UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title

Endogenous Econometric Models and Multi-Stage Estimation in High-Dimensional Settings: Theory and Applications

Permalink

https://escholarship.org/uc/item/9vw1524p

Author

Zhu, Ying

Publication Date

2015

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Endogenous Econometric Models and Multi-Stage Estimation in High-Dimensional Settings: Theory and Applications

by

Ying Zhu

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Business Administration

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor James Powell, Co-Chair
Professor J. Miguel Villas-Boas, Co-Chair
Professor Ganesh Iyer
Assistant Professor Przemyslaw Jeziorski
Assistant Professor Demian Pouzo
Professor Martin Wainwright

Spring 2015

Endogenous Econometric Models and Multi-Stage Estimation in High-Dimensional Settings: Theory and Applications

Copyright 2015
by
Ying Zhu

Abstract

Endogenous Econometric Models and Multi-Stage Estimation in High-Dimensional Settings: Theory and Applications

by

Ying Zhu

Doctor of Philosophy in Business Administration

University of California, Berkeley

Professor James Powell, Co-Chair
Professor J. Miguel Villas-Boas, Co-Chair

Econometric models based on observational data are often endogenous due to measurement error, autocorrelated errors, simultaneity and omitted variables, non-random sampling, self-selection, etc. Parameter estimates of these models without corrective measures may be inconsistent. The potential high-dimensional feature of these models (where the dimension of the parameters of interests is comparable to or even larger than the sample size) further complicates the statistical estimation and inference. My dissertation studies two different types of high-dimensional endogenous econometrics problems in depth and develops statistical tools together with their theoretical guarantees.

The first essay in this dissertation explores the validity of the two-stage regularized least squares estimation procedure for sparse linear models in high-dimensional settings with possibly many endogenous regressors. The second essay is focused on the semiparametric sample selection model in high-dimensional settings under a weak nonparametric restriction on the form of the selection correction, for which a multi-stage projection-based regularized procedure is proposed. The number of regressors in the main equation, p, and the number of regressors in the first-stage equation, d, can grow with and exceed the sample size n in the respective models. The analysis considers the *sparsity* case where the number of non-zero components in the vectors of coefficients is bounded above by some integer which is allowed to grow with n but slowly compared to n, or the vectors of coefficients can be approximated by exactly sparse vectors. Simulations are conducted to gain insight on the small-sample performance of these high-dimensional multi-stage estimators. The proposed estimators in the second essay

are also applied to study the pricing decisions of the gasoline retailers in the Greater Saint Louis area.

The main theoretical results of both essays are finite-sample bounds from which sufficient scaling conditions on the sample size for estimation consistency and variable-selection consistency (i.e., the multi-stage high-dimensional estimation procedures correctly select the non-zero coefficients in the main equation with high probability) are established. A technical issue regarding the so-called "restricted eigenvalue (RE) condition" for estimation consistency and the "mutual incoherence (MI) condition" for selection consistency arises in these multi-stage estimation procedures from allowing the number of regressors in the main equation to exceed n and this paper provides analysis to verify these RE and MI conditions. In particular, for the semiparametric sample selection model, these verifications also provide a finite-sample guarantee of the population identification condition required by the semiparametric sample selection models.

In the second essay, statistical efficiency of the proposed estimators is studied via lower bounds on minimax risks and the result shows that, for a family of models with exactly sparse structure on the coefficient vector in the main equation, one of the proposed estimators attains the smallest estimation error up to the $(n, d, p)-$scaling among a class of procedures in worst-case scenarios. Inference procedures for the coefficients of the main equation, one based on a pivotal Dantzig selector to construct non-asymptotic confidence sets and one based on a post-selection strategy (when perfect or near-perfect selection of the high-dimensional coefficients is achieved), are discussed. Other theoretical contributions of this essay include establishing the non-asymptotic counterpart of the familiar asymptotic "oracle" type of results from previous literature: the estimator of the coefficients in the main equation behaves as if the unknown nonparametric component were known, provided the nonparametric component is sufficiently smooth.

Acknowledgments

I would like to express my deepest appreciation and thanks to my co-advisor Professor James Powell, who introduced me to theoretical econometrics, and who has been a very precious mentor and friend for me. I would like to thank him for encouraging my research and for allowing me to grow as a researcher in theoretical econometrics. His advice on both research as well as on my life have been priceless. Without his supervision and constant help this dissertation would not have been possible.

I would also like to thank my other co-advisor, Professor J. Miguel Villas-Boas, who has been unwavering in his personal and professional support throughout my PhD education at Berkeley, who has always been very patient with whatever I ask and no matter how worried I am when I walk in or call, and who always seems to calm my fears during my toughest times. To me, Professor Villas-Boas is almost like a father, stubborn but extremely caring.

I would like to thank my committee member, Professor Martin Wainwright, who led me into theoretical statistics and particularly, high-dimensional statistics, empirical processes, and concentration of measure, whose papers and books on high-dimensional statistics have a great impact on my research and help me build a very solid theoretical foundation.

I am also grateful to Professor Demian Pouzo for his constant support and help on my research and presentations. In addition, a thank you to Professor Ganesh Iyer, who provided me the data set for my dissertation and useful comments on my empirical application.

I take this opportunity to express gratitude to other Berkeley faculty members including Professor Aditya Guntuboyina, Professor David Brillinger, Professor Michael Jansson, Professor Bryan Graham, and Professor Przemyslaw Jeziorski. I also want to thank Haas School of Business and the Department of Economics at UC Berkeley for their financial support granted through various fellowship and teaching assistantship.

A special thanks to my family. Words cannot express how grateful I am to my mother and father for all of the sacrifices that they have made on me. I would also like to thank my other family members (especially my aunt) and my friends (especially Xiangyu Chang, Patty Glidden, Melissa Hacker, Kim Guilfoyle, Bradley Jong, Yuchen Zhang, Carla Trujillo, etc.) who always make themselves available for me.

Chapter 1

Introduction

The past decade has witnessed research activities in high-dimensional statistics that considers estimation of models in which the dimension of the parameters of interests is comparable to or even larger than the sample size. The rapid advance of data collection technology is a major driving force of the development of high-dimensional statistics: it allows for not only more observations but also more explanatory variables to be collected. A great deal of attention in the literature has been given to high-dimensional sparse linear regression models and the l_1-penalized least squares. In particular, the Lasso and the Dantzig selector are the most studied techniques (see, e.g., Tibshirani, 1996; Candès and Tao, 2007; Bickel, Ritov, and Tsybakov, 2009; Belloni, Chernozhukov, and Wang, 2011; Belloni and Chernozhukov, 2011b; Loh and Wainwright, 2012; Negahban, Ravikumar, Wainwright, and Yu, 2012). Variable selection when the dimension of the problem is larger than the sample size has also been studied in the likelihood method setting with penalty functions other than the l_1-norm (see, e.g., Fan and Li, 2001; Fan and Lv, 2011). Lecture notes by Koltchinskii (2011), as well as recent books by Bühlmann and van de Geer (2011) and Wainwright (2015) have given a more comprehensive introduction to high-dimensional statistics.

Recently, these l_1-penalized techniques have been applied in a number of econometric papers. Caner (2009) studies a Lasso-type GMM estimator. Rosenbaum and Tsybakov (2010) study the high-dimensional errors-in-variables problem where the non-random regressors are observed with additive error and they present an application to hedge fund portfolio replication. Belloni and Chernozhukov (2011a) study the l_1-penalized quantile regression and illustrate its use on an international economic growth application. Fan, Lv, and Li (2011) review the literature on sparse high-dimensional econometric models including the vector autoregressive model for measuring the effects of monetary policy, panel data model for forecasting home price, and volatility matrix estimation in finance. Their discussion is not restricted to l_1-based regularization methods. Manresa (2014) considers settings where outcomes depend on an agent's own characteristics and on the characteristics of other agents in the data and applies a Lasso type estimator to study individuals generating spillovers and their strength using panel data on outcomes and characteristics. Bonaldi, Hortacsu, and Kastl (2014) propose a new measure of systemic risk based on estimating spillovers between funding costs of individual banks with a Lasso type procedure applied to the panel of each individual bank to recover the financial network. Belloni, Chen, Chernozhukov, and Hansen (2012) estimate the optimal instruments using the Lasso and in an empirical example dealing with the effect of judicial eminent domain decisions on economic outcomes, they find the Lasso-based instrumental variable estimator outperforms an intuitive benchmark. Belloni, Chernozhukov, and Hansen (2014) propose robust methods for inference on the effect of a treatment variable on a scalar outcome in the presence of many controls with an application to abortion and crime.

Belloni and Chernozhukov (2011b) discuss the l_1-based penalization methods with various econometric problems including earning regressions and instrumental selection in Angrist and Krueger data (1991). In particular, they consider the following linear instrumental variable model which has a single endogenous regressor but many instruments:

$$\begin{aligned} y_i &= \theta_0 + \theta_1 x_{1i} + \mathbf{x}_{2i}^T \gamma + \epsilon_i \\ x_{1i} &= \mathbf{z}_i^T \beta + \mathbf{x}_{2i}^T \delta + \eta_i, \end{aligned}$$

with $\mathbb{E}(\epsilon_i|\mathbf{x}_{2i}, \mathbf{z}_i) = \mathbb{E}(\eta_i|\mathbf{x}_{2i}, \mathbf{z}_i) = 0$. Here y_i, x_{1i}, and $\mathbf{x}_{2i}$ denote wage, education (the endogenous regressor), and a vector of other explanatory variables (the exogenous regressors) respectively, and $\mathbf{z}_i$ denotes a vector of instrumental variables that have direct effect on education but are uncorrelated with the unobservables (i.e., ϵ_i) such as innate abilities in the wage equation. They show the instruments selected by the Lasso technique in the first-stage regression can produce an efficient estimator with a small bias at the same time.

In many applications, the number of endogenous regressors is also large relative to the sample size. The case of many endogenous regressors and many instrumental variables has been studied in the context of Method of Moments by Gautier and Tsybakov (2011) and Fan and Liao (2014). Fan and Liao show that the penalized GMM and penalized empirical likelihood are consistent in both estimation and selection. Gautier and Tsybakov propose a new estimation procedure called the Self Tuning Instrumental Variables (STIV) estimator based on the moment conditions $\mathbb{E}(\mathbf{z}_i\epsilon_i) = \mathbf{0}$. To the best of my knowledge, the two-stage regularized least squares estimation of a linear instrumental variable model that has a large number of endogenous regressors relative to the sample size has not been studied in the literature, even though the two-stage least squares estimator is one of the most popular tools used to correct for endogeneity in empirical economic research. Chapter 2 in this dissertation aims to explore the validity of these two-step estimation procedures for the triangular simultaneous linear equation models in the high-dimensional setting under the sparsity scenario. In particular, the number of endogenous regressors in the main equation and the instruments in the first-stage equations can grow with and exceed the sample size n. The analysis concerns the *exact sparsity* case, i.e., the maximum number of non-zero components in the vectors of parameters in the first-stage equations, k_1, and the number of non-zero components in the vector of parameters in the second-stage equation, k_2, are allowed to grow with n but slowly compared to n. I consider the high-dimensional version of the two-stage least squares estimator where one obtains the fitted regressors from the first-stage regression by a least squares estimator with l_1- regularization (the Lasso or Dantzig selector) when the first-stage regression concerns a large number of instruments relative to n, and then construct a similar estimator using these fitted regressors in the second-stage regression.

Another important class of endogenous econometric models that have not been considered in high-dimensional settings is the sample selection model. Selection models are central in many marketing applications. For example, on the demand side, consumers often face choosing a service or brand followed by the amount of utilization or the number of quantities to purchase conditional on the chosen service or brand. On the supply side, firms first decide on the product positioning and then a pricing scheme based on the chosen product type. Selection models are also seen in the auction literature. In estimating the underlying selection models to study these empirical problems, only a low-dimensional set of explanatory variables has been considered even though the actual information available to these empirical problems can be far richer than what has been used by the researchers. The lack of estimation methods that deal with these "data-rich" selection problems may have limited the use of high-dimensional techniques in many economics and marketing problems. Chapter 3 in this dissertation aims to provide estimation tools together with their theoretical guarantees for this important but little studied topic. In particular, the number of regressors in the outcome equation, p, and the number of regressors in the selection equation, d, can grow with and exceed the sample size n. The analysis considers the *exact sparsity* case where the number of non-zero components in the vectors of coefficients is bounded above by some integer which is allowed to grow with n but slowly compared to n, and also considers the *approximate sparsity* case, where the vectors of coefficients can be approximated by exactly sparse vectors.

1.1 An Overview of the Lasso

I will begin with an overview of the Lasso for estimating the vector of coefficients in the following high-dimensional sparse linear models

$$y_i = x_i^T \beta^* + \epsilon_i = \sum_{j=1}^{p} x_{ij}\beta_j^* + \epsilon_i, \quad i = 1, ..., n, \tag{1.1}$$

where $\mathbb{E}(x_i \epsilon_i) = \mathbf{0}$ for $i = 1, ..., n$. Assume p, the number of regressors, in the above equation grows with and exceeds the sample size n. Let us first consider the class of models where β^* has at most k non-zero parameters, where k is also allowed to increase to infinity with n but slowly compared to n. The Lasso procedure is a combination of the residual sum of squares and a l_1-regularization defined by the following program

$$\hat{\beta}_{Las} \in \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} |y - X\beta|_2^2 + \lambda_n |\beta|_1 \right\}, \tag{1.2}$$

where $\lambda_n > 0$ is some regularization or tuning parameter. Denote the minimizer to the above program by $\hat{\beta}_{Las}$. A necessary and sufficient condition of $\hat{\beta}_{Las}$ is that 0 belongs to the subdifferential of the convex function $\beta \mapsto \frac{1}{2n}|y - X\beta|_2^2 + \lambda_n|\beta|_1$. This implies that the Lasso solution $\hat{\beta}_{Las}$ satisfies the constraint

$$\left| \frac{1}{2n} X^T (y - X\hat{\beta}_{Las}) \right|_\infty \leq \lambda_n.$$

The Dantzig selector of the linear regression function is defined as a vector having the smallest l_1-norm among all β satisfying the above constraint, i.e.,

$$\hat{\beta}_{Dan} \in \arg\min \left\{ |\beta|_1 : \left| \frac{1}{2n} X^T (y - X\beta) \right|_\infty \leq \lambda_n \right\}.$$

Under the exact sparsity assumption, Bickel et al., 2009 shows that the Lasso and the Dantzig selector exhibit similar behavior.

Note that, by Lagrangian duality theory, the following constrained version of the Lasso

$$\hat{\beta}_{Las} \in \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2n} |Y - X\beta|_2^2 \quad \text{such that } |\beta|_1 \leq R.$$

is equivalent to (1.2). For example, for any choice of radius $R > 0$ in the constrained variant of the Lasso, there is a tuning parameter $\lambda_n(R) \geq 0$ such that solving the Lagrangian form of the Lasso is equivalent to solving the constrained version. Consider the constrained Lasso program above with radius $R = |\beta^*|_1$. With this setting, the true parameter vector β^* is feasible for the problem. By definition, the estimate $\hat{\beta}_{Las}$ minimizes the quadratic loss function $\mathcal{L}(\beta;\,(y,\,X)) = \frac{1}{2n}|y - X\beta|_2^2$ over the l_1-ball of radius R. As n increases, we expect that β^* should become a near-minimizer of the same loss, so that $\mathcal{L}(\hat{\beta}_{Las};\,(y,\,X)) \approx \mathcal{L}(\beta^*;\,(y,\,X))$. But when does closeness in the loss imply that the error vector $v := \hat{\beta}_{Las} - \beta^*$ is also small? The link between the excess loss $\mathcal{L}(\hat{\beta}_{Las}) - \mathcal{L}(\beta^*)$ and the size of the error $v = \hat{\beta}_{Las} - \beta^*$ is the Hessian of the loss function, $\nabla^2 L(\beta) = \frac{1}{n}X^TX$, which captures the curvature of the loss function. In the low-dimensional setting where $p < n$, as long as $\text{rank}(X) = p$, we are guaranteed that the Hessian matrix, $\hat{\Sigma} = \frac{1}{n}X^TX$, of the loss function is positive definite, i.e., $v^T\hat{\Sigma}v \geq \delta > 0$ for $v \in \mathbb{R}^p \backslash \{0\}$. In the high-dimensional setting with $p > n$, the Hessian is a $p \times p$ matrix with rank at most n, so that it is impossible to guarantee that it has a positive curvature in all directions.

In the high-dimensional setting, a sufficient condition for the l_2- consistency of the Lasso estimator $\hat{\beta}_{Las}$ is the *restricted eigenvalue* (RE) condition related to the positive definiteness of the Gram matrix $\frac{X^TX}{n}$ over a restricted set (see, e.g., Bickel, et. al., 2009; Meinshausen and Yu, 2009; Raskutti, et al., 2010; Bühlmann and van de Geer, 2011; Loh and Wainwright 2012; Negahban, et. al., 2012; etc.). The restricted eigenvalue (RE) condition is one of the plausible ways to relax the stringency of the uniform curvature condition. The RE condition assumes that the Hessian matrix, $\hat{\Sigma} = \frac{1}{n}X^TX$, of the loss function is positive definite on a restricted set (the choice of this set is associated with the l_1-penalty and to be explained shortly). In this dissertation, I will use the following definition (see, Negahban, et. al., 2012; Wainwright, 2015).

Definition 1.1 (RE): The matrix $X \in \mathbb{R}^{n\times p}$ satisfies the RE condition over a subset $S \subseteq \{1,\,2,\dots,p\}$ with parameter $(\delta,\,\gamma)$ if

$$\frac{\frac{1}{n}|Xv|_2^2}{|v|_2^2} \geq \delta > 0 \qquad \text{for all } v \in \mathbb{C}(S;\,\gamma)\backslash\{\mathbf{0}\},$$

where

$$\mathbb{C}(S;\,\gamma) := \{v \in \mathbb{R}^p \,|\, |v_{S^c}|_1 \leq \gamma|v_S|_1\} \quad \text{for some constant } \gamma \geq 1$$

with v_S denoting the vector in $\mathbb{R}^p$ that has the same coordinates as v on S and zero coordinates on the complement S^c of S.

When the unknown vector $\beta^* \in \mathbb{R}^p$ is exactly sparse, a natural choice of S is the support set of β^*, i.e., $J(\beta^*)$. RE is a weaker condition than other restrictions in the literature including the pairwise incoherence condition (Donoho, 2006; Gautier and Tsybakov, 2011, Proposition 4.2) and the restricted isometry property (Candès and Tao, 2007). As shown by Bickel et al., 2009, the restricted isometry property implies the RE condition but not vice versa. Additionally, Raskutti et al., 2010 give examples of matrix families for which the RE condition holds, but the restricted isometry constants tend to infinity as $(n, |S|)$ grow. Furthermore, they show that even when a matrix exhibits a high amount of dependency among the covariates, it might still satisfy RE. To be more precise, they show that, if $X \in \mathbb{R}^{n \times p}$ is formed by independently sampling each row $X_i \sim N(0, \Sigma)$, then there are strictly positive constants (κ_1, κ_2), depending only on the positive definite matrix Σ, such that

$$\frac{|Xv|_2^2}{n} \geq \kappa_1 |v|_2^2 - \kappa_2 \frac{\log p}{n} |v|_1^2, \qquad \text{for all } v \in \mathbb{R}^p,$$

with probability at least $1 - c_1 \exp(-c_2 n)$ for some universal constants c_1 and c_2. The bound above ensures the RE condition holds with $\delta = \frac{\kappa_1}{2}$ and $\gamma = 3$ as long as $n > 32 \frac{\kappa_2}{\kappa_1} k \log p$. To see this, note that for any $v \in \mathbb{C}(J(\beta^*), 3)$, we have $|v|_1^2 \leq 16 |v_{J(\beta^*)}|_1^2 \leq 16k |v_{J(\beta^*)}|_2^2$. Given the lower bound above, for any $v \in \mathbb{C}(J(\beta^*); 3)$, we have the lower bound

$$\frac{|Xv|_2^2}{n} \geq \left(\kappa_1 - 16\kappa_2 \frac{k \log p}{n} \right) |v|_2^2 \geq \frac{\kappa_1}{2} |v|_2^2,$$

where the final inequality follows as long as $n > 32 (\frac{\kappa_2}{\kappa_1})^2 k \log p$. An appropriate choice of the tuning parameter λ_n in the Lasso program ensures $\hat{v} := \hat{\beta}_{Las} - \beta^* \in \mathbb{C}(J(\beta^*); 3)$. Rudelson and Zhou (2011) as well as Loh and Wainwright (2012) extend this type of RE analysis from the case of Gaussian designs to the case of sub-Gaussian designs. The sub-Gaussian assumption says that the explanatory variables need to be drawn from distributions with well-behaved tails like Gaussian. In contrast to the Gaussian assumption, sub-Gaussian variables constitute a more general family of distributions. In this dissertation, I make use of the following definition for a sub-Gaussian matrix.

Definition 1.2: A random variable X with mean $\mu = \mathbb{E}[X]$ is sub-Gaussian if there is a positive number σ such that

$$\mathbb{E}[\exp(t(X - \mu))] \leq \exp(\sigma^2 t^2 / 2) \qquad \text{for all } t \in \mathbb{R},$$

and a random matrix $A \in \mathbb{R}^{n \times p}$ is sub-Gaussian with parameters (Σ_A, σ_A^2) if (a) each row $A_i^T \in \mathbb{R}^p$ is sampled independently from a zero-mean distribution with covariance Σ_A, (b) for any unit vector $u \in \mathbb{R}^p$, the random variable $u^T A_i^T$ is sub-Gaussian with parameter at most σ_A^2.

Chapter 2

High-Dimensional Linear Models with Endogeneity and Sparsity

2.1 Introduction

The objective of this chapter is consistent estimation and selection of regression coefficients in models with a large number of endogenous regressors relative to the sample size. One example concerns the nonparametric regression model with endogenous explanatory variables. Consider the model $y_i = f(x_i) + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ and $f(\cdot)$ is an unknown function of interest. Assume $\mathbb{E}(\epsilon_i|X_i) \neq 0$ for all i. Suppose we want to approximate $f(x_i)$ by linear combinations of some set of basis functions, i.e., $f(x_i) = \sum_{j=1}^{p} \beta_j \phi_j(x_i)$, where $\{\phi_1, ..., \phi_p\}$ are some known functions. Then, we end up with a linear regression model with many endogenous regressors.

Empirical examples of many endogenous regressors can be found in hedonic price regressions of consumer products (e.g., personal computers, automobiles, pharmaceutical drugs, residential housing, etc.), where the number of explanatory variables formed by the characteristics (and the transformations of these characteristics) of the products can be very large. For example, in the study of hedonic price index analysis in personal computers, the data considered by Benkard and Bajari involved 65 product characteristics (Benkard and Bajari, 2005). Together with the various transformations of these characteristics, the number of the potential regressors can be very large. On the other hand, it is plausible that only a few of these variables matter to the underlying prices but which variables constitute the relevant regressors are unknown to the researchers. Housing data also tends to exhibit a similar high-dimensional but sparse pattern in terms of the underlying explanatory variables (e.g., Lin and Zhang, 2006; Ravikumar, et. al, 2009). Moreover, product characteristics are likely to be endogenous because just like price, product characteristics are typically choice variables of firms, and it is possible that they are correlated with unobserved components of price (Ackerberg and Crawford, 2009). An alternative is to use product characteristics of other firms (other markets) as instruments. In demand estimation literature, this type of instruments are sometimes referred to as BLP instruments, e.g., Berry, et. al., 1995 (respectively, Hausman instruments, e.g., Nevo, 2001). Another empirical example of many endogenous regressors concerns the study of network or community influence. For example, Manresa (2014) looks at how a firm's production output is influenced by the investment of other firms. As a future extension, she suggests an alternative model that looks at the network influence in terms of the output of the other firms rather than their investment:

$$y_{it} = \alpha_i + \zeta_t + \mathbf{x}_{it}^T\theta + \sum_{j\in\{1,...,n\},\, j\neq i} \beta_{ji} y_{jt} + \epsilon_{it}, \qquad i = 1, ..., n,\ t = 1, ..., T$$

$\mathbf{x}_{it}$ denotes a vector of exogenous regressors specific to firm i (e.g., investment) at period t. α_i and ζ_t are the fixed effects of firm i and period t, respectively. Notice that y_{jt}, the output of other firms enters the right-hand-side of the equations above as additional regressors and β_{ji}, $j = 1, ..., n$, and $j \neq i$ are interpreted as the network influence arising from other firms' output on firm i's output. Furthermore, the influence on firm i from firm j is allowed to differ from the influence on firm j from firm i. Endogeneity arises from the simultaneity of the output variables when $\text{cov}(\epsilon_{it}, \epsilon_{jt}) \neq 0$ (e.g., presence of unobserved network characteristics that are common to all firms). As a result, the number of endogenous regressors in the model above is of the order $O(n)$, which exceeds the number of periods T in the application considered by Manresa (2014).

The following sets up the models of interests and highlights the major contributions made by this chapter. We consider the linear model

$$y_i = \mathbf{x}_i^T\beta^* + \epsilon_i = \sum_{j=1}^{p} x_{ij}\beta_j^* + \epsilon_i, \quad i = 1, ..., n \tag{2.1}$$

where ϵ_i is a zero-mean random error possibly correlated with $\mathbf{x}_i$ and β^* is an unknown vector of parameters of our main interests. The j^{th} component of β^* is denoted by β_j^*. A component in the $p-$dimensional vector $\mathbf{x}_i$ is said to be *endogenous* if it is correlated with ϵ_i (i.e., $\mathbb{E}(\mathbf{x}_i\epsilon_i) \neq \mathbf{0}$) and *exogenous* otherwise (i.e., $\mathbb{E}(\mathbf{x}_i\epsilon_i) = \mathbf{0}$). Without loss of generality, I will assume all regressors are endogenous throughout the rest of this chapter for notational convenience (a modification to allow mix of endogenous and exogenous regressors is trivial.). When endogenous regressors are present, the classical least squares estimator will be inconsistent for β^* (i.e., $\hat{\beta}_{OLS} \overset{p}{\nrightarrow} \beta^*$) even when the dimension p of β^* is small relative to the sample size n. The classical solution to this problem of endogenous regressors supposes that there is some L-dimensional vector of instrumental variables, denoted by $\mathbf{z}_i$, which is observable and satisfies $\mathbb{E}(\mathbf{z}_i\epsilon_i) = \mathbf{0}$ for all i. In particular, the two-step estimation procedures including the two-stage least square (2SLS) estimation and the control function approach play an important role in accounting for endogeneity that comes from individual choice or market equilibrium (e.g., Wooldridge, 2002). Consider the following “first stage” equations for the components of $\mathbf{x}_i$

$$x_{ij} = \mathbf{z}_{ij}^T\pi_j^* + \eta_{ij} = \sum_{l=1}^{d_j} z_{ijl}\pi_{jl}^* + \eta_{ij}, \qquad i = 1, ..., n, \ j = 1, ..., p. \tag{2.2}$$

For each $j = 1, ..., p$, $\mathbf{z}_{ij}$ is a $d_j \times 1$ vector of instrumental variables, and η_{ij} a zero-mean random error which is uncorrelated with $\mathbf{z}_{ij}$, and π_j^* is an unknown vector of nuisance parameters. I will refer to equation (2.1) as the main equation (or second-stage equation) and the equations in (2.2) as the first-stage equations. Throughout the rest of this chapter, I will impose the following assumption. Without loss of generality, this assumption implies a triangular simultaneous equations model structure.

Assumption 2.1.1: The data $\{y_i, \mathbf{x}_i, \mathbf{z}_i\}_{i=1}^n$ are *i.i.d.* with finite second moments; $\mathbb{E}(\mathbf{z}_{ij}\epsilon_i) = \mathbb{E}(\mathbf{z}_{ij}\eta_{ij}) = \mathbf{0}$ for all $j = 1, ..., p$ and $\mathbb{E}(\mathbf{z}_{ij}\eta_{ij'}) = \mathbf{0}$ for all $j \neq j'$.

High dimensionality arises in the triangular simultaneous equations structure (2.1) and (2.2) when the dimension p of β^* is large relative to the sample size n (namely, $p \gg n$) or when the dimension d_j of π_j^* is large relative to the sample size n (namely, $d_j \gg n$) for at least one j. The linear instrumental variable model with a single or a few endogenous regressors and many instruments (the case where $d_j \gg n$ for at least one j but $p \ll n$) has been studied in the econometrics literature on high dimensional models (e.g., Belloni and Chernozhukov, 2011b). In this chapter, I consider the scenario where $p \gg n$ and $d_j \ll n$ for all j, or $p \gg n$ and $d_j \gg n$ for at least one j, and the number of non-zero coefficients in β^* and π_j^* is small relative to n (i.e., β^* and $\pi_j^*$for $j = 1, ..., p$ are *exactly sparse*). To the best of my knowledge, the case where $p \gg n$ and $d_j \ll n$ for all j, or the case where $p \gg n$ and $d_j \gg n$ for at least one j in the context of triangular simultaneous equations with a 2SLS type of estimation procedure has not been studied in the literature. In the presence of endogenous regressors, the direct implementation of the Lasso or Dantzig selector fails as sparsity of coefficients in equation (2.1) does not correspond to sparsity of linear projection coefficients. Nevertheless, one can still use the ideas of the 2SLS estimation together with the Lasso technique. For instance, in the case where $p \gg n$ and $d_j \ll n$ for all j, one can obtain the fitted regressors by a standard least squares estimation on each of the first-stage equations separately as usual and then apply a Lasso-type technique with these fitted regressors in the second-stage regression. Similarly, in the case where $p \gg n$ and $d_j \gg n$ for all j, one can obtain the fitted regressors by performing a regression with a Lasso-type estimator on each of the first-stage equations separately and then apply another Lasso-type estimator with these fitted regressors in the second-stage regression.

Compared to existing two-stage techniques which limit the number of regressors entering the first-stage equations or the second-stage equation or both, the two-stage estimation procedures with l_1-regularization in both stages are more flexible and particularly powerful for applications in which the vector of parameters of interests is sparse and there is lack of information about the relevant explanatory variables and instruments. In terms of practical implementations, these above-mentioned high-dimensional two-stage estimation procedures are intuitive and can be easily implemented using existing software packages for the standard Lasso-type technique for linear models without endogeneity. In analyzing the statistical properties of these estimators, the extension from models with a few endogenous regressors to models with many endogenous regressors ($p \gg n$) in the context of triangular simultaneous equations with two-stage estimation is not obvious. Chapter 2 aims to explore the validity of these two-step estimation procedures for the triangular simultaneous linear equation models in the high-dimensional setting under the sparsity scenario.

An important contribution of this chapter is to introduce analysis that is suitable for showing estimation consistency and selection consistency of the two-step type of high-dimensional estimators. When endogeneity is absent from model (2.1), there is a well-developed theory on what conditions on the design matrix $X \in \mathbb{R}^{n \times p}$ are sufficient (sufficient and necessary) for an l_1-based regularized estimator to consistently estimate (respectively, select) β^*. In some situations one can impose these conditions directly as an assumption on the underlying design matrix. However, when employing a regularized 2SLS estimator in the context of triangular simultaneous linear equation models in the high-dimensional setting, namely, (2.1) and (2.2), there is no guarantee that the random matrix $\hat{X}^T\hat{X}$ (with $\hat{X}$ obtained from regressing X on the instrumental variables) would automatically satisfy these previously established conditions for estimation or selection consistency. This chapter explicitly shows that these conditions for estimation consistency indeed hold for $\hat{X}^T\hat{X}$ with high probability under a broad class of sub-Gaussian design matrices formed by the instrumental variables allowing for correlations among the covariates. It also establishes the sample size required for $\hat{X}^T\hat{X}$ to satisfy these conditions. Furthermore, with an additional stronger assumption on the structure of the design matrices formed by the instrumental variables, this chapter shows $\hat{X}^T\hat{X}$ also satisfies the conditions for selection consistency under a stronger sample size requirement.

I begin in Section 2.2 with notation used in this chapter. Results regarding the estimation consistency and selection consistency of the high-dimensional 2SLS procedure under the sparsity scenario are established in Section 2.3. Section 2.4 presents simulation results. Section 2.5 concludes this chapter and discusses future extensions. All the proofs are collected in Section 2.6.

2.2 Notation

Notation. For the convenience of the reader, I summarize here notations to be used throughout this chapter. The l_q norm of a vector $v \in m \times 1$ is denoted by $|v|_q$, $1 \leq q \leq \infty$ where $|v|_q := \left(\sum_{i=1}^m |v_i|^q\right)^{1/q}$ when $1 \leq q < \infty$ and $|v|_q := \max_{i=1,...,m} |v_i|$ when $q = \infty$. For a matrix $A \in \mathbb{R}^{m \times m}$, write $|A|_\infty := \max_{i,j} |a_{ij}|$ to be the elementwise l_∞- norm of A. The l_2-operator norm, or spectral norm of the matrix A corresponds to its maximum singular value; i.e., it is defined as $||A||_2 := \sup_{v \in S^{m-1}} |Av|_2$, where $S^{m-1} = \{v \in \mathbb{R}^m \,|\, |v|_2 = 1\}$. The l_∞ matrix norm (maximum absolute row sum) of A is denoted by $||A||_\infty := \max_i \sum_j |a_{ij}|$ (note the difference between $|A|_\infty$ and $||A||_\infty$). I make use of the bound $||A||_\infty \leq \sqrt{m}||A||_2$ for any symmetric matrix $A \in \mathbb{R}^{m \times m}$. For a square matrix A, denote its minimum eigenvalue and maximum eigenvalue by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$, respectively. For functions $f(n)$ and $g(n)$, write $f(n) \succsim g(n)$ to mean that $f(n) \geq cg(n)$ for a universal constant $c \in (0, \infty)$ and similarly, $f(n) \precsim g(n)$ to mean that $f(n) \leq c^{'}g(n)$ for a universal constant $c^{'} \in (0, \infty)$. $f(n) \asymp g(n)$ when $f(n) \succsim g(n)$ and $f(n) \precsim g(n)$ hold simultaneously. For some integer $s \in \{1, 2, ..., m\}$, the l_0-ball of radius s is given by $\mathbb{B}_0^m(s) := \{v \in \mathbb{R}^m \,|\, |v|_0 \leq s\}$ where $|v|_0 := \sum_{i=1}^m 1\{v_i \neq 0\}$. Similarly, the l_2-ball of radius r is given by $\mathbb{B}_2^m(r) := \{v \in \mathbb{R}^m \,|\, |v|_2 \leq r\}$. Also, write $\mathbb{K}(s, m) := \mathbb{B}_0^m(s) \cap \mathbb{B}_2^m(1)$ and $\mathbb{K}^2(s, m) := \mathbb{K}(s, m) \times \mathbb{K}(s, m)$. For a vector $v \in \mathbb{R}^p$, let $J(v) = \{j \in \{1, ..., p\} \,|\, v_j \neq 0\}$ be its support, i.e., the set of indices corresponding to its non-zero components v_j. The cardinality of a set $J \subseteq \{1, ..., p\}$ is denoted by $|J|$.

2.3 High-dimensional 2SLS estimation

Suppose from performing a first-stage regression on each of the equations in (2.2) separately, we obtain estimates $\hat{\pi}_j$ and let $\hat{\mathbf{x}}_j := Z_j\hat{\pi}_j$ for $j = 1, ..., p$. Denote the fitted regressors from the first-stage estimation by $\hat{X}$, where $\hat{X} = (\hat{\mathbf{x}}_1, ..., \hat{\mathbf{x}}_p)$. For the second-stage regression, consider the following Lasso program:

$$\hat{\beta}_{H2SLS} \in \text{argmin}_{\beta \in \mathbb{R}^p} : \frac{1}{2n}|y - \hat{X}\beta|_2^2 + \lambda_n|\beta|_1. \tag{2.3}$$

The following is a standard assumption in the literature on sparsity for high-dimensional linear models.

Assumption 2.3.1: The numbers of regressors $p(= p_n)$ and $d_j(= d_{jn})$ for every $j = 1, ..., p$ in (2.1) and (2.2) can grow with and exceed the sample size n. The number of non-zero components in π_j^* is at most $k_1(= k_{1n})$ for all $j = 1, ..., p$, and the number of non-zero components in β^* is at most $k_2(= k_{2n})$. Both k_1 and k_2 can increase to infinity with n but slowly compared to n.

