) \le \lambda_{\max}(I^*) \le c_{\max} < \infty$. For any subset $S \subset \{1, 2, .., p\}$ with $|S| \le \eta n$, $0 < \eta < 1$, and $\forall j \in S$, define the partial information as $I_{j|S} = I_{jj} - I_{jS_j} I^{-1}_{S_jS_j} I_{S_jj}$, where $S_j = S \setminus j$. Denote its empirical estimator as $\widehat{I}_{j|S}$. Then with probability at least $1 - 2\exp(-\frac{\varepsilon^2\eta}{C_K}n)$,*

$$\frac{c_{\min}}{2} \le \widehat{I}_{j|S} \le c_{\max} + \frac{1 + c_{\min}}{2}$$

where $\varepsilon = \min(\frac{1}{2}, \frac{c_{\min}}{2})$.

Corollary B.2. *Given model (3.1,3.2) and assumptions (A1,A2), consider the one-time estimator as defined in (3.4,3.5). If* $|S| \leq \eta n$, $0 < \eta < 1$, *then with probability at least* $1 - 2\exp(-\frac{\varepsilon^2 \eta}{C_K} n)$,

$$\widetilde{\beta}_j \leq C_\beta n^{c_1},$$

where C_β *depends on* $c_{\min}, c_{\max}, c_\beta$.

APPENDIX C

Chapter IV Supplementary Material

Lemmas and Proofs

We first introduce two useful lemmas, before proving the main theorem.

Lemma C.1. *Assuming the quantile regression model (4.1) with true parameter vector* $\beta^*(\tau)$, $\tau \in (0, \tau_U]$, *and observed data* $D = \{(X_i, \Delta_i, \mathbf{Z}_i), i = 1, 2, .., n\}$, *where* $X = \min\{T, C\}, \Delta = I(T \leq C)$ *and* C *is conditionally independent given* $\mathbf{Z}$. *For a fixed subset* S *satisfying* $S^* \subset S \subset [p]$, *and* $|S| \leq K_1 n^{c_1}$ *for some* $0 \leq c_1 < 1/2$ *and* $K_1 \leq 1$, *denote the partial data* $D^{(S)} = \{(X_i, \Delta_i, \mathbf{Z}_{S,i}), i = 1, 2, .., n\}$ *that includes only covariates* $j \in S$. *Let* $\acute{\beta}_S(\tau), \tau \in [\tau_L, \tau_U]$ *be the estimator from Peng and Huang (2008) of fitting the censored quantile regression on* $D^{(S)}$ *with the* τ *grid* Γ_m. *Under assumptions (A1)-(A4),(A6),(A7),(A8), then for any* $j \in S$,

$$\sqrt{n}\left(\acute{\beta}_j(\tau) - \beta_j^*(\tau)\right), \tau \in [\tau_L, \tau_U]$$

converges weakly to a mean zero Gaussian process.

Proof. The estimator $\acute{\beta}_j(\tau)$ is the entry in $\acute{\beta}_S(\tau)$ for variable Z_j, which is fitted from

$$Q_Y(\tau|Z_S) = \mathbf{Z}_S^{\mathrm{T}}\beta_S(\tau).$$

Denote $\theta_{iS}(\tau) = \mathbf{Z}_{iS}^{\mathrm{T}}\beta_S(\tau)$ and $\theta_{iS}^*(\tau) = \mathbf{Z}_{iS}^{\mathrm{T}}\beta_S^*(\tau)$ for subject i in $D^{(S)}$ and the fixed set S satisfying conditions in the lemma, then $\acute{\beta}_S(\tau)$ is the solution to the following estimating equation as in *Peng and Huang* (2008),

$$n^{1/2}\mathbf{U}_n(\beta_S, \tau) = 0,$$

where

$$\mathbf{U}_n(\beta_S, \tau) = n^{-1}\sum_{i=1}^{n}\mathbf{Z}_i\left(N_i\left(\theta_{iS}(\tau)\right) - \int_0^\tau I[\log X_i \geq \theta_{iS}(u)]dH(u)\right).$$

Let $\mathbf{u}(\beta_S, \tau) = \mathbf{E}\left\{\mathbf{U}_n(\beta_S, \tau)\right\}$, the martingale property gives $\mathbf{u}(\beta_S^*, \tau) = 0,\ \tau \in \Gamma_m$. We further define $\boldsymbol{\mu}_S(\mathbf{b}) = \mathbf{E}\left[\mathbf{Z}_S N(\mathbf{Z}_S^{\mathrm{T}}\mathbf{b})\right]$, $\mathbf{B}_S(\mathbf{b}) = \mathbf{E}\left[\mathbf{Z}_S^{\otimes 2} \times g\{\mathbf{Z}_S^{\mathrm{T}}\mathbf{b}|\mathbf{Z}_S\}(\mathbf{Z}_S^{\mathrm{T}}\mathbf{b})\right]$, and $\mathbf{J}_S(\mathbf{b}) = \mathbf{E}\left[\mathbf{Z}_S^{\otimes 2} \times f\{\mathbf{Z}_S^{\mathrm{T}}\mathbf{b}|\mathbf{Z}_S\}(\mathbf{Z}_S^{\mathrm{T}}\mathbf{b})\right]$ for vector $\mathbf{b}$ of length $|S|$. By Theorem 2 of *Peng and Huang* (2008), $-n^{-1/2}\mathbf{U}_n(\beta_S^*, \tau)$ converges weakly to a tight Gaussian process, $\mathbf{G}_S(\tau)$, with mean 0 and covariance $\boldsymbol{\Sigma}_S(s,t)$ for $\tau \in [\tau_L, \tau_U]$, where $\boldsymbol{\Sigma}_S(s,t) = \mathbf{E}\left\{\imath_{iS}(s)\imath_{iS}(t)^{\mathrm{T}}\right\}$ with