I first present a general bound on the statistical error measured by the quantity $|\hat{\beta}_{H2SLS} - \beta^*|_2$.

Lemma 2.3.1 (General upper bound on the l_2-error). Let $\hat{\Gamma} = \frac{1}{n}\hat{X}^T\hat{X}$ and $e = (X - \hat{X})\beta^* + \boldsymbol{\eta}\beta^* + \epsilon$. Suppose the random matrix $\hat{\Gamma}$ satisfies the RE condition in Definition 1.1 from Chapter 1 with $\gamma = 3$ and the vector β^* is supported on a subset $J(\beta^*) \subseteq \{1, 2, ...p\}$ with its cardinality $|J(\beta^*)| \leq k_2$. If a solution $\hat{\beta}_{H2SLS}$, defined in (2.3) has λ_n satisfying

$$\lambda_n \geq 2|\frac{1}{n}\hat{X}^T e|_\infty > 0,$$

for any given n, then there is a constant $c > 0$ such that

$$|\hat{\beta}_{H2SLS} - \beta^*|_2 \leq \frac{c}{\delta}\sqrt{k_2}\lambda_n.$$

The proof for Lemma 2.3.1 is provided in Section 2.6.1.

In order to apply Lemma 2.3.1 to prove consistency, we need to show (i) $\hat{\Gamma} = \frac{1}{n}\hat{X}^T\hat{X}$ satisfies the RE condition in Definition 1.1 from Chapter 1 with $\gamma = 3$ and (ii) the term $|\frac{1}{n}\hat{X}^T e|_\infty \precsim f(k_1, k_2, d_1, ..., d_p, p, n)$ with high probability, and then we can show

$$|\hat{\beta}_{H2SLS} - \beta^*|_2 \precsim \sqrt{k_2} f(k_1, k_2, d_1, ..., d_p, p, n)$$

by choosing $\lambda_n \asymp f(k_1, k_2, d_1, ..., d_p, p, n)$. The assumption

$$\sqrt{k_2} f(k_1, k_2, d_1, ...d_p, p, n) = o(1)$$

will therefore imply the l_2-consistency of $\hat{\beta}_{H2SLS}$. Applying Lemma 2.3.1 to the triangular simultaneous equations model (2.1) and (2.2) requires additional work to establish conditions (i) and (ii) discussed above, which depends on the specific first-stage estimator for $\hat{X}$. It is worth mentioning that, while in many situations one can impose the RE condition as an assumption on the design matrix (e.g., Belloni and Chernozhukov, 2011b; Belloni, Chen, Chernozhukov, and Hansen, 2012) in analyzing the consistency property of the Lasso, appropriate analysis is needed in this chapter to verify that $\frac{1}{n}\hat{X}^T\hat{X}$ satisfies the RE condition because $\hat{X}$ is obtained from a first-stage estimation and there is no guarantee that the random matrix $\frac{1}{n}\hat{X}^T\hat{X}$ would automatically satisfy the RE condition. To the best of my knowledge, previous literature has not dealt with this issue directly. Consequently, the RE analysis introduced in this chapter is particularly useful for analyzing the statistical properties of the two-step type of high-dimensional estimators in the simultaneous equations model context. As discussed previously, this chapter focuses on the case where $p \gg n$ and $d_j \ll n$ for all j and the case where $p \gg n$ and $d_j \gg n$ for at least one j. The following two subsections present results concerning estimation consistency and variable-selection consistency for the exact sparsity case.

2.3.1 Estimation consistency for the sparsity case

To derive the non-asymptotic bounds and asymptotic properties (i.e., estimation consistency and selection consistency) for $\hat{\beta}_{H2SLS}$, I impose the following regularity conditions.

Assumption 2.3.2: The error terms ϵ and η_j for $j = 1, ..., p$ are *i.i.d.* zero-mean sub-Gaussian vectors with parameters σ_ϵ^2 and σ_η^2, respectively. The random matrix $Z_j \in \mathbb{R}^{n\times d_j}$ is sub-Gaussian with parameters $(\Sigma_{Z_j},\, \sigma_Z^2)$ for $j = 1, ..., p$.

Assumption 2.3.3: For every $j = 1, ..., p$, $\mathbf{x}_j^* := Z_j\pi_j^*$. The matrix $X^* \in \mathbb{R}^{n\times p}$ is sub-Gaussian with parameters $(\Sigma_{X^*},\, \sigma_{X^*}^2)$ where the jth column of X^* is $\mathbf{x}_j^*$.

Assumption 2.3.4: For every $j = 1, ..., p$, $\mathbf{w}_j := Z_jv_j$ where $v_j \in \mathbb{K}(k_1,\, d_j) := \mathbb{B}_0^{d_j}(k_1) \cap \mathbb{B}_2^{d_j}(1)$. The matrix $W \in \mathbb{R}^{n\times p}$ is sub-Gaussian with parameters $(\Sigma_W,\, \sigma_W^2)$ where the jth column of W is $\mathbf{w}_j$.

Assumption 2.3.5: The first-stage estimator $\hat{\pi}^T \in \mathbb{R}^{p\times d}$ satisfies the bound

$$\max_{j=1,...,p} |\hat{\pi}_j - \pi_j^*|_2 \leq \frac{c\sigma_Z\sigma_\eta}{\lambda_{\min}(\Sigma_Z)}\sqrt{\frac{k_1 \log \max(d,\, p)}{n}}$$

with probability at least $1 - c_1 \exp(-c_2 \log \max(d,\, p,\, n))$ for some universal constants c_1 and c_2, where $d = \max_{j=1,...,p} d_j$ and $\lambda_{\min}(\Sigma_Z) = \min_{j=1,...,p} \lambda_{\min}(\Sigma_{Z_j})$.

Assumption 2.3.6: For every $j = 1, ..., p$, the first-stage estimator $\hat{\pi}_j$ achieves the selection consistency (i.e., it recovers the true support $J(\pi_j^*)$) or has at most k_j^* components that are different from the components in $J(\pi_j^*)$ where $k_j^* \ll n$, with probability at least $1 - c_1 \exp(-c_2 \log \max(d,\, p,\, n))$ for some universal constants c_1 and c_2, where $d = \max_{j=1,...,p} d_j$. For simplicity, we consider the case where the first-stage estimator recovers the true support $J(\pi_j^*)$ for every $j = 1, ..., p$.

Remarks

Assumption 2.3.2 is common in the literature (see, Loh and Wainwright, 2012; Negahban, et. al 2012; Rosenbaum and Tsybakov, 2013). The assumption that $Z_j \in \mathbb{R}^{n \times d_j}$ is sub-Gaussian with parameters $(\Sigma_{Z_j},\, \sigma_Z^2)$ for all j provides a primitive condition which guarantees that the random matrix formed by the instrumental variables satisfies the RE condition with high probability.

Based on the second part of Assumption 2.3.2 that $Z_j \in \mathbb{R}^{n \times d_j}$ is sub-Gaussian with parameters $(\Sigma_{Z_j},\, \sigma_Z^2)$ for all j, we have that $Z_j \pi_j^* := \mathbf{x}_j^*$ and $Z_j v_j := \mathbf{w}_j$ are sub-Gaussian vectors where $v_j \in \mathbb{K}(k_1,\, d_j) := \mathbb{B}_0^{d_j}(k_1) \cap \mathbb{B}_2^{d_j}(1)$. Therefore, the conditions that $X^* \in \mathbb{R}^{n \times p}$ is a sub-Gaussian matrix with parameters $(\Sigma_{X^*},\, \sigma_{X^*}^2)$ where the jth column of X^* is $\mathbf{x}_j^*$(Assumption 2.3.3) and $W \in \mathbb{R}^{n \times p}$ is a sub-Gaussian matrix with parameters $(\Sigma_W,\, \sigma_W^2)$ where the jth column of W is $\mathbf{w}_j$ (Assumption 2.3.4) are mild extensions. In terms of the instrumental variables and their linear combinations, Assumptions 2.3.2-2.3.4 together with Assumption 2.3.5 on the first-stage estimation error provide primitive conditions which guarantee that the random matrix $\frac{1}{n}\hat{X}^T\hat{X}$ formed by the fitted regressors $\hat{\mathbf{x}}_j (:= Z_j \hat{\pi}_j)$ for $j = 1, ..., p$ satisfies the RE condition with high probability.

Many existing high-dimensional estimation procedures such as the Lasso or Dantzig selector (see, e.g., Candès and Tao, 2007; Bickel, et. al, 2009; Negahban, et. al. 2012; Loh and Wainwright, 2012) are shown to satisfy the error bound in Assumption 2.3.5 with high probability. It is worth noting that while the l_2-error from applying the Lasso on a single first-stage equation should be of the order $O\left(\sqrt{\frac{k_1 \log d}{n}}\right)$ with probability at least $1 - c_1 \exp(-c_2 \log \max(d,\, n))$, the extra term $\log p$ in the error $\max_{j=1,...,p} |\hat{\pi}_j - \pi_j^*|_2$ and the probability guarantee $1 - c_1 \exp(-c_2 \log \max(d,\, p,\, n))$ with which these errors hold comes from the application of a union bound which takes into account the fact that there are p endogenous regressors in the main equation and hence, p equations to estimate in the first-stage. As a result, it is not hard to see that the sample size required for consistently estimating p equations simultaneously when a Lasso-type procedure is applied on each of the first-stage equations separately should satisfy $\sqrt{\frac{k_1 \log \max(d,p)}{n}} = o(1)$ as opposed to the condition $\sqrt{\frac{k_1 \log d}{n}} = o(1)$ for the case where a single equation is estimated with a Lasso-type procedure.

Assumption 2.3.6 says that the first-stage estimators correctly select the non-zero coefficients with probability close to 1. In analogy to the various sparsity assumptions on the true parameters in the high-dimensional statistics literature (including the case of *exact sparsity* assumption meaning that the true parameter vector has only a few non-zero components, or *approximate sparsity* assumption based on imposing a certain decay rate on the ordered entries of the true parameter vector), Assumption 2.3.6 can be interpreted as an *exact sparsity* constraint on the first-stage estimate $\hat{\pi}_j$ for $j = 1, ..., p$, in terms of the l_0- ball, given by

$$\mathbb{B}_0^{d_j}(k_1) := \left\{ \hat{\pi}_j \in \mathbb{R}^{d_j} \mid \sum_{l=1}^{d_j} 1\{\hat{\pi}_{jl} \neq 0\} \leq k_1 \right\} \text{ for } j = 1, ..., p.$$

It is known that under some stringent conditions such as the "irrepresentable condition" (Zhao and Yu, 2006; Bühlmann and van de Geer, 2011) or the "mutual incoherence condition" (Wainwright, 2009) together with the "beta-min condition" (Bühlmann and van de Geer, 2011), Lasso and Dantzig types of selectors can recover the support of the true parameter vector with high probability. The "irrepresentable condition", as discussed in Bühlmann and van de Geer, 2011, is in fact a sufficient and necessary condition to achieve variable-selection consistency with the Lasso. Furthermore, they show that the "irrepresentable condition" implies the RE condition. Assumption 2.3.6 is the key condition that differentiates the upper bounds in the two theorems to be presented immediately. Similar to the problem of estimating p equations as in the discussion of Assumption 2.3.5, the sample size required for consistently selecting the coefficients in each of the p equations simultaneously when a Lasso-type selector is applied on each of the first-stage equations separately should satisfy $\sqrt{\frac{k_1 \log \max(d,p)}{n}} = O(1)$ as opposed to the condition $\sqrt{\frac{k_1 \log d}{n}} = O(1)$ for the case where a single equation is estimated with a Lasso-type selector. In addition, the "beta-min" condition for consistent selection in the p-equation problem needs to satisfy $\min_{j=1,...,p} \min_{l \in J(\pi_j^*)} |\pi_{jl}^*| \geq O\left(\sqrt{\frac{\log \max(d,p)}{n}}\right)$ as opposed to $\min_{j=1,...,p} \min_{l \in J(\pi_j^*)} |\pi_{jl}^*| \geq O\left(\sqrt{\frac{\log d}{n}}\right)$ for the consistent selection in a single equation problem.

First, I present two results for the case where $p \gg n$ and $d_j \gg n$ for at least one j. The key difference between the two theorems is that the bound in the second theorem hinges on the additional assumption that the first-stage estimators correctly select the non-zero coefficients with probability close to 1, i.e., Assumption 2.3.6. With this assumption, the scaling of the sample size required for estimation consistency is guaranteed to be no greater (and in some cases strictly smaller) than that without Assumption 2.3.6.

Theorem 2.3.2: Suppose Assumptions 2.1.1, 2.3.1-2.3.3, and 2.3.5 hold. Then, if

$$\frac{k_1 k_2^2 \log \max(d,\, p)}{n} = O(1),$$

and the tuning parameter λ_n satisfies

$$\lambda_n \asymp k_2 \sqrt{\frac{k_1 \log \max(d,\, p)}{n}},$$

we have

$$|\hat{\beta}_{H2SLS} - \beta^*|_2 \precsim \max\{\varphi_1 \sqrt{k_2} \sqrt{\frac{k_1 \log \max(d,\, p)}{n}},\, \varphi_2 \sqrt{\frac{k_2 \log p}{n}}\},$$

where

$$\begin{aligned} \varphi_1 &= \frac{\max\left\{\sigma_\eta,\, \sigma_{X^*},\, \sigma_\epsilon\right\} \sigma_Z \sigma_\eta \sqrt{\lambda_{\max}(\Sigma_Z)} |\beta^*|_1}{\lambda_{\min}(\Sigma_Z) \lambda_{\min}(\Sigma_{X^*})}, \\ \varphi_2 &= \max\left\{\frac{\sigma_{X^*} \sigma_\eta |\beta^*|_1}{\lambda_{\min}(\Sigma_{X^*})},\, \frac{\sigma_{X^*} \sigma_\epsilon}{\lambda_{\min}(\Sigma_{X^*})}\right\}, \end{aligned}$$

with probability at least $1 - c_1 \exp(-c_2 \log \max(\min(p,\, d),\, n))$ for some universal positive constants c_1 and c_2. If we also have $k_2 \sqrt{\frac{k_1 k_2 \log \max(d,\, p)}{n}} = o(1)$, then the two-stage estimator $\hat{\beta}_{H2SLS}$ is l_2-consistent for β^*.

Theorem 2.3.3: Suppose Assumptions 2.1.1, 2.3.1-2.3.6 hold. Then, if

$\min\left\{k_1 k_2^2 \log \max(d,\, p),\, \min_{r \in [0,\,1]} \max\left\{k_1^{3-2r} \log d,\, k_1^{3-2r} \log p,\, k_1^r k_2 \log d,\, k_1^r k_2 \log p\right\}\right\}$
$= O(n)$

and the tuning parameter λ_n satisfies

$$\lambda_n \asymp k_2 \sqrt{\frac{k_1 \log \max(d,\, p)}{n}},$$

we have the same upper bound on $|\hat{\beta}_{H2SLS} - \beta^*|_2$ as in Theorem 2.3.2 with probability at least $1 - c_1 \exp(-c_2 \log \max(\min(p,\, d),\, n))$ for some universal positive constants c_1 and c_2.

The proofs for Theorems 2.3.2 and 2.3.3 are provided in Sections 2.6.2 and 2.6.3, respectively. The proofs for Theorems 2.3.2 and 2.3.3 each consist of two parts. The first part is to show $\frac{1}{n}\hat{X}^T\hat{X}$ satisfies the RE condition in Definition 1.1 from Chapter 1 and the second part is to bound the term $|\frac{1}{n}\hat{X}^T e|_\infty$ from above. Based on Lemma 2.3.1, the upper bound on $|\frac{1}{n}\hat{X}^T e|_\infty$ pins down the scaling requirement of λ_n.

From Theorems 2.3.2 and 2.3.3, in the simple case of $\sigma_\eta = 0$ (for example, $\boldsymbol{\eta} = \mathbf{0}$ with probability 1 as in a high-dimensional linear regression model without endogeneity), the l_2-errors in Theorems 2.3.2 and 2.3.3 reduce to $|\hat{\beta}_{H2SLS} - \beta^*|_2 \precsim \frac{\sigma_{X^*}\sigma_\epsilon}{\lambda_{\min}(\Sigma_{X^*})}\sqrt{\frac{k_2 \log p}{n}}$, where the factor $\frac{\sigma_{X^*}\sigma_\epsilon}{\lambda_{\min}(\Sigma_{X^*})}$ has a natural interpretation of an inverse signal-to-noise ratio. For instance, when X^* is a zero-mean Gaussian matrix with covariance $\Sigma_{X^*} = \sigma^2_{X^*} I$, one has $\lambda_{\min}(\Sigma_{X^*}) = \sigma^2_{X^*}$, so

$$\frac{\sigma_{X^*}\sigma_\epsilon}{\lambda_{\min}(\Sigma_{X^*})} = \frac{\sigma_\epsilon}{\sigma_{X^*}},$$

which measures the inverse signal-to-noise ratio of the regressors in a high-dimensional linear regression model without endogeneity. Hence, the statistical error of the parameters of interests in the main equation matches the scaling of the upper bound for the Lasso in the context of the high-dimensional linear regression model without endogeneity, i.e., $\sqrt{\frac{k_2 \log p}{n}}$.

Under the assumption that the first-stage estimators correctly select the non-zero coefficients with high probability (Assumption 2.3.6), the scaling of the sample size required in Theorem 2.3.3 is guaranteed to be no greater (and in some cases strictly smaller) than that in Theorem 2.3.2. For instance, if $p \leq d$, then letting $r = 1$ yields

$$\begin{aligned}\max\{k_1 \log d,\, k_1 \log p,\, k_1 k_2 \log d,\, k_1 k_2 \log p\} &= k_1 k_2 \log d \\ &\leq k_1 k_2^2 \log \max(d,\, p) = k_1 k_2^2 \log d.\end{aligned}$$

In this example, Theorem 2.3.2 suggests that the choice of sample size needs to satisfy $\frac{k_1 k_2^2 \log d}{n} = O(1)$ while Theorem 2.3.3 suggests that the choice of sample size only needs to satisfy $\frac{k_1 k_2 \log d}{n} = O(1)$.

The estimation error of the parameters of interests in the main equation can be bounded by the maximum of a term involving the first-stage estimation error and a term involving the second-stage estimation error, which partially confirms the speculation in Gautier and Tsybakov (2011) (Section 7.2) that the two-stage estimation procedure can achieve the estimation error of an order $\sqrt{\frac{\log p}{n}}$. My results show that $\sqrt{\frac{\log p}{n}}$ is achieved either when the second-stage estimation error dominates the first-stage estimation error, or when p is large relative to d. In the case where the second-stage estimation error dominates the first-stage estimation error, the statistical error of the parameters of interests in the main equation matches (up to a factor of $|\beta^*|_1$) the order of the upper bound for the Lasso estimate in the context of the high-dimensional linear regression model without endogeneity, i.e., $\sqrt{\frac{k_2 \log p}{n}}$. An example of the second case where p is large relative to d is when the first-stage estimation concerns regressions in low-dimensional settings and the result for this specific example is formally stated in Corollary 2.3.4 below.

Corollary 2.3.4 (First-stage estimation in low-dimensional settings): Suppose Assumptions 2.1.1, 2.3.2, and 2.3.3 hold. Assume the number of regressors $p(= p_n)$ in (1) can grow with and exceed the sample size n; the number of non-zero components in β^* is at most k_2, which is allowed to increase to infinity with n but slowly compared to n; also $d = \max_{j=1,...,p} d_j \ll n$ and does *not* grow with n. Suppose that the first-stage estimator $\hat{\pi}$ satisfies the bound $\max_{j=1,...,p} |\hat{\pi}_j - \pi_j^*|_2 \precsim \sqrt{\frac{\log p}{n}}$ with probability at least $1 - O(\frac{1}{\max(p, n)})$. Then, if

$$\frac{k_2 \log p}{n} = O(1),$$

and the tuning parameter λ_n satisfies

$$\lambda_n \asymp k_2 \sqrt{\frac{\log p}{n}},$$

we have

$$|\hat{\beta}_{H2SLS} - \beta^*|_2 \precsim \max\{\varphi_1 \sqrt{k_2} \sqrt{\frac{\log p}{n}}, \varphi_2 \sqrt{\frac{k_2 \log p}{n}}\},$$

with probability at least $1 - O(\frac{1}{\max(p, n)})$, where φ_1 and φ_2 are defined in Theorem 2.3.2. If we also have $k_2 \sqrt{\frac{k_2 \log p}{n}} = o(1)$, then the two-stage estimator $\hat{\beta}_{H2SLS}$ is l_2-consistent for β^*.

Note that Corollary 2.3.4 is a special case of Theorem 2.3.3 and hence the result is obvious from Theorem 2.3.3.

Under the condition that the first-stage estimators correctly select the non-zero coefficients with probability close to 1, we can also compare the high-dimensional two-stage estimator $\hat{\beta}_{H2SLS}$ with another type of multi-stage procedure. These multi-stage procedures include three steps. In the first step, one carries out the same first-stage estimation as before such as applying the Lasso or Dantzig selector. Under some stringent conditions that guarantee the selection-consistency of these first-stage estimators (such as the "irrepresentable condition" or the "mutual incoherence condition" described earlier), we can recover the supports of the true parameter vectors with high probability. In the second step, we apply OLS with the regressors in the estimated support set to obtain $\hat{\pi}_j^{OLS}$ for $j = 1, ..., p$. In the third step, we apply a Lasso technique to the main equation with these fitted regressors based on the second-stage OLS estimates. This type of procedure is in the similar spirit as the literature on sparsity in high-dimensional linear models without endogeneity (see, e.g., Candès and Tao, 2007; Belloni and Chernozhukov, 2013).

Under this three-stage procedure, Corollary 2.3.4 above tells us that the statistical error of the parameters of interests in the main equation is of the order $O\left(|\beta^*|_1\sqrt{\frac{k_2 \log p}{n}}\right)$, which is at least as good as $\hat{\beta}_{H2SLS}$. Nevertheless, this improved statistical error is at the expense of imposing stringent conditions that ensure the first-stage estimators to achieve selection consistency. These assumptions only hold in a rather narrow range of problems, excluding many cases where the design matrices exhibit strong (empirical) correlations. If these stringent conditions in fact do not hold, then the three-stage procedure may not work. On the other hand, even in the absence of the selection-consistency in the first-stage estimation, $\hat{\beta}_{H2SLS}$ is still a valid procedure and the bound as well as the consistency result in Theorem 2.3.2 still hold. Therefore, $\hat{\beta}_{H2SLS}$ may be more appealing in the sense that it works for a broader range of problems in which the first-stage design matrices (formed by the instruments) $Z_j \in \mathbb{R}^{n \times d_j}$ for $j = 1, ..., p$ exhibit a relatively high amount of dependency among the covariates.

For Theorems 2.3.2 and 2.3.3, the results are derived for the case where each of the first-stage equations is estimated separately with a Lasso-type procedure. Depending on the specific structures of the first-stage equations, other methods that take into account the interrelationships between these equations might yield a smaller first-stage estimation error and consequently a potential improvement on the l_2-error of $\hat{\beta}_{H2SLS}$. This chapter does not pursue these more efficient first-stage estimators but rather considers the extension of Theorem 2.3.2 in the following manner. Notice that for Theorem 2.3.2, we give an explicit form of the first-stage estimation error in Assumption 2.3.5 and as discussed earlier, Lasso type of techniques yield this estimation error. However, the estimation error of the parameters of interests in the main equation can be bounded by the maximum of a term involving the first-stage related error and a term involving the second-stage related error, which holds for general first-stage estimation errors. This claim is formally stated in Theorem 2.3.5 below.

Theorem 2.3.5: Suppose Assumptions 2.1.1 and 2.3.1-2.3.3 hold. Also, assume the first-stage estimator $\hat{\pi}$ satisfies the bound

$$\max_{j=1,\ldots,p} |\hat{\pi}_j - \pi_j^*|_2 \leq M(d,\, p,\, k_1,\, n)$$

with probability $1-\alpha$. Then, if

$$\max\left\{k_2^2 M^2(d,\, p,\, k_1,\, n),\; \frac{k_2 \log p}{n}\right\} = O(1),$$

and the tuning parameter λ_n satisfies

$$\lambda_n \asymp k_2 \max\left\{M(d,\, p,\, k_1,\, n),\; \sqrt{\frac{\log p}{n}}\right\},$$

we have

$$|\hat{\beta}_{H2SLS} - \beta^*|_2 \precsim \max\{\varphi_1 \sqrt{k_2} M(d,\, p,\, k_1,\, n),\, \varphi_2 \sqrt{\frac{k_2 \log p}{n}}\},$$

where

$$\begin{aligned}
\varphi_1 &= \frac{\max\{\sigma_\eta,\, \sigma_{X^*},\, \sigma_\epsilon\} \sqrt{\lambda_{\max}(\Sigma_Z)} |\beta^*|_1}{\lambda_{\min}(\Sigma_{X^*})}, \\
\varphi_2 &= \max\left\{\frac{\sigma_{X^*}\sigma_\eta |\beta^*|_1}{\lambda_{\min}(\Sigma_{X^*})},\, \frac{\sigma_{X^*}\sigma_\epsilon}{\lambda_{\min}(\Sigma_{X^*})}\right\},
\end{aligned}$$

with probability at least $1-\alpha-c_1\exp(-c_2 \log \max(p,\, n))$ for some universal positive constants c_1 and c_2. If we also have $k_2 \max\left\{\sqrt{k_2} M(d,\, p,\, k_1,\, n),\, \sqrt{\frac{k_2 \log p}{n}}\right\} = o(1)$, then the two-stage estimator $\hat{\beta}_{H2SLS}$ is l_2-consistent for β^*.

2.3.2 Variable-selection consistency

This section addresses the following question: given an optimal two-stage Lasso solution $\hat{\beta}_{H2SLS}$, when do we have $\mathbb{P}[J(\hat{\beta}_{H2SLS}) = J(\beta^*)] \to 1$? The property $\mathbb{P}[J(\hat{\beta}_{H2SLS}) = J(\beta^*)] \to 1$ is referred to as *variable-selection consistency*. For consistent variable selection with the standard Lasso in the context of linear models without endogeneity, it is known that the so-called "neighborhood stability condition" (Meinshausen and

Bühlmann, 2006) for the design matrix, re-formulated in a nicer form as the "irrepresentable condition" by Zhao and Yu, 2006, is sufficient and necessary. A further refined analysis is given in Wainwright (2009), which presents under a certain "incoherence condition" the smallest sample size needed to recover a sparse signal. In this chapter, I adopt the analysis by Wainwright (2009), Ravikumar, Wainwright, and Lafferty (2010), and Wainwright (2015) to analyze the selection consistency of $\hat{\beta}_{H2SLS}$. In particular, I need the following assumption.

Assumption 2.3.7: $\left\| \mathbb{E}\left[X^{*T}_{1,J(\beta^*)^c} X^*_{1,J(\beta^*)}\right] \left[\mathbb{E}(X^{*T}_{1,J(\beta^*)} X^*_{1,J(\beta^*)})\right]^{-1} \right\|_\infty \leq 1-\phi$ for some $\phi \in (0,\, 1]$.

Remarks
Assumption 2.3.7, the so-called "mutual incoherence condition" originally formalized by Wainwright (2009), captures the intuition that the large number of irrelevant covariates cannot exert an overly strong effect on the subset of relevant covariates. In the most desirable case, the columns indexed by $j \in J(\beta^*)^c$ would all be orthogonal to the columns indexed by $j \in J(\beta^*)$ and then we would have $\phi = 1$. In the high-dimensional setting, this perfect orthogonality is not possible, but one can still hope for a type of "near orthogonality" to hold.

Assumptions 2.1.1 and 2.3.3 ensure that the left-hand-side of the inequality in Assumption 2.3.7 falls in $[0,\, 1)$. Under Assumptions 2.1.1 and 2.3.3, we know that each column X^*_j, $j = 1, ..., p$ is consisted of *i.i.d.* sub-Gaussian variables. Without loss of generality, we can assume $\mathbb{E}(X^*_{1j}) = 0$ for all $j = 1, ..., p$. Consequently, from a standard bound for the norms of zero-mean sub-Gaussian vectors and a union bound,

$$\mathbb{P}\left[\max_{j=1,\ldots,p} \frac{|X^*_j|_2}{\sqrt{n}} \leq \kappa_c\right] \geq 1 - 2\exp(-cn + \log p) \geq 1 - 2\exp(-c^{'}n),$$

where the last inequality follows from $n \gg \log p$. For example, if X^* has a Gaussian design, then we have

$$\max_{j=1,\ldots,p} \frac{|X^*_j|_2}{\sqrt{n}} \leq \max_{j=1,\ldots,p} \Sigma_{jj} \left(1 + \sqrt{\frac{32 \log p}{n}}\right),$$

where $\max_{j=1,..,p} \Sigma_{jj}$ corresponds to the maximal variance of any element of X^*(see Raskutti, et. al, 2011).

Theorem 2.3.6 (Selection consistency): Suppose Assumptions 2.1.1, 2.3.1-2.3.3, 2.3.5, and 2.3.7 hold. If

$$\begin{aligned}\frac{1}{n}k_2^3\log p &= O(1),\\ \frac{1}{n}k_1k_2^2\log\max(d,\,p) &= o(1),\end{aligned}$$

and the tuning parameter λ_n satisfies

$$\lambda_n \asymp k_2\sqrt{\frac{k_1\log\max(d,\,p)}{n}},$$

then, we have: (a) The Lasso has a unique optimal solution $\hat{\beta}_{H2SLS}$, (b) the support $J(\hat{\beta}_{H2SLS}) \subseteq J(\beta^*)$,
(c) $\quad |\hat{\beta}_{H2SLS,\,J(\beta^*)} - \beta^*_{H2SLS,\,J(\beta^*)}|_\infty \leq c\max\left\{\varphi_1\sqrt{\frac{k_1k_2\log\max(d,p)}{n}},\ \varphi_2\sqrt{\frac{k_2\log p}{n}}\right\} := B$

where

$$\begin{aligned}\varphi_1 &= \frac{\max\left\{\sigma_\eta,\,\sigma_{X^*},\,\sigma_\epsilon\right\}\sigma_\eta\sigma_Z\sqrt{\lambda_{\max}(\Sigma_Z)}|\beta^*|_1}{\lambda_{\min}(\Sigma_Z)\lambda_{\min}\left(\mathbb{E}\left[X^{*T}_{1,J(\beta^*)}X^*_{1,J(\beta^*)}\right]\right)},\\ \varphi_2 &= \max\left\{\frac{\sigma_{X^*}\sigma_\eta|\beta^*|_1}{\lambda_{\min}\left(\mathbb{E}\left[X^{*T}_{1,J(\beta^*)}X^*_{1,J(\beta^*)}\right]\right)},\ \frac{\sigma_{X^*}\sigma_\epsilon}{\lambda_{\min}\left(\mathbb{E}\left[X^{*T}_{1,J(\beta^*)}X^*_{1,J(\beta^*)}\right]\right)}\right\},\end{aligned}$$

with probability at least $1 - c_1\exp(-c_2\log\max(\min(p,\,d),\,n))$, (d) if $\min_{j\in J(\beta^*)}|\beta^*_j| > B$, then $J(\hat{\beta}_{H2SLS}) \supseteq J(\beta^*)$ and hence $\hat{\beta}_{H2SLS}$ is variable-selection consistent, i.e., $J(\hat{\beta}_{H2SLS}) = J(\beta^*)$.

The proof for Theorem 2.3.6 is provided in Section 2.6.6. The proof for Theorems 2.3.6 hinges on an intermediate result that shows the "mutual incoherence" assumption on $\mathbb{E}[X_1^{*T}X_1^*]$ (the population version of $\frac{1}{n}X^{*T}X^*$) guarantees that, with high probability, analogous conditions hold for the estimated quantity $\frac{1}{n}\hat{X}^T\hat{X}$, formed by the fitted regressors from the first-stage regression. This result is established in Lemma 2.6.4 in Section 2.6.5.

Theorem 2.3.6 includes four parts. Part (a) guarantees the uniqueness of the optimal solution of the two-stage Lasso procedure, $\hat{\beta}_{H2SLS}$. Based on this uniqueness claim, one can then talk unambiguously about the support of the two-stage Lasso estimate. Part (b) guarantees that the Lasso does not falsely include elements that are not in the support of β^*. Part (c) ensures that $\hat{\beta}_{H2SLS,\,J(\beta^*)}$ is uniformly close to $\beta^*_{J(\beta^*)}$ in the

$l_\infty-$norm. The last claim is a consequence of this uniform norm bound: as long as the minimum value of $|\beta_j^*|$ over $j \in J(\beta^*)$ is not too small, then the two-stage Lasso does not falsely exclude elements that are in the support of β^* with high probability. The minimum value requirement of $|\beta_j^*|$ over $j \in J(\beta^*)$ is comparable to the so-called "beta-min" condition in Bühlmann and van de Geer (2011). Combining the claims from (b) and (d), the two-stage Lasso is variable-selection consistent with high probability.

2.4 Simulations

In this section, simulations are conducted to gain insight on the finite sample performance of the regularized two-stage estimators. I consider the triangular simultaneous equations model (2.1) and (2.2) from Section 2.1 where $d_j = d$ for all $j = 1, ..., p$, $(y_i, \mathbf{x}_i^T, \mathbf{z}_i^T, \epsilon_i, \boldsymbol{\eta}_i)$ are *i.i.d.*, and $(\epsilon_i, \boldsymbol{\eta}_i)$ have the following joint normal distribution

$$(\epsilon_i, \boldsymbol{\eta}_i) \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\epsilon^2 & \rho\sigma_\epsilon\sigma_\eta & \cdots & \cdots & \rho\sigma_\epsilon\sigma_\eta \\ \rho\sigma_\epsilon\sigma_\eta & \sigma_\eta^2 & 0 & \cdots & 0 \\ \vdots & 0 & \sigma_\eta^2 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \rho\sigma_\epsilon\sigma_\eta & 0 & \cdots & 0 & \sigma_\eta^2 \end{pmatrix}\right).$$

The matrix $\mathbf{z}_i^T$ is a $p \times d$ matrix of normal random variables with identical variances σ_z, and $\mathbf{z}_{ij}^T$ is independent of $(\epsilon_i, \eta_{i1}, ..., \eta_{ip})$ for all $j = 1, ..., p$. With this setup, I simulate 1000 sets of $(y_i, \mathbf{x}_i^T, \mathbf{z}_i^T, \epsilon_i, \boldsymbol{\eta}_i)_{i=1}^n$ where n is the sample size (i.e., the number of data points) in each set, and perform 14 Monte Carlo simulation experiments constructed from various combinations of model parameters (d, k_1, p, k_2, β^*, σ_ϵ, and σ_η), the design of $\mathbf{z}_i$, the random matrix formed by the instrumental variables, as well as the types of first-stage and second-stage estimators employed (Lasso vs. OLS). For each replication $t = 1, ..., 1000$, I compute the estimates $\hat{\beta}^t$ of the main-equation parameters β^*, l_2-errors of these estimates, $|\hat{\beta}^t - \beta^*|_2$, and selection percentages of $\hat{\beta}^t$ (computed by the number of the elements in $\hat{\beta}^t$ sharing the same sign as their corresponding elements in β^*, divided by the total number of elements in β^*). Table 4.1 displays the designs of the 14 experiments. For Experiment 1 and Experiments 3-14, I set the number of parameters in each first-stage equation $d = 100$, the number of parameters in the main equation $p = 50$, the number of non-zero parameters in each first-stage equation $k_1 = 4$, the number of non-zero parameters in the main equation $k_2 = 5$. Also, choose $(\pi_{j1}^*, ..., \pi_{j4}^*) = \mathbf{1}$, $(\pi_{j5}^*, ..., \pi_{j100}^*) = \mathbf{0}$ for all $j = 1, ..., 50$; and $(\beta_1^*, ..., \beta_5^*) = \mathbf{1}$, $(\beta_6^*, ..., \beta_{50}^*) = \mathbf{0}$. For convenience, in the following discussion, I will refer to those non-zero parameters as "relevant" parameters and those zero parameters as "irrelevant" parameters. Experiment 2 sets $d = 4$, $p = 5$, $(\pi^*_{j1}, ..., \pi^*_{j4}) = \mathbf{1}$, and $(\beta^*_1, ..., \beta^*_5) = \mathbf{1}$. The motivations of these experiments are explained in the following discussion.

Table 4.1: Designs of the Monte-Carlo simulation experiments, 1000 replications

Experiment #	1	2	3	4	5	6	7	8	9	10	11	12	13	14
d	100	4	NA	100	100	100	100	100	100	100	100	100	100	100
k_1	4	4	NA	4	4	4	4	4	4	4	4	4	4	4
p	50	5	50	50	50	50	50	50	50	50	50	50	50	50
k_2	5	5	5	5	5	5	5	5	5	5	5	5	5	5
$(\beta_1, ..., \beta_5)$	$\mathbf{1}$	$\mathbf{1}$	$\mathbf{1}$	$\mathbf{1}$	$\mathbf{1}$	$\mathbf{1}$	$\mathbf{1}$	$\mathbf{1}$	$\mathbf{1}$	$\mathbf{1}$	$\mathbf{1}$	$\mathbf{1}$	$\mathbf{1}$	**0.01**
$(\beta_6, ..., \beta_{50})$	$\mathbf{0}$	NA	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{0}$
$(\pi_{j1}, ..., \pi_{j4})$	$\mathbf{1}$	$\mathbf{1}$	NA	$\mathbf{1}$	$\mathbf{1}$	$\mathbf{1}$	$\mathbf{1}$	$\mathbf{1}$	$\mathbf{1}$	$\mathbf{1}$	$\mathbf{1}$	$\mathbf{1}$	$\mathbf{1}$	$\mathbf{1}$
$(\pi_{j5}, ..., \pi_{j100})$	$\mathbf{0}$	NA	NA	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{0}$
σ_ϵ	0.4	0.4	0.4	0.4	0.4	0.4	1	0.4	0.4	0.4	1	0.4	0.4	0.4
σ_η	0.4	0.4	NA	0.4	0.4	0.4	0.4	1	0.4	0.4	0.4	1	0.4	0.4
σ_z	1	1	NA	1	1	1	1	1	0.4	1	1	1	0.4	1
Row corr. in $\mathbf{z}_i^T$	No	No	NA	No	No	No	No	No	No	Yes	Yes	Yes	Yes	No
1st-stg. est.	Las	OLS	NA	OLS	Las	OLS	Las	Las	Las	Las	Las	Las	Las	Las
2nd-stg. est.	Las	OLS	Las	Las	OLS	OLS	Las	Las	Las	Las	Las	Las	Las	Las

The baseline experiment (Experiment 1) applies the two-stage Lasso procedure to the endogenous sparse linear model with a triangular simultaneous equations structure (2.1) and (2.2). For each data point $i = 1, ..., n$, the instruments $\mathbf{z}_i^T$ is a $p \times d$ matrix of independent standard normal random variables. As a benchmark for Experiment 1, Experiment 2 concerns the classical 2SLS procedure when both stage equations are in the low-dimensional setting and the supports of the true parameters in both stages are known *a priori*. As another benchmark for Experiment 1, Experiment 3 applies a one-step Lasso procedure (without instrumenting the endogenous regressors) to the same main equation model (2.1) as in Experiment 1.

Experiments 4-6 concern, in a relatively large sample size setting with sparsity, the performance of alternative "partially" regularized or non-regularized estimators: first-stage-OLS-second-stage-Lasso (Experiment 4), first-stage-Lasso-second-stage-OLS (Experiment 5), and first-stage-OLS-second-stage-OLS (Experiment 6). Experiments 7-14 return to the two-stage Lasso procedure with changes applied to the model parameters that generate the data. Experiment 7 (Experiment 8) increases the standard deviation of the "noise" in the main equation, σ_ϵ (respectively, the standard deviation of the "noise" in the first-stage equations, σ_η); Experiment 9 reduces σ_z, the standard deviation of the "signal", i.e., the instrumental variables; Experiment 10 introduces correlations between the rows of the design matrix $\mathbf{z}_i^T$. Notice that each row of $\mathbf{z}_i^T \in \mathbb{R}^{p\times d}$ is associated with each of the endogenous regressors and the row-wise correlation in $\mathbf{z}_i^T$ hence introduces correlations between the "purged" regressors X_j^* and $X_{j'}^*$ for all $j \neq j'$. The level of the correlation is set to 0.5, i.e., $\text{corr}(z_{ijl},\, z_{ij'l}) = 0.5$ for $j \neq j'$ and $l = 1, ..., d$ (notice that we still have $\text{corr}(z_{ijl},\, z_{ijl'}) = 0$ for $l \neq l'$ and $j = 1, ..., p$; i.e., there is no column-wise correlation in $\mathbf{z}_i^T$). Experiment 11 (Experiment 12) increases the "noise" level in the main equation (respectively, the "noise" level in the first-stage equations) and introduces the correlations between the "purged" regressors X_j^* and $X_{j'}^*$ for all $j \neq j'$ simultaneously. Experiment 13 reduces the "signal" level of the instrumental variables and introduces the correlations between the "purged" regressors X_j^* and $X_{j'}^*$ for all $j \neq j'$ simultaneously. Experiment 14 reduces the magnitude of $(\beta_1^*, \ldots, \beta_5^*)$ from $(1, \ldots, 1)$ to $(0.01, \ldots, 0.01)$.

The tuning parameters λ_{1n} in the first-stage Lasso estimation (in Experiments 1, 5, 7-14) are chosen according to the standard Lasso theory of high-dimensional estimation techniques (e.g., Bickel, 2009); in particular, $\lambda_{1n} = 0.4\sqrt{\frac{\log d}{n}}$. The tuning parameters λ_{2n} in the second-stage Lasso estimation (in Experiments 1, 3, 4, 7-14) are chosen according to the scaling condition in Theorem 2.3.3; in particular, $\lambda_{2n} = 0.1 \cdot k_2 \max\left\{\sqrt{\frac{k_1 \log d}{n}},\, \sqrt{\frac{\log p}{n}}\right\}$ in Experiments 1, 3, 4, 7-13 and $\lambda_{2n} =$

$0.001 \cdot k_2 \max\left\{\sqrt{\frac{k_1 \log d}{n}}, \sqrt{\frac{\log p}{n}}\right\}$ in Experiment 14. The value of λ_{2n} in Experiments 1, 3, 4, 7-13 exceeds the value of λ_{2n} in Experiments 14 by a factor of 0.01. This adjustment reflects the fact that the non-zero parameters $(\beta_1, ..., \beta_5) = (1, \ldots, 1)$ in Experiments 1, 3, 4, 7-13 exceed the non-zero parameters $(\beta_1, ..., \beta_5) = (0.01, \ldots, 0.01)$ in Experiment 14 by a factor of 0.01.

Figure 4.1a plots (in ascending values) the 1000 estimates of β_5^* when the sample size $n = 47$. The estimates of other "relevant" main-equation parameters behave similarly as the estimates of β_5^*. Figure 4.1b plots (in ascending values) the 1000 estimates of β_6^* when the sample size $n = 47$. The estimates of other "irrelevant" main-equation parameters behave similarly as the estimates of β_6^*. The sample size 47 satisfies the scaling condition in Theorem 2.3.3. With the choice of $d = 100$, $k_1 = 4$, $p = 50$, $k_2 = 5$ in Experiments 1 and 3, the sample size $n = 47$ represents a high-dimensional setting with sparsity. Figure 4.1c (Figure 4.1d) is similar to Figure 4.1a (Figure 4.1b) except that the sample size $n = 4700$.

With the 1000 estimates of the main-equation parameters from Experiments 1-3, Table 4.2 shows the mean of the l_2-errors of these estimates (computed as $\frac{1}{1000}\sum_{t=1}^{1000}|\hat{\beta}^t - \beta^*|_2$), the mean of the selection percentages (computed in a similar fashion as the mean of the l_2-errors of the estimates of β^*), the mean of the squared l_2-errors (i.e., the *sample mean squared error*, SMSE, computed as $\frac{1}{1000}\sum_{t=1}^{1000}|\hat{\beta}^t - \beta^*|_2^2$), and the sample squared bias $\sum_{j=1}^{50}(\bar{\hat{\beta}}_j - \beta_j^*)^2$ (where $\bar{\hat{\beta}}_j = \frac{1}{1000}\sum_{t=1}^{1000}\hat{\beta}_j^t$ for $j = 1, ..., 50$). To provide a sense of how well the first-stage estimates behave, Table 4.2 also displays the "averaged" mean of the l_2-errors of the first-stage estimates (computed as $\frac{1}{50}\sum_{j=1}^{50}\frac{1}{1000}\sum_{t=1}^{1000}|\hat{\pi}_j^t - \pi_j^*|_2$), the "averaged" mean of the selection percentages of the first-stage estimates (computed in a similar fashion as the "averaged" mean of the l_2-errors of the first-stage estimates), the "averaged" mean of the squared l_2-errors (i.e., the "averaged" SMSE, computed as $\frac{1}{50}\sum_{j=1}^{50}\frac{1}{1000}\sum_{t=1}^{1000}|\hat{\pi}_j^t - \pi_j^*|_2^2$), and the "averaged" sample squared bias $\frac{1}{50}\sum_{j=1}^{50}\sum_{l=1}^{100}(\bar{\hat{\pi}}_{jl} - \pi_{jl}^*)^2$ (where $\bar{\hat{\pi}}_{jl} = \frac{1}{1000}\sum_{t=1}^{1000}\hat{\pi}_{jl}^t$ for $j = 1, ..., 50$ and $l = 1, ..., 100$).

Compared to the two-stage Lasso procedure, in estimating the "relevant" main-equation parameters with both sample sizes $n = 47$ and $n = 4700$, Figures 4.1a and 4.1c show that the classical 2SLS procedure where the supports of the true parameters in both stages are known *a priori* produces larger estimates while the one-step Lasso procedure (without instrumenting the endogenous regressors) produces smaller estimates. The two-stage Lasso outperforms the classical 2SLS above the 60^{th} percentile of the estimates while underestimates the "relevant" main-equation parameters below the 60^{th} percentile relative to the classical 2SLS procedure. The one-step Lasso procedure (without instrumenting the endogenous regressors) produces the poorest estimates of the "relevant" main-equation parameters. The mean $\bar{\hat{\beta}}_5$ of the 1000 estimates $\hat{\beta}_5$ from the two-stage Lasso is 0.931 (respectively, 1.000 from the classical 2SLS and 0.826 from the one-step Lasso) when $n = 47$ and 0.997 (respectively, 1.000 from the classical 2SLS and 0.988 from the one-step Lasso) when $n = 4700$. The fact that the two-stage Lasso yields smaller estimates of the "relevant" main-equation parameters relative to the classical 2SLS for both sample sizes is most likely due to the shrinkage effect from the l_1-penalization in the second-stage estimation of the two-stage Lasso procedure.

In estimating the "irrelevant" main-equation parameters, the estimates of $(\beta_6^*, ..., \beta_{50}^*)$ from both the two-stage Lasso and the one-step Lasso are exactly 0 at the 5^{th} percentile, the median, and the 95^{th}percentile when $n = 47$ and $n = 4700$. The mean statistics of the estimates of $(\beta_6^*, ..., \beta_{50}^*)$ range from -0.001 (-2.938×10^{-4}) to 0.001 (3.403×10^{-4}) when $n = 47$, and -6.926×10^{-5} (0) to 5.593×10^{-5} (3.076×10^{-6}) when $n = 4700$ for the two-stage Lasso (respectively, the one-step Lasso). Table 4.2 shows that the selection percentages of the main-equation estimates from the two-stage Lasso and the one-step Lasso are high for the designs considered. Figures 4.1b and 4.1d show that, in estimating the "irrelevant" main-equation parameters, the one-step Lasso performs slightly better relative to the two-stage Lasso procedure below the 2^{nd} percentile and above the 98^{th} percentile.

In terms of estimation errors and sample bias, from Table 4.2 we see that the mean of the l_2-errors of the estimates $\hat{\beta}_{H2SLS}$ of β^* (or the "averaged" mean of the l_2-errors of the first-stage estimates) from the two-stage Lasso are greater than those of $\hat{\beta}_{2SLS}$ (respectively, of the first-stage estimates) from the classical 2SLS procedure for both $n = 47$ and $n = 4700$. As n increases, the mean of the l_2-errors of $\hat{\beta}_{H2SLS}$ and the mean of the l_2-errors of $\hat{\beta}_{2SLS}$ become very close to each other as in the case when $n = 4700$. Also, the sample bias of $\hat{\beta}_{H2SLS}$ (or, the "averaged" sample bias of the first-stage estimates) from the two-stage Lasso are greater by a magnitude of $100 \sim 1000$ (respectively, $1000 \sim 10^4$) than those of $\hat{\beta}_{2SLS}$ (respectively, of the first-stage estimates) from the classical 2SLS procedure for both sample sizes.

For more investigation on how the l_2-error and sample bias of $\hat{\beta}_{H2SLS}$ compare to those of $\hat{\beta}_{2SLS}$, I have also considered designs where σ_ϵ and/or σ_η are increased or decreased while everything else in Experiments 1 and 2 remains the same. In these modified designs except for those with very large values of σ_ϵ under $n = 47$, the mean of the l_2-errors of $\hat{\beta}_{H2SLS}$ are generally greater than those of $\hat{\beta}_{2SLS}$. The sample bias of $\hat{\beta}_{H2SLS}$ are consistently greater by a magnitude of $10 \sim 10^5$ than those of $\hat{\beta}_{2SLS}$ for both sample sizes. This suggests that the shrinkage effect from the l_1-penalization in both the first and second stage estimations of the two-stage Lasso procedure might have made its bias term converge to zero at a slower rate relative to the classical 2SLS for the designs considered here. Whether this conjecture holds true for general designs is an interesting question for further research. Compared to the two-stage Lasso and the classical 2SLS, the one-step Lasso procedure without instrumenting the endogenous regressors yields the largest l_2-errors as well as sample bias of the main-equation estimates for both $n = 47$ and $n = 4700$, which is expected. Finally notice that the l_2-errors (and the sample bias) shrink as the sample size increases. $|\hat{\hat{\beta}}_{2SLS} - \beta^*|_2$ being proportional to $\frac{1}{\sqrt{n}}$ is a known fact in low-dimensional settings. From Section 2.3.1, we also have that the upper bounds for $|\hat{\beta}_{H2SLS} - \beta^*|_2$ are proportional to $\frac{1}{\sqrt{n}}$ up to factors involving $\log d$, $\log p$, k_1, and k_2.

Figure 4.1a: Estimates of the fifth parameter (n = 47)

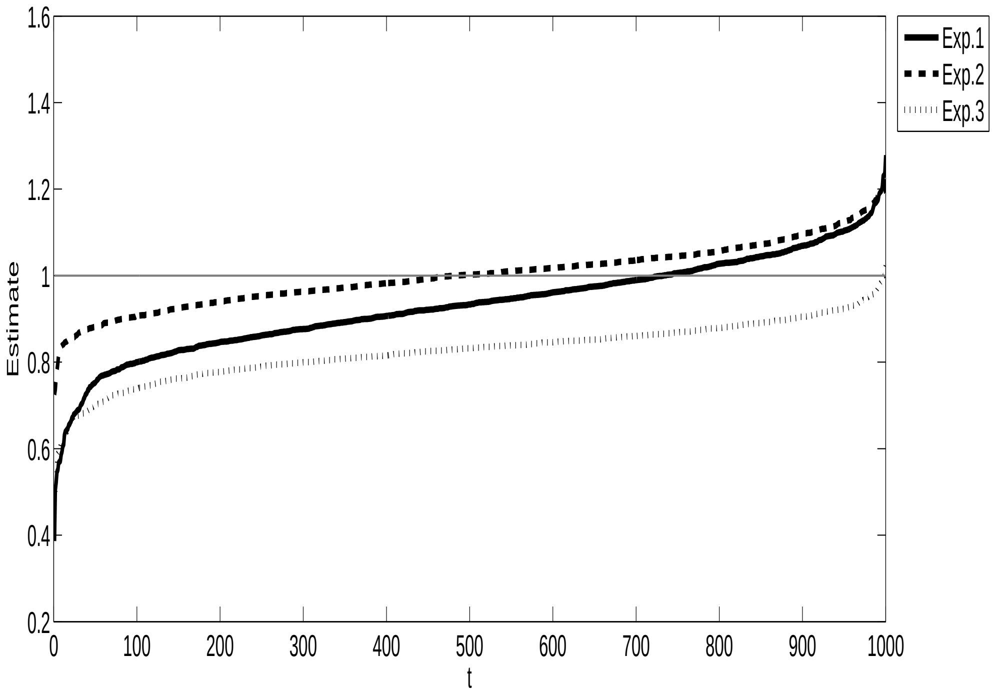


Figure 4.1b: Estimates of the sixth parameter (n = 47)

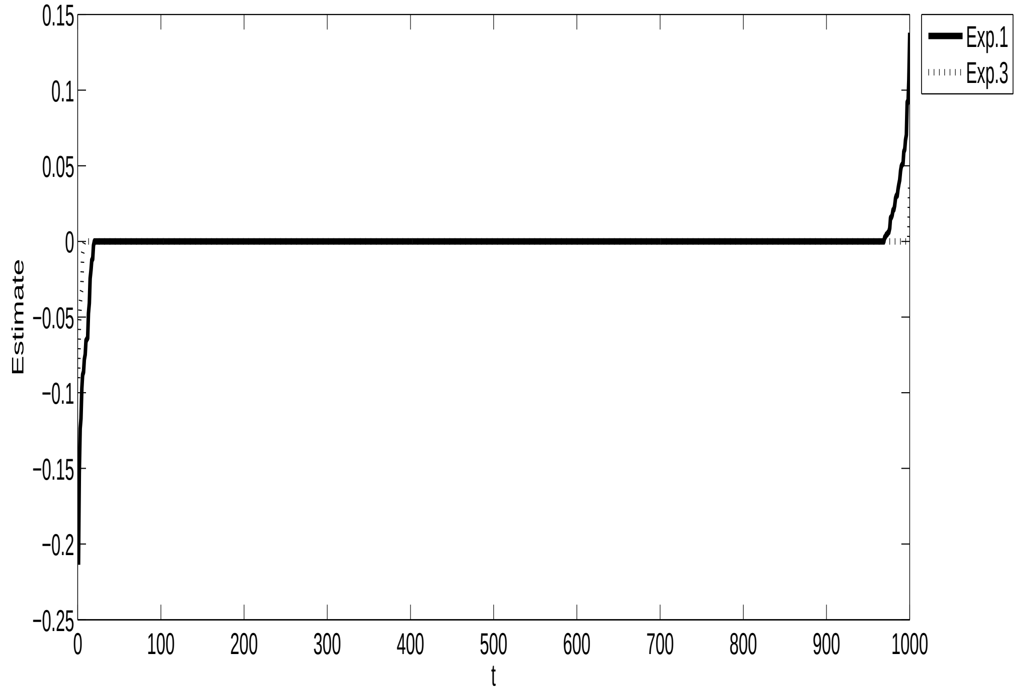
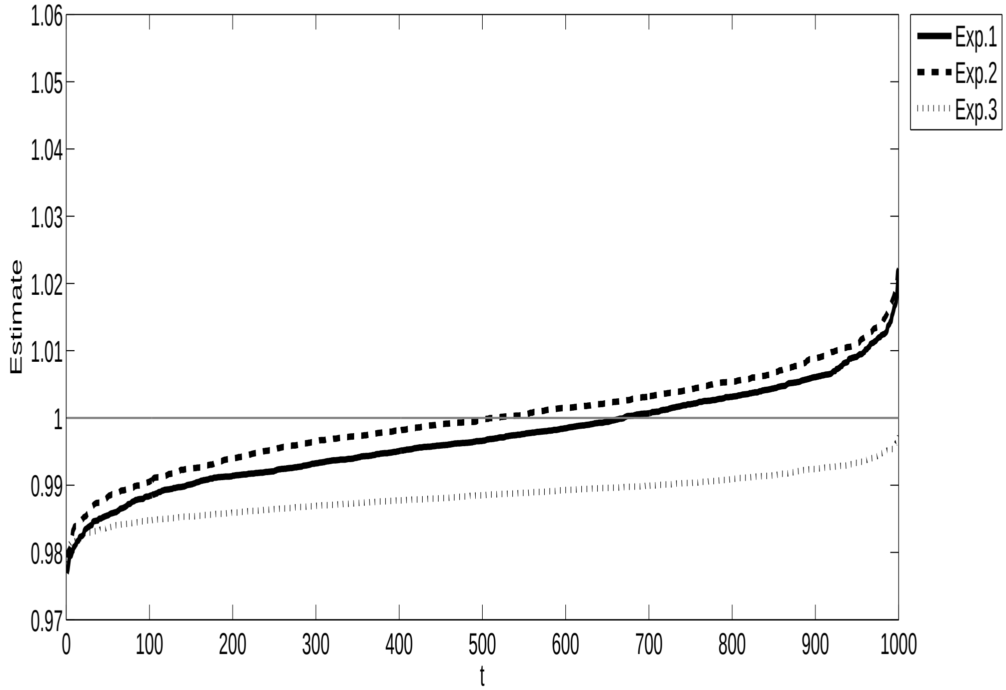


Figure 4.1c: Estimates of the fifth parameter (n = 4700)



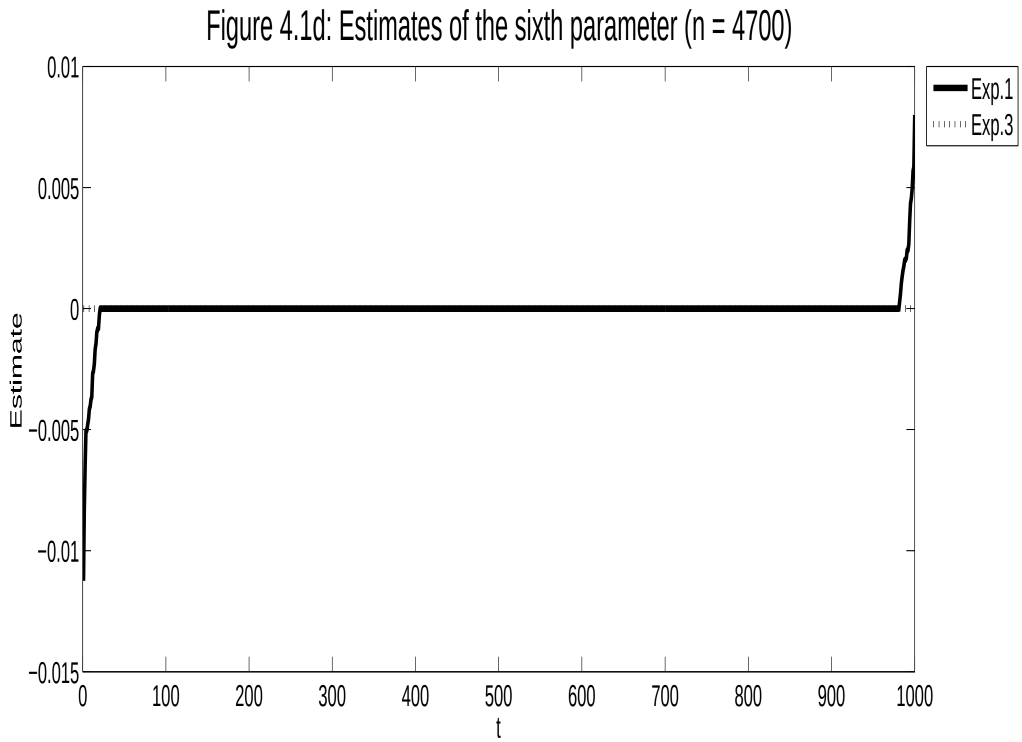


Table 4.2: l_2-errors, SMSE, bias, and selection (Exp. 1-3)

Mean	$n = 47$			$n = 4700$		
Exp. #	1	2	3	1	2	3
$2^{nd}-$stage select %	97.3	NA	99.5	98.1	NA	100
$2^{nd}-$stage l_2-error	0.288	0.156	0.412	0.018	0.015	0.026
$2^{nd}-$stage SMSE	0.099	0.028	0.179	3.38×10^{-4}	2.43×10^{-4}	7.07×10^{-4}
$2^{nd}-$stage squared bias	0.024	1.49×10^{-5}	0.154	5.04×10^{-5}	1.84×10^{-7}	6.66×10^{-4}
1^{st}-stage select %	0.977	NA	NA	0.985	NA	NA
$1^{st}-$stage l_2-error	0.349	0.115	NA	0.028	0.011	NA
$1^{st}-$stage SMSE	0.132	0.003	NA	7.92×10^{-4}	2.78×10^{-5}	NA
$1^{st}-$stage squared bias	0.097	1.07×10^{-5}	NA	6.31×10^{-4}	1.23×10^{-7}	NA

In the following relatively large sample size setting (i.e., $n = 4700$) with sparsity, I compare the performance of the two-stage Lasso estimator with the performances of the alternative "partially" regularized or non-regularized estimators as mentioned earlier. Figure 4.2a plots (in ascending values) the 1000 estimates of β_5^* when the sample size $n = 4700$. Figure 4.2b plots (in ascending values) the 1000 estimates of β_6^* when the sample size $n = 4700$. With the 1000 estimates of the "relevant" ("irrelevant") main-equation parameters from Experiment 1 and Experiments 4-6, Figures 4.2c-4.2f (respectively, Figures 4.2g-4.2j) display the 5^{th} percentile, the median, the 95^{th} percentile, and the mean of these estimates. The mean of the l_2-errors and the mean of the selection percentages of the main-equation estimates together with the "averaged" mean of the l_2-errors and the "averaged" mean of the selection percentages of the first-stage estimates from these "partially" regularized or non-regularized estimators are displayed in Table 4.3.

Figure 4.2a and Figures 4.2c-4.2f show that, compared to the two-stage Lasso procedure, in estimating the "relevant" main-equation parameters when $n = 4700$, the first-stage-Lasso-second-stage-OLS estimator and the first-stage-OLS-second-stage-OLS estimator produce larger estimates while the first-stage-OLS-second-stage-Lasso estimator produces smaller estimates. In estimating the "irrelevant" main-equation parameters when $n = 4700$, Figure 4.2b and Figures 4.2g-4.2j show that the two-stage Lasso and the first-stage-OLS-second-stage-Lasso estimator perform well while the first-stage-Lasso-second-stage-OLS and the first-stage-OLS-second-stage-OLS do poorly (also see Table 4.3 for a comparison between the selection percentages of these estimators). This suggests that employing regularization in the second-stage estimation helps selecting the "relevant" main-equation parameters.

Turning to the comparison with "partially" regularized or non-regularized estimators in terms of l_2-errors, from Table 4.3 we see that the two-stage Lasso estimator achieves the smallest l_2-error of the main-equation estimates among all the estimators considered here. The fact that the l_2-error (of the main-equation estimates) of the two-stage Lasso estimator is smaller than the l_2-errors of the first-stage-OLS-second-stage-Lasso estimator and the first-stage-OLS-second-stage-OLS estimator could be attributed to the following. Based on the first-stage estimation results from these experiments, the first-stage Lasso estimator outperforms the first-stage OLS estimator in both estimation errors and variable selections even in the relatively large sample size setting with sparsity. Recall in Section 2.3, we have seen that, the estimation error of the parameters of interests in the main equation can be bounded by the maximum of a term involving the first-stage estimation error and a term involving the second-stage estimation error. Given the choices of p, d, k_1, and k_2 in Experiment 1 and Experiments 4-6, these results agree with the theorems in Section 2.3.1. Additionally, compared to the first-stage-OLS-second-stage-OLS estimator, the fact that the l_2-error (of the main-equation estimates) of the two-stage Lasso estimator is smaller than the l_2-error of the first-stage-OLS-second-stage-OLS estimator can also be explained by the fact that the two-stage Lasso reduces the l_2-error of the first-stage-OLS-second-stage-OLS estimates from $O\left(\sqrt{\frac{\max(p,d)}{n}}\right)$ to $O\left(\sqrt{\frac{\log\max(p,d)}{n}}\right)$, as we have seen in Section 2.3. Similarly, the fact that the l_2-error (of the main-equation estimates) of the two-stage Lasso estimator is smaller than the l_2-error of the first-stage-Lasso-second-stage-OLS estimator can be explained by the fact that the two-stage Lasso reduces the l_2-error of the first-stage-Lasso-second-stage-OLS estimates from $O\left(\sqrt{\frac{\max(p,\log d)}{n}}\right)$ to $O\left(\sqrt{\frac{\log\max(p,d)}{n}}\right)$.

Figure 4.2a: Estiamtes of the fifth parameter (n = 4700)

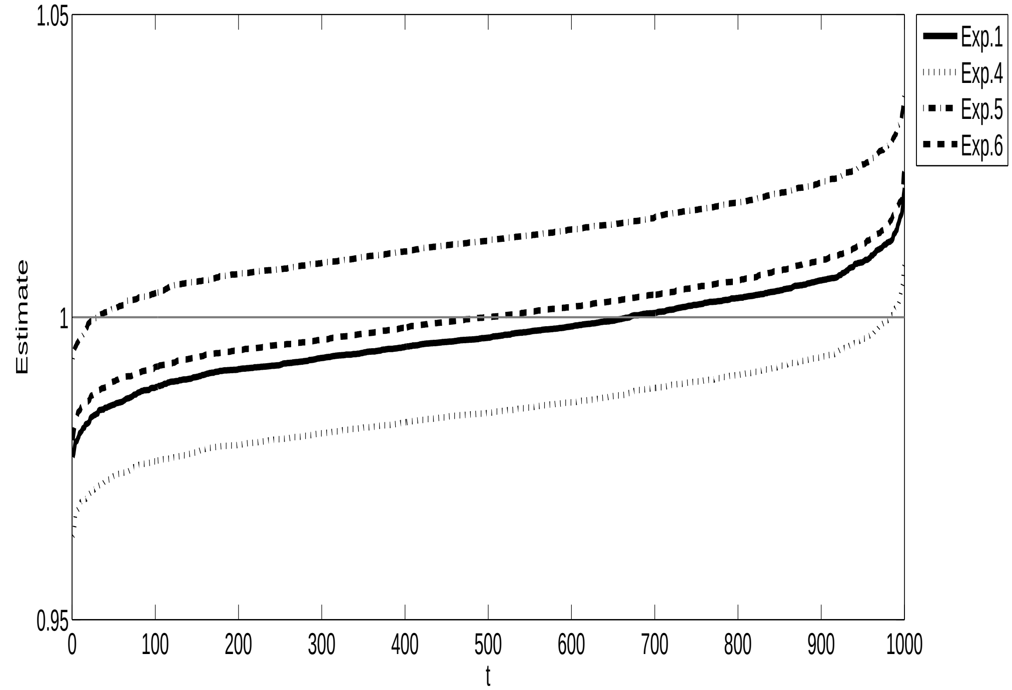


Figure 4.2b: Estimates of the sixth parameter (n = 4700)

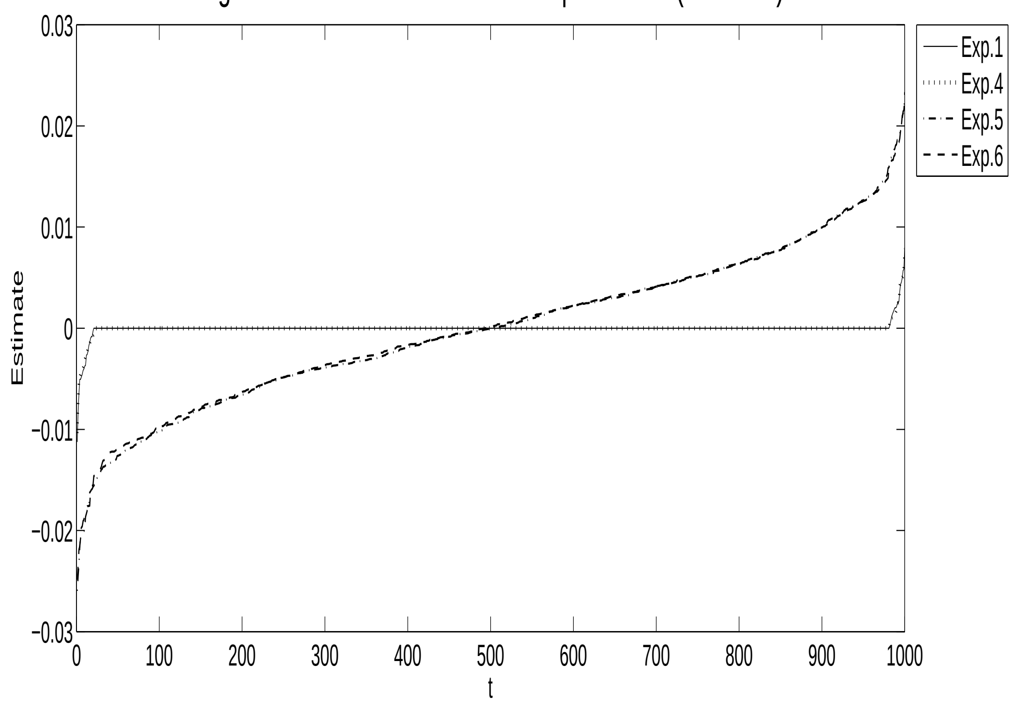


Figure 4.2c: Estimates of "relevant" parameters (5th percentile)

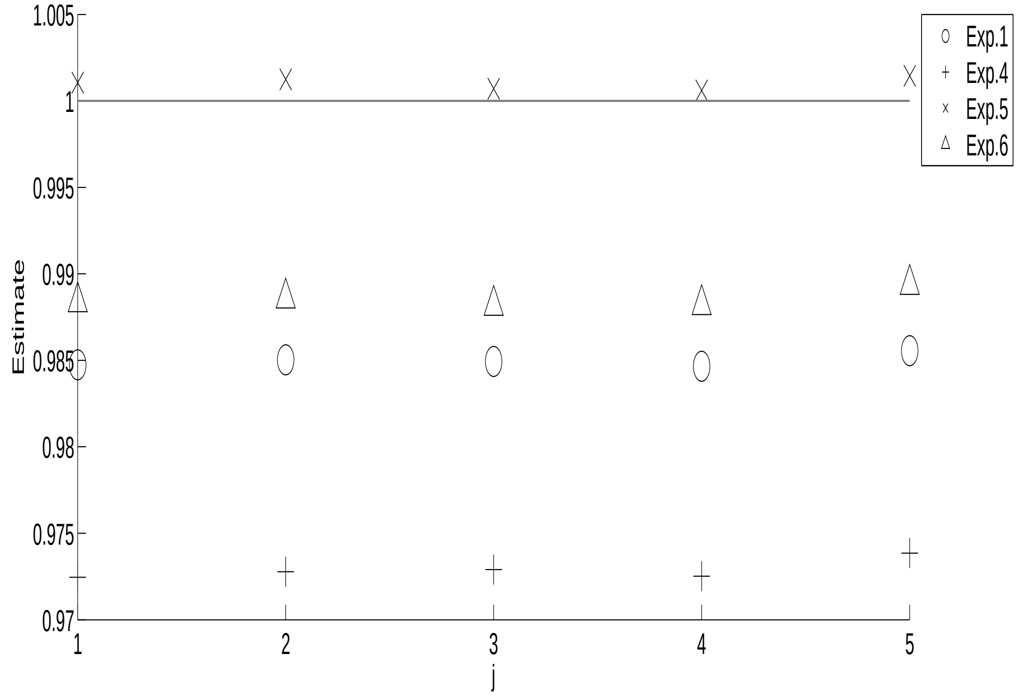


Figure 4.2d: Estimates of "relevant" parameters (median)