$$\imath_{iS}(\tau) = \mathbf{Z}_{iS}\left(N_i\left(\theta_{iS}^*(\tau)\right) - \int_0^\tau I[\log X_i \geq \theta_{iS}^*(\tau)]du\right).$$

Because given $S \supset S^*$, $\theta_{iS}^*(\tau) = \theta_{iS^*}^*(\tau)$, $\boldsymbol{\Sigma}_S(s,t)$ depends on the set S only through $\mathbf{Z}_S$ as in the expression of $\imath_{iS}(\tau)$. The restricted eigenvalue condition implies that $\left[\mathbf{B}_S\{\beta_S^*(\tau)\}\right]^{-1}$ is bounded uniformly for $\tau \in [\tau_L, \tau_U]$. By the Taylor expansion technique and the continuous mapping theorem, for $\tau \in [\tau_L, \tau_U]$, $\sqrt{n}\left(\acute{\beta}_S(\tau) - \beta_S^*(\tau)\right)$ converges weakly to $\mathbf{B}_S\{\beta_S^*(\tau)\}^{-1}\boldsymbol{\phi}\{\mathbf{G}_S(\tau)\}$, which is also Gaussian, where $\boldsymbol{\phi}$ is de-

fined in *Peng and Huang* (2008) and we reiterate here that $\boldsymbol{\phi}$ is a map from $\mathcal{F}$ to $\mathcal{F}$ such that for $\mathbf{g} \in \mathcal{F}$,

$$\boldsymbol{\phi}(\mathbf{g})(\tau) = \int_0^\tau \mathcal{I}(s,\tau) d\mathbf{g}(s),$$

with $\mathcal{I}(s,t) = \Pi_{u\in[s,t]}[\mathbf{I}_p + \mathbf{J}\{\beta^*(u)\}\mathbf{B}\{\beta^*(u)\}^{-1}dH(u)]$ and

$$\mathcal{F} = \{\mathbf{g} : [0,\tau_U] \to \mathbf{R}^p, \mathbf{g} \text{ is left-continuous with right limit, } \mathbf{g}(0) = \mathbf{0}\}.$$

Consequently, $\acute{\beta}_j(\tau)$, as the component of $\acute{\beta}_S(\tau)$, satisfies that $\sqrt{n}\left(\acute{\beta}_j(\tau) - \beta_j^*(\tau)\right)$ converges weakly to $\mathbf{e}_j^{\mathrm{T}}\mathbf{B}_S\{\beta_S^*(\tau)\}^{-1}\boldsymbol{\phi}\{\mathbf{G}_S(\tau)\}$, with $\mathbf{e}_j = (0,..,1,..,0)^{\mathrm{T}}$ the basis vector for variable $\mathbf{Z}_j$, which is still mean 0 Gaussian.

□

Lemma C.2. ***Bound of coefficient estimates*** *Given data* $D = \{(X_i, \Delta_i, \mathbf{Z}_i), i = 1,2,..,n\}$, *and a subset* $S \subset [p]$ *with* $|S| \leq K_1 n^{c_1}$ *for some* $0 \leq c_1 < 1/2$ *and* $K_1 \leq 1$, *denote the partial data* $D^{(S)} = \{(X_i, \Delta_i, \mathbf{Z}_{S,i}), i = 1,2,..,n\}$ *that includes only covariates* $j \in S$ *as predictors. Let* $\acute{\beta}_S(\tau), \tau \in [\tau_L, \tau_U]$ *be the estimator from Peng and Huang (2008) of fitting model (4.1) on* $D^{(S)}$. *Under assumptions (A1)-(A3),(A6),(A7),(A8), there exists constant* $M_0 > 0$, *such that* $\sup_{j\in S,\tau\in[\tau_L,\tau_U]} |\acute{\beta}_j(\tau)| < M_0$.

Proof. From *Peng and Huang* (2008), $\acute{\beta}_S(\tau)$ is sequentially estimated for $\tau_k \in \Gamma_m$, $k = 0,1,..,m$ by solving the following minimization problem of an L_1-type convex objective function for $\mathbf{h}$ at k,

$$\begin{aligned} L_k(\mathbf{h}) = &\sum_{i=1}^n \Delta_i \left|\log X_i - \mathbf{h}^{\mathrm{T}}\mathbf{Z}_i\right| + \left|M^* - \mathbf{h}^{\mathrm{T}} \sum_{i=1}^n (-\Delta_i \mathbf{Z}_i)\right| \\ &+ \left|M^* - \mathbf{h}^{\mathrm{T}} \sum_{i=1}^n 2\mathbf{Z}_i \left(\sum_{r=0}^{k-1} \int_{\tau_r}^{\tau_{r+1}} I[\log X_i \geq \mathbf{Z}_i^{\mathrm{T}} \acute{\beta}_j(\tau_r)] dH(u) + \tau_0\right)\right|, \end{aligned}$$

where M^* is a large positive number. Since $\acute{\beta}_S(\tau)$ is defined as a right-continuous

function on the grid Γ_m, to show the boundedness of $\acute{\beta}_j(\tau)$'s, we only need to argue at the grid points τ_k's. We first show $L_k(\mathbf{h})$ is a coercive function in $\mathbf{h}$, that is $L_k(\mathbf{h}) \to \infty$ whenever $\|\mathbf{h}\| \to \infty$. Since $L_k(\mathbf{h}) \geq \sum_{i=1}^n \Delta_i \left|\log X_i - \mathbf{h}^{\mathrm{T}}\mathbf{Z}_i\right|$, which does not depend on τ or k, it is sufficient to show $L(\mathbf{h}) = \sum_{i=1}^n \Delta_i \left|\log X_i - \mathbf{h}^{\mathrm{T}}\mathbf{Z}_i\right|$ is coercive. By Proposition 12.3.1 in *Lange* (2004), a sufficient and necessary condition is that $L(\mathbf{h})$ is coercive along all nontrivial rays $\{\mathbf{h} : \mathbf{h} = t\mathbf{v}, t \geq 0\}$. The condition is met because $\forall \mathbf{v} \in \mathbf{R}^{|S|}$, $L(t\mathbf{v}) = \sum_{i=1}^n \Delta_i \left|\log X_i - t\mathbf{v}^{\mathrm{T}}\mathbf{Z}_i\right|$ is an absolute value function in t, and thus goes to infinity as $t \to \infty$. Now let $L_0 = L_k(\mathbf{0})$, which does not depend on k and is bounded, then the set $\{\mathbf{h} : L_k(\mathbf{h}) \leq L_0\}$ is compact and contains the minimizer $\acute{\beta}_S(\tau_k)$. Thus there exists a uniform bound $M_0 > 0$ depending on L_0, such that $\sup_{j\in S, \tau\in[\tau_L,\tau_U]} |\acute{\beta}_j(\tau)| < M_0$. □

Now we are equipped to prove Theorem IV.3.

Proof of Theorem IV.3. We first introduce the oracle estimators of $\beta_j^*(\tau)$'s assuming the true active set S^* is known. Let $\check{\beta}_{S^*_{+j}}(\tau)$ be the oracle estimator by fitting the following CQR on the full data,

$$Q_Y(\tau|\mathbf{Z}_{S^*_{+j}}) = \mathbf{Z}^{\mathrm{T}}_{S^*_{+j}}\beta_{S^*_{+j}}(\tau),$$

where $S^*_{+j} = \{j\} \cup S^*$. Then the oracle estimator $\check{\beta}_j(\tau) = \left(\check{\beta}_{S^*_{+j}}(\tau)\right)_j$ is the entry that is the coefficient for variable Z_j. Analogically, let $\check{\beta}_j^b(\tau)$ denote the oracle estimator fitted on the b-th sub-sample D_1^b in the Fused-HDCQR procedure.

The objective can be decomposed as below,

$$
\begin{aligned}
&\sqrt{n}\left(\widehat{\beta}_j(\tau)-\beta_j^*(\tau)\right)\\
=&\sqrt{n}\left(\check{\beta}_j(\tau)-\beta_j^*(\tau)\right)+\sqrt{n}\left(\widehat{\beta}_j(\tau)-\check{\beta}_j(\tau)\right)\\
=&\sqrt{n}\left(\check{\beta}_j(\tau)-\beta_j^*(\tau)\right)+\sqrt{n}\left(\frac{1}{B}\sum_{b=1}^{B}\widetilde{\beta}_j^b(\tau)-\check{\beta}_j(\tau)\right)\\
=&\underbrace{\sqrt{n}\left(\check{\beta}_j(\tau)-\beta_j^*(\tau)\right)}_{\text{I}}+\underbrace{\sqrt{n}\left(\frac{1}{B}\sum_{b=1}^{B}\check{\beta}_j^b(\tau)-\check{\beta}_j(\tau)\right)}_{\text{II}}+\underbrace{\sqrt{n}\left(\frac{1}{B}\sum_{b=1}^{B}\left\{\widetilde{\beta}_j^b(\tau)-\check{\beta}_j^b(\tau)\right\}\right)}_{\text{III}}.
\end{aligned}
$$

We will argue the asymptotic behavior of the three terms separately, as the first two terms do not involve the selections $\widehat{S}^b$'s, instead deal with the oracle estimators and the true active set S^*.