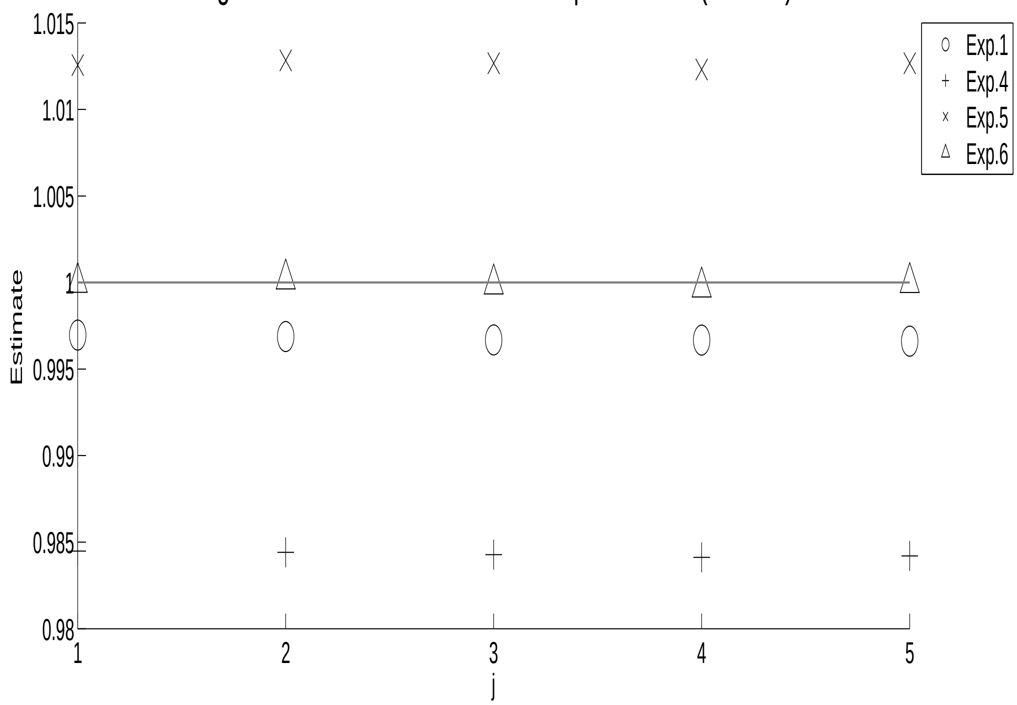


Figure 4.2e: Estimates of "relevant" parameters (95th percentile)

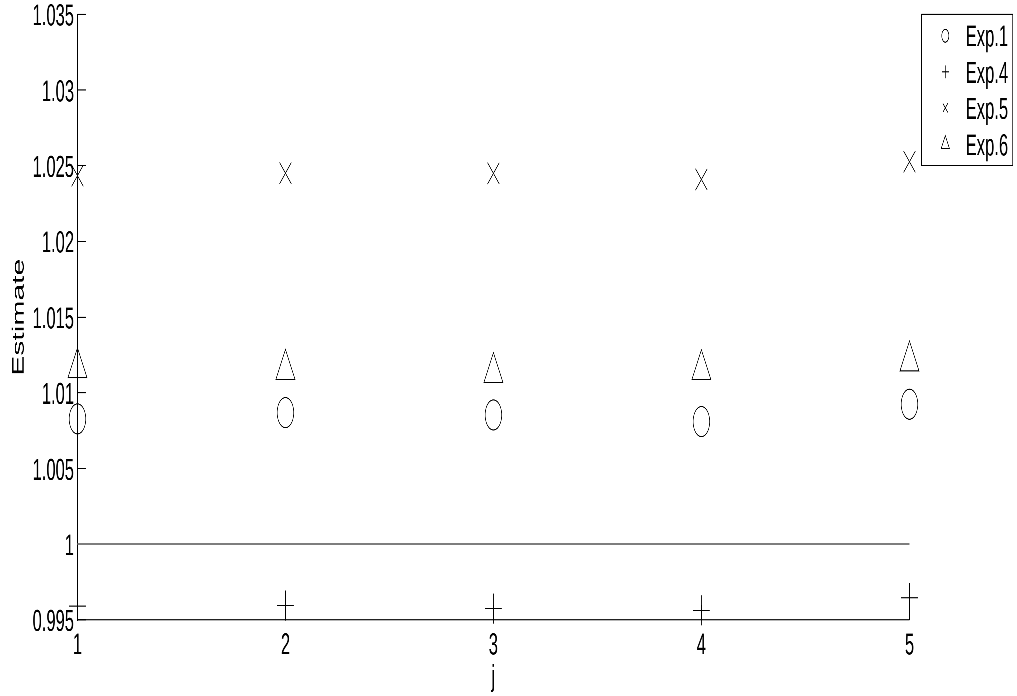


Figure 4.2f: Estimates of "relevant" parameters (mean)

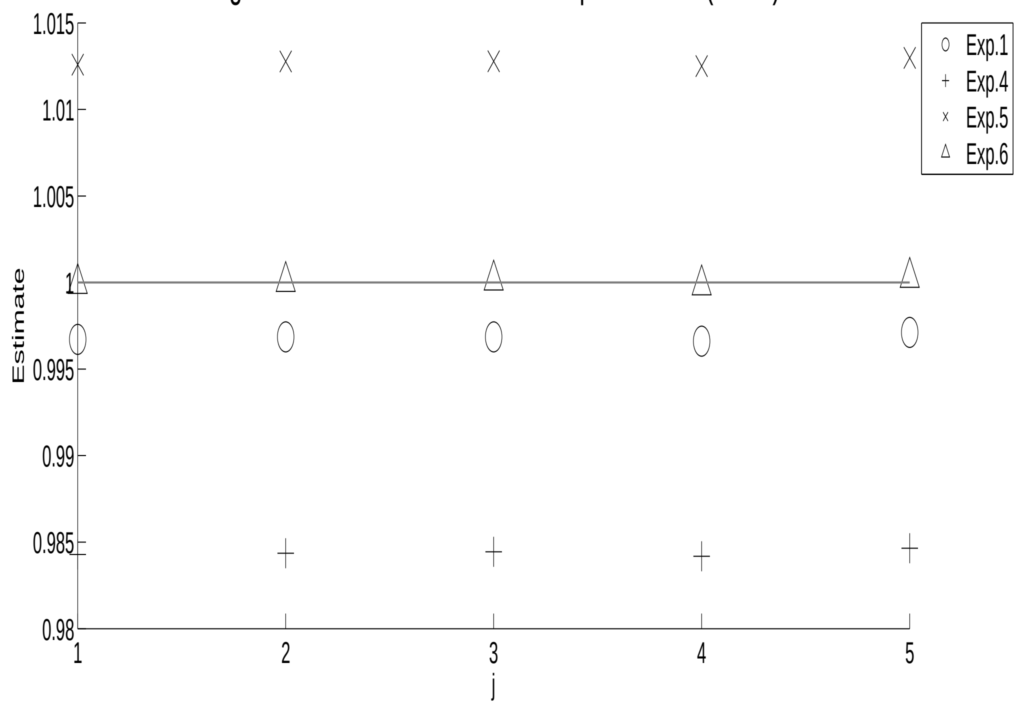


Figure 4.2g: Estimates of "irrelevant" parameters (5th percentile)

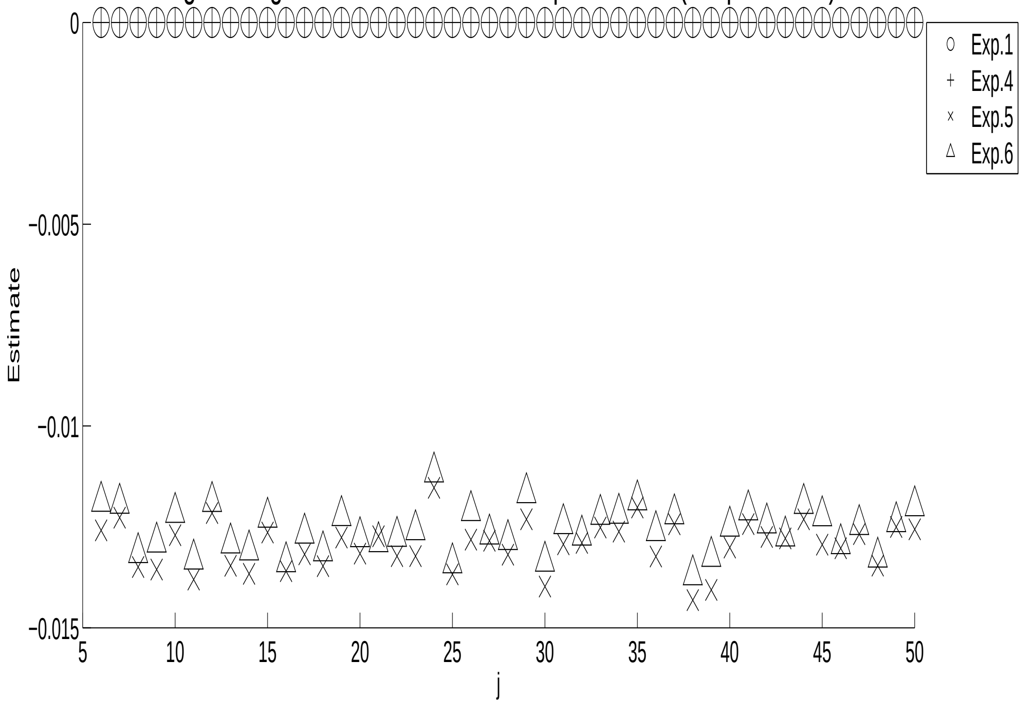
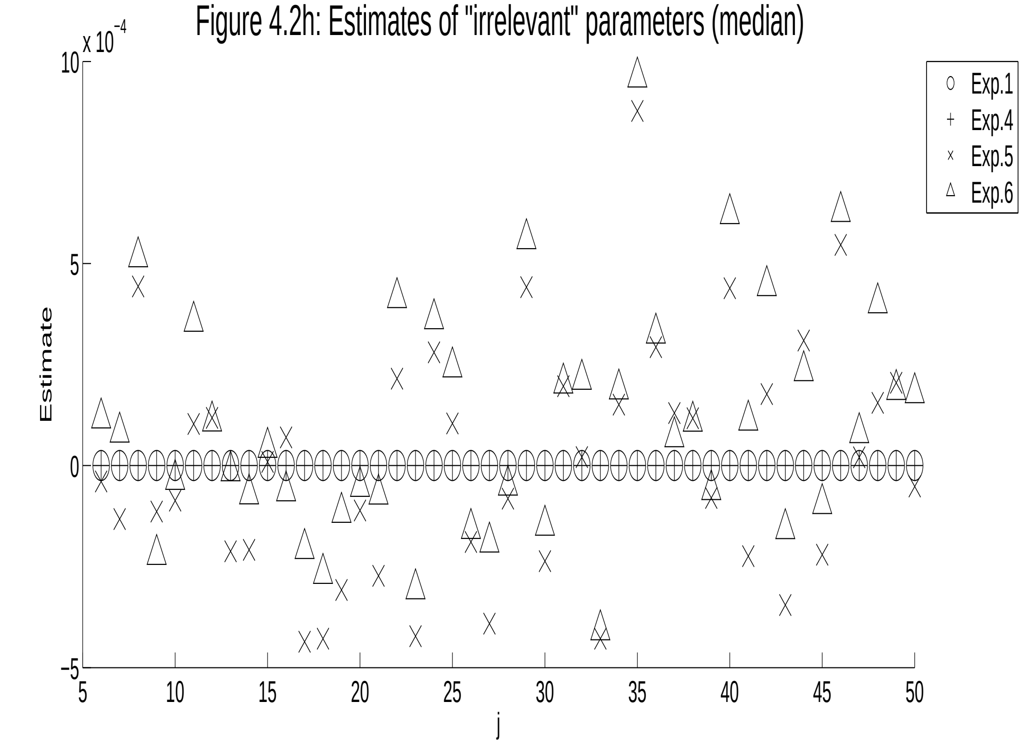


Figure 4.2h: Estimates of "irrelevant" parameters (median)



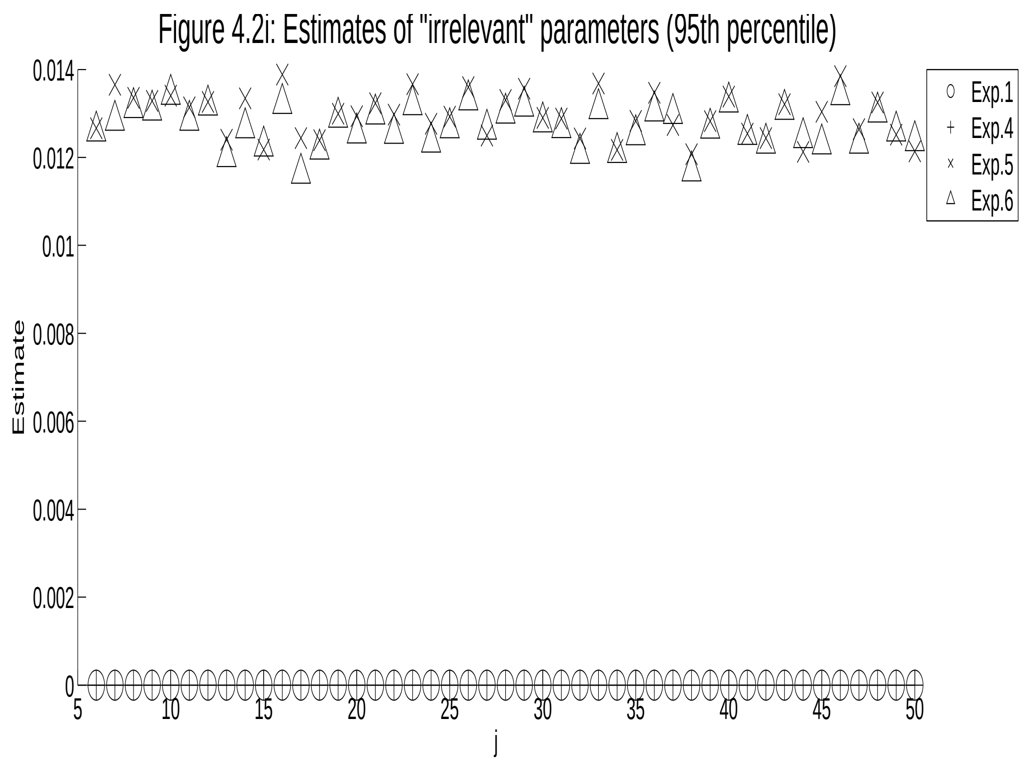


Figure 4.2i: Estimates of "irrelevant" parameters (95th percentile)
Estimate
j
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
5
10
15
20
25
30
35
40
45
50
Exp.1
Exp.4
Exp.5
Exp.6

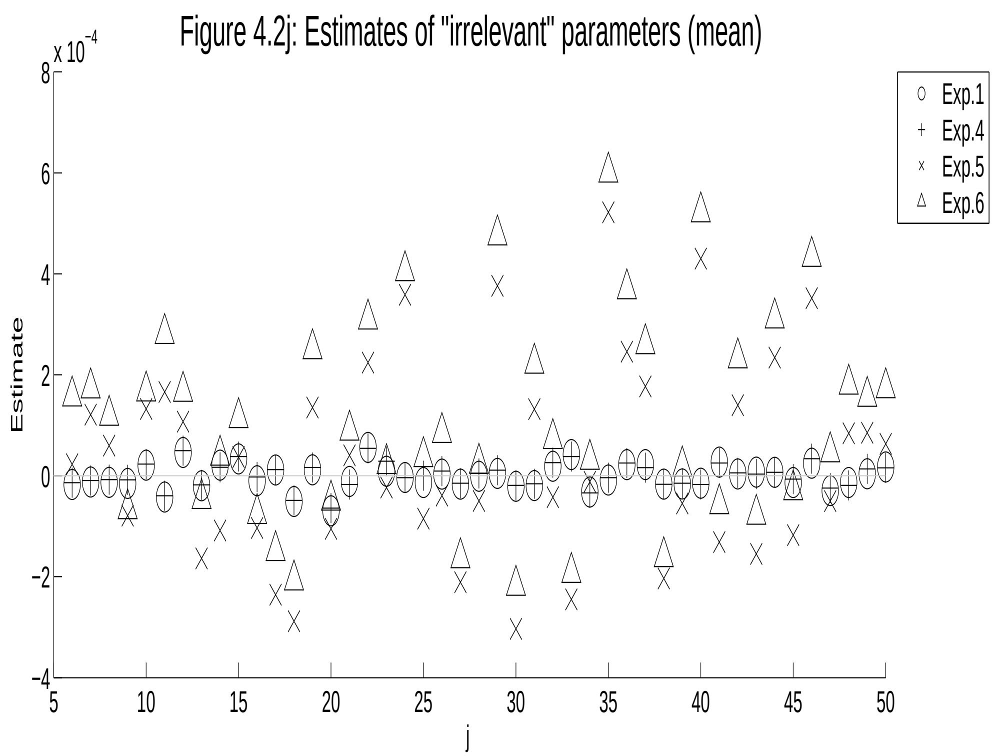


Figure 4.2j: Estimates of "irrelevant" parameters (mean)
x 10⁻⁴
Estimate
j
8
6
4
2
0
-2
-4
5
10
15
20
25
30
35
40
45
50
Exp.1
Exp.4
Exp.5
Exp.6

Table 4.3: l_2-errors and selection (Exp. 1, 4-6)

Mean	$n = 4700$			
Exp. #	1	4	5	6
$2^{nd}-$stage select %	98.1	98.1	55.0	54.5
$2^{nd}-$stage l_2-err	0.018	0.038	0.062	0.054
1^{st}-stage select %	98.5	52.0	98.5	52.0
$1^{st}-$stage l_2-err	0.028	0.059	0.028	0.059

In the next group of experiments which explore the sensitivity of the results for the two-stage Lasso estimator to design parameters, changes are applied to σ_ϵ, σ_η, σ_z, and the correlations between the rows of the design matrix $\mathbf{z}_i^T \in \mathbb{R}^{p\times d}$ for all $i = 1, ..., n$. Figure 4.3a plots (in ascending values) the 1000 estimates of β_5^* when the sample size $n = 47$. Figure 4.3b plots (in ascending values) the 1000 estimates of β_6^* when the sample size $n = 47$. Figures 4.3c-4.3f (Figures 4.3g-4.3j) displays the 5^{th} percentile, the median, the 95^{th} percentile, and the mean of the estimates of the "relevant" ("irrelevant") main-equation parameters from Experiment 1 and Experiments 7-13. The mean of the l_2-errors and the mean of the selection percentages of the main-equation estimates together with the "averaged" mean of the l_2-errors and the "averaged" mean of the selection percentages of the first-stage estimates from these experiments are displayed in Table 4.4.

Overall, we see from Table 4.4 that, relative to the baseline experiment (Experiment 1), the mean of the l_2-errors of the estimates of the main-equation parameters increase in Experiments 7-13; the mean of the selection percentages of the estimates of the main-equation parameters decrease in Experiments 7, 8, 10-13. The "averaged" mean of the l_2-errors of the first-stage estimates increase the most in Experiments 8, 9, 12, and 13 while those first-stage statistics in Experiment 10 are comparable to those in Experiment 1. This makes sense since Experiments 8, 9, 12, and 13 involve increasing the noise level σ_η or decreasing the signal level σ_z of the instruments in the first-stage model while introducing correlations between the rows of the design matrix $\mathbf{z}_i^T$ (for $i = 1, ..., n$) (Experiment 10) should have little impact on the first-stage estimates, which are obtained by performing the Lasso procedure on each of the 50 first-stage equations separately. Note that since Experiment 7 (Experiment 11) has exactly the same first-stage set up as Experiment 1 (respectively, Experiment 10), there is no need to look at the behavior of their first-stage estimates separately.

From Figure 4.3a, we see that, below the 10^{th} percentile, compared to the baseline experiment (Experiment 1), introducing correlations between the rows of the design matrix $\mathbf{z}_i^T$ (for $i = 1, ..., n$) improves the estimates of the "relevant" main-equation parameters while the other changes to the data generating process yield worse estimates; above the 80^{th} percentile, Figure 4.3a shows that, any changes made to the data generating process yield worse estimates of the "relevant" main-equation parameters; at the median (or mean), Figures 4.3a and 4.3d (respectively, Figure 4.3f) show that, increasing the standard deviation of the "noise" in the main equation, σ_ϵ, and reducing the standard deviation of the "signal", σ_z, yield worse estimates of the "relevant" main-equation parameters. Figures 4.3b and 4.3j show that, in estimating the "irrelevant" main-equation parameters, any changes to the data generating process yield worse estimates; in particular, those that involve introducing correlations between the rows of the design matrix $\mathbf{z}_i^T$ (for $i = 1, ..., n$) (Experiments 10-13) yield the poorest estimates of the "irrelevant" main-equation parameters (also see Table 4.4 for a comparison between the selection percentages of these experiments). Recall that each row of $\mathbf{z}_i^T \in \mathbb{R}^{p \times d}$ is associated with each of the endogenous regressors and row-wise correlations in $\mathbf{z}_i^T$ hence introduce correlations between the "purged" regressors X_j^* and $X_{j'}^*$ for all $j \neq j'$. As seen in Section 2.3.2, selection consistency hinges on Assumption 2.3.7 (the "mutual incoherence condition"), whose violation can lead the Lasso to falsely include elements that are not in the support of β^*, namely, the violation of Part (b) in Theorems 2.3.7 and 3.8. For the baseline experiment (Experiment 1), the quantity $\left\| \mathbb{E}\left[X_{1,J(\beta^*)^c}^{*T} X_{1,J(\beta^*)}^*\right] \left[\mathbb{E}(X_{1,J(\beta^*)}^{*T} X_{1,J(\beta^*)}^*)\right]^{-1} \right\|_\infty$ in Assumption 2.3.7 equals 0 (because $\mathbf{z}_i^T$ is a $p \times d$ matrix of independent standard normal random variables) and Assumption 2.3.7 is easily satisfied. For Experiments 10-13, this quantity increases, and therefore in these experiments, the estimates of the "irrelevant" main-equation parameters are worse relative to the baseline experiment.

Figure 4.3a: Estimates of the fifth parameter (n = 47)

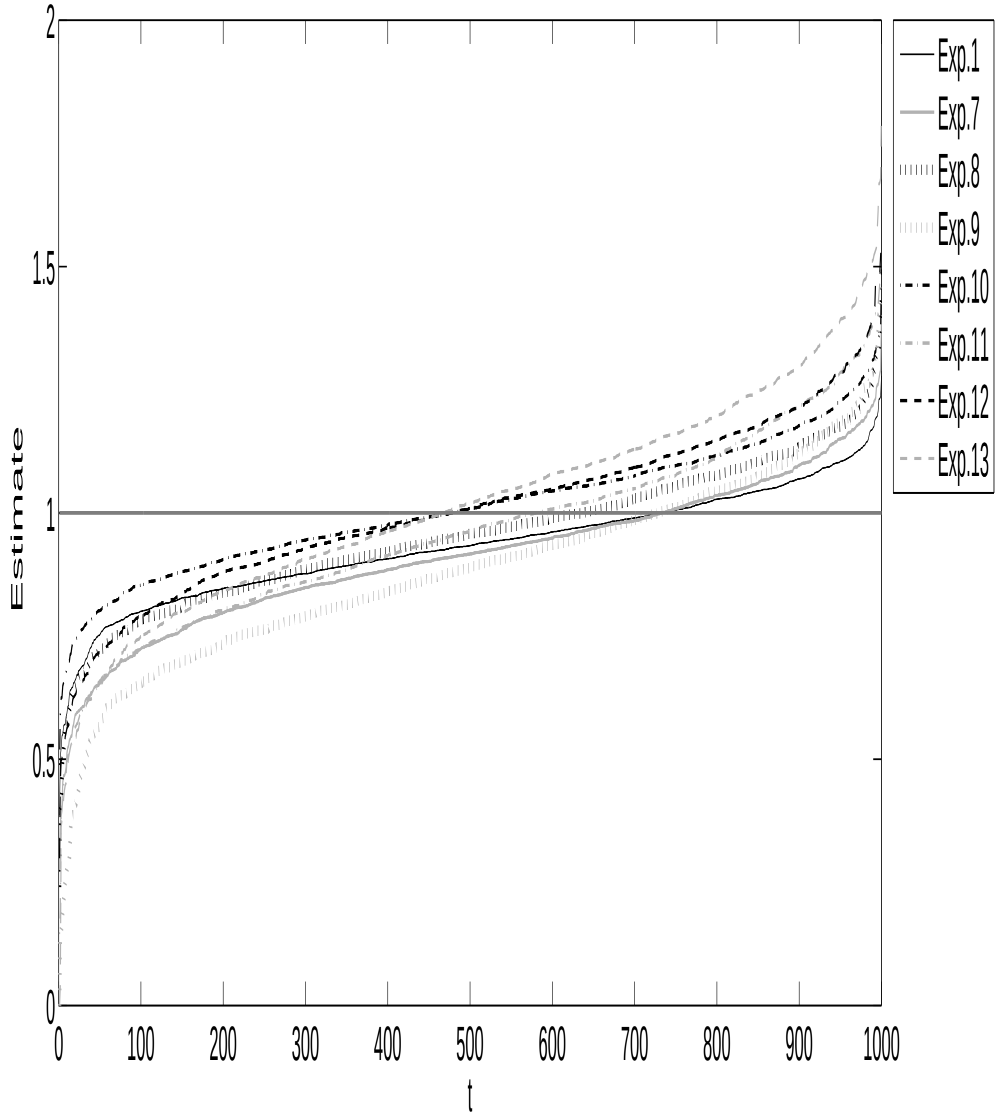


Figure 4.3b: Estiamtes of the sixth parameter (n = 47)

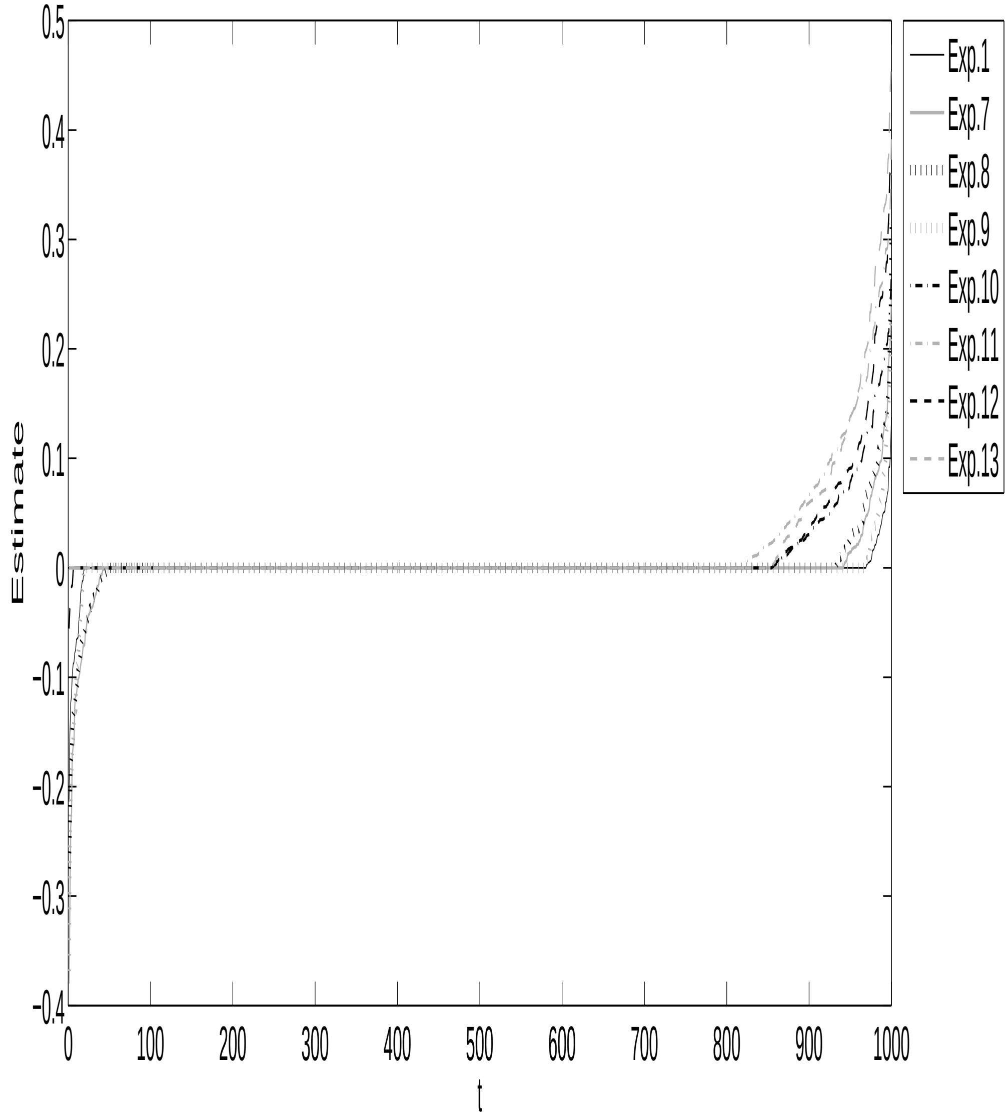


Figure 4.3c: Estimates of "relevant" parameters (5th percentile)

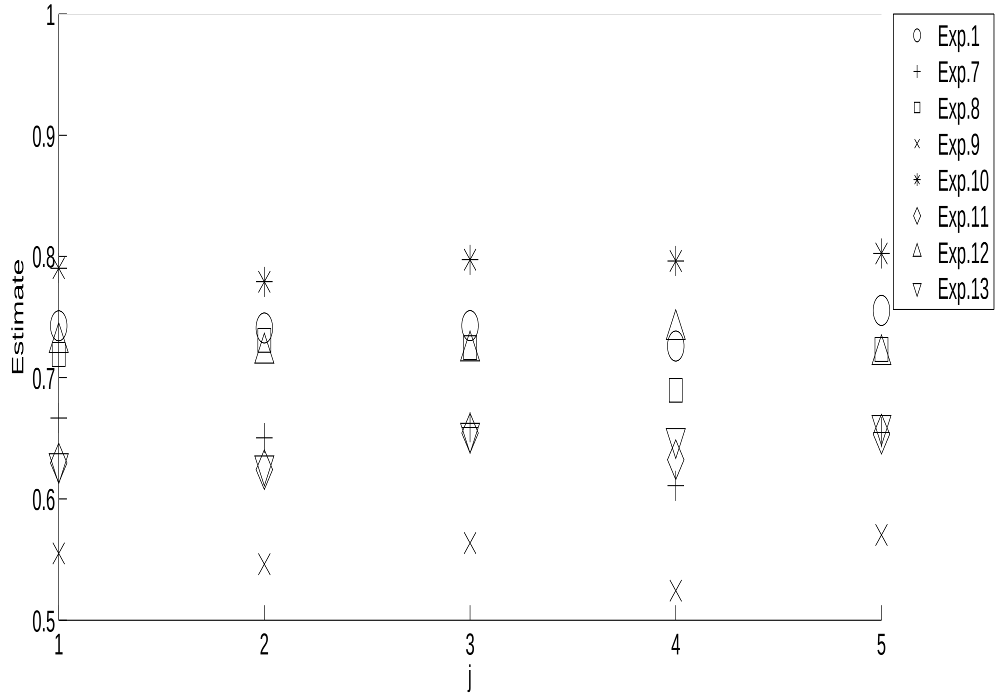


Figure 4.3d: Estimates of the "relevant" parameters (median)

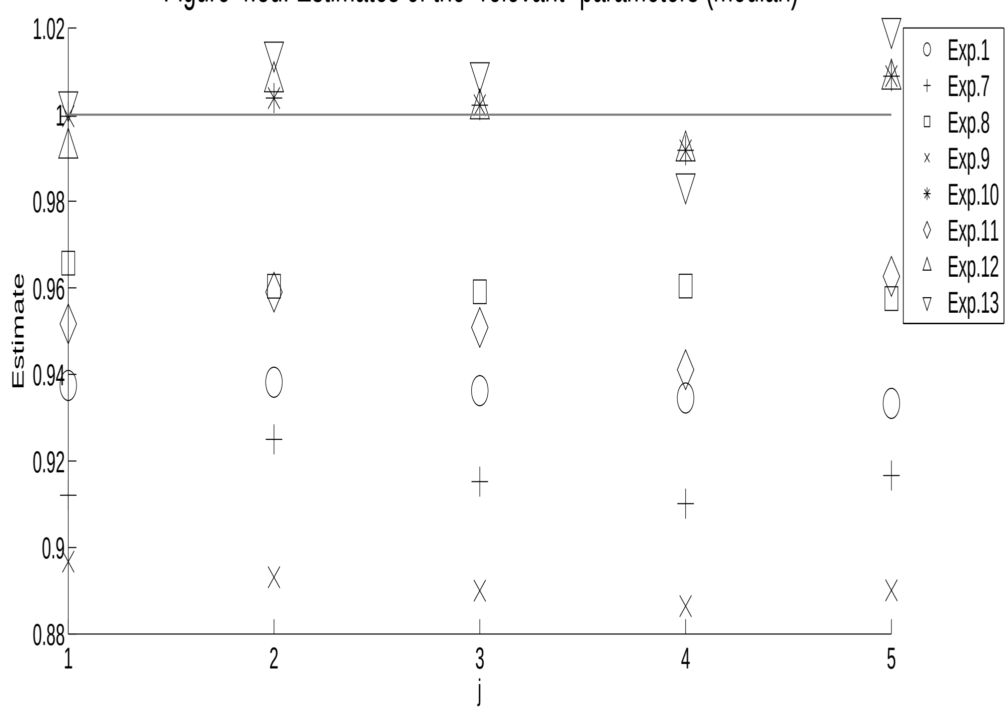


Figure 4.3e: Estimates of the "relevant" parameters (95th percentile)

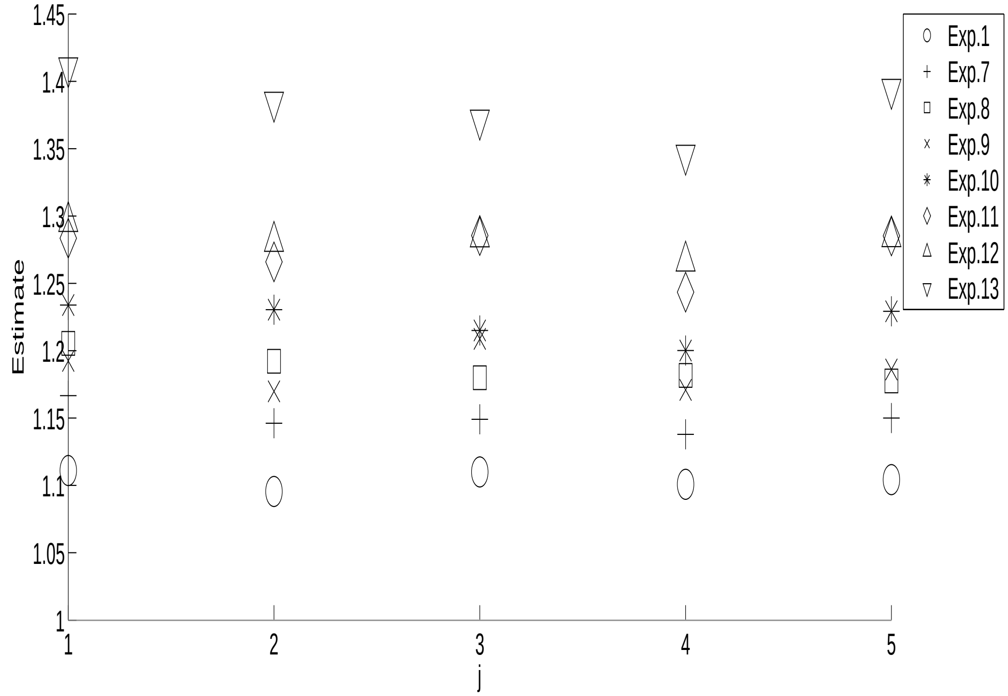


Figure 4.3f: Estimates of the "relevant" parameters (mean)