- $\text{I}=\sqrt{n}\left(\check{\beta}_j(\tau)-\beta_j^*(\tau)\right)$ converges weakly to a mean zero Gaussian process;
- $\text{II}=\sqrt{n}\left(\frac{1}{B}\sum_{b=1}^{B}\check{\beta}_j^b(\tau)-\check{\beta}_j(\tau)\right)=o_p(1)$, uniformly in $\tau\in[\tau_L,\tau_U]$;
- $\text{III}=\sqrt{n}\left(\frac{1}{B}\sum_{b=1}^{B}\left\{\widetilde{\beta}_j^b(\tau)-\check{\beta}_j^b(\tau)\right\}\right)=o_p(1)$, uniformly in $\tau\in[\tau_L,\tau_U]$.

By Slutsky's theorem for random processes (Theorem 18.10 in *Van der Vaart* (2000)), if the above statements all hold, we would conclude that $\sqrt{n}\left(\widehat{\beta}_j(\tau)-\beta_j^*(\tau)\right),\tau\in[\tau_L,\tau_U]$ converges weakly to a mean zero Gaussian process.

a) Let $S=S_{+j}^*$ for each $j\in[p]$, and by Lemma (C.1), $\text{I}=\sqrt{n}\left(\check{\beta}_j(\tau)-\beta_j^*(\tau)\right),\tau\in[\tau_L,\tau_U]$ converges weakly to a mean zero Gaussian process $\mathbf{e}_j^{\mathrm{T}}\mathbf{B}_S\{\beta_S^*(\tau)\}^{-1}\boldsymbol{\phi}\{\mathbf{G}_S(\tau)\}$, with $\mathbf{e}_j=(0,..,1,..,0)^{\mathrm{T}}\in\mathbf{R}^{|S|}$ the basis vector for variable $\mathbf{Z}_j$, and $\mathbf{B}_S(\cdot),\boldsymbol{\phi}(\cdot),\mathbf{G}_S(\cdot)$ defined in the proof of Lemma (C.1). Denote its covariance as $\sigma_j^*(s,t)$, which is uniformly bounded for $s,t\in[\tau_L,\tau_U]$.

b) To show $\text{II}=o_p(1)$, we first denote $\xi_{b,n}(\tau)=\sqrt{n}\left(\check{\beta}_j^b(\tau)-\check{\beta}_j(\tau)\right)$, then $\text{II}=\left(\sum_{b=1}^{B}\xi_{b,n}(\tau)\right)/B$. Since D_1^b's are random sub-samples, $\xi_{b,n}(\tau)$'s are i.i.d. conditional on data. By Appendix C of *Peng and Huang* (2008), the conditional distribution

of $\sqrt{n}\left(\check{\beta}_j^b(\tau) - \check{\beta}_j(\tau)\right)$ given the observed data is asymptotically the same as the unconditional distribution of $\mathrm{I} = \sqrt{n}\left(\check{\beta}_j(\tau) - \beta_j^*(\tau)\right)$, which is mean zero Gaussian from part a). Thus $\mathbf{E}\left(\xi_{b,n}(\tau)|D\right) \to \mathbf{E}\left(\mathrm{I}\right) \to 0$ and $\mathrm{Var}\left(\xi_{b,n}(\tau)|D\right) \to \sigma_j^*(\tau,\tau) \doteq \sigma_j^2(\tau)$, as $n \to \infty$. Denote $\sigma_j^2 = \sup_{\tau\in[\tau_L,\tau_U]} \sigma_j^2(\tau) < \infty$, then $\mathbf{E}\left(\mathrm{II}|D\right) \to 0$ uniform in $\tau \in [\tau_L, \tau_U]$, and for n large enough,

$$\mathrm{Var}\left(\mathrm{II}|D\right) = \frac{1}{B^2}\sum_{b=1}^{B} \mathrm{Var}\left(\xi_{b,n}|D\right) \leq \frac{2\sigma_j^2(\tau)}{B} \leq \frac{2\sigma_j^2}{B},\ \tau \in [\tau_L, \tau_U].$$

Now $\forall \delta, \zeta > 0$, $\exists N_0, B_0 > 0$ such that $\forall \tau \in [\tau_L, \tau_U], n > N_0, B > B_0$,

$$\begin{aligned}
&\mathrm{P}(|\mathrm{II}| \geq \delta) \\
\leq &\int_{D\in\Omega_n} \mathrm{P}\left(|\mathrm{II}| \geq \delta \,|\, D\right) \mathrm{dP}(D) \\
\leq &\int_{\Omega_n} \mathrm{P}\left(|\mathrm{II} - \mathbf{E}\left(\mathrm{II}\right)| \geq \delta/2 \,|\, D\right) \mathrm{dP}(D) \\
\leq &\int_{\Omega_n} \frac{\mathrm{Var}\ \left(\mathrm{II}\,|\,D\right)}{\delta^2/4} \mathrm{dP}(D) \\
\leq &\frac{2\sigma_j^2}{B_0\delta^2/4} \int_{\Omega_n} \mathrm{dP}(D) \\
\leq &\zeta.
\end{aligned}$$

Thus, $\mathrm{II} = o_p(1)$ uniformly in $\tau \in [\tau_L, \tau_U]$.

c) Each subsample D_2^b can be regarded as a random sample of $\lceil n/2 \rceil$ i.i.d. observations from the population distribution for which assumption (A5) hold, that is $|\widehat{S}^b| \leq K_1 n^{c_1}$ and $\mathrm{P}\left(\widehat{S}^b = S^*\right) \geq 1 - K_2(p \vee n)^{-1}$.

Notice that whenever $\widehat{S}^b = S^*$, the estimators based on the respective selections are equivalent, $\widetilde{\beta}_j^b(\tau) = \check{\beta}_j^b(\tau), \forall \tau$. Define $\eta_b(\tau) = I(\widehat{S}^b \neq S^*)\sqrt{n}\left\{\widetilde{\beta}_j^b(\tau) - \check{\beta}_j^b(\tau)\right\}$, while omitting subscripts j for simplicity, then $\mathrm{III} = \left(\sum_{b=1}^{B} \eta_b(\tau)\right)/B$.

By Lemma (C.2), there exists $M_0 > 0$ such that $\sup_{\tau\in[\tau_L,\tau_U]} \left|\widetilde{\beta}_j^b(\tau) - \check{\beta}_j^b(\tau)\right| \leq 2M_0$

for any $\widehat{S}^b$ with $|\widehat{S}^b| \leq K_1 n^{c_1}$. Therefore, by (A4) that $n/p = o(1)$,

$$\mathbf{E}\left(\eta_b(\tau)\right) \leq \mathrm{P}\left(\widehat{S}^b \neq S^*\right) \sqrt{n} \sup_{b\in[B],\tau\in[\tau_L,\tau_U]} \left|\widetilde{\beta}_j^b(\tau) - \check{\beta}_j^b(\tau)\right| \leq 2M_0\sqrt{n}K_2(p\vee n)^{-1} \to 0;$$
$$\mathrm{Var}\left(\eta_b(\tau)\right) \leq \mathrm{P}\left(\widehat{S}^b \neq S^*\right) n \sup_{b\in[B],\tau\in[\tau_L,\tau_U]} \left|\widetilde{\beta}_j^b(\tau) - \check{\beta}_j^b(\tau)\right|^2 \leq 4M_0^2 nK_2(p\vee n)^{-1} \to 0.$$

Although $\eta_b(\tau)$'s are dependent, we further have

$$\mathbf{E}\left(\mathrm{III}\right) = \mathbf{E}\left\{\left(\sum_{b=1}^{B}\eta_b(\tau)\right)/B\right\} \leq 2M_0\sqrt{n}K_2(p\vee n)^{-1} \to 0;$$
$$\mathrm{Var}\left(\mathrm{III}\right) = \frac{1}{B^2}\sum_{b=1}^{B}\sum_{b'=1}^{B}\mathrm{Cov}\left(\eta_b(\tau),\eta_{b'}(\tau)\right) \leq 4M_0^2 nK_2(p\vee n)^{-1} \to 0.$$

Thus $\mathrm{III} = o_p(1)$ uniformly in $\tau \in [\tau_L, \tau_U]$ by definition, as $\forall \delta, \zeta > 0$, $\exists N_0 > 0$ such that $\forall \tau \in [\tau_L, \tau_U], n > N_0$,

$$\begin{aligned}
&\mathrm{P}(|\mathrm{III}| \geq \delta)\\
\leq&\mathrm{P}\left(|\mathrm{III} - \mathbf{E}\left(\mathrm{III}\right)| \geq \delta/2\right)\\
\leq&\frac{\mathrm{Var}\ \left(\mathrm{III}\right)}{\delta^2/4}\\
\leq&\frac{16M_0^2K_2 n}{\delta^2 p}\\
\leq&\zeta.
\end{aligned}$$

□

BIBLIOGRAPHY

BIBLIOGRAPHY

Bach, F. R. (2008), Bolasso: model consistent lasso estimation through the bootstrap, in *Proceedings of the 25th international conference on Machine learning*, ACM.

Beer, D. G., S. L. Kardia, C.-C. Huang, T. J. Giordano, A. M. Levin, D. E. Misek, et al. (2002), Gene-expression profiles predict survival of patients with lung adenocarcinoma, *Nature Medicine*, *8*(8), 816–824.

Belloni, A., and V. Chernozhukov (2011), ℓ1-penalized quantile regression in high-dimensional sparse models, *The Annals of Statistics*, *39*(1), 82–130.

Belloni, A., V. Chernozhukov, and Y. Wei (2013), Honest confidence regions for a regression parameter in logistic regression with a large number of controls, *Tech. rep.*, cemmap working paper, Centre for Microdata Methods and Practice.

Belloni, A., V. Chernozhukov, and C. Hansen (2014), Inference on treatment effects after selection among high-dimensional controls, *The Review of Economic Studies*, *81*(2), 608–650.

Belloni, A., V. Chernozhukov, and Y. Wei (2016), Post-selection inference for generalized linear models with many controls, *Journal of Business & Economic Statistics*, *34*(4), 606–619.