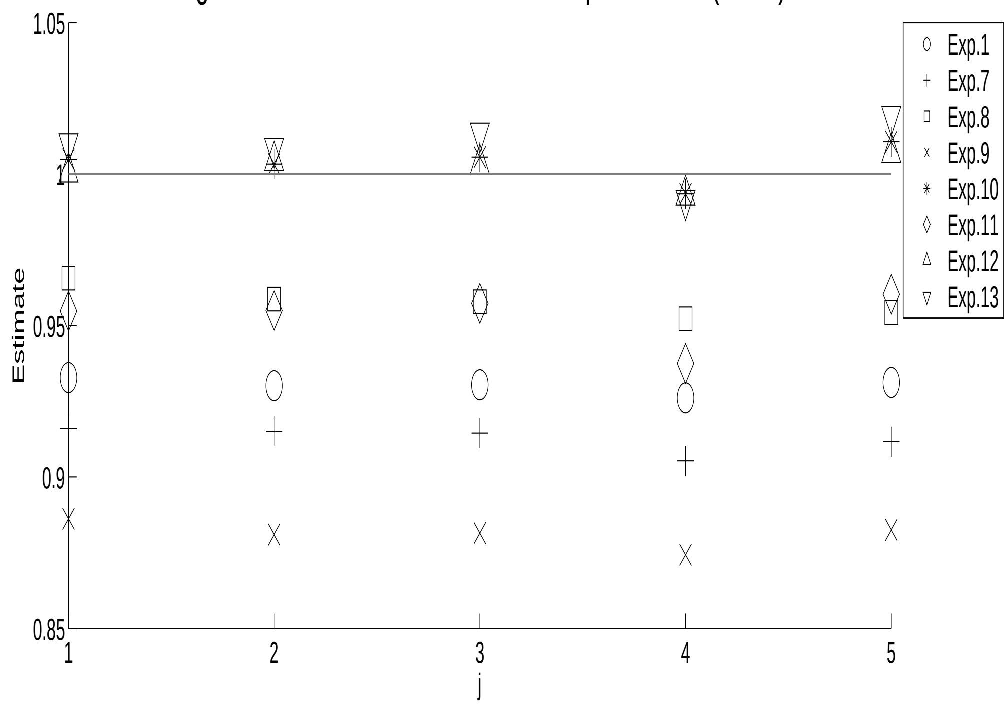


Figure 4.3g: Estimates of the "irrelevant" parameters (5th percentile)

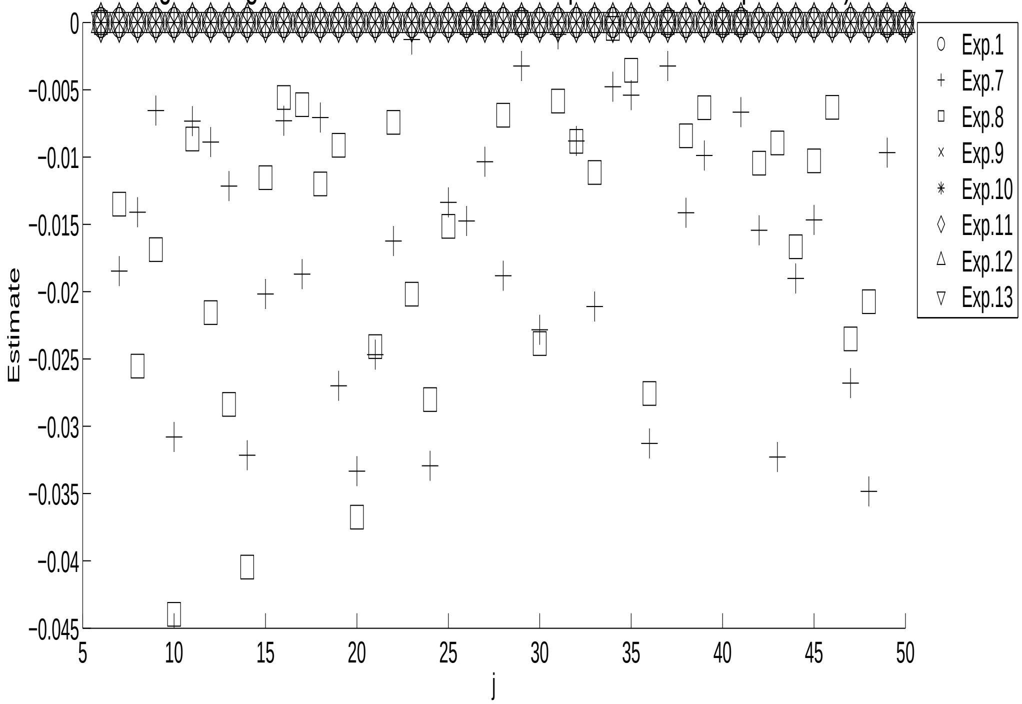


Figure 4.3h: Estimates of the "irrelevant" parameters (median)

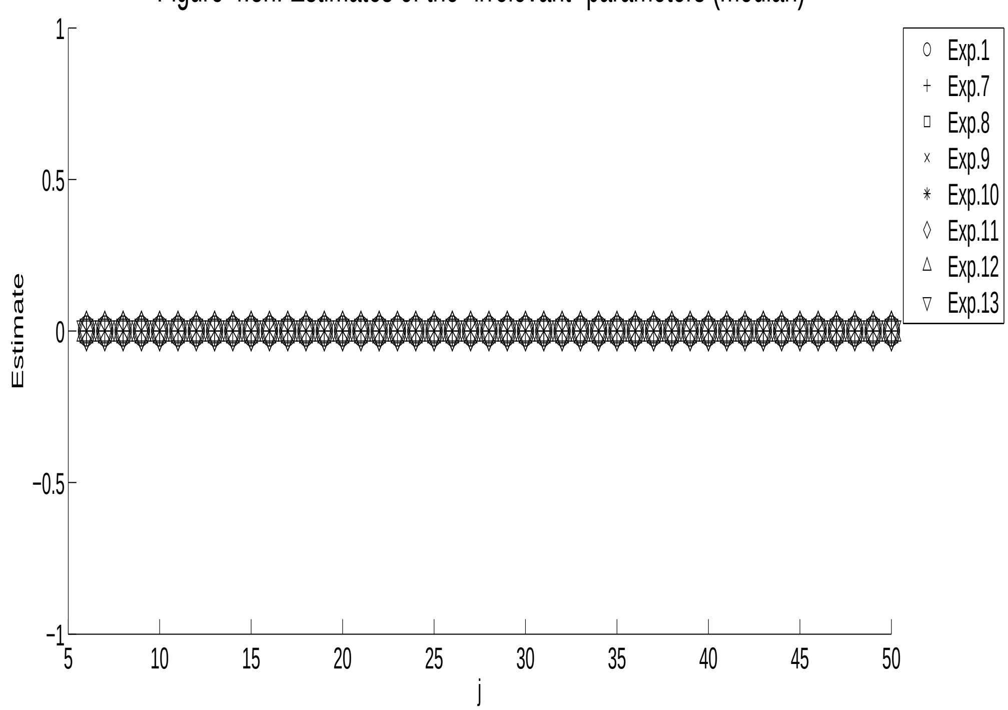


Figure 4.3i: Estimates of the "irrelevant" parameters (95th percentile)

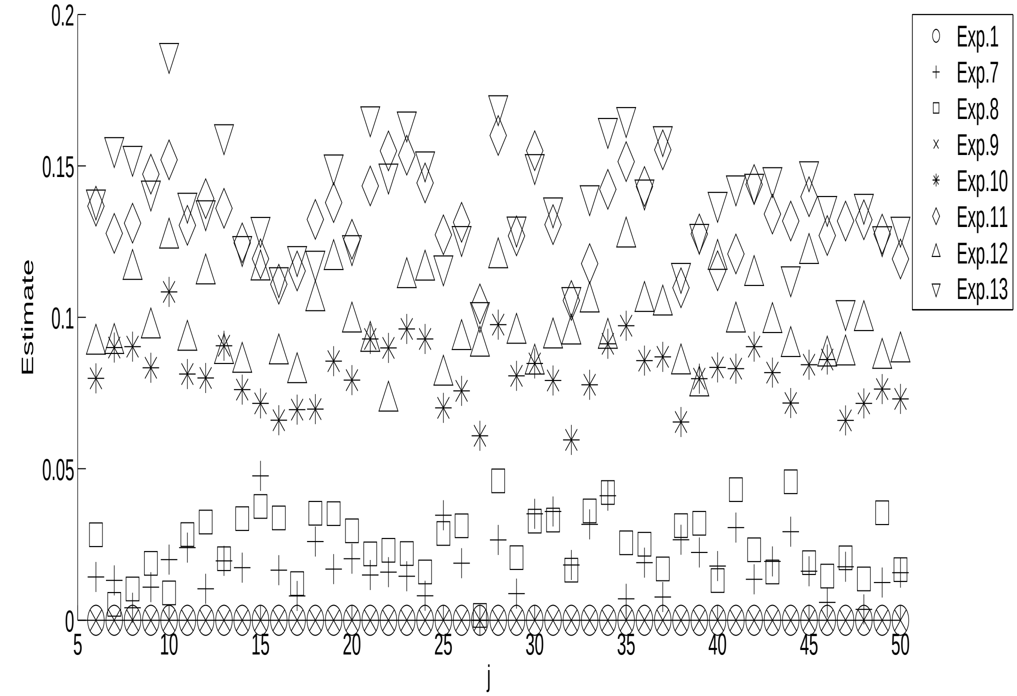


Figure 4.3j: Estimates of the "irrelevant" parameters (mean)

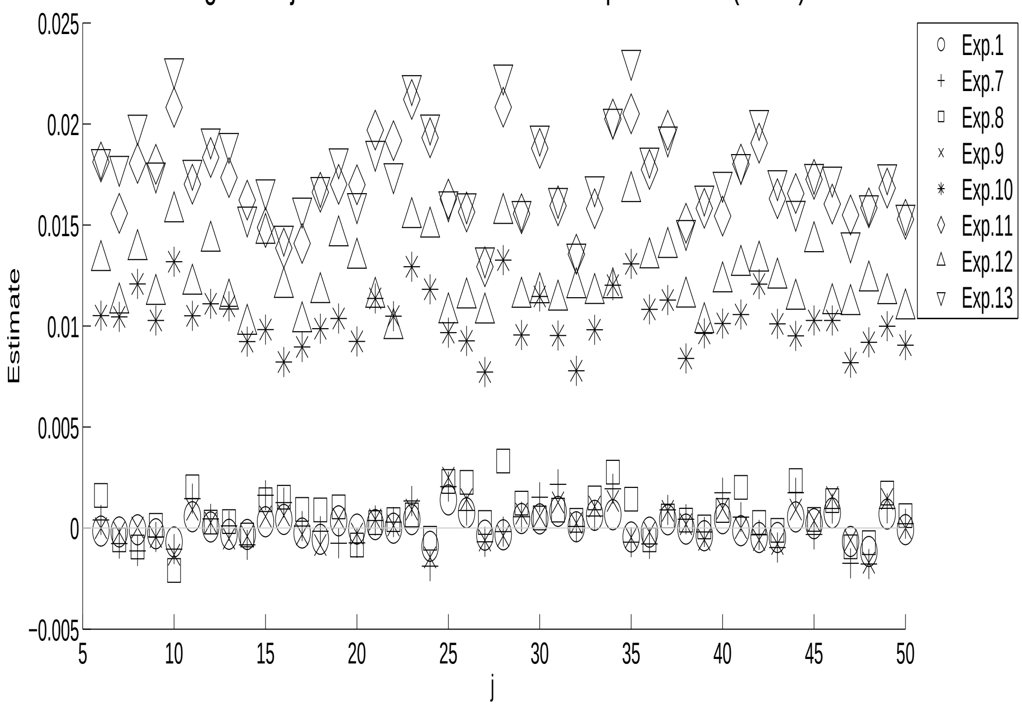


Table 4.4: l_2-errors and selection (Exp. 1, 7-13)

Mean	$n = 47$							
Exp. #	1	7	8	9	10	11	12	13
$2^{nd}-$stage select %	97.3	94.3	93.8	97.3	87.0	85.4	87.8	87.1
2^{nd}-stage l_2-err	0.288	0.422	0.376	0.497	0.365	0.557	0.471	0.626
$1^{st}-$stage select %	97.7	97.7	90.8	97.7	97.7	97.7	90.8	97.7
$1^{st}-$stage l_2-err	0.349	0.349	0.789	0.552	0.352	0.352	0.793	0.557

In the final experiment 14 where $(\beta_1^*, \ldots, \beta_5^*) = (0.01, \ldots, 0.01)$ (as opposed to $(\beta_1^*, \ldots, \beta_5^*) = (1, \ldots, 1)$ in the previous experiments), based on the estimates obtained from the two-stage Lasso procedure, I count the number of occurrences that each estimate $\hat{\beta}_{H2SLS,1}, \ldots, \hat{\beta}_{H2SLS,5}$ equals exactly 0, respectively, over the 1000 replications (Table 4.5). Because the "relevant" main-equation parameters are reduced by a factor of 100, it is clearly more difficult for the two-stage Lasso procedure to distinguish the "relevant" coefficients from the "irrelevant" coefficients and Table 4.5 verifies this. Recall in Experiments 10-13, by introducing correlations between the "purged" regressors X_j^* and $X_{j'}^*$ for all $j \neq j'$, the estimates of the "irrelevant" main-equation parameters become worse. On the other hand, making the "relevant" main-equation parameters sufficiently smaller results in worse estimates of the "relevant" main-equation parameters. This observation confirms Part (d) of Theorems 2.3.7 and 2.3.8; i.e., the violation of the "beta-min" condition can lead the Lasso to mistake the "relevant" coefficients for the "irrelevant" coefficients. In terms of the l_2-errors and overall selection percentages, from Table 4.5 we see that poorer estimation of the "relevant" parameters also results in larger l_2-errors[1] and worse selection percentages, as expected. The significant drop in the overall selection percentages suggests that not only the estimation of the "relevant" coefficients becomes less accurate in Experiment 14 but also the estimation of the "irrelevant" coefficients.

[1] To compensate for the fact that $(\beta_1, ..., \beta_5) = (1, \ldots, 1)$ in Experiment 1 exceeds the parameters $(\beta_1, ..., \beta_5) = (0.01, \ldots, 0.01)$ in Experiments 14 by a factor of 0.01, the l_2-error in Experiment 14 is adjusted as $\left[\frac{\sum_{j=1}^{5}(\hat{\beta}_j - \beta_j^*)^2}{0.01^2} + \sum_{j=6}^{j=50}(\hat{\beta}_j - \beta_j^*)^2\right]^{1/2}$. The unadjusted l_2-error in Experiment 14 is 0.388.

Table 4.5: Exp. 14

$n = 47$	# of zeros					$2^{nd}-$stg select %	$2^{nd}-$stg l_2-err
	β_1	β_2	β_3	β_4	β_5		
Exp. 1	0	0	0	0	0	97.3	0.288
Exp. 14	187	187	218	194	193	57.7	11.5

2.5 Conclusion and extensions

This chapter has explored the validity of the two-stage estimation procedure for sparse linear models in high-dimensional settings with possibly many endogenous regressors. In particular, the number of endogenous regressors in the main equation and the number of instruments in the first-stage equations are permitted to grow with and exceed n. Sufficient scaling conditions on the sample size for estimation consistency in l_2-norm and variable-selection consistency of the high-dimensional two-stage estimators have been established. I provide theoretical justifications to a technical issue (regarding the RE condition and the MI condition) that arises in the two-stage estimation procedure from allowing the number of regressors in the main equation to grow with and exceed n. Depending on the underlying assumptions that are imposed, the scaling of the sample size required to obtain these consistency results differs by factors involving the sparsity parameter k_2. Simulations are conducted to gain insight on the small-sample performance of the high-dimensional two-stage estimator. The approach and results of this chapter suggest a number of possible extensions including the ones listed in the following.

***1. The approximate sparsity* case**. It is useful to extend the analysis for the high-dimensional 2SLS estimator to the *approximate sparsity* case, i.e., most of the coefficients in the main equation and/or the first-stage equations are too small to matter. One can have the approximate sparsity assumption in the first-stage equations only (and assume the main equation parameters are sparse), the main equation only (and assume the first-stage equations parameters are sparse) or both-stage equations. Results on this extension are established in Zhu (2014).

2. Minimax lower bounds for the high-dimensional linear models with endogeneity. It would be worthwhile to establish the minimax lower bounds on the parameters in the main equation for the linear models in high-dimensional settings with endogeneity. In particular, the goal is to derive lower bounds on the estimation error achievable by any estimator, regardless of its computational complexity. Obtaining lower bounds of this type is useful because on one hand, if the lower bound matches the upper bound up to some constant factors, then there is no need to search for estimators with a lower statistical error (although it might still be useful to study estimators with lower computational costs). On the other hand, if the lower bound does not match the best known upper bounds, then it is worthwhile to search for new estimators that potentially achieve the lower bound. A recent result of Zhu (2014) shows that, for the class of triangular simultaneous equations models (2.1) and (2.2) with exactly sparse structure on the coefficient vector in the main equation (2.1), the estimator $\hat{\beta}_{H2SLS}$ is minimax optimal up to the $(n,\, d,\, p)-$scaling and the upper bound on the $\sqrt{MSE}$ of $\hat{\beta}_{H2SLS}$ exceeds the minimax lower bound by a factor of k_2.

3. Control function approach in high-dimensional settings. As an alternative to the "two-stage" estimation proposed here, it would be interesting to explore the validity of the high-dimensional two-stage estimators based on the "control function" approach in the high-dimensional setting. When both the first and second-stage equations are in low-dimensional settings (i.e., $p \ll n$ and $d_j \ll n$ for all $j = 1, ..., p$) and the supports of the true parameters in both stages are known *a priori*, the 2SLS procedure is algebraically equivalent to a "control function" estimator of β^* that includes first-stage residuals $\hat{\eta}_{ij} = x_{ij} - \mathbf{z}_{ij}^T \hat{\pi}_j$ as "control variables" in the regression of y_i on $\mathbf{x}_i$ (e.g., Garen, 1984). Such algebraic equivalence no longer holds for regularized estimators because the regularization employed destroys the projection algebra. The extension for the 2SLS estimator from low-dimensional settings to high-dimensional settings is somewhat more natural than the extension for the two-stage estimator based on the control function approach. In particular, it is useful to explore under what conditions one can translate the sparsity or approximate sparsity assumption on the coefficients β^* in the triangular simultaneous equations model (2.1) and (2.2) to the sparsity or approximate sparsity assumption on the coefficients β^* and α^* in the model $y_i = \mathbf{x}_i^T \beta^* + \boldsymbol{\eta}_i \alpha^* + \xi_i$ where $\mathbb{E}(\boldsymbol{\eta}_i \xi_i) = \mathbb{E}(\mathbf{x}_i \xi_i) = \mathbf{0}$. A simple sufficient condition for such a translation is to impose the joint normality assumption of the error terms ϵ_i and $\boldsymbol{\eta}_i = (\eta_{i1}, ..., \eta_{ip})$. Then, by the property of multivariate normal distributions, we have $\mathbb{E}(\epsilon_i | \boldsymbol{\eta}_i) = \Sigma_{\epsilon\eta} \Sigma_{\eta\eta}^{-1} \boldsymbol{\eta}_i^T$. If we further assume only a few of the correlation coefficients $(\rho_{\epsilon_i \eta_{i1}}, ..., \rho_{\epsilon_i \eta_{ip}})$ (associated with the covariance matrix $\Sigma_{\epsilon\eta}$) are non-zero or most of these correlation coefficients are too small to matter, the sparsity or approximate sparsity can be carried to the model $y_i = \mathbf{x}_i^T \beta^* + \boldsymbol{\eta}_i \alpha^* + \xi_i$. Then, we can obtain consistent estimates of η, $\hat{\eta}$, from the first-stage regression by either a standard least square estimator when the first-stage regression concerns a small number of regressors relative to n, or a least square estimator with l_1- regularization (the Lasso or Dantzig selector) when the first-stage regression concerns a large number of regressors relative to n, and then apply a Lasso technique in the second stage as follows

$$\hat{\beta}_{HCF} \in \text{argmin}_{\beta,\alpha\in\mathbb{R}^p}: \frac{1}{2n}|y - X\beta - \hat{\eta}\alpha|_2^2 + \lambda_n \left(|\beta|_1 + |\alpha|_1\right).$$

The statistical properties of $\hat{\beta}_{HCF}$ can be analyzed in the same way as those of $\hat{\beta}_{H2SLS}$. How this argument can be extended to non-Gaussian error settings is an interesting question for future research.

2.6 Proofs for results in Chapter 2

For technical simplifications, in the following proofs, I assume without loss of generality that the first moment of $(y_i, \mathbf{x}_i, \mathbf{z}_i)$ is zero for all i (if it is not the case, we can simply subtract their population mean). Also, for notational simplicity, assume $d_j = d$ for all $j = 1, ..., p$; additionally, as in most high-dimensional statistics literature, I assume the regime of interest is $p \geq n$ and $d \geq n$ (except for Corollary 2.3.4 where $d \ll n$ is assumed). The modification to allow $p < n$ or $d < n$ or $d_j \neq d_{j'}$ for some j and j' is trivial. Also, as a general rule for the proofs, b constants denote positive constants that do not involve n, p, d, k_1 and k_2 but possibly the sub-Gaussian parameters defined in Assumptions 2.3.2-2.3.4; c constants denote universal positive constants that are independent of both n, p, d, k_1 and k_2 as well as the sub-Gaussian parameters. The specific values of these constants may change from place to place.

2.6.1 Lemma 2.3.1

Proof. First, write

$$\begin{aligned} y &= X\beta^* + \epsilon = X^*\beta^* + (X\beta^* - X^*\beta^* + \epsilon) \\ &= X^*\beta^* + (\boldsymbol{\eta}\beta^* + \epsilon) \\ &= \hat{X}\beta^* + (X^* - \hat{X})\beta^* + \boldsymbol{\eta}\beta^* + \epsilon \\ &= \hat{X}\beta^* + e, \end{aligned}$$

where $e := (X^* - \hat{X})\beta^* + \boldsymbol{\eta}\beta^* + \epsilon$. Define $\hat{v}^0 = \hat{\beta}_{H2SLS} - \beta^*$ and the Lagrangian $L(\beta;\lambda_n) = \frac{1}{2n}|y - \hat{X}\beta|_2^2 + \lambda_n|\beta|_1$. Since $\hat{\beta}_{H2SLS}$ is optimal, we have

$$L(\hat{\beta}_{H2SLS};\lambda_n) \leq L(\beta^*;\lambda_n) = \frac{1}{2n}|e|_2^2 + \lambda_n|\beta^*|_1,$$

Some algebraic manipulation of the *basic inequality* above yields

$$
\begin{aligned}
0 &\leq \frac{1}{2n}|\hat{X}\hat{v}^0|_2^2 \leq \frac{1}{n}e^T\hat{X}\hat{v}^0 + \lambda_n\left\{|\beta^*_{J(\beta^*)}|_1 - |(\beta^*_{J(\beta^*)} + \hat{v}^0_{J(\beta^*)},\, \hat{v}^0_{J(\beta^*)^c})|_1\right\} \\
&\leq |\hat{v}^0|_1|\frac{1}{n}\hat{X}^Te|_\infty + \lambda_n\left\{|\hat{v}^0_{J(\beta^*)}|_1 - |\hat{v}^0_{J(\beta^*)^c}|_1\right\} \\
&\leq \frac{\lambda_n}{2}\left\{3|\hat{v}^0_{J(\beta^*)}|_1 - |\hat{v}^0_{J(\beta^*)^c}|_1\right\},
\end{aligned}
$$

where the last inequality holds as long as $\lambda_n \geq 2|\frac{1}{n}\hat{X}^Te|_\infty > 0$. Consequently, $|\hat{v}^0|_1 \leq 4|\hat{v}^0_{J(\beta^*)}|_1 \leq 4\sqrt{k_2}|\hat{v}^0_{J(\beta^*)}|_2 \leq 4\sqrt{k_2}|\hat{v}^0|_2$. Note that we also have

$$
\begin{aligned}
\frac{1}{2n}|\hat{X}\hat{v}^0|_2^2 &\leq |\hat{v}^0|_1|\frac{1}{n}\hat{X}^Te|_\infty + \lambda_n\left\{|\hat{v}^0_{J(\beta^*)}|_1 - |\hat{v}^0_{J(\beta^*)^c}|_1\right\} \\
&\leq 4\sqrt{k_2}|\hat{v}^0|_2\lambda_n.
\end{aligned}
$$

Since we assume in Lemma 2.3.1 that the random matrix $\hat{\Gamma} = \hat{X}^T\hat{X}$ satisfies the RE condition in Definition 1.1 from Chapter 1 with $\gamma = 3$, we have

$$
|\hat{\beta}_{H2SLS} - \beta^*|_2 \leq \frac{c'}{\delta}\sqrt{k_2}\lambda_n.
$$

2.6.2 Theorem 2.3.2

As discussed in Section 2.3, the l_2-consistency of $\hat{\beta}_{H2SLS}$ requires verifications of two conditions: (i) $\hat{\Gamma} = \hat{X}^T\hat{X}$ satisfies the RE condition in Definition 1.1 from Chapter 1 with $\gamma = 3$, and (ii) the term $|\frac{1}{n}\hat{X}^Te|_\infty \precsim f(k_1,\, k_2,\, d,\, p,\, n)$ with high probability. This is done via Lemmas 2.6.1 and 2.6.2.

Lemma 2.6.1 (RE condition): Under Assumptions 2.1.1, 2.3.1, 2.3.3, 2.3.5 and the condition

$$
n \succsim \max\{k_1 \log d,\, k_1 \log p\},
$$

we have, for some universal constants c, c_1, and c_2,

$$
\frac{|\hat{X}v^0|_2^2}{n} \geq \kappa_1|v^0|_2^2 - c\kappa_2\sqrt{\frac{k_1 \log \max(p,\, d)}{n}}|v^0|_1^2, \qquad \text{for all } v^0 \in \mathbb{R}^p,
$$

with probability at least $1 - c_1 \exp(-c_2 \log \max(p,\, d))$, where

$$\begin{aligned}
\kappa_1 &= \frac{\lambda_{\min}(\Sigma_{X^*})}{2},\ \kappa_2 = \max\left\{b_0,\ b_1\right\}, \\
b_0 &= \lambda_{\min}(\Sigma_{X^*}) \max\left\{\frac{\sigma_{X^*}^4}{\lambda_{\min}^2(\Sigma_{X^*})},\ 1\right\}, \\
b_1 &= \frac{\sigma_\eta \sigma_Z \sqrt{\lambda_{\max}(\Sigma_Z)}}{\lambda_{\min}(\Sigma_Z)}
\end{aligned}$$

Proof. We have

$$\left| v^{0T} \frac{\hat{X}^T \hat{X}}{n} v^0 \right| + \left| v^{0T} \left(\frac{X^{*T} X^* - \hat{X}^T \hat{X}}{n} \right) v^0 \right| \geq \left| v^{0T} \frac{X^{*T} X^*}{n} v^0 \right|,$$

which implies

$$\begin{aligned}
\left| v^{0T} \frac{\hat{X}^T \hat{X}}{n} v^0 \right| &\geq \left| v^{0T} \frac{X^{*T} X^*}{n} v^0 \right| - \left| v^{0T} \left(\frac{X^{*T} X^* - \hat{X}^T \hat{X}}{n} \right) v^0 \right| \\
&\geq \left| v^{0T} \frac{X^{*T} X^*}{n} v^0 \right| - \left| \frac{X^{*T} X^* - \hat{X}^T \hat{X}}{n} \right|_\infty \left| v^0 \right|_1^2 \\
&\geq \left| v^{0T} \frac{X^{*T} X^*}{n} v^0 \right| - \left(\left| \frac{X^{*T}(\hat{X} - X^*)}{n} \right|_\infty + \left| \frac{(\hat{X} - X^*)^T \hat{X}}{n} \right|_\infty \right) \left| v^0 \right|_1^2 \\
&\geq \left| v^{0T} \frac{X^{*T} X^*}{n} v^0 \right| - \left| \frac{X^{*T}(\hat{X} - X^*)}{n} \right|_\infty \left| v^0 \right|_1^2 \\
&\quad - \left| \frac{(\hat{X} - X^*)^T X^*}{n} \right|_\infty \left| v^0 \right|_1^2 - \left| \frac{(\hat{X} - X^*)^T (\hat{X} - X^*)}{n} \right|_\infty \left| v^0 \right|_1^2.
\end{aligned}$$

To bound the term $\left| \frac{X^{*T}(\hat{X} - X^*)}{n} \right|_\infty$, first note that by Lemma 2.13 in Loh and Wainwright (2012), under Assumption 2.3.5 and the condition $n \gtrsim \max\{k_1 \log d,\ k_1 \log p\}$, we have

$$\max_{j=1,\ldots,p} \sqrt{\frac{1}{n} \sum_{i=1}^n \left[\mathbf{z}_{ij} (\hat{\pi}_j - \pi_j^*) \right]^2} \leq \frac{c \sigma_\eta \sigma_Z \sqrt{\lambda_{\max}(\Sigma_Z)}}{\lambda_{\min}(\Sigma_Z)} \sqrt{\frac{k_1 \log \max(d,\ p)}{n}}$$

with probability at least $1 - c_1 \exp(-c_2 \log \max(d,\, p))$ where

$$\begin{aligned}
\lambda_{\min}(\Sigma_Z) &= \min_{j=1,\ldots,p} \lambda_{\min}(\Sigma_{Z_j}), \\
\lambda_{\max}(\Sigma_Z) &= \max_{j=1,\ldots,p} \lambda_{\max}(\Sigma_{Z_j}).
\end{aligned}$$

As a consequence,

$$\begin{aligned}
\max_{j',j}\left|\frac{1}{n}\mathbf{x}_{j'}^{*T}(\hat{\mathbf{x}}_j-\mathbf{x}_j^*)\right| &= \left|\frac{1}{n}\sum_{i=1}^n x_{ij'}^*\mathbf{z}_{ij}(\hat{\pi}_j-\pi_j^*)\right| \\
&\le \sqrt{\frac{1}{n}\sum_{i=1}^n x_{ij'}^{*2}}\sqrt{\frac{1}{n}\sum_{i=1}^n\left[\mathbf{z}_{ij}(\hat{\pi}_j-\pi_j^*)\right]^2} \\
&\le c'\frac{\sigma_{X^*}\sigma_\eta\sigma_Z\sqrt{\lambda_{\max}(\Sigma_Z)}}{\lambda_{\min}(\Sigma_Z)}\sqrt{\frac{k_1\log\max(d,\,p)}{n}}
\end{aligned}$$

where the last inequality follows from an application of Lemma 2.6.7 and a union bound.

To bound the term $\left|\frac{(\hat{X}-X^*)^T(\hat{X}-X^*)}{n}\right|_\infty$, we have,

$$\left|\frac{(\hat{X}-X^*)^T(\hat{X}-X^*)}{n}\right|_\infty \le c''\frac{\sigma_\eta^2\sigma_Z^2\lambda_{\max}(\Sigma_Z)}{\lambda_{\min}^2(\Sigma_Z)}\frac{k_1\log\max(d,\,p)}{n}$$

with probability at least $1-c_1\exp(-c_2\log\max(p,\,d))$.

Putting everything together, under the condition $n\gtrsim\max\{k_1\log d,\,k_1\log p\}$ and applying Lemma 2.6.9 with $r=0$, we have

$$\begin{aligned}
\left|v^{0T}\frac{\hat{X}^T\hat{X}}{n}v^0\right| &\ge \left|v^{0T}\frac{X^{*T}X^*}{n}v^0\right| - c_0\frac{\sigma_{X^*}\sigma_\eta\sigma_Z\sqrt{\lambda_{\max}(\Sigma_Z)}}{\lambda_{\min}(\Sigma_Z)}\sqrt{\frac{k_1\log\max(d,\,p)}{n}}\left|v^0\right|_1^2 \\
&\ge \frac{\lambda_{\min}(\Sigma_{X^*})}{2}\left|v^0\right|_2^2 - c_0\frac{\sigma_{X^*}\sigma_\eta\sigma_Z\sqrt{\lambda_{\max}(\Sigma_Z)}}{\lambda_{\min}(\Sigma_Z)}\sqrt{\frac{k_1\log\max(d,\,p)}{n}}\left|v^0\right|_1^2
\end{aligned}$$

with probability at least $1-c_1\exp(-c_2\log\max(p,\,d))$ (given $d>n$ and $p>n$ is the regime of our interests). Notice the last inequality can be written in the form

$$\frac{|\hat{X}v^0|_2^2}{n}\ge\kappa_1|v^0|_2^2-c\kappa_2\sqrt{\frac{k_1\log\max(p,\,d)}{n}}|v^0|_1^2,\qquad\text{for all } v^0\in\mathbb{R}^p.$$

□

In proving Lemma 2.3.1, upon our choice of λ_n, we have shown

$$\hat{v}=\hat{\beta}_{H2SLS}-\beta^*\in\mathbb{C}(J(\beta^*),\,3),$$

which implies $|\hat{v}^0|_1^2\le 16|\hat{v}^0_{J(\beta^*)}|_1^2\le 16k_2|\hat{v}^0_{J(\beta^*)}|_2^2$. Therefore, if we have the scaling

$$\frac{1}{n}k_1k_2^2\log\max(p,\,d)=O(1),$$

so that

$$\kappa_2 k_2 \sqrt{\frac{k_1 \log \max(p,\, d)}{n}} < \kappa_1,$$

then,

$$\left|\hat{v}^{0T} \frac{\hat{X}^T \hat{X}}{n} \hat{v}^0\right| \geq c_0 \lambda_{\min}(\Sigma_{X^*}) \left|\hat{v}^0\right|_2^2,$$

provided σ_η, σ_Z, σ_{X^*}, and $\lambda_{\max}(\Sigma_Z)$ are bounded from above while $\lambda_{\min}(\Sigma_Z)$ and $\lambda_{\min}(\Sigma_{X^*})$ are bounded away from 0. The above inequality implies RE (3).

Lemma 2.6.2 (Upper bound on $|\frac{1}{n}\hat{X}^T e|_\infty$): Under Assumptions 2.1.1, 2.3.1-2.3.3, 2.3.5, and the condition $\frac{\max\{k_1 \log d,\, k_1 \log p\}}{n} = O(1)$, we have

$$|\frac{1}{n}\hat{X}^T e|_\infty \precsim \max\left\{\varphi_1 \sqrt{\frac{k_1 \log \max(p,\, d)}{n}},\, \varphi_2 \sqrt{\frac{\log p}{n}}\right\},$$

where

$$\begin{aligned} \varphi_1 &= \frac{\max\left\{\sigma_\eta,\, \sigma_{X^*},\, \sigma_\epsilon\right\} \sigma_Z \sigma_\eta \sqrt{\lambda_{\max}(\Sigma_Z)} |\beta^*|_1}{\lambda_{\min}(\Sigma_Z)\lambda_{\min}(\Sigma_{X^*})}, \\ \varphi_2 &= \max\left\{\frac{\sigma_{X^*}\sigma_\eta |\beta^*|_1}{\lambda_{\min}(\Sigma_{X^*})},\, \frac{\sigma_{X^*}\sigma_\epsilon}{\lambda_{\min}(\Sigma_{X^*})}\right\}, \end{aligned}$$

with probability at least $1 - c_1 \exp(-c_2 \log \min(p,\, d))$ for some universal constants c_1 and c_2.