Belloni, A., V. Chernozhukov, and K. Kato (2018), Valid post-selection inference in high-dimensional approximately sparse quantile regression models, *Journal of the American Statistical Association*, (just-accepted), 1–33.

Berk, R., L. Brown, A. Buja, K. Zhang, and L. Zhao (2013), Valid post-selection inference, *The Annals of Statistics*, *41*(2), 802–837.

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

Brose, M. S., P. Volpe, M. Feldman, M. Kumar, I. Rishi, R. Gerrero, et al. (2002), BRAF and RAS mutations in human lung cancer and melanoma, *Cancer research*, *62*(23), 6997–7000.

Bühlmann, P., and B. Yu (2002), Analyzing bagging, *The Annals of Statistics*, *30*(4), 927–961.

Bühlmann, P., M. Kalisch, and L. Meier (2014), High-dimensional statistics with a view toward applications in biology, *Annual Review of Statistics and Its Application*, *1*, 255–278.

Burrell, R. A., N. McGranahan, J. Bartek, and C. Swanton (2013), The causes and consequences of genetic heterogeneity in cancer evolution, *Nature*, *501*(7467), 338–345.

Cai, C., et al. (2015), Mir-195 inhibits tumor progression by targeting rps6kb1 in human prostate cancer, *Clinical Cancer Research*, *21*(21), 4922–4934.

Candès, E., and T. Tao (2007), The Dantzig selector: Statistical estimation when p is much larger than n, *The Annals of Statistics*, *35*(6), 2313–2351.

Carlson, M. (2015), *hgu133plus2.db: Affymetrix Human Genome U133 Plus 2.0 Array annotation data (chip hgu133plus2)*, r package version 3.2.2.

Christiani, D. C. (2017), The Boston lung cancer survival cohort, `http://grantome.com/grant/NIH/U01-CA209414-01A1`, [Online; accessed November 27, 2018].

Dezeure, R., P. Bühlmann, L. Meier, and N. Meinshausen (2015), High-dimensional inference: confidence intervals, p-values and r-software hdi, *Statistical Science*, *30*(4), 533–558.

Efron, B. (2014), Estimation and accuracy after model selection, *Journal of the American Statistical Association*, *109*(507), 991–1007.

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

Fan, J., and J. Lv (2008), Sure independence screening for ultrahigh dimensional feature space, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(5), 849–911.

Fan, J., and R. Song (2010), Sure independence screening in generalized linear models with np-dimensionality, *The Annals of Statistics*, *38*(6), 3567–3604.

Fan, J., R. Samworth, and Y. Wu (2009), Ultrahigh dimensional feature selection: beyond the linear model, *Journal of Machine Learning Research*, *10*, 2013–2038.

Fan, J., Y. Fan, and E. Barut (2014), Adaptive robust variable selection, *Annals of statistics*, *42*(1), 324.

Fei, Z., J. Zhu, M. Banerjee, and Y. Li (2018), Drawing inferences for high-dimensional linear models: A selection-assisted partial regression and smoothing approach, *Biometrics*, in press, https://doi.org/10.1111/biom.13013.

Fisher, R., L. Pusztai, and C. Swanton (2013), Cancer heterogeneity: implications for targeted therapeutics, *British Journal of Cancer*, *108*(3), 479–485.

Freedman, D. A., et al. (1981), Bootstrapping regression models, *The Annals of Statistics*, *9*(6), 1218–1228.

Friedman, J., T. Hastie, and R. Tibshirani (2010), Regularization paths for generalized linear models via coordinate descent, *Journal of Statistical Software*, *33*(1), 1–22.

Friedman, J. H., and P. Hall (2007), On bagging and nonlinear estimation, *Journal of Statistical Planning and Inference*, *137*(3), 669–683.

He, X., L. Wang, H. G. Hong, et al. (2013), Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data, *The Annals of Statistics*, *41*(1), 342–369.

Huang, J., S. Ma, and C.-H. Zhang (2008), Adaptive lasso for sparse high-dimensional regression models, *Statistica Sinica*, pp. 1603–1618.

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

Javanmard, A., and A. Montanari (2018), Debiasing the lasso: Optimal sample size for gaussian designs, *The Annals of Statistics*, *46*(6A), 2593–2622.

Kelley, M. J., S. Li, and D. H. Harpole (2001), Genetic analysis of the β-tubulin gene, tubb, in non-small-cell lung cancer, *Journal of the National Cancer Institute*, *93*(24), 1886–1888.

Koenker, R., and G. Bassett Jr (1978), Regression quantiles, *Econometrica: journal of the Econometric Society*, pp. 33–50.

Kunst, F., et al. (1997), The complete genome sequence of the gram-positive bacterium bacillus subtilis, *Nature*, *390*(6657), 249–256.

Lange, K. (2004), *Optimization*, 2 ed., Springer, analysis of Convergence.

Larsen, J. E., and J. D. Minna (2011), Molecular biology of lung cancer: clinical implications, *Clinics in Chest Medicine*, *32*(4), 703–740.

Lee, J. D., and J. E. Taylor (2014), Exact post model selection inference for marginal screening, in *Advances in Neural Information Processing Systems*, pp. 136–144.

Lee, J. D., D. L. Sun, Y. Sun, and J. E. Taylor (2016), Exact post-selection inference, with application to the lasso, *The Annals of Statistics*, *44*(3), 907–927.

Li, G., and X. Wang (2018), Prediction accuracy measures for a nonlinear model and for right-censored time-to-event data, *Journal of the American Statistical Association*, dOI: 10.1080/01621459.2018.1515079.

Lord, R. V., J. Brabender, D. Gandara, V. Alberola, C. Camps, M. Domine, et al. (2002), Low ERCC1 expression correlates with prolonged survival after cisplatin plus gemcitabine chemotherapy in non-small cell lung cancer, *Clinical Cancer Research*, *8*(7), 2286–2291.

Lv, J., and Y. Fan (2009), A unified approach to model selection and sparse recovery using regularized least squares, *The Annals of Statistics*, *37*(6A), 3498–3528.

Mander, L., and H.-W. Liu (2010), *Comprehensive Natural Products II: Chemistry and Biology*, vol. 1, Newnes.

Meinshausen, N., L. Meier, and P. Bühlmann (2009), P-values for high-dimensional regression, *Journal of the American Statistical Association*, *104*(488), 1671–1681.

Monzó, M., R. Rosell, J. J. Sánchez, J. S. Lee, A. O'Brate, J. L. González-Larriba, et al. (1999), Paclitaxel resistance in non–small-cell lung cancer associated with beta-tubulin gene mutations, *Journal of Clinical Oncology*, *17*(6), 1786–1786.

Neocleous, T., K. V. Branden, and S. Portnoy (2006), Correction to censored regression quantiles by S. Portnoy, 98 (2003), 1001–1012, *Journal of the American Statistical Association*, *101*(474), 860–861.

Niemiro, W. (1992), Asymptotics for m-estimators defined by convex minimization, *The Annals of Statistics*, *20*(3), 1514–1533.

Ning, Y., and H. Liu (2017), A general theory of hypothesis tests and confidence regions for sparse high dimensional models, *The Annals of Statistics*, *45*(1), 158–195.

Parkin, D. M., F. Bray, J. Ferlay, and P. Pisani (2005), Global cancer statistics, 2002, *CA: A Cancer Journal for Clinicians*, *55*(2), 74–108.

Peng, L., and Y. Huang (2008), Survival analysis with quantile regression models, *Journal of the American Statistical Association*, *103*(482), 637–649.

Portnoy, S. (2003), Censored regression quantiles, *Journal of the American Statistical Association*, *98*(464), 1001–1012.

Portnoy, S., and G. Lin (2010), Asymptotics for censored regression quantiles, *Journal of Nonparametric Statistics*, *22*(1), 115–130.

Powell, J. L. (1986), Censored regression quantiles, *Journal of econometrics*, *32*(1), 143–155.

Rahal, R., et al. (2014), Pharmacological and genomic profiling identifies nf-[kappa] b-targeted treatment strategies for mantle cell lymphoma, *Nature Medicine*, *20*(1), 87–92.

Rosell, R., M. Tarón, and A. O'brate (2001), Predictive molecular markers in non–small cell lung cancer, *Current Opinion in Oncology*, *13*(2), 101–109.

Rosell, R., M. A. Molina, C. Costa, S. Simonetti, A. Gimenez-Capitan, J. Bertran-Alamillo, et al. (2011), Pretreatment EGFR T790M mutation and BRCA1 mRNA expression in erlotinib-treated advanced non–small-cell lung cancer patients with EGFR mutations, *Clinical Cancer Research*, *17*(5), 1160–1168.

Saleem, M., M. I. Qadir, N. Perveen, B. Ahmad, U. Saleem, and T. Irshad (2013), Inhibitors of apoptotic proteins: new targets for anticancer therapy, *Chemical Biology & Drug Design*, *82*(3), 243–251.

Sasaki, H., S. Shimizu, K. Endo, M. Takada, M. Kawahara, H. Tanaka, et al. (2006), EGFR and erbB2 mutation status in japanese lung cancer patients, *International journal of cancer*, *118*(1), 180–184.

Schallmey, M., A. Singh, and O. P. Ward (2004), Developments in the use of bacillus species for industrial production, *Canadian Journal of Microbiology*, *50*(1), 1–17.

Shows, J. H., W. Lu, and H. H. Zhang (2010), Sparse estimation and inference for censored median regression, *Journal of statistical planning and inference*, *140*(7), 1903–1917.

Sinclair, C. S., M. Rowley, A. Naderi, and F. J. Couch (2003), The 17q23 amplicon and breast cancer, *Breast Cancer Research and Treatment*, *78*(3), 313–322.

Slattery, M. L., A. Lundgreen, J. S. Herrick, and R. K. Wolff (2011), Genetic variation in rps6ka1, rps6ka2, rps6kb1, rps6kb2, and pdk1 and risk of colon or rectal cancer, *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, *706*(1), 13–20.