Proof. We have

$$\begin{aligned} \frac{1}{n}\hat{X}^T e &= \frac{1}{n}\hat{X}^T \left[(X^* - \hat{X})\beta^* + \boldsymbol{\eta}\beta^* + \epsilon\right] \\ &= \frac{1}{n}X^{*T} \left[(X^* - \hat{X})\beta^* + \boldsymbol{\eta}\beta^* + \epsilon\right] + \frac{1}{n}(\hat{X} - X^*)^T \left[(X^* - \hat{X})\beta^* + \boldsymbol{\eta}\beta^* + \epsilon\right]. \end{aligned}$$

Hence,

$$\begin{aligned} |\frac{1}{n}\hat{X}^T e|_\infty \leq\ & |\frac{1}{n}X^{*T}(\hat{X} - X^*)\beta^*|_\infty + |\frac{1}{n}X^{*T}\boldsymbol{\eta}\beta^*|_\infty + |\frac{1}{n}X^{*T}\epsilon|_\infty \\ +\ & |\frac{1}{n}(\hat{X} - X^*)^T(\hat{X} - X^*)\beta^*|_\infty + |\frac{1}{n}(\hat{X} - X^*)^T\boldsymbol{\eta}\beta^*|_\infty + |\frac{1}{n}(\hat{X} - X^*)^T\epsilon|_\infty. \end{aligned} \tag{2.4}$$

We need to bound each of the terms on the right-hand-side of the above inequality. Let us first bound $|\frac{1}{n}X^{*T}(\hat{X}-X^*)\beta^*|_\infty$. We have

$$\frac{1}{n}X^{*T}(\hat{X}-X^*)\beta^* = \begin{bmatrix} \sum_{j=1}^{p}\beta_j^*\frac{1}{n}\sum_{i=1}^{n}x_{i1}^*(\hat{x}_{ij}-x_{ij}^*) \\ \vdots \\ \sum_{j=1}^{p}\beta_j^*\frac{1}{n}\sum_{i=1}^{n}x_{ip}^*(\hat{x}_{ij}-x_{ij}^*) \end{bmatrix}.$$

For any $j^{'}=1,...,p$, we have

$$\begin{aligned} |\sum_{j=1}^{p}\beta_j^*\frac{1}{n}\sum_{i=1}^{n}x_{ij^{'}}^*(\hat{x}_{ij}-x_{ij}^*)| &\leq \max_{j^{'},j}|\frac{1}{n}\sum_{i=1}^{n}x_{ij^{'}}^*(\hat{x}_{ij}-x_{ij}^*)||\beta^*|_1 \\ &= \left|\frac{X^{*T}(\hat{X}-X^*)}{n}\right|_\infty |\beta^*|_1. \end{aligned}$$

In proving Lemma 2.6.1, under the condition $\frac{\log\max(p,d)}{n}=O(1)$, we have,

$$\left|\frac{X^{*T}(\hat{X}-X^*)}{n}\right|_\infty \leq c\frac{\sigma_{X^*}\sigma_\eta\sigma_Z\sqrt{\lambda_{\max}(\Sigma_Z)}}{\lambda_{\min}(\Sigma_Z)}\sqrt{\frac{k_1\log\max(d,\,p)}{n}},$$

with probability at least $1-c_1\exp(-c_2\log\max(p,\,d))$. Therefore,

$$|\frac{1}{n}X^{*T}(\hat{X}-X^*)\beta^*|_\infty \leq c|\beta^*|_1\frac{\sigma_{X^*}\sigma_\eta\sigma_Z\sqrt{\lambda_{\max}(\Sigma_Z)}}{\lambda_{\min}(\Sigma_Z)}\sqrt{\frac{k_1\log\max(d,\,p)}{n}}.$$

The term $|\frac{1}{n}(\hat{X}-X^*)^T(\hat{X}-X^*)\beta^*|_\infty$ can be bounded using a similar argument and we have,

$$|\frac{1}{n}(\hat{X}-X^*)^T(\hat{X}-X^*)\beta^*|_\infty \leq c|\beta^*|_1\frac{\sigma_\eta^2\sigma_Z^2\lambda_{\max}(\Sigma_Z)}{\lambda_{\min}^2(\Sigma_Z)}\frac{k_1\log\max(d,\,p)}{n},$$

with probability at least $1-c_1\exp(-c_2\log\max(p,\,d))$. For the term $|\frac{1}{n}X^{*T}\boldsymbol{\eta}\beta^*|_\infty$, we have

$$\begin{aligned} |\frac{1}{n}X^{*T}\boldsymbol{\eta}\beta^*|_\infty &\leq \max_{j^{'},j}|\frac{1}{n}\sum_{i=1}^{n}x_{ij^{'}}^*\eta_{ij}||\beta^*|_1 \\ &\leq c\sigma_{X^*}\sigma_\eta|\beta^*|_1\sqrt{\frac{\log p}{n}}, \end{aligned}$$

with probability at least $1 - c_1 \exp(-c_2 \log p)$. The last inequality follows from Lemma 2.6.7 and Assumption 2.1.1 that $\mathbb{E}(\mathbf{z}_{ij'}\eta_{ij}) = \mathbf{0}$ for all $j^{'}$, j as well as Assumption 2.3.2 that η_j is an *i.i.d.* zero-mean sub-Gaussian vector with parameter σ_η^2 for $j = 1, ..., p$, and the random matrix $Z_j \in \mathbb{R}^{n\times d_j}$ is sub-Gaussian with parameters $(\Sigma_{Z_j},\, \sigma_Z^2)$ for $j = 1, ..., p$. For the term $|\frac{1}{n}(X^* - \hat{X})^T \boldsymbol{\eta}\beta^*|_\infty$, we have,

$$|\frac{1}{n}(X^* - \hat{X})^T \boldsymbol{\eta}\beta^*|_\infty \leq c|\beta^*|_1 \frac{\sigma_\eta^2 \sigma_Z \sqrt{\lambda_{\max}(\Sigma_Z)}}{\lambda_{\min}(\Sigma_Z)} \sqrt{\frac{k_1 \log \max(d,\, p)}{n}}$$

with probability at least $1 - c_1 \exp(-c_2 \log \max(p,\, d))$.

To bound the term $|\frac{1}{n}X^{*T}\epsilon|_\infty$, note under Assumptions 2.3.2 and 2.3.3 as well as Assumption 2.1.1 that $\mathbb{E}(\mathbf{z}_{ij}\epsilon_i) = \mathbf{0}$ for all $j = 1, ..., p$, again by Lemma 2.6.7,

$$|\frac{1}{n}X^{*T}\epsilon|_\infty \leq c\sigma_{X^*}\sigma_\epsilon \sqrt{\frac{\log p}{n}},$$

with probability at least $1 - c_1 \exp(-c_2 \log p)$.

For the term $|\frac{1}{n}(X^* - \hat{X})^T\epsilon|_\infty$, we have

$$|\frac{1}{n}(X^* - \hat{X})^T\epsilon|_\infty \leq c\frac{\sigma_\eta\sigma_Z\sigma_\epsilon\sqrt{\lambda_{\max}(\Sigma_Z)}}{\lambda_{\min}(\Sigma_Z)} \sqrt{\frac{k_1 \log \max(d,\, p)}{n}}$$

with probability at least $1 - c_1 \exp(-c_2 \log \max(p,\, d))$.

Putting everything together, under the condition $\frac{\max\{k_1 \log d,\, k_1 \log p\}}{n} = O(1)$, the claim in Lemma 2.6.2 follows. □

Combining Lemmas 2.3.1, 2.6.1, and 2.6.2, we have

$$|\hat{\beta}_{H2SLS} - \beta^*|_2 \precsim \max\{\varphi_1\sqrt{k_2}\sqrt{\frac{k_1 \log \max(d,\, p)}{n}},\, \varphi_2\sqrt{\frac{k_2 \log p}{n}}\},$$

with probability at least $1 - c_1 \exp(-c_2 \log \min(p,\, d))$ for some universal positive constants c_1 and c_2, which proves Theorem 2.3.2. □

2.6.3 Theorem 2.3.3

The verification of the RE condition in Definition 1.1 from Chapter 1 is done via Lemma 2.6.3. Combining Lemmas 2.6.2 and 2.6.3 yields the desired result.

Lemma 2.6.3 (RE condition): Let $r \in [0, 1]$. Under Assumptions 2.1.1, 2.3.1, 2.3.3-2.3.6, and the condition $n \gtrsim k_1^{3-2r} \log \max(p, d)$, we have, for some universal constants c, $c^{'}$, c_1, and c_2,

$$\frac{|\hat{X}v^0|_2^2}{n} \geq \left(\kappa_1 - c\kappa_2 k_1^{3/2-r}\sqrt{\frac{\log \max(p,\, d)}{n}}\right)|v^0|_2^2 - c^{'}\kappa_3 \frac{k_1^r \log \max(p,\, d)}{n}|v^0|_1^2,$$

for all $v^0 \in \mathbb{R}^p$, with probability at least $1 - c_1 \exp(-c_2 \log \max(p, d))$, where

$$\begin{aligned}
\kappa_1 &= \frac{\lambda_{\min}(\Sigma_{X^*})}{2}, \quad \kappa_2 = \max(b_2 b_1^{-1},\, b_3 b_1^{-2}), \quad \kappa_3 = \max\left\{b_0,\, b_2 b_1^{-1},\, b_3 b_1^{-2}\right\}, \\
b_0 &= \lambda_{\min}(\Sigma_{X^*}) \max\left\{\frac{\sigma_{X^*}^4}{\lambda_{\min}^2(\Sigma_{X^*})},\, 1\right\}, \\
b_1 &= \frac{\lambda_{\min}(\Sigma_Z)}{\sigma_\eta}, \\
b_2 &= \max\left\{\sigma_{X^*}\sigma_W,\, \sup_{v \in \mathbb{K}(2s,\,p)\times\mathbb{K}(k_1,\,d_1)\times\ldots\times\mathbb{K}(k_1,\,d_p)} \left|v^{0T}\left[\mathbb{E}(x_{1j'}^* \mathbf{z}_{1j} v^j)\right] v^0\right|\right\}, \\
b_3 &= \max\left\{\sigma_W^2,\, \sup_{v \in \mathbb{K}(2s,\,p)\times\mathbb{K}^2(k_1,\,d_1)\times\ldots\times\mathbb{K}^2(k_1,\,d_p)} \left|v^{0T}\left[\mathbb{E}(v^{j^{'}} \mathbf{z}_{1j'}^T \mathbf{z}_{1j} v^j)\right] v^0\right|\right\}.
\end{aligned}$$

Proof. Again,

$$\begin{aligned}
\left|v^{0T}\frac{\hat{X}^T\hat{X}}{n}v^0\right| &\geq \left|v^{0T}\frac{X^{*T}X^*}{n}v^0\right| - \left|v^{0T}\left(\frac{X^{*T}X^* - \hat{X}^T\hat{X}}{n}\right)v^0\right| \\
&\geq \left|v^{0T}\frac{X^{*T}X^*}{n}v^0\right| - \left(\left|v^{0T}\frac{X^{*T}(\hat{X}-X^*)}{n}v^0\right| + \left|v^{0T}\frac{(\hat{X}-X^*)^T\hat{X}}{n}v^0\right|\right) \\
&\geq \left|v^{0T}\frac{X^{*T}X^*}{n}v^0\right| - \left|v^{0T}\frac{X^{*T}(\hat{X}-X^*)}{n}v^0\right| \\
&\quad - \left|v^{0T}\frac{(\hat{X}-X^*)^T X^*}{n}v^0\right| - \left|v^{0T}\frac{(\hat{X}-X^*)^T(\hat{X}-X^*)}{n}v^0\right|.
\end{aligned}$$

To bound the above terms, I apply a discretization argument motivated by the idea in Loh and Wainwright (2012). This type of argument is often used in statistical problems requiring manipulating and controlling collections of random variables indexed by sets with an infinite number of elements. For the particular problem in this paper, I work with the product space $\mathbb{K}(2s,\, p) \times \mathbb{K}(k_1,\, d_1) \times \ldots \times \mathbb{K}(k_1,\, d_p)$ and $\mathbb{K}(2s,\, p) \times$

$\mathbb{K}^2(k_1, d_1) \times ... \times \mathbb{K}^2(k_1, d_p)$. For $s \geq 1$ and $L \geq 1$, recall the notation $\mathbb{K}(s, L) := \{v \in \mathbb{R}^L \,|\, |v|_2 \leq 1, |v|_0 \leq s\}$. Given $V^j \subseteq \{1, ..., d_j\}$ and $V^0 \subseteq \{1, ..., p\}$, define $S_{V^j} = \{v \in \mathbb{R}^{d_j} : |v|_2 \leq 1, J(v) \subseteq V^j\}$ and $S_{V^0} = \{v \in \mathbb{R}^p : |v|_2 \leq 1, J(v) \subseteq V^0\}$. Note that $\mathbb{K}(k_1, d_j) = \cup_{|V^j| \leq k_1} S_{V^j}$ and $\mathbb{K}(2s, p) = \cup_{|V^0| \leq 2s} S_{V^0}$ with $s = s(r) := \frac{1}{c} \frac{n}{k_1^r \log \max(p, d)} \min\left\{\frac{\lambda^2_{\min}(\Sigma_{X^*})}{\sigma^4_{X^*}}, 1\right\}$, $r \in [0, 1]$. The choice of s is explained in the proof for Lemma 2.6.9. If $\mathcal{V}^j = \{t_1^j, ..., t_{m_j}^j\}$ is a $\frac{1}{9}$-cover of S_{V^j} ($\mathcal{V}^0 = \{t_1^0, ..., t_{m_0}^0\}$ is a $\frac{1}{9}$-cover of S_{V^0}), for every $v^j \in S_{V^j}$ ($v^0 \in S_{V^0}$), we can find some $t_i^j \in \mathcal{V}^j$ ($t_{i'}^0 \in \mathcal{V}^0$) such that $|\triangle v^j|_2 \leq \frac{1}{9}$ ($|\triangle v^0|_2 \leq \frac{1}{9}$), where $\triangle v^j = v^j - t_i^j$ (respectively, $\triangle v^0 = v^0 - t_{i'}^0$). By Ledoux and Talagrand (1991), we can construct $\mathcal{V}^j$ with $|\mathcal{V}^j| \leq 81^{k_1}$ and $|\mathcal{V}^0| \leq 81^{2s}$. Therefore, for $v^0 \in \mathbb{K}(2s, p)$, there is some S_{V^0} and $t_{i'}^0 \in \mathcal{V}^0$ such that

$$
\begin{aligned}
v^{0T} \frac{X^{*T}(\hat{X} - X^*)}{n} v^0 &= (t_{i'}^0 + v^0 - t_{i'}^0)^T \frac{X^{*T}(\hat{X} - X^*)}{n} (t_{i'}^0 + v^0 - t_{i'}^0) \\
&= t_{i'}^{0T} \frac{X^{*T}(\hat{X} - X^*)}{n} t_{i'}^0 + 2\triangle v^{0T} \frac{X^{*T}(\hat{X} - X^*)}{n} t_{i'}^0 \\
&\quad + \triangle v^{0T} \frac{X^{*T}(\hat{X} - X^*)}{n} \triangle v^0
\end{aligned}
$$

with $|\triangle v^0|_2 \leq \frac{1}{9}$. Recall for the (j', j) element of the matrix $\frac{X^{*T}(\hat{X} - X^*)}{n}$, we have

$$
\frac{1}{n} \mathbf{x}_{j'}^{*T} (\hat{\mathbf{x}}_j - \mathbf{x}_j^*) = \left(\frac{1}{n} \sum_{i=1}^n x_{ij'}^* \mathbf{z}_{ij} \right) (\hat{\pi}_j - \pi_j^*).
$$

Let $\frac{\lambda_{\min}(\Sigma_Z)}{c\sigma_\eta} = b_1$. Notice that, under Assumptions 2.3.5 and 2.3.6,

$$
|\hat{\pi}_j - \pi_j^*|_2 b_1 \sqrt{\frac{n}{k_1 \log \max(p, d)}} \leq 1
$$

and $|\mathrm{supp}(\hat{\pi}_j - \pi_j^*)| \leq k_1$ for every $j = 1, ..., p$. Define $\bar{\pi}_j = (\hat{\pi}_j - \pi_j^*) b_1 \sqrt{\frac{n}{k_1 \log \max(p, d)}}$ and hence, $\bar{\pi}_j \in \mathbb{K}(k_1, d_j) = \cup_{|V^j| \leq k_1} S_{V^j}$. Therefore, there is some S_{V^j} with $|V^j| \leq k_1$ and $t_i^j \in \mathcal{V}^j$ (where $\mathcal{V}^j = \{t_1^j, ..., t_{m_j}^j\}$ is a $\frac{1}{9}$-cover of S_{V^j}) such that

$$
\begin{aligned}
\frac{1}{n} x_{j'}^{*T} \mathbf{z}_j (\hat{\pi}_j - \pi_j^*) &= \frac{1}{n} x_{j'}^{*T} \mathbf{z}_j (t_i^j + \bar{\pi}_j - t_i^j) b_1^{-1} \sqrt{\frac{k_1 \log \max(p, d)}{n}} \\
&= b_1^{-1} \sqrt{\frac{k_1 \log \max(p, d)}{n}} \left(\frac{1}{n} x_{j'}^{*T} \mathbf{z}_j t_i^j + \frac{1}{n} x_{j'}^{*T} \mathbf{z}_j \triangle v^j \right)
\end{aligned}
$$

with $|\triangle v^j|_2 \leq \frac{1}{9}$. Denote a matrix A by $\left[A_{j'j}\right]$, where the $(j',\, j)$ element of A is $A_{j'j}$. Define $v = (v^0,\, v^1,\, ...,\, v^p) \in S_V := S_{V^0} \times S_{V^1} \times ... \times S_{V^p}$. Hence,

$$\left| v^{0T} \frac{X^{*T}(\hat{X} - X^*)}{n} v^0 - \mathbb{E}(v^{0T} \frac{X^{*T}(\hat{X} - X^*)}{n} v^0) \right|$$

$$\begin{aligned}
&\leq \sup_{v \in S_V} b_1^{-1} \sqrt{\frac{k_1 \log \max(p,\, d)}{n}} \left| v^{0T} \left[\frac{1}{n} x_{j'}^{*T} \mathbf{z}_j v^j - \mathbb{E}(x_{1j'}^{*} \mathbf{z}_{1j} v^j) \right] v^0 \right| \\
&\leq b_1^{-1} \sqrt{\frac{k_1 \log \max(p,\, d)}{n}} \{ \max_{i',\, i} \left| t_{i'}^{0T} \left[\frac{1}{n} x_{j'}^{*T} \mathbf{z}_j t_i^j - \mathbb{E}(x_{1j'}^{*} \mathbf{z}_{1j} t_i^j) \right] t_{i'}^0 \right| \\
&\quad + \sup_{v \in S_V} \left| t_{i'}^{0T} \left[\frac{1}{n} x_{j'}^{*T} \mathbf{z}_j \triangle v^j - \mathbb{E}(x_{1j'}^{*} \mathbf{z}_{1j} \triangle v^j) \right] t_{i'}^0 \right| \\
&\quad + \sup_{v \in S_V} 2 \left| \triangle v^{0T} \left[\frac{1}{n} x_{j'}^{*T} \mathbf{z}_j t_i^j - \mathbb{E}(x_{1j'}^{*} \mathbf{z}_{1j} t_i^j) \right] t_{i'}^0 \right| \\
&\quad + \sup_{v \in S_V} 2 \left| \triangle v^{0T} \left[\frac{1}{n} x_{j'}^{*T} \mathbf{z}_j \triangle v^j - \mathbb{E}(x_{1j'}^{*} \mathbf{z}_{1j} \triangle v^j) \right] t_{i'}^0 \right| \\
&\quad + \sup_{v \in S_V} \left| \triangle v^{0T} \left[\frac{1}{n} x_{j'}^{*T} \mathbf{z}_j t_i^j - \mathbb{E}(x_{1j'}^{*} \mathbf{z}_{1j} t_i^j) \right] \triangle v^0 \right| \\
&\quad + \sup_{v \in S_V} \left| \triangle v^{0T} \left[\frac{1}{n} x_{j'}^{*T} \mathbf{z}_j \triangle v^j - \mathbb{E}(x_{1j'}^{*} \mathbf{z}_{1j} \triangle v^j) \right] \triangle v^0 \right| \} \\
&\leq b_1^{-1} \sqrt{\frac{k_1 \log \max(p,\, d)}{n}} \{ \max_{i',\, i} \left| t_{i'}^{0T} \left[\frac{1}{n} x_{j'}^{*T} \mathbf{z}_j t_i^j - \mathbb{E}(x_{1j'}^{*} \mathbf{z}_{1j} t_i^j) \right] t_{i'}^0 \right| \\
&\quad + \sup_{v \in S_V} \frac{1}{9} \left| v^{0T} \left[\frac{1}{n} x_{j'}^{*T} \mathbf{z}_j v^j - \mathbb{E}(x_{1j'}^{*} \mathbf{z}_{1j} v^j) \right] v^0 \right| \\
&\quad + \sup_{v \in S_V} \frac{2}{9} \left| v^{0T} \left[\frac{1}{n} x_{j'}^{*T} \mathbf{z}_j v^j - \mathbb{E}(x_{1j'}^{*} \mathbf{z}_{1j} v^j) \right] v^0 \right| \\
&\quad + \sup_{v \in S_V} \frac{2}{81} \left| v^{0T} \left[\frac{1}{n} x_{j'}^{*T} \mathbf{z}_j v^j - \mathbb{E}(x_{1j'}^{*} \mathbf{z}_{1j} v^j) \right] v^0 \right| \\
&\quad + \sup_{v \in S_V} \frac{1}{81} \left| v^{0T} \left[\frac{1}{n} x_{j'}^{*T} \mathbf{z}_j v^j - \mathbb{E}(x_{1j'}^{*} \mathbf{z}_{1j} v^j) \right] v^0 \right| \\
&\quad + \sup_{v \in S_V} \frac{1}{729} \left| v^{0T} \left[\frac{1}{n} x_{j'}^{*T} \mathbf{z}_j v^j - \mathbb{E}(x_{1j'}^{*} \mathbf{z}_{1j} v^j) \right] v^0 \right| \},
\end{aligned}$$

where the last inequality uses the fact that $9\triangle v^j \in S_{V^j}$ and $9\triangle v^0 \in S_{V^0}$. Therefore,

$$\sup_{v\in S_V} b_1^{-1}\sqrt{\frac{k_1\log\max(p,\, d)}{n}}\left|v^{0T}\left[\frac{1}{n}x_{j'}^{*T}\mathbf{z}_j v^j - \mathbb{E}(x_{1j'}^{*}\mathbf{z}_{1j}v^j)\right]v^0\right|$$

$$\begin{aligned} &\leq \frac{729}{458}b_1^{-1}\sqrt{\frac{k_1\log\max(p,\, d)}{n}}\max_{i',\,i} t_{i'}^{0T}\left[\frac{1}{n}x_{j'}^{*T}\mathbf{z}_j t_i^j - \mathbb{E}(x_{1j'}^{*}\mathbf{z}_{1j}t_i^j)\right]t_{i'}^0 \\ &\leq 2b_1^{-1}\sqrt{\frac{k_1\log\max(p,\, d)}{n}}\max_{i',\,i} t_{i'}^{0T}\left[\frac{1}{n}x_{j'}^{*T}\mathbf{z}_j t_i^j - \mathbb{E}(x_{1j'}^{*}\mathbf{z}_{1j}t_i^j)\right]t_{i'}^0. \end{aligned}$$

Under Assumptions 2.3.3 and 2.3.4, $x_{j'}^{*}$ is a sub-Gaussian vector with parameter at most σ_{X^*} for every $j^{'} = 1, ..., p$, and $Z_j t_i^j := \mathbf{w}_j$ is a sub-Gaussian vector with parameter at most σ_{W^*}. An application of Lemma 2.6.7 and a union bound yields

$$\mathbb{P}\left(\sup_{v\in S_V}\left|v^{0T}\left[\frac{1}{n}x_{j'}^{*T}\mathbf{z}_j v^j\right]v^0 - v^{0T}\left[\mathbb{E}(x_{1j'}^{*}\mathbf{z}_{1j}v^j)\right]v^0\right| \geq t\right) \leq$$

$$81^{2sk_1}81^{2s}2\exp(-cn\min(\frac{t^2}{\sigma_{X^*}^2\sigma_W^2},\ \frac{t}{\sigma_{X^*}\sigma_W})),$$

where the exponent $2sk_1$ in 81^{2sk_1} uses the fact that there are at most $2s$ non-zero components in $v^0 \in S_{V^0}$ and hence only $2s$ out of p entries of $v^1, ..., v^p$ will be multiplied by a non-zero scalar, which leads to a reduction of dimensions. Let

$$\Omega = \mathbb{K}(2s,\, p)\times\mathbb{K}(k_1,\, d_1)\times ...\times\mathbb{K}(k_1,\, d_p).$$

A second application of a union bound over the $\begin{pmatrix} d_j \\ \lfloor k_1 \rfloor \end{pmatrix} \leq d^{k_1}$ choices of V^j and respectively, the $\begin{pmatrix} p \\ \lfloor 2s \rfloor \end{pmatrix} \leq p^{2s}$ choices of V^0 yields

$$\mathbb{P}\left(\sup_{v\in\Omega}\left|v^{0T}\left[\frac{1}{n}x_{j'}^{*T}\mathbf{z}_j v^j\right]v^0 - v^{0T}\left[\mathbb{E}(x_{1j'}^{*}\mathbf{z}_{1j}v^j)\right]v^0\right| \geq t\right)$$

$$\begin{aligned} &\leq p^{2s}d^{2sk_1}\cdot 2\exp(-cn\min(\frac{t^2}{\sigma_{X^*}^2\sigma_W^2},\ \frac{t}{\sigma_{X^*}\sigma_W})) \\ &\leq 2\exp(-cn\min(\frac{t^2}{\sigma_{X^*}^2\sigma_W^2},\ \frac{t}{\sigma_{X^*}\sigma_W}) + 2sk_1\log d + 2s\log p). \end{aligned}$$

With the choice of $s = s(r) := \frac{1}{c}\frac{n}{k_1^r \log\max(p,d)} \min\left\{\frac{\lambda^2_{\min}(\Sigma_{X^*})}{\sigma^4_{X^*}},\, 1\right\}$, $r \in [0, 1]$ from the proof for Lemma 6.9 and $t = c^{'}k_1^{1-r}\sigma_{X^*}\sigma_W$ for some universal constant $c^{'} \geq 1$, we have

$$\left| v^{0T}\frac{X^{*T}(\hat{X}-X^*)}{n}v^0 - \mathbb{E}[v^{0T}\frac{X^{*T}(\hat{X}-X^*)}{n}v^0]\right|$$

$$\begin{aligned}
&\leq \left(\sup_{v\in\Omega}\left|v^{0T}\left[\frac{1}{n}x^{*T}_{j^{'}}\mathbf{z}_j v^j - \mathbb{E}(x^*_{1j^{'}}\mathbf{z}_{1j}v^j)\right]v^0\right|\right) b_1^{-1}\sqrt{\frac{k_1\log\max(p,\, d)}{n}} \\
&\leq c^{'}b_1^{-1}k_1^{1-r}\sqrt{\frac{k_1\log\max(p,\, d)}{n}}\sigma_{X^*}\sigma_W
\end{aligned}$$

with probability at least

$$1 - c_1^{'}\exp(-c_2^{'}nk_1^{1-r}) - c_1^{''}\exp(-c_2^{''}\log\max(p,\, d)) = 1 - c_1\exp(-c_2\log\max(p,\, d))$$

(given $d > n$ and $p > n$ is the regime of our interests). Therefore, we have

$$\begin{aligned}
\left|v^{0T}\frac{X^{*T}(\hat{X}-X^*)}{n}v^0\right| &\leq \left(\sup_{v\in\Omega}\left|v^{0T}\left[\mathbb{E}(x^*_{1j^{'}}\mathbf{z}_{1j}v^j)\right]v^0\right|\right) b_1^{-1}\sqrt{\frac{k_1\log\max(p,\, d)}{n}} \\
&+ c^{'}b_1^{-1}k_1^{3/2-r}\sqrt{\frac{\log\max(p,\, d)}{n}}\sigma_{X^*}\sigma_W \\
&\leq cb_2b_1^{-1}k_1^{3/2-r}\sqrt{\frac{\log\max(p,\, d)}{n}},
\end{aligned}$$

where $b_2 = \max\left\{\sigma_{X^*}\sigma_W,\, \sup_{v\in\Omega}\left|v^{0T}\left[\mathbb{E}(x^*_{1j^{'}}\mathbf{z}_{1j}v^j)\right]v^0\right|\right\}$. Notice that the term

$$\sup_{v\in\Omega}\left|v^{0T}\left[\mathbb{E}(x^*_{1j^{'}}\mathbf{z}_{1j}v^j)\right]v^0\right|$$

is bounded above by the spectral norm of the matrix $\left[\mathbb{E}(x^*_{1j^{'}}\mathbf{z}_{1j}v^j)\right]$ for some $v^1\times...\times v^p \in \mathbb{K}(k_1,\, d_1)\times ...\times \mathbb{K}(k_1,\, d_p)$.

The term $\left|v^{0T}\frac{(\hat{X}-X^*)^T(\hat{X}-X^*)}{n}v^0\right|$ can be bounded using a similar argument. In particular, for the (j', j) element of the matrix $\frac{(\hat{X}-X^*)^T(\hat{X}-X^*)}{n}$, we have

$$\begin{aligned}
\frac{1}{n}(\hat{\mathbf{x}}_{j'}-\mathbf{x}^*_{j'})^T(\hat{\mathbf{x}}_j-\mathbf{x}^*_j) &= (\hat{\pi}_{j'}-\pi^*_{j'})^T\left(\frac{1}{n}\sum_{i=1}^n \mathbf{z}_{ij'}\mathbf{z}_{ij}\right)(\hat{\pi}_j-\pi^*_j) \\
&= \frac{1}{n}(t^{j'}_{i'}+\bar{\pi}_{j'}-t^{j'}_{i'})^T\mathbf{z}^T_{j'}\mathbf{z}_j(t^j_i+\bar{\pi}_j-t^j_i)b_1^{-2}\frac{k_1\log\max(p,\, d)}{n} \\
&= b_1^{-2}\frac{k_1\log\max(p,\, d)}{n}\{\frac{1}{n}t^{j'T}_{i'}\mathbf{z}^T_{j'}\mathbf{z}_j t^j_i+\frac{1}{n}\triangle v^{j'T}\mathbf{z}^T_{j'}\mathbf{z}_j t^j_i \\
&+ \frac{1}{n}t^{j'T}_{i'}\mathbf{z}^T_{j'}\mathbf{z}_j\triangle v^j+\frac{1}{n}\triangle v^{j'T}\mathbf{z}^T_{j'}\mathbf{z}_j\triangle v^j\}
\end{aligned}$$

Combining with

$$\begin{aligned}
v^{0T}\frac{(\hat{X}-X^*)^T(\hat{X}-X^*)}{n}v^0 &= t^{0T}_{i''}\frac{(\hat{X}-X^*)^T(\hat{X}-X^*)}{n}t^0_{i''} \\
&+2\triangle v^{0T}\frac{(\hat{X}-X^*)^T(\hat{X}-X^*)}{n}t^0_{i''} \\
&+\triangle v^{0T}\frac{(\hat{X}-X^*)^T(\hat{X}-X^*)}{n}\triangle v^0,
\end{aligned}$$

Define $S_V := S_{V^0}\times S^2_{V^1}\times ...\times S^2_{V^p}$. After some tedious algebra, we obtain

$$\left|v^{0T}\frac{(\hat{X}-X^*)^T(\hat{X}-X^*)}{n}v^0-\mathbb{E}(v^{0T}\frac{(\hat{X}-X^*)^T(\hat{X}-X^*)}{n}v^0)\right|$$

$$\begin{aligned}
&\leq \sup_{v\in S_V} b_1^{-2}\frac{k_1\log\max(p,\, d)}{n}\left|v^{0T}\left[\frac{1}{n}v^{j'}\mathbf{z}^T_{j'}\mathbf{z}_j v^j - v^{j'}\mathbb{E}(\mathbf{z}^T_{1j'}\mathbf{z}_{1j})v^j\right]v^0\right| \\
&\leq b_1^{-2}\frac{k_1\log\max(p,\, d)}{n}\{\max_{i'',i',i}\left|t^{0T}_{i''}\left[\frac{1}{n}t^{j'T}_{i'}\mathbf{z}^T_{j'}\mathbf{z}_j t^j_i-\mathbb{E}(t^{j'}_{i'}\mathbf{z}^T_{1j'}\mathbf{z}_{1j}t^j_i)\right]t^0_{i''}\right| \\
&+\frac{3439}{6561}\sup_{v\in S_V}\left|v^{0T}\left[\frac{1}{n}v^{j'T}\mathbf{z}^T_{j'}\mathbf{z}_j v^j-\mathbb{E}(v^{j'}\mathbf{z}^T_{1j'}\mathbf{z}_{1j}v^j)\right]v^0\right|\}.
\end{aligned}$$

Hence,

$$\sup_{v\in S_V} b_1^{-2}\frac{k_1\log\max(p,\, d)}{n}\left|v^{0T}\left[\frac{1}{n}v^{j'}\mathbf{z}^T_{j'}\mathbf{z}_j v^j-\mathbb{E}(v^{j'}\mathbf{z}^T_{1j'}\mathbf{z}_{1j}v^j)\right]v^0\right|$$

$$\leq \frac{6561}{3122} b_1^{-2} \frac{k_1 \log \max(p,\, d)}{n} \max_{i'',i',i} \left| t_{i''}^{0T} \left[\frac{1}{n} t_{i'}^{j'T} \mathbf{z}_{j'}^T \mathbf{z}_j t_i^j - \mathbb{E}(t_{i'}^{j'} \mathbf{z}_{1j'}^T \mathbf{z}_{1j} t_i^j) \right] t_{i''}^0 \right|$$
$$\leq 3 b_1^{-2} \frac{k_1 \log \max(p,\, d)}{n} \max_{i'',i',i} \left| t_{i''}^{0T} \left[\frac{1}{n} t_{i'}^{j'T} \mathbf{z}_{j'}^T \mathbf{z}_j t_i^j - \mathbb{E}(t_{i'}^{j'} \mathbf{z}_{1j'}^T \mathbf{z}_{1j} t_i^j) \right] t_{i''}^0 \right|.$$

Let

$$\Omega_1 = \mathbb{K}(2s,\, p) \times \mathbb{K}^2(k_1,\, d_1) \times \ldots \times \mathbb{K}^2(k_1,\, d_p).$$

An application of Lemma 2.6.7 and a sequence of union bounds yields

$$\mathbb{P}\left(\sup_{v \in \Omega_1} \left| v^{0T} \left[\frac{1}{n} v^{j'} \mathbf{z}_{j'}^T \mathbf{z}_j v^j \right] v^0 - v^{0T} \left[\mathbb{E}(v^{j'} \mathbf{z}_{1j'}^T \mathbf{z}_{1j} v^j) \right] v^0 \right| \geq t \right)$$
$$\leq 2 \exp(-cn \min(\frac{t^2}{\sigma_W^4},\, \frac{t}{\sigma_W^2}) + 4 s k_1 \log d + 2 s \log p).$$

Under the choice of $s = s(r) := \frac{1}{c} \frac{n}{k_1^r \log \max(p, d)} \min \left\{ \frac{\lambda_{\min}^2(\Sigma_{X^*})}{\sigma_{X^*}^4},\, 1 \right\}$, $r \in [0,\, 1]$ from the proof for Lemma 2.6.9 and $t = c^{''} k_1^{1-r} \sigma_W^2$ for some universal constant $c^{''} \geq 1$, we have,

$$\left| v^{0T} \frac{(\hat{X} - X^*)^T (\hat{X} - X^*)}{n} v^0 - \mathbb{E}[v^{0T} \frac{(\hat{X} - X^*)^T (\hat{X} - X^*)}{n} v^0] \right|$$

$$\leq \left(\sup_{v \in \Omega_1} \left| v^{0T} \left[\frac{1}{n} v^{j'} \mathbf{z}_{j'}^T \mathbf{z}_j v^j \right] v^0 - v^{0T} \left[\mathbb{E}(v^{j'} \mathbf{z}_{1j'}^T \mathbf{z}_{1j} v^j) \right] v^0 \right| \right) b_1^{-2} \frac{k_1 \log \max(p,\, d)}{n}$$
$$\leq c^{''} b_1^{-2} \frac{k_1^{2-r} \log \max(p,\, d)}{n} \sigma_W^2$$

with probability at least

$$1 - c_1^{'} \exp(-c_2^{'} n k_1^{1-r}) - c_1^{''} \exp(-c_2^{''} \log \max(p,\, d)) = 1 - c_1 \exp(-c_2 \log \max(p,\, d))$$

(given $d > n$ and $p > n$ is the regime of our interests). Therefore, we have

$$\left| v^{0T} \frac{(\hat{X} - X^*)^T (\hat{X} - X^*)}{n} v^0 \right| \leq$$

$$\left(\sup_{v \in \Omega_1} \left| v^{0T} \left[\mathbb{E}(v^{j'} \mathbf{z}_{1j'}^T \mathbf{z}_{1j} v^j) \right] v^0 \right| \right) b_1^{-2} \frac{k_1 \log \max(p,\, d)}{n} + c^{''} b_1^{-2} \frac{k_1^{2-r} \log \max(p,\, d)}{n} \sigma_W^2$$

$$\leq cb_3 b_1^{-2} \frac{k_1^{2-r} \log \max(p,\, d)}{n},$$

where $b_3 = \max\left\{\sigma_W^2,\ \sup_{v\in\Omega_1}\left|v^{0T}\left[\mathbb{E}(v^{j'}\mathbf{z}_{1j'}^T \mathbf{z}_{1j} v^j)\right] v^0\right|\right\}$. Notice that the term

$$\sup_{v\in\Omega_1}\left|v^{0T}\left[\mathbb{E}(v^{j'}\mathbf{z}_{1j'}^T \mathbf{z}_{1j} v^j)\right] v^0\right|$$

is bounded above by the spectral norm of the matrix $\left[\mathbb{E}(v^{j'}\mathbf{z}_{1j'}^T \mathbf{z}_{1j} v^j)\right]$ for some $(v^1 \times ... \times v^p) \times (v^1 \times ... \times v^p) \in \mathbb{K}^2(k_1,\, d_1) \times ... \times \mathbb{K}^2(k_1,\, d_p)$.