Stephens, P., C. Hunter, G. Bignell, S. Edkins, H. Davies, J. Teague, et al. (2004), Lung cancer: intragenic erbb2 kinase mutations in tumours, *Nature*, *431*(7008), 525–526.

Takeuchi, K., M. Soda, Y. Togashi, R. Suzuki, S. Sakata, S. Hatano, et al. (2012), RET, ROS1 and ALK fusions in lung cancer, *Nature medicine*, *18*(3), 378.

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

Tojo, S., M. Matsunaga, T. Matsumoto, C.-M. Kang, H. Yamaguchi, K. Asai, Y. Sadaie, K.-i. Yoshida, and Y. Fujita (2003), Organization and expression of the bacillus subtilis sigy operon, *Journal of Biochemistry*, *134*(6), 935–946.

Toyooka, S., T. Tsuda, and A. F. Gazdar (2003), The TP53 gene, tobacco exposure, and lung cancer, *Human mutation*, *21*(3), 229–239.

US Department of Health and Human Services (2004), The health consequences of smoking: a report of the surgeon general.

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

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

Van der Hage, J., L. van den Broek, C. Legrand, P. Clahsen, C. Bosch, E. Robanus-Maandag, C. van de Velde, and M. Van de Vijver (2004), Overexpression of p70 s6 kinase protein is associated with increased risk of locoregional recurrence in node-negative premenopausal early breast cancer patients, *British Journal of Cancer*, *90*(8), 1543–1550.

Van der Vaart, A. W. (2000), *Asymptotic statistics*, vol. 3, Cambridge university press.

Vershynin, R. (2010), Introduction to the non-asymptotic analysis of random matrices, *arXiv preprint arXiv:1011.3027*.

Wager, S., and S. Athey (2017), Estimation and inference of heterogeneous treatment effects using random forests, *Journal of the American Statistical Association*, *113*(523), 1228–1242.

Wager, S., and S. Athey (2018), Estimation and inference of heterogeneous treatment effects using random forests, *Journal of the American Statistical Association*, *113*(523), 1228–1242.

Wager, S., T. Hastie, and B. Efron (2014), Confidence intervals for random forests: The jackknife and the infinitesimal jackknife, *Journal of Machine Learning Research*, *15*(1), 1625–1651.

Wang, L., Y. Wu, and R. Li (2012), Quantile regression for analyzing heterogeneity in ultra-high dimension, *Journal of the American Statistical Association*, *107*(497), 214–222.

Wang, T., C. C. Pan, J. R. Yu, Y. Long, X. H. Cai, X. De Yin, et al. (2013), Association between tyms expression and efficacy of pemetrexed–based chemotherapy in advanced non-small cell lung cancer: A meta-analysis, *PLoS One*, *8*(9), e74,284.

Wang, Y., P. Broderick, E. Webb, X. Wu, J. Vijayakrishnan, A. Matakidou, et al. (2008), Common 5p15. 33 and 6p21. 33 variants influence lung cancer risk, *Nature Genetics*, *40*(12), 1407–1409.

Wang, Y., P. Broderick, A. Matakidou, T. Eisen, and R. S. Houlston (2009), Role of 5p15. 33 (TERT-CLPTM1L), 6p21. 33 and 15q25. 1 (CHRNA5-CHRNA3) variation and lung cancer risk in never-smokers, *Carcinogenesis*, *31*(2), 234–238.

Wang, Y., Q. Dong, Q. Zhang, Z. Li, E. Wang, and X. Qiu (2010), Overexpression of yes-associated protein contributes to progression and poor prognosis of non-small-cell lung cancer, *Cancer Science*, *101*(5), 1279–1285.

Wasserman, L., and K. Roeder (2009), High dimensional variable selection, *The Annals of Statistics*, *37*(5A), 2178–2201.

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

Zhang, C.-H., and S. S. Zhang (2014), 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.

Zhang, Y., H.-J. Ni, and D.-Y. Cheng (2013), Prognostic value of phosphorylated mtor/rps6kb1 in non-small cell lung cancer, *Asian Pacific Journal of Cancer Prevention*, *14*(6), 3725–3728.

Zhao, P., and B. Yu (2006), On model selection consistency of lasso, *Journal of Machine Learning Research*, *7*, 2541–2563.

Zheng, Q., C. Gallagher, and K. Kulasekera (2013), Adaptive penalized quantile regression for high dimensional data, *Journal of Statistical Planning and Inference*, *143*(6), 1029–1038.

Zheng, Q., L. Peng, and X. He (2015), Globally adaptive quantile regression with ultra-high dimensional data, *Annals of statistics*, *43*(5), 2225.

Zheng, Q., L. Peng, and X. He (2018), High dimensional censored quantile regression, *The Annals of Statistics*, *46*(1), 308–343.

Zheng, Z., T. Chen, X. Li, E. Haura, A. Sharma, and G. Bepler (2007), DNA synthesis and repair genes RRM1 and ERCC1 in lung cancer, *New England Journal of Medicine*, *356*(8), 800–808.

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

Zou, H., and T. Hastie (2005), Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320.