By Lemma 2.6.8, the bound

$$\left|v^{0T} \frac{X^{*T}(\hat{X} - X^*)}{n} v^0\right| \leq cb_2 b_1^{-1} k_1^{3/2-r} \sqrt{\frac{\log \max(p,\, d)}{n}} \qquad \forall v^0 \in \mathbb{K}(2s,\, p)$$

implies

$$\left|v^{0T} \frac{X^{*T}(\hat{X} - X^*)}{n} v^0\right| \leq 27cb_2 b_1^{-1} k_1^{3/2-r} \sqrt{\frac{\log \max(p,\, d)}{n}} (|v^0|_2^2 + \frac{1}{s}|v^0|_1^2) \qquad \forall v^0 \in \mathbb{R}^p. \tag{2.5}$$

Similarly, the bound

$$\left|v^{0T} \frac{(\hat{X} - X^*)^T(\hat{X} - X^*)}{n} v^0\right| \leq c^{''} b_3 b_1^{-2} \frac{k_1^{2-r} \log \max(p,\, d)}{n}$$

for all $v^0 \in \mathbb{K}(2s,\, p)$ implies

$$\left|v^{0T} \frac{(\hat{X} - X^*)^T(\hat{X} - X^*)}{n} v^0\right| \leq 27c^{''} b_3 b_1^{-2} \frac{k_1^{2-r} \log \max(p,\, d)}{n} (|v^0|_2^2 + \frac{1}{s}|v^0|_1^2) \qquad \forall v^0 \in \mathbb{R}^p. \tag{2.6}$$

Therefore, applying Lemma 2.6.9 by choosing

$$s = s(r) := \frac{1}{c} \frac{n}{k_1^r \log \max(p,\, d)} \min\left\{\frac{\lambda_{\min}^2(\Sigma_{X^*})}{,} \sigma_{X^*}^4,\, 1\right\}$$

$r \in [0,\, 1]$, under the condition $n \succsim k_1^{3-2r} \log \max(p,\, d)$, we have

$$\frac{|\hat{X} v^0|_2^2}{n} \geq \left(\kappa_1 - c\kappa_2 k_1^{3/2-r} \sqrt{\frac{\log \max(p,\, d)}{n}}\right) |v^0|_2^2 - c^{'} \kappa_3 \frac{k_1^r \log \max(p,\, d)}{n} |v^0|_1^2,$$

for all $v^0 \in \mathbb{R}^p$, with probability at least $1 - c_1 \exp(-c_2 \log \max(p,\, d))$, where κ_1, κ_2, and κ_2 are defined in the statement of Lemma 2.6.3.□

Again, recalling in proving Lemma 2.3.1, upon our choice λ_n, we have shown

$$\hat{v} = \hat{\beta}_{H2SLS} - \beta^* \in \mathbb{C}(J(\beta^*),\, 3),$$

and $|\hat{v}^0|_1^2 \leq 16|\hat{v}^0_{J(\beta^*)}|_1^2 \leq 16k_2|\hat{v}^0_{J(\beta^*)}|_2^2$. Therefore, if we have the scaling

$$\frac{\min_{r\in[0,\,1]} \max\left\{k_1^{3-2r}\log d,\, k_1^{3-2r}\log p,\, k_1^r k_2 \log d,\, k_1^r k_2 \log p\right\}}{n} = O(1),$$

so that

$$c\kappa_2 k_1^{3/2-r}\sqrt{\frac{\log \max(p,\, d)}{n}} + c^{'}\kappa_3 \frac{k_2 k_1^r \log \max(p,\, d)}{n} < \kappa_1,$$

then,

$$\left|\hat{v}^{0T}\frac{\hat{X}^T\hat{X}}{n}\hat{v}^0\right| \geq c_0 \lambda_{\min}(\Sigma_{X^*}) \left|\hat{v}^0\right|_2^2.$$

The above inequality implies the RE condition in Definition 1.1 from Chapter 1. Because the argument for showing Lemma 2.6.1 and that it implies the RE condition in Definition 1.1 from Chapter 1 also works under the assumptions of Lemma 2.6.3, we can combine the scaling $\frac{k_1 k_2^2 \log \max(p,\, d)}{n} = O(1)$ from the proof for Lemma 2.6.1 with the scaling $\frac{\min_{r\in[0,\,1]} \max\left\{k_1^{3-2r}\log d,\, k_1^{3-2r}\log p,\, k_1^r k_2 \log d,\, k_1^r k_2 \log p\right\}}{n} = O(1)$ from above to obtain a more optimal scaling of the required sample size

$\min\left\{k_1 k_2^2 \log \max(p,\, d),\, \min_{r\in[0,\,1]} \max\left\{k_1^{3-2r}\log d,\, k_1^{3-2r}\log p,\, k_1^r k_2 \log d,\, k_1^r k_2 \log p\right\}\right\} = O(n)$.

2.6.4 Corollary 2.3.4 and Theorem 2.3.5

Corollary 2.3.4 is obvious from Theorem 2.3.3. The proof for Theorem 2.3.5 is completely identical to that for Theorem 2.3.2 except we replace the inequality in Assumption 2.3.5 by

$$\max_{j=1,\ldots,p} |\hat{\pi}_j - \pi_j^*|_2 \leq M(d,\, p,\, k_1,\, n).$$

2.6.5 Lemma 2.6.4

Lemma 2.6.4: Suppose Assumption 2.3.7 holds. Let $J(\beta^*) = K$, $\Sigma_{K^cK} := \mathbb{E}\left[X^{*T}_{1,K^c} X^*_{1,K}\right]$, $\hat{\Sigma}_{K^cK} := \frac{1}{n} X^{*T}_{K^c} X^*_K$, and $\tilde{\Sigma}_{K^cK} := \frac{1}{n}\hat{X}^T_{K^c}\hat{X}_K$. Similarly, let $\Sigma_{KK} := \mathbb{E}\left[X^{*T}_{1,K} X^*_{1,K}\right]$, $\hat{\Sigma}_{KK} := \frac{1}{n} X^{*T}_K X^*_K$, and $\tilde{\Sigma}_{KK} := \frac{1}{n}\hat{X}^T_K\hat{X}_K$. If the assumptions in Lemmas 2.6.1 and 2.6.2 hold, then under the condition

$$\begin{aligned} \frac{1}{n}k_2^3 \log p &= O(1), \\ \frac{1}{n}k_1 k_2^2 \log \max(p,\, d) &= o(1), \end{aligned}$$

the sample matrix $\frac{1}{n}\hat{X}^T\hat{X}$ satisfies an analogous version of the "mutual incoherence" assumption with high probability, i.e.,

$$\mathbb{P}\left[\left\|\frac{1}{n}\hat{X}^T_{K^c}\hat{X}_K\left(\frac{1}{n}\hat{X}^T_K\hat{X}_K\right)^{-1}\right\|_\infty \geq 1 - \frac{\phi}{4}\right] \leq O\left(\frac{1}{\min(p,\, d)}\right).$$

Proof. I use the following decomposition similar to the method used in Ravikumar, et. al. (2010)

$$\tilde{\Sigma}_{K^cK}\tilde{\Sigma}^{-1}_{KK} - \Sigma_{K^cK}\Sigma^{-1}_{KK} = R_1 + R_2 + R_3 + R_4 + R_5 + R_6,$$

where

$$\begin{aligned} R_1 &= \Sigma_{K^cK}[\hat{\Sigma}^{-1}_{KK} - \Sigma^{-1}_{KK}], \\ R_2 &= [\hat{\Sigma}_{K^cK} - \Sigma_{K^cK}]\Sigma^{-1}_{KK}, \\ R_3 &= [\hat{\Sigma}_{K^cK} - \Sigma_{K^cK}][\hat{\Sigma}^{-1}_{KK} - \Sigma^{-1}_{KK}], \\ R_4 &= \hat{\Sigma}_{K^cK}[\tilde{\Sigma}^{-1}_{KK} - \hat{\Sigma}^{-1}_{KK}], \\ R_5 &= [\tilde{\Sigma}_{K^cK} - \hat{\Sigma}_{K^cK}]\hat{\Sigma}^{-1}_{KK}, \\ R_6 &= [\tilde{\Sigma}_{K^cK} - \hat{\Sigma}_{K^cK}][\tilde{\Sigma}^{-1}_{KK} - \hat{\Sigma}^{-1}_{KK}]. \end{aligned}$$

By Assumption 2.3.7, we have

$$\left\|\Sigma_{K^cK}\Sigma^{-1}_{KK}\right\|_\infty \leq 1 - \phi.$$

It suffices to show that $||R_i||_\infty \leq \frac{\phi}{6}$ for $i = 1, ..., 3$ and $||R_i||_\infty \leq \frac{\phi}{12}$ for $i = 4, ..., 6$.

For the first term R_1, we have

$$R_1 = -\Sigma_{K^cK}\Sigma^{-1}_{KK}[\hat{\Sigma}_{KK} - \Sigma_{KK}]\hat{\Sigma}^{-1}_{KK},$$

Using the sub-multiplicative property $||AB||_\infty \leq ||A||_\infty ||B||_\infty$ and the elementary inequality $||A||_\infty \leq \sqrt{a}||A||_2$ for any symmetric matrix $A \in \mathbb{R}^{a\times a}$, we can bound R_1 as follows:

$$\begin{aligned} ||R_1||_\infty &\leq \left\|\Sigma_{K^cK}\Sigma_{KK}^{-1}\right\|_\infty \left\|\hat{\Sigma}_{KK} - \Sigma_{KK}\right\|_\infty \left\|\hat{\Sigma}_{KK}^{-1}\right\|_\infty \\ &\leq (1-\phi)\left\|\hat{\Sigma}_{KK} - \Sigma_{KK}\right\|_\infty \sqrt{k_2}\left\|\hat{\Sigma}_{KK}^{-1}\right\|_2, \end{aligned}$$

where the last inequality follows from Assumption 2.3.7. Using bound (2.13) from the proof for Lemma 2.6.10, we have

$$\left\|\hat{\Sigma}_{KK}^{-1}\right\|_2 \leq \frac{2}{\lambda_{\min}(\Sigma_{KK})}$$

with probability at least $1 - c_1 \exp(-c_2 n)$. Next, applying bound (2.8) from Lemma 2.6.10 with $\varepsilon = \frac{\phi\lambda_{\min}(\Sigma_{KK})}{12(1-\phi)\sqrt{k_2}}$, we have

$$\mathbb{P}\left[\left\|\hat{\Sigma}_{KK} - \Sigma_{KK}\right\|_\infty \geq \frac{\phi\lambda_{\min}(\Sigma_{KK})}{12(1-\phi)\sqrt{k_2}}\right] \leq 2\exp(-bn\min\{\frac{1}{k_2^3}, \frac{1}{k_2^{3/2}}\} + 2\log k_2).$$

Then, we are guaranteed that

$$\mathbb{P}[||R_1||_\infty \geq \frac{\phi}{6}] \leq 2\exp(-bn\min\{\frac{1}{k_2^3}, \frac{1}{k_2^{3/2}}\} + 2\log k_2).$$

For the second term R_2, we first write

$$\begin{aligned} ||R_2||_\infty &\leq \sqrt{k_2}\left\|\Sigma_{KK}^{-1}\right\|_2 \left\|\hat{\Sigma}_{K^cK} - \Sigma_{K^cK}\right\|_\infty \\ &\leq \frac{\sqrt{k_2}}{\lambda_{\min}(\Sigma_{KK})}\left\|\hat{\Sigma}_{K^cK} - \Sigma_{K^cK}\right\|_\infty. \end{aligned}$$

An application of bound (2.7) from Lemma 2.6.10 with $\varepsilon = \frac{\phi}{6}\frac{\lambda_{\min}(\Sigma_{KK})}{\sqrt{k_2}}$ to bound the term $\left\|\hat{\Sigma}_{K^cK} - \Sigma_{K^cK}\right\|_\infty$ yields

$$\mathbb{P}[||R_2||_\infty \geq \frac{\phi}{6}] \leq 2\exp(-bn\min\{\frac{1}{k_2^3}, \frac{1}{k_2^{3/2}}\} + \log(p-k_2) + \log k_2).$$

For the third term R_3, by applying bounds (2.7) from Lemma 2.6.10 with $\varepsilon = \frac{\phi\lambda_{\min}(\Sigma_{KK})}{6}$ to bound the term $\left\|\hat{\Sigma}_{K^cK} - \Sigma_{K^cK}\right\|_\infty$ and (2.9) from Lemma 2.6.10 to bound the term $\left\|\hat{\Sigma}_{KK}^{-1} - \Sigma_{KK}^{-1}\right\|_\infty$, we have

$$\mathbb{P}[||R_3||_\infty \geq \frac{\phi}{6}] \leq 2\exp(-bn\min\{\frac{1}{k_2^2},\, \frac{1}{k_2}\} + \log(p-k_2) + \log k_2).$$

Putting everything together, we conclude that

$$\mathbb{P}[||\hat{\Sigma}_{K^cK}\hat{\Sigma}_{KK}^{-1}||_\infty \geq 1 - \frac{\phi}{2}] \leq O\left(\exp(-bn\min\{\frac{1}{k_2^3},\, \frac{1}{k_2^{3/2}}\} + 2\log p)\right).$$

For the fourth term R_4, we have, with probability at least $1 - c\exp(-bn\min\{\frac{1}{k_2^3},\, \frac{1}{k_2^{3/2}}\} + 2\log p)$,

$$\begin{aligned} ||R_4||_\infty &\leq \left\|\hat{\Sigma}_{K^cK}\hat{\Sigma}_{KK}^{-1}\right\|_\infty \left\|\tilde{\Sigma}_{KK} - \hat{\Sigma}_{KK}\right\|_\infty \left\|\tilde{\Sigma}_{KK}^{-1}\right\|_\infty \\ &\leq (1 - \frac{\phi}{2}) \left\|\tilde{\Sigma}_{KK} - \hat{\Sigma}_{KK}\right\|_\infty \sqrt{k_2} \left\|\tilde{\Sigma}_{KK}^{-1}\right\|_2, \end{aligned}$$

where the last inequality follows from the bound on $||\hat{\Sigma}_{K^cK}\hat{\Sigma}_{KK}^{-1}||_\infty$ established previously. Using bounds (2.18) from the proof for Lemma 2.6.11, we have

$$\left\|\tilde{\Sigma}_{KK}^{-1}\right\|_2 \leq \frac{4}{\lambda_{\min}(\Sigma_{KK})}$$

with probability at least $1 - c_1\exp(-c_2\log\max(p,\, d))$. Next, applying bound (2.15) from Lemma 2.6.11 with $\varepsilon = \frac{\phi\lambda_{\min}(\Sigma_{KK})}{48(1-\frac{\phi}{2})\sqrt{k_2}}$ to bound the term $\left\|\tilde{\Sigma}_{KK} - \hat{\Sigma}_{KK}\right\|_\infty$ yields,

$$\begin{aligned} \mathbb{P}[||R_4||_\infty &\geq \frac{\phi}{12}] \leq 6\exp(\min\{\frac{-bn^2}{k_1k_2^3\log\max(p,\, d)},\, \frac{-bn^{3/2}}{k_2^{3/2}\sqrt{k_1\log\max(p,\, d)}}\} + 2\log k_2) \\ &+ c_1\exp(-c_2\log\max(p,\, d)), \end{aligned}$$

For the fifth term R_5, using bound (2.13) from the proof for Lemma 2.6.10, we have

$$\begin{aligned} ||R_5||_\infty &\leq \sqrt{k_2}\left\|\hat{\Sigma}_{KK}^{-1}\right\|_2 \left\|\tilde{\Sigma}_{K^cK} - \hat{\Sigma}_{K^cK}\right\|_\infty \\ &\leq \frac{2\sqrt{k_2}}{\lambda_{\min}(\Sigma_{KK})} \left\|\tilde{\Sigma}_{K^cK} - \hat{\Sigma}_{K^cK}\right\|_\infty. \end{aligned}$$

An application of bound (2.14) from Lemma 2.6.11 with $\varepsilon = \frac{\phi\lambda_{\min}(\Sigma_{KK})}{24\sqrt{k_2}}$ to bound the term $\left\|\tilde{\Sigma}_{K^cK} - \hat{\Sigma}_{K^cK}\right\|_\infty$ yields

$$\begin{aligned} \mathbb{P}[||R_5||_\infty &\geq \frac{\phi}{12}] \leq 6\exp(\min\{\frac{-bn^2}{k_1k_2^3\log\max(p,\, d)},\, \frac{-bn^{3/2}}{k_2^{3/2}\sqrt{k_1\log\max(p,\, d)}}\} + 2\log p) \\ &+ c_1\exp(-c_2\log\max(p,\, d)), \end{aligned}$$

For the sixth term R_6, by applying bounds (2.14) and (2.16) to bound the terms $\left\|\tilde{\Sigma}_{K^cK} - \hat{\Sigma}_{K^cK}\right\|_\infty$ and $\left\|\tilde{\Sigma}_{KK}^{-1} - \hat{\Sigma}_{KK}^{-1}\right\|_\infty$ respectively, with $\varepsilon = \frac{\phi}{12}\frac{\lambda_{\min}(\Sigma_{KK})}{8}$ for (2.14), we are guaranteed that

$$\begin{aligned}\mathbb{P}[||R_6||_\infty \quad \geq \quad & \frac{\phi}{12}] \leq 6\exp(\min\{\frac{-bn^2}{k_1 k_2^2 \log\max(p,\, d)}, \frac{-bn^{3/2}}{k_2\sqrt{k_1 \log\max(p,\, d)}}\} + 2\log p) \\ & + c_1 \exp(-c_2 \log\max(p,\, d)),\end{aligned}$$

Under the assumptions in Lemmas 2.6.1 and 2.6.2 and the condition

$$\begin{aligned}\frac{1}{n}k_2^3 \log p &= O(1), \\ \frac{1}{n}k_1 k_2^2 \log\max(p,\, d) &= o(1),\end{aligned}$$

putting the bounds on $R_1 - R_6$ together, we conclude that

$$\mathbb{P}[||\tilde{\Sigma}_{K^cK}\tilde{\Sigma}_{KK}^{-1}||_\infty \geq 1 - \frac{\phi}{4}] \leq O\left(\frac{1}{\min(p,\, d)}\right).$$

□

2.6.6 Theorem 2.3.6

The proof for Theorem 2.3.6 is based on a construction called Primal-Dual Witness (PDW) method developed by Wainwright (2009) (also see Wainwright, 2015). This method constructs a pair $(\hat{\beta},\, \hat{\mu})$. When this procedure succeeds, the constructed pair is primal-dual optimal, and acts as a witness for the fact that the Lasso has a unique optimal solution with the correct signed support. The procedure is described in the following.

(i) Set $\hat{\beta}_{J(\beta^*)^c} = 0$.

(ii) Obtain $(\hat{\beta}_{J(\beta^*)},\, \hat{\mu}_{J(\beta^*)})$ by solving the oracle subproblem

$$\hat{\beta}_{J(\beta^*)} \in \arg\min_{\beta_{J(\beta^*)}\in\mathbb{R}^{k_2}} \{\frac{1}{2n}|y - \hat{X}_{J(\beta^*)}\beta_{J(\beta^*)}|_2^2 + \lambda_n|\beta_{J(\beta^*)}|_1\},$$

and choose $\hat{\mu}_{J(\beta^*)} \in \partial|\hat{\beta}_{J(\beta^*)}|_1$, where $\partial|\hat{\beta}_{J(\beta^*)}|_1$ denotes the set of subgradients at $\hat{\beta}_{J(\beta^*)}$ for the function $|\cdot|_1 : \mathbb{R}^{k_2} \to \mathbb{R}$.

(iii) Solve for $\hat{\mu}_{J(\beta^*)^c}$ via the zero-subgradient equation

$$\frac{1}{n}\hat{X}^T(y - \hat{X}\hat{\beta}) + \lambda_n \hat{\mu} = 0,$$

and check whether or not the *strict dual feasibility* condition $|\hat{\mu}_{J(\beta^*)^c}|_\infty < 1$ holds.

The proof for the first claim in Theorem 2.3.6 is established in Lemma 2.6.5, which shows that $\hat{\beta}_{H2SLS} = (\hat{\beta}_K, \mathbf{0})$ where $\hat{\beta}_K$ is the solution obtained in step 2 of the PDW construction (recall we let $J(\beta^*) := K$ and $J(\beta^*)^c := K^c$ for notational convenience in Lemma 2.6.4). The second and third claims are proved using Lemma 2.6.6. The last claim is a consequence of the third claim.

Lemma 2.6.5: If the PDW construction succeeds and if $\lambda_{\min}(\Sigma_{KK}) \geq C_{\min} > 0$, then the vector $(\hat{\beta}_K, \mathbf{0}) \in \mathbb{R}^p$ is the unique optimal solution of the Lasso.

Proof. The proof for Lemma 2.6.5 adopts the proof for Lemma 1 from Chapter 6.4.2 of Wainwright (2015). If the PDW construction succeeds, then $\hat{\beta} = (\hat{\beta}_K, \mathbf{0})$ is an optimal solution with associated subgradient vector $\hat{\mu} \in \mathbb{R}^p$ satisfying $|\hat{\mu}_{K^c}|_\infty < 1$, and $\left\langle \hat{\mu},\, \hat{\beta} \right\rangle = |\hat{\beta}|_1$. Suppose $\tilde{\beta}$ is another optimal solution. Letting $F(\beta) = \frac{1}{2n}|y - \hat{X}\beta|_2^2$, then $F(\hat{\beta}) + \lambda_n \left\langle \hat{\mu},\, \hat{\beta} \right\rangle = F(\tilde{\beta}) + \lambda_n |\tilde{\beta}|_1$, and hence $F(\hat{\beta}) - \lambda_n \left\langle \hat{\mu},\, \tilde{\beta} - \hat{\beta} \right\rangle = F(\tilde{\beta}) + \lambda_n \left(|\tilde{\beta}|_1 - \left\langle \hat{\mu},\, \tilde{\beta} \right\rangle \right)$. However, by the zero-subgradient[2] conditions for optimality, we have $\lambda_n \hat{\mu} = -\nabla F(\hat{\beta})$, which implies that $F(\hat{\beta}) + \left\langle \nabla F(\hat{\beta}),\, \tilde{\beta} - \hat{\beta} \right\rangle - F(\tilde{\beta}) = \lambda_n \left(|\tilde{\beta}|_1 - \left\langle \hat{\mu},\, \tilde{\beta} \right\rangle \right)$. By convexity of F, the left-hand side is non-positive, which implies that $|\tilde{\beta}|_1 \leq \left\langle \hat{\mu},\, \tilde{\beta} \right\rangle$. But since we also have $\left\langle \hat{\mu},\, \tilde{\beta} \right\rangle \leq |\hat{\mu}|_\infty |\tilde{\beta}|_1$, we must have $|\tilde{\beta}|_1 = \left\langle \hat{\mu},\, \tilde{\beta} \right\rangle$. Since $|\hat{\mu}_{K^c}|_\infty < 1$, this equality can only occur if $\tilde{\beta}_j = 0$ for all $j \in K^c$. Thus, all optimal solutions are supported only on K, and hence can be obtained by solving the oracle subproblem in the PDW procedure described in Section 2.3.2. If $\lambda_{\min}(\Sigma_{KK}) \geq C_{\min} > 0$, this subproblem is strictly convex, and hence it has a unique minimizer. □

[2]Given a convex function $g : \mathbb{R}^p \mapsto \mathbb{R}$, $\mu \in \mathbb{R}^p$ is a subgradient at β, denoted by $\mu \in \partial g(\beta)$, if $g(\beta + \triangle) \geq g(\beta) + \langle \mu,\, \triangle \rangle$ for all $\triangle \in \mathbb{R}^p$. When $g(\beta) = |\beta|_1$, notice that $\mu \in \partial |\beta|_1$ if and only if $\mu_j = \text{sign}(\beta_j)$ for all $j = 1, ..., p$, where sign(0) is allowed to be any number in $[-1,\, 1]$.

Lemma 2.6.6: Suppose Assumptions 1.1, 3.1-3.3, 3.5 and 3.7 hold. Let

$$\begin{aligned}\varphi_1 &= \frac{\sigma_\eta\sqrt{\lambda_{\max}(\Sigma_Z)}|\beta^*|_1}{\lambda_{\min}(\Sigma_Z)\lambda_{\min}(\Sigma_{KK})}, \\ \varphi_2 &= \max\left\{\frac{\sigma_{X^*}\sigma_\eta|\beta^*|_1}{\lambda_{\min}(\Sigma_{KK})},\ \frac{\sigma_{X^*}\sigma_\epsilon}{\lambda_{\min}(\Sigma_{KK})}\right\}, \\ \varphi_3 &= \frac{\max\left\{\sigma_\eta,\ \sigma_{X^*},\ \sigma_\epsilon\right\}\sigma_\eta\sigma_Z\sqrt{\lambda_{\max}(\Sigma_Z)}|\beta^*|_1}{\lambda_{\min}(\Sigma_Z)}\end{aligned}$$

With the choice of the tuning parameter

$$\begin{aligned}\lambda_n &\geq c'\frac{48(2-\frac{\phi}{4})}{\phi}\max\left\{\varphi_3\sqrt{\frac{k_1\log\max(p,\ d)}{n}},\ \sigma_{X^*}\sigma_\eta|\beta^*|_1\sqrt{\frac{\log p}{n}},\ \sigma_{X^*}\sigma_\epsilon\sqrt{\frac{\log p}{n}}\right\} \\ &\asymp k_2\sqrt{\frac{k_1\log\max(p,\ d)}{n}},\end{aligned}$$

and under the conditions

$$\begin{aligned}\frac{1}{n}k_2^3\log p &= O(1), \\ \frac{1}{n}k_1k_2^2\log\max(p,\ d) &= o(1),\end{aligned}$$

we have $|\hat{\mu}_{K^c}|_\infty \leq 1-\frac{\phi}{8}$ with probability at least $1-c_1\exp(-c_2\log\min(p,\ d))$. Furthermore,

$$|\hat{\beta}_K-\beta^*_K|_\infty \leq c\max\left\{\varphi_1 k_2\sqrt{\frac{k_1\log\max(p,\ d)}{n}},\ \varphi_2\sqrt{\frac{k_2\log p}{n}}\right\},$$

with probability at least $1-c_1\exp(-c_2\log\min(p,\ d))$.

Proof. By construction, the sub-vectors $\hat{\beta}_K$, $\hat{\mu}_K$, and $\hat{\mu}_{K^c}$ satisfy the zero-subgradient condition in the PDW construction. Recall $e := (X-\hat{X})\beta^* + \boldsymbol{\eta}\beta^* + \epsilon$ from Lemma 2.3.1. With the fact that $\hat{\beta}_{K^c} = \beta^*_{K^c} = 0$, we have

$$\begin{aligned}\frac{1}{n}\hat{X}_K^T\hat{X}_K\left(\hat{\beta}_K-\beta^*_K\right)+\frac{1}{n}\hat{X}_K^Te+\lambda_n\hat{\mu}_K &= 0, \\ \frac{1}{n}\hat{X}_{K^c}^T\hat{X}_K\left(\hat{\beta}_K-\beta^*_K\right)+\frac{1}{n}\hat{X}_{K^c}^Te+\lambda_n\hat{\mu}_{K^c} &= 0.\end{aligned}$$

From the equations above, by solving for the vector $\hat{\mu}_{K^c} \in \mathbb{R}^{p-k_2}$, we obtain

$$\begin{aligned} \hat{\mu}_{K^c} &= -\frac{1}{n\lambda_n}\hat{X}_{K^c}^T\hat{X}_K\left(\hat{\beta}_K - \beta_K^*\right) - \hat{X}_{K^c}^T\frac{e}{n\lambda_n}, \\ \hat{\beta}_K - \beta_K^* &= -\left(\frac{1}{n}\hat{X}_K^T\hat{X}_K\right)^{-1}\frac{\hat{X}_K^T e}{n} - \lambda_n\left(\frac{\hat{X}_K^T\hat{X}_K}{n}\right)^{-1}\hat{\mu}_K, \end{aligned}$$

which yields

$$\hat{\mu}_{K^c} = \left(\tilde{\Sigma}_{K^cK}\tilde{\Sigma}_{KK}^{-1}\right)\hat{\mu}_K + \left(\hat{X}_{K^c}^T\frac{e}{n\lambda_n}\right) - \left(\tilde{\Sigma}_{K^cK}\tilde{\Sigma}_{KK}^{-1}\right)\hat{X}_K^T\frac{e}{n\lambda_n}.$$

By the triangle inequality, we have

$$|\hat{\mu}_{K^c}|_\infty \le \left\|\tilde{\Sigma}_{K^cK}\tilde{\Sigma}_{KK}^{-1}\right\|_\infty + \left|\hat{X}_{K^c}^T\frac{e}{n\lambda_n}\right|_\infty + \left\|\tilde{\Sigma}_{K^cK}\tilde{\Sigma}_{KK}^{-1}\right\|_\infty\left|\hat{X}_K^T\frac{e}{n\lambda_n}\right|_\infty,$$

where the fact that $|\hat{\mu}_K|_\infty \le 1$ is used in the inequality above. By Lemma 2.6.4, we have $\left\|\tilde{\Sigma}_{K^cK}\tilde{\Sigma}_{KK}^{-1}\right\|_\infty \le 1 - \frac{\phi}{4}$ with probability at least $1 - c\exp(-\log\min(p,\, d))$. Hence,

$$\begin{aligned} |\hat{\mu}_{K^c}|_\infty &\le 1 - \frac{\phi}{4} + \left|\hat{X}_{K^c}^T\frac{e}{n\lambda_n}\right|_\infty + \left\|\tilde{\Sigma}_{K^cK}\tilde{\Sigma}_{KK}^{-1}\right\|_\infty\left|\hat{X}_K^T\frac{e}{n\lambda_n}\right|_\infty \\ &\le 1 - \frac{\phi}{4} + \left(2 - \frac{\phi}{4}\right)\left|\hat{X}^T\frac{e}{n\lambda_n}\right|_\infty. \end{aligned}$$

Therefore, it suffices to show that $\left(2 - \frac{\phi}{4}\right)\left|\hat{X}^T\frac{e}{n\lambda_n}\right|_\infty \le \frac{\phi}{8}$ with high probability. This result is established in Lemma 2.6.12. Thus, we have $|\hat{\mu}_{K^c}|_\infty \le 1 - \frac{\phi}{8}$ with high probability.

It remains to establish a bound on the $l_\infty-$norm of the error $\hat{\beta}_K - \beta_K^*$. By the triangle inequality, we have

$$\begin{aligned} |\hat{\beta}_K - \beta_K^*|_\infty &\le \left|\left(\frac{\hat{X}_K^T\hat{X}_K}{n}\right)^{-1}\frac{\hat{X}_K^T e}{n}\right|_\infty + \lambda_n\left\|\left(\frac{\hat{X}_K^T\hat{X}_K}{n}\right)^{-1}\right\|_\infty \\ &\le \left\|\left(\frac{\hat{X}_K^T\hat{X}_K}{n}\right)^{-1}\right\|_\infty\left|\frac{\hat{X}_K^T e}{n}\right|_\infty + \lambda_n\left\|\left(\frac{\hat{X}_K^T\hat{X}_K}{n}\right)^{-1}\right\|_\infty, \end{aligned}$$

Using bound (2.18) from Lemma 2.6.11, we have

$$\left\|\left(\frac{\hat{X}_K^T\hat{X}_K}{n}\right)^{-1}\right\|_\infty \le \frac{2\sqrt{k_2}}{\lambda_{\min}(\hat{\Sigma}_{KK})} \le \frac{4\sqrt{k_2}}{\lambda_{\min}(\Sigma_{KK})}.$$

By Lemma 2.6.2, we have, with probability at least $1 - c_1 \exp(-c_2 \log\min(p,\, d))$,

$$|\tfrac{1}{n}\hat{X}^T e|_\infty \precsim \max\left\{\varphi_3\sqrt{\tfrac{k_1 \log\max(p,d)}{n}},\; \sigma_{X^*}\sigma_\eta|\beta^*|_1\sqrt{\tfrac{\log p}{n}},\; \sigma_{X^*}\sigma_\epsilon\sqrt{\tfrac{\log p}{n}}\right\}.$$

Putting everything together, with the choice of λ_n given in the statement of Lemma 2.6.6, we obtain

$$|\hat{\beta}_K - \beta^*_K|_\infty \leq c\max\left\{\varphi_1\sqrt{\frac{k_1 k_2 \log\max(p,\, d)}{n}},\; \varphi_2\sqrt{\frac{k_2 \log p}{n}}\right\},$$

with probability at least $1 - c_1 \exp(-c_2 \log\min(p,\, d)))$, as claimed. □

2.6.7 Lemmas 2.6.7-2.6.12

Lemma 2.6.7: If $X \in \mathbb{R}^{n\times p_1}$ is a zero-mean sub-Gaussian matrix with parameters $(\Sigma_X,\, \sigma_X^2)$, then for any fixed (unit) vector $v \in \mathbb{R}^{p_1}$, we have

$$\mathbb{P}(\left||Xv|_2^2 - \mathbb{E}[|Xv|_2^2]\right| \geq nt) \leq 2\exp(-cn\min\{\frac{t^2}{\sigma_X^4},\, \frac{t}{\sigma_X^2}\}).$$

Moreover, if $Y \in \mathbb{R}^{n\times p_2}$ is a zero-mean sub-Gaussian matrix with parameters $(\Sigma_Y,\, \sigma_Y^2)$, then

$$\mathbb{P}(|\frac{Y^T X}{n} - \mathrm{cov}(\mathbf{y}_i,\, \mathbf{x}_i)|_\infty \geq t) \leq 6p_1p_2\exp(-cn\min\{\frac{t^2}{\sigma_X^2\sigma_Y^2},\, \frac{t}{\sigma_X\sigma_Y}\}),$$

where $\mathbf{x}_i$ and $\mathbf{y}_i$ are the i^{th} rows of X and Y, respectively. In particular, if $n \succsim \log p$, then

$$\mathbb{P}(|\frac{Y^T X}{n} - \mathrm{cov}(\mathbf{y}_i,\, \mathbf{x}_i)|_\infty \geq c_0\sigma_X\sigma_Y\sqrt{\frac{\log(\max\{p_1,\, p_2\})}{n}}) \leq c_1\exp(-c_2\log(\max\{p_1,\, p_2\})).$$

Remark. Lemma 2.6.7 is Lemma 14 in Loh and Wainwright (2012).

Lemma 2.6.8: For a fixed matrix $\Gamma \in \mathbb{R}^{p\times p}$, parameter $s \geq 1$, and tolerance $\tau > 0$, suppose we have the deviation condition

$$|v^T\Gamma v| \leq \tau \qquad \forall v \in \mathbb{K}(2s,\, p).$$

Then,

$$|v^T\Gamma v| \leq 27\tau\left(|v|_2^2 + \frac{1}{s}|v|_1^2\right) \qquad \forall v \in \mathbb{R}^p.$$

Remark. Lemma 2.6.8 is Lemma 12 in Loh and Wainwright (2012).

Lemma 2.6.9: Under Assumptions 2.1.1 and 2.3.3, we have

$$\frac{|X^* v^0|_2^2}{n} \geq \kappa_1 |v^0|_2^2 - \kappa_2 \frac{k_1^r \log \max(p,\, d)}{n} |v^0|_1^2, \qquad \text{for all } v^0 \in \mathbb{R}^p,\, r \in [0,\, 1]$$

with probability at least $1 - c_1 \exp(-c_2 n)$, where

$$\kappa_1 = \frac{\lambda_{\min}(\Sigma_{X^*})}{2}$$

and $\kappa_2 = c_0 \lambda_{\min}(\Sigma_{X^*}) \max\left\{\frac{\sigma_{X^*}^4}{\lambda_{\min}^2(\Sigma_{X^*})},\, 1\right\}$.

Proof. First, we show

$$\sup_{v^0 \in \mathbb{K}(2s,\, p)} \left| v^{0T} \left(\frac{X^{*T} X^*}{n} - \Sigma_{X^*} \right) v^0 \right| \leq \frac{\lambda_{\min}(\Sigma_{X^*})}{54}$$

with high probability, where $\Sigma_{X^*} = \mathbb{E}(X^{*T} X^*)$. Under Assumption 2.3.3, we have that X^* is sub-Gaussian with parameters $(\Sigma_{X^*},\, \sigma_{X^*})$. Therefore, by Lemma 2.6.8 and a discretization argument similar to those in Lemma 2.6.3, we have

$$\mathbb{P}\left(\sup_{v^0 \in \mathbb{K}(2s,\, p)} \left| v^{0T} \left(\frac{X^{*T} X^*}{n} - \Sigma_{X^*} \right) v^0 \right| \geq t \right) \leq 2 \exp(-cn \min(\frac{t^2}{\sigma_{X^*}^4},\, \frac{t}{\sigma_{X^*}^2}) + 2s \log p),$$

for some universal constants $c > 0$. By choosing $t = \frac{\lambda_{\min}(\Sigma_{X^*})}{54}$ and

$$s = s(r) := \frac{1}{c'} \frac{n}{k_1^r \log \max(p,\, d)} \min\left\{ \frac{\lambda_{\min}^2(\Sigma_{X^*})}{\sigma_{X^*}^4},\, 1 \right\}, \quad r \in [0,\, 1],$$

where c' is chosen sufficiently small so that $s \geq 1$, we get

$$\mathbb{P}\left(\sup_{v^0 \in \mathbb{K}(2s,\, p)} \left| v^{0T} \left(\frac{X^{*T} X^*}{n} - \Sigma_{X^*} \right) v^0 \right| \geq \frac{\lambda_{\min}(\Sigma_{X^*})}{54} \right) \leq 2 \exp(-c_2 n \min(\frac{\lambda_{\min}^2(\Sigma_{X^*})}{\sigma_{X^*}^4},\, 1)).$$

Now, by Lemma 2.6.8 and the following substitutions

$$\Gamma = \frac{X^{*T} X^*}{n} - \Sigma_{X^*}, \quad \text{and} \quad \tau := \frac{\lambda_{\min}(\Sigma_{X^*})}{54},$$

we obtain
$$\left|v^{0T}\left(\frac{X^{*T}X^{*}}{n}-\Sigma_{X^{*}}\right)v^{0}\right|\leq\frac{\lambda_{\min}(\Sigma_{X^{*}})}{2}\left(|v^{0}|_{2}^{2}+\frac{1}{s}|v^{0}|_{1}^{2}\right),$$
which implies
$$v^{0T}\frac{X^{*T}X^{*}}{n}v^{0}\geq v^{0T}\Sigma_{X^{*}}v^{0}-\frac{\lambda_{\min}(\Sigma_{X^{*}})}{2}\left(|v^{0}|_{2}^{2}+\frac{1}{s}|v^{0}|_{1}^{2}\right).$$
Recalling the choice of
$$s=s(r):=\frac{1}{c^{'}}\frac{n}{k_{1}^{r}\log\max(p,\,d)}\min\left\{\frac{\lambda_{\min}^{2}(\Sigma_{X^{*}})}{\sigma_{X^{*}}^{4}},\,1\right\},\quad r\in[0,\,1],$$
where $c^{'}$ is chosen sufficiently small so $s\geq1$, the claim follows. □

Lemma 2.6.10: Suppose Assumptions 2.1.1 and 2.3.3 hold. For any $\varepsilon>0$ and constant c, we have

$$\mathbb{P}\left\{\left\|\hat{\Sigma}_{K^{c}K}-\Sigma_{K^{c}K}\right\|_{\infty}\geq\varepsilon\right\}\leq(p-k_{2})k_{2}\cdot2\exp(-cn\min\{\frac{\varepsilon^{2}}{4k_{2}^{2}\sigma_{X^{*}}^{4}},\,\frac{\varepsilon}{2k_{2}\sigma_{X^{*}}^{2}}\}),\tag{2.7}$$

$$\mathbb{P}\left\{\left\|\hat{\Sigma}_{KK}-\Sigma_{KK}\right\|_{\infty}\geq\varepsilon\right\}\leq k_{2}^{2}\cdot2\exp(-cn\min\{\frac{\varepsilon^{2}}{4k_{2}^{2}\sigma_{X^{*}}^{4}},\,\frac{\varepsilon}{2k_{2}\sigma_{X^{*}}^{2}}\}).\tag{2.8}$$

Furthermore, under the scaling $n\gtrsim k_{2}\log p$, for constants b_1 and b_2, we have
$$\left\|\hat{\Sigma}_{KK}^{-1}-\Sigma_{KK}^{-1}\right\|_{\infty}\leq\frac{1}{\lambda_{\min}(\Sigma_{KK})},\tag{2.9}$$
with probability at least $1-c_{1}\exp(-c_{2}n\min\{\frac{\lambda_{\min}^{2}(\Sigma_{KK})}{4k_{2}\sigma_{X^{*}}^{4}},\,\frac{\lambda_{\min}(\Sigma_{KK})}{2\sqrt{k_{2}}\sigma_{X^{*}}^{2}}\})$.

Proof. Denote the element $(j^{'},\,j)$ of the matrix difference $\hat{\Sigma}_{K^{c}K}-\Sigma_{K^{c}K}$ by $u_{j^{'}j}$. By the definition of the l_∞matrix norm, we have

$$\begin{aligned}
\mathbb{P}\left\{\left\|\hat{\Sigma}_{K^cK}-\Sigma_{K^cK}\right\|_\infty \geq \varepsilon\right\} &= \mathbb{P}\left\{\max_{j'\in K^c}\sum_{j\in K}|u_{j'j}|\geq\varepsilon\right\} \\
&\leq (p-k_2)\mathbb{P}\left\{\sum_{j\in K}|u_{j'j}|\geq\varepsilon\right\} \\
&\leq (p-k_2)\mathbb{P}\left\{\exists j\in K \mid |u_{j'j}|\geq\frac{\varepsilon}{k_2}\right\} \\
&\leq (p-k_2)k_2\mathbb{P}\left\{|u_{j'j}|\geq\frac{\varepsilon}{k_2}\right\} \\
&\leq (p-k_2)k_2\cdot 2\exp(-cn\min\{\frac{\varepsilon^2}{k_2^2\sigma_{X^*}^4},\frac{\varepsilon}{k_2\sigma_{X^*}^2}\}),
\end{aligned}$$

where the last inequality follows the deviation bound for sub-exponential random variables, i.e., Lemma 2.6.7. Bound (2.8) can be obtained in a similar way except that the pre-factor $(p-k_2)$ is replaced by k_2. To prove the last bound (2.9), write

$$\begin{aligned}
\left\|\hat{\Sigma}_{KK}^{-1}-\Sigma_{KK}^{-1}\right\|_\infty &= \left\|\Sigma_{KK}^{-1}\left[\Sigma_{KK}-\hat{\Sigma}_{KK}\right]\hat{\Sigma}_{KK}^{-1}\right\|_\infty \\
&= \sqrt{k_2}\left\|\Sigma_{KK}^{-1}\left[\Sigma_{KK}-\hat{\Sigma}_{KK}\right]\hat{\Sigma}_{KK}^{-1}\right\|_2 \\
&= \sqrt{k_2}\left\|\Sigma_{KK}^{-1}\right\|_2\left\|\Sigma_{KK}-\hat{\Sigma}_{KK}\right\|_2\left\|\hat{\Sigma}_{KK}^{-1}\right\|_2 \\
&\leq \frac{\sqrt{k_2}}{\lambda_{\min}(\Sigma_{KK})}\left\|\Sigma_{KK}-\hat{\Sigma}_{KK}\right\|_2\left\|\hat{\Sigma}_{KK}^{-1}\right\|_2. \qquad (2.10)
\end{aligned}$$

To bound the term $\left\|\Sigma_{KK}-\hat{\Sigma}_{KK}\right\|_2$ in (2.10), applying Lemma 2.6.7 with $X^TX=\hat{\Sigma}_{KK}$ and $t=\frac{\lambda_{\min}(\Sigma_{KK})}{2\sqrt{k_2}}$ yields

$$\left\|\hat{\Sigma}_{KK}-\Sigma_{KK}\right\|_2\leq\frac{\lambda_{\min}(\Sigma_{KK})}{2\sqrt{k_2}},$$

with probability at least $1-c_1\exp(-c_2n\min\{\frac{\lambda_{\min}^2(\Sigma_{KK})}{4k_2\sigma_{X^*}^4},\frac{\lambda_{\min}(\Sigma_{KK})}{2\sqrt{k_2}\sigma_{X^*}^2}\})$.

To bound the term$\left\|\hat{\Sigma}_{KK}^{-1}\right\|_2$ in (2.10), note that we can write

$$\begin{aligned}
\lambda_{\min}(\Sigma_{KK}) &= \min_{||h'||_2=1} h'^T\Sigma_{KK}h' \\
&= \min_{||h'||_2=1}\left[h'^T\hat{\Sigma}_{KK}h'+h'^T(\Sigma_{KK}-\hat{\Sigma}_{KK})h'\right] \\
&\leq h^T\hat{\Sigma}_{KK}h+h^T(\Sigma_{KK}-\hat{\Sigma}_{KK})h \qquad (2.11)
\end{aligned}$$

where $h \in \mathbb{R}^{k_2}$ is a unit-norm minimal eigenvector of $\hat{\Sigma}_{KK}$. Applying Lemma 2.6.7 yields

$$\left|h^T\left(\Sigma_{KK} - \hat{\Sigma}_{KK}\right)h\right| \leq \frac{\lambda_{\min}(\Sigma_{KK})}{2}$$

with probability at least $1 - c_1 \exp(-c_2 n)$. Therefore,

$$\begin{aligned} \lambda_{\min}(\Sigma_{KK}) &\leq \lambda_{\min}(\hat{\Sigma}_{KK}) + \frac{\lambda_{\min}(\Sigma_{KK})}{2} \\ \Longrightarrow \lambda_{\min}(\hat{\Sigma}_{KK}) &\geq \frac{\lambda_{\min}(\Sigma_{KK})}{2}, \end{aligned} \tag{2.12}$$

and consequently,

$$\left\|\hat{\Sigma}_{KK}^{-1}\right\|_2 \leq \frac{2}{\lambda_{\min}(\Sigma_{KK})}. \tag{2.13}$$

Putting everything together, we have

$$\left\|\hat{\Sigma}_{KK}^{-1} - \Sigma_{KK}^{-1}\right\|_\infty \leq \frac{\sqrt{k_2}}{\lambda_{\min}(\Sigma_{KK})} \frac{\lambda_{\min}(\Sigma_{KK})}{2\sqrt{k_2}} \frac{2}{\lambda_{\min}(\Sigma_{KK})} = \frac{1}{\lambda_{\min}(\Sigma_{KK})}.$$

with probability at least $1 - c_1 \exp(-c_2 n \min\{\frac{\lambda_{\min}^2(\Sigma_{KK})}{4k_2\sigma_{X^*}^4}, \frac{\lambda_{\min}(\Sigma_{KK})}{2\sqrt{k_2}\sigma_{X^*}^2}\})$. □

Lemma 2.6.11: Suppose the assumptions in Lemmas 2.6.1 and 6.2 hold. For any $\varepsilon > 0$, under the condition $\frac{k_1 k_2^2 \log\max(p,\, d)}{n} = o(1)$, we have

$$\begin{aligned} &\mathbb{P}\left\{\left\|\tilde{\Sigma}_{K^cK} - \hat{\Sigma}_{K^cK}\right\|_\infty \geq \varepsilon\right\} \leq \\ &6(p-k_2)k_2 \cdot \exp(-cn\min\{\frac{n\varepsilon^2}{k_1k_2^2\log\max(p,\, d)}, \frac{\sqrt{n}\varepsilon}{k_2\sqrt{k_1\log\max(p,\, d)}}\}) \\ &+c_1\exp(-c_2\log\max(p,\, d)), \end{aligned} \tag{2.14}$$

$$\begin{aligned} &\mathbb{P}\left\{\left\|\tilde{\Sigma}_{KK} - \hat{\Sigma}_{KK}\right\|_\infty \geq \varepsilon\right\} \leq \\ &6k_2^2 \cdot \exp(-cn\min\{\frac{n\varepsilon^2}{k_1k_2^2\log\max(p,\, d)}, \frac{\sqrt{n}\varepsilon}{k_2\sqrt{k_1\log\max(p,\, d)}}\}) \\ &+c_1\exp(-c_2\log\max(p,\, d)). \end{aligned} \tag{2.15}$$

Furthermore, we have

$$\left\|\tilde{\Sigma}_{KK}^{-1} - \hat{\Sigma}_{KK}^{-1}\right\|_\infty \leq \frac{8}{\lambda_{\min}(\Sigma_{KK})} \quad \text{with probability at least } 1-c_1\exp(-c_2\log\max(p,\, d)). \tag{2.16}$$

Proof. Denote the element $(j',\, j)$ of the matrix difference $\tilde{\Sigma}_{K^cK} - \hat{\Sigma}_{K^cK}$ by $w_{j'j}$. Using the same argument as in Lemma 2.6.10, we have

$$\mathbb{P}\left\{\left\|\tilde{\Sigma}_{K^cK} - \hat{\Sigma}_{K^cK}\right\|_\infty \geq \varepsilon\right\} \leq (p-k_2)k_2\mathbb{P}\left\{|w_{j'j}| \geq \frac{\varepsilon}{k_2}\right\}.$$

Following the derivation of the upper bounds on $\left|\frac{(\hat{X}-X^*)^TX^*}{n}\right|_\infty$ and $\left|\frac{(\hat{X}-X^*)^T(\hat{X}-X^*)}{n}\right|_\infty$ in the proof for Lemma 2.6.2 and the identity

$$\frac{1}{n}\left(\tilde{\Sigma}_{K^cK} - \hat{\Sigma}_{K^cK}\right) = \frac{1}{n}X_{K^c}^{*T}(\hat{X}_K - X_K^*) + \frac{1}{n}(\hat{X}_{K^c} - X_{K^c}^*)^T X_K^* + \frac{1}{n}(\hat{X}_{K^c} - X_{K^c}^*)^T(\hat{X}_K - X_K^*),$$

we notice that to upper bound $|w_{j'j}|$, it suffices to upper bound $3 \cdot \left|\frac{1}{n}\mathbf{x}_{j'}^{*T}(\hat{\mathbf{x}}_j - \mathbf{x}_j^*)\right|$. From the proof for Lemma 2.6.2, we have

$$\begin{aligned}\left|\frac{1}{n}\mathbf{x}_{j'}^{*T}(\hat{\mathbf{x}}_j - \mathbf{x}_j^*)\right| &\leq \sqrt{\frac{1}{n}\sum_{i=1}^n x_{ij'}^{*2}}\sqrt{\frac{1}{n}\sum_{i=1}^n\left[\mathbf{z}_{ij}(\hat{\pi}_j - \pi_j^*)\right]^2} \\ &\leq c\sqrt{\frac{1}{n}\sum_{i=1}^n x_{ij'}^{*2}}\,\frac{\sigma_\eta\sigma_Z\sqrt{\lambda_{\max}(\Sigma_Z)}}{\lambda_{\min}(\Sigma_Z)}\sqrt{\frac{k_1\log\max(d,\, p)}{n}}\end{aligned}$$

and

$$\mathbb{P}\left[\max_{j'}\frac{1}{n}\sum_{i=1}^n x_{ij'}^{*2} \geq \sigma_{X^*}^2 + t\right] \leq 6(p-k_2)\exp(-cn\min\{\frac{t^2}{\sigma_{X^*}^2\sigma_Z^2},\, \frac{t}{\sigma_{X^*}\sigma_Z}\}).$$

Under the condition $\frac{k_1k_2^2\log\max(p,d)}{n} = o(1)$, setting $t = \varepsilon\frac{\sigma_{X^*}^2\lambda_{\min}(\Sigma_Z)}{c\sigma_\eta\sigma_Z\sqrt{\lambda_{\max}(\Sigma_Z)}}\sqrt{\frac{n}{k_1k_2^2\log\max(p,d)}}$ for any $\varepsilon > 0$ yields

$$\begin{aligned}\mathbb{P}\left[|w_{j'j}| \geq \frac{\varepsilon}{k_2}\right] &\leq \\ 6\exp(-cn\min\{\frac{n\varepsilon^2}{k_1k_2^2\log\max(p,\, d)},\, \frac{\sqrt{n}\varepsilon}{k_2\sqrt{k_1\log\max(p,\, d)}}\}) \\ +c_1\exp(-c_2\log\max(p,\, d)).\end{aligned}$$

Therefore,

$$\mathbb{P}\left\{\left\|\tilde{\Sigma}_{K^cK}-\hat{\Sigma}_{K^cK}\right\|_\infty \geq \varepsilon\right\} \leq$$
$$6(p-k_2)k_2\cdot\exp(-cn\min\{\frac{n\varepsilon^2}{k_1k_2^2\log\max(p,\,d)},\,\frac{\sqrt{n}\varepsilon}{k_2\sqrt{k_1\log\max(p,\,d)}}\})$$
$$+c_1\exp(-c_2\log\max(p,\,d)).$$

Bound (2.15) can be obtained in a similar way except that the pre-factor $(p-k_2)$ is replaced by k_2.

To prove bound (2.16), by applying the same argument as in Lemma 2.6.10, we have

$$\begin{aligned}\left\|\tilde{\Sigma}_{KK}^{-1}-\hat{\Sigma}_{KK}^{-1}\right\|_\infty &\leq \frac{\sqrt{k_2}}{\lambda_{\min}(\hat{\Sigma}_{KK})}\left\|\hat{\Sigma}_{KK}-\tilde{\Sigma}_{KK}\right\|_2\left\|\tilde{\Sigma}_{KK}^{-1}\right\|_2\\ &\leq \frac{2\sqrt{k_2}}{\lambda_{\min}(\Sigma_{KK})}\left\|\hat{\Sigma}_{KK}-\tilde{\Sigma}_{KK}\right\|_2\left\|\tilde{\Sigma}_{KK}^{-1}\right\|_2,\end{aligned}$$

where the last inequality comes from bound (2.12).

To bound the term $\left\|\hat{\Sigma}_{KK}-\tilde{\Sigma}_{KK}\right\|_2$, applying bound (2.15) with $\varepsilon=\frac{\lambda_{\min}(\Sigma_{KK})}{\sqrt{k_2}}$ yields

$$\begin{aligned}\left\|\hat{\Sigma}_{KK}-\tilde{\Sigma}_{KK}\right\|_2 &\leq \left\|\hat{\Sigma}_{KK}-\tilde{\Sigma}_{KK}\right\|_\infty\\ &\leq \frac{\lambda_{\min}(\Sigma_{KK})}{\sqrt{k_2}},\end{aligned}\tag{2.17}$$

with probability at least

$$1-6k_2^2\cdot\exp(-cn\min\{\frac{n}{k_1k_2^3\log\max(p,\,d)},\,\frac{\sqrt{n}}{k_2^{3/2}\sqrt{k_1\log\max(p,\,d)}}\})$$
$$-c_1\exp(-c_2\log\max(p,\,d))\geq 1-O\left(\frac{1}{\max(p,\,d)}\right)$$

if $\frac{k_1^{1/2}k_2^2\log\max(p,\,d)}{n}=O(1)$.

To bound the term $\left\|\tilde{\Sigma}_{KK}^{-1}\right\|_2$, again we have,

$$\begin{aligned}\lambda_{\min}(\hat{\Sigma}_{KK}) &\leq h^T\tilde{\Sigma}_{KK}h+h^T(\hat{\Sigma}_{KK}-\tilde{\Sigma}_{KK})h\\ &\leq h^T\tilde{\Sigma}_{KK}h+k_2\left|\hat{\Sigma}_{KK}-\tilde{\Sigma}_{KK}\right|_\infty\\ &\leq h^T\tilde{\Sigma}_{KK}h+bk_2\sqrt{\frac{k_1\log\max(p,\,d)}{n}},\end{aligned}$$

where $h \in \mathbb{R}^{k_2}$ is a unit-norm minimal eigenvector of $\tilde{\Sigma}_{KK}$. The last inequality follows from the bounds on $\left|\frac{(\hat{X}-X^*)^T X^*}{n}\right|_\infty$ and $\left|\frac{(\hat{X}-X^*)^T(\hat{X}-X^*)}{n}\right|_\infty$ from the proof for Lemma 2.6.1 with probability at least $1 - c_1\exp(-c_2 \log\max(p,\, d))$. Therefore, if $\frac{k_1 k_2^2 \log\max(p,\, d)}{n} = o(1)$, then we have

$$\lambda_{\min}(\tilde{\Sigma}_{KK}) \geq \frac{\lambda_{\min}(\hat{\Sigma}_{KK})}{2}$$

$$\begin{aligned} \Longrightarrow \left\|\tilde{\Sigma}_{KK}^{-1}\right\|_2 &\leq \frac{2}{\lambda_{\min}(\hat{\Sigma}_{KK})} \\ &\leq \frac{4}{\lambda_{\min}(\Sigma_{KK})}, \end{aligned} \tag{2.18}$$

where the last inequality follows from bound (2.12) from the proof for Lemma 2.6.10. Putting everything together, we have

$$\left\|\hat{\Sigma}_{KK}^{-1} - \tilde{\Sigma}_{KK}^{-1}\right\|_\infty \leq \frac{2\sqrt{k_2}}{\lambda_{\min}(\Sigma_{KK})} \frac{\lambda_{\min}(\Sigma_{KK})}{\sqrt{k_2}} \frac{4}{\lambda_{\min}(\Sigma_{KK})} = \frac{8}{\lambda_{\min}(\Sigma_{KK})}.$$

with probability at least $1 - c_1\exp(-c_2 \log\max(p,\, d))$.

Lemma 2.6.12: Suppose the conditions in Lemma 2.6.2 hold. With the choice of the tuning parameter

$$\begin{aligned} \lambda_n &\geq c' \frac{8(2-\frac{\phi}{4})}{\phi} \max\left\{\varphi_3\sqrt{\frac{k_1 \log\max(p,\, d)}{n}},\ \sigma_{X^*}\sigma_\eta|\beta^*|_1\sqrt{\frac{\log p}{n}},\ \sigma_{X^*}\sigma_\epsilon\sqrt{\frac{\log p}{n}}\right\} \\ &\asymp k_2\sqrt{\frac{k_1 \log\max(p,\, d)}{n}}, \end{aligned}$$

for some sufficiently large constant $c > 0$, then, we have

$$\left(2 - \frac{\phi}{4}\right)\left|\hat{X}^T \frac{e}{n\lambda_n}\right|_\infty \leq \frac{\phi}{8},$$

with probability at least $1 - c_1\exp(-c_2 \log\min(p,\, d))$.

Proof. From the proof for Lemma 2.6.2, given the choice of λ_n in Lemma 2.6.12 for some sufficiently large constant $c^{'} > 0$, we have

$$\left(2 - \frac{\phi}{4}\right) \left| \hat{X}^T \frac{e}{n\lambda_n} \right|_\infty \leq \frac{\phi}{8},$$

with probability at least $1 - c_1 \exp(-c_2 \log \min(p,\, d))$.□

Chapter 3

High-Dimensional Semiparametric Sample Selection Models

3.1 Introduction

The object of interest of this chapter is a class of *high-dimensional* selection models under a weak nonparametric restriction on the form of the selection correction. Observational studies are rarely based on pure random samples. When a sample, intentionally or unintentionally, is based in part on values taken by a dependent variable (e.g., Gronau, 1973; Heckman, 1974), parameter estimates without corrective measures may be inconsistent. Such samples can be broadly defined as selected samples. Selection may be due to self-selection, with the outcome of interest determined in part by individual choice of whether or not to participate in the activity of interest. It can also result from endogenous stratification, with those who participate in the activity of interest deliberately oversampled - an extreme case being sampling only participants.

In the classical low-dimensional selection models, parameter estimates obtained from OLS may be inconsistent unless corrective measures are taken. For the parametric case where the error terms are jointly normally distributed and homoskedastic, the most well-known estimator is Heckman's two-step procedure (1974, 1976). For semiparametric estimation of the parameters of selection models when the joint distribution of the error terms is of unknown form, many estimators have adopted the two-step estimation strategy similar to Heckman's, under an additional "single-index" restriction on the form of the selection equation. Several estimators of the parameters of the selection equation have been proposed in the literature on semiparametric estimation; while some of these methods sidestep estimation of the unknown distribution function of the errors (e.g., Manski, 1975 and 1985; Han, 1987), others use nonparametric regression methods to estimate this distribution function along with the parameters of the underlying regression function (e.g., Cosslett, 198l; Ichimura, 1987; Klein and Spady, 1987). Similarly, the methods for estimation of the parameters in the second stage also involve nonparametric regression methods, which are applied either to estimation of the selection correction function directly (Lee, 1982; Cosslett, 1991; Gallant and Nychka, 1987; Ichimura and Lee, 1991; Newey, 1991) or to estimation of other regression functions which depend upon the estimated single index (Powell, 1989; Ahn and Powell, 1993).

Selection models have not been considered in high-dimensional settings even though many economic applications actually fit into this setup. On the demand side, selection models are used in the context where a consumer faces choosing a service (such as electricity, cell phone service, etc.) or brand followed by the amount of utilization or the number of quantities to purchase conditional on the chosen service or brand (e.g., Krishnamurth and Raj, 1988; Chintagunta, 1993; Fox, Kim, and Yang, 2013). On the firms' side, selection models are useful for situations where a firm first decides on its product positioning and then a pricing scheme based on the chosen product type. For example, a grocery store sometimes needs to choose which products to put on sales or promotions and then the amount of discount on these chosen products; a gas station first chooses to be either a two-product station offering both self-service and full-service gasoline or a single-product station offering only full-service or self-service gasoline, and then decides on a pricing scheme conditional on the choice of the station type (Iyer and Seetharaman, 2003). Selection models are also seen in the auction literature (e.g., Roberts and Sweeting, 2011, 2012); in particular, by estimating a Heckman selection model with the exclusion restriction that potential competition affects a bidder's decision to enter an auction, but has no direct effect on the values of the bids, Roberts and Sweeting (2011) presents reduced form evidence that the auction data are best explained by a selection model.

In estimating the underlying selection models to study these empirical problems, analysis has been restricted to only a low-dimensional set of explanatory variables in both the selection equation and the main equation. However, the actual information available to these empirical problems can be far richer than what has been used by the researchers. More importantly, economic theory is not always explicit about the variables that belong to the true model (e.g., Sala-i-Martin, 1997 concerning development economics). In the selection models used for consumer demand estimation, the number of explanatory variables formed by the characteristics (and the transformations of these characteristics) of a service or brand can be very large. In the grocery store example, when choosing whether to put a product on sale and the amount of discount, the store often considers not only the own characteristics of this product but also characteristics of other products. All these characteristics can potentially exceed the number of products chosen to be on sale (namely, the sample size of the observations in the main-equation), which makes it a high-dimensional selection problem. Similarly, in the bidder example, when deciding whether to enter an auction, a bidder considers potential competition from other bidders; when deciding on the values of its bid upon the entry decision, the bidder may still consider competition from the set of other "enters". Consequently, the number of explanatory variables entering the selection equation and the main equation may grow with the number of bidders.

In the gasoline example mentioned above, besides the large number of station characteristics and demographic characteristics which amount to approximately 400 regressors with only 700 gas stations in the data studied by Iyer and Seetharaman (2008), more explanatory variables can be obtained by utilizing the geographic information and spatial data. In particular, geographic information may be used to analyze the interaction between different gas stations and identify the competitive market structure, as will be shown in Section 3.6. Despite that the explanatory variables in the examples above are high-dimensional, it is plausible that only a small set of these variables (relative to the sample size) matter to the underlying response variables but which variables constitute the relevant regressors are unknown to the researchers.

The following sets up the models of interests and highlights the major contributions made by this chapter. In particular, we consider estimation and selection of regression coefficients in the class of selection models captured by the following system: for all $i = 1, ..., n$,

$$\begin{aligned} y_{1i} &= \mathbb{I}\left\{w_i^T\theta^* + \epsilon_{1i} > 0\right\}, \\ y_{2i} &= y_{1i}\left(x_i^T\beta^* + \epsilon_{2i}\right), \\ \mathbb{E}\left(\epsilon_{2i}|w_i,\ x_i,\ y_{1i}=1\right) &= g(w_i^T\theta^*), \end{aligned} \tag{3.1}$$

where x_i is a $p-$dimensional vector of explanatory variables and the dimension p of β^* is large relative to the sample size n (namely, $p \asymp n$ or even $p \gg n$). Furthermore, $g(\cdot)$ is an unknown function and w_i is a $d-$dimensional vector of explanatory variables with an unknown coefficient vector θ^*. Note here the dimension d of θ^* can also be large relative the sample size n (namely, $d \asymp n$ or even $d \gg n$). This chapter considers the *exact sparsity* case of β^* and θ^* where the number of non-zero components in the vectors of coefficients is bounded above by some integer which is allowed to grow with n but slowly compared to n, and also considers the *approximate sparsity* case, where the vectors of coefficients can be approximated by exactly sparse vectors. The definition of *approximate sparsity* is based on the notion of $l_{q_2}-$"balls" where $q_2 \in [0,\ 1]$ and to be introduced formally in Section 3.2. The third equation in (3.1) is known as the "single-index" restriction used in Powell (1989), Newey (1991), and Ahn and Powell (1993). Newey (1991) and Powell (1994) discuss sufficient conditions for this restriction. In particular, it is implied by independence of the errors $(\epsilon_{1i},\, \epsilon_{2i})$ and the regressors $(w_i,\, x_i)$. Note that the second equation of model (3.1) implies

$$y_{2i} = x_i^T \beta^* + g(w_i^T \theta^*) + \eta_i \quad \text{whenever } y_{1i} = 1, \tag{3.2}$$

where by construction $\mathbb{E}\left[\eta_i | w_i,\, x_i,\, y_{1i} = 1\right] = 0$. Throughout the rest of this chapter, when it is clear from the context that only the selected sample is of our interests, the notation $y_{1i} = 1$ will be suppressed. In addition, the values of n will vary according to whether we are working with the whole sample (the observations in the selection equation) or the selected sample (the observations in the main equation). Motivated by the Frisch-Waugh Theorem, applying a projection idea used in Engle, Granger, Rice, and Weiss (1986) and Robinson's semilinear models (1988) yields the following equivalent model

$$v_{i0} = v_i \beta^* + \eta_i, \tag{3.3}$$

where

$$\begin{aligned} v_i &= \left(x_{i1} - \mathbb{E}\left(x_{i1} | w_i^T \theta^*\right),\, \ldots,\, x_{ip} - \mathbb{E}\left(x_{ip} | w_i^T \theta^*\right)\right), \\ v_{i0} &= y_{2i} - \mathbb{E}\left(y_{2i} | w_i^T \theta^*\right). \end{aligned}$$

For convenience, the first equation in (3.1) is referred to as the selection equation and the second equation in (3.1) as the main equation.

High dimensionality arises in selection model (3.1) when the dimension p of β^* is large relative to the sample size n (namely, $p \asymp n$ or $p \gg n$) in the main equation. In addition, a weak nonparametric restriction is imposed on the form of the selection correction. Specifically, the selection effect is assumed to depend on the linear combination of some observable selection variables. The selection model under this nonparametric restriction on the form of the selection correction when $p \geq n$ has apparently not been studied in the literature. As in classical low-dimensional selection models where parameter estimates obtained from OLS may be inconsistent, direct implementation of the Lasso or the Dantzig selector fails as sparsity of coefficients in the main equation in (3.1) may not correspond to sparsity of linear projection coefficients and "bias" from parameter estimates by the direct Lasso procedure without corrective measures is found to be only exacerbated in the high-dimensional setting. This evidence is given by the Monte-Carlo simulation results in Section 3.5. The selection equation in (3.1) is a linear latent variable model and the selection bias $g(\cdot)$ is assumed to be an unknown function of the single index $w_i^T\theta^*$. This setup allows us to consider special cases where the dimension d of θ^* is also large relative to the sample size n (namely, $d \asymp n$ or $d \gg n$) in the selection equation described by some of the most popular binary response models. It is worth noting that the general results provided by this chapter also hold for the more general structure where $\mathbb{E}\left(\epsilon_{2i}|w_i,\, y_{1i}=1\right) = g(h(w_i^T,\, \theta^*))$ and $h(w_i^T,\, \theta^*)$ is a scalar unobservable index, under appropriate identification assumptions.

The proposed estimation procedure for the high-dimensional linear coefficients in the main equation in this chapter is the penalized version of a projection-type strategy. In the first-stage, given consistent estimates $\hat{\theta}$ of θ^* in the selection equation obtained using one of several methods recently proposed in the high-dimensional statistics literature, estimates $w_i^T\hat{\theta}$ of the "single index" variables $w_i^T\theta^*$ are formed. In the second-stage, nonparametric regression is performed to obtain estimate $\hat{\mathbb{E}}\left(x_{ij}|w_i^T\hat{\theta}\right)$ of $\mathbb{E}\left(x_{ij}|w_i^T\theta^*\right)$ for $j=1,...,p$ and $\hat{\mathbb{E}}\left(y_{2i}|w_i^T\hat{\theta}\right)$ of $\mathbb{E}\left(y_{2i}|w_i^T\theta^*\right)$; then the estimated residuals $\hat{v}_i = \left(x_{i1} - \hat{\mathbb{E}}\left(x_{i1}|w_i^T\hat{\theta}\right),\, ...\, , x_{ip} - \hat{\mathbb{E}}\left(x_{ip}|w_i^T\hat{\theta}\right)\right)$ of v_i and $\hat{v}_{i0} = y_{2i} - \hat{\mathbb{E}}\left(y_{2i}|w_i^T\hat{\theta}\right)$ of v_{i0} are formed. This step is motivated by the estimator of Robinson (1988) for semilinear models. The second-stage estimation in this chapter involves $p+1$ nonparametric regressions where $p \asymp n$ or $p \gg n$, and in contrast to the classical low-dimensional settings (e.g., Robinson 1988), a more careful control for the noise from the $p+1$ nonparametric regressions is required. In particular, the prediction errors of the nonparametric procedures are shown in this chapter to satisfy

$$\mathbb{P}\left\{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\mathbb{E}}\left(z_{ij}|w_i^T\hat{\theta}\right) - \mathbb{E}\left(z_{ij}|w_i^T\hat{\theta}\right)\right]^2} \geq t\right\} \leq c\exp\left(-nt^2\right)$$

where $z_{ij} = x_{ij}$ for $j = 1, ..., p$ and $z_{i0} = y_{2i}$, and as a consequence,

$$\begin{aligned}\mathbb{P}\left\{\max_{j=0,...,p}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\mathbb{E}}\left(z_{ij}|w_i^T\hat{\theta}\right) - \mathbb{E}\left(z_{ij}|w_i^T\hat{\theta}\right)\right]^2} \geq t\right\} &\leq c\exp\left(-nt^2 + \log p\right) \\ &= O\left(\frac{1}{p}\right)\end{aligned}$$

where the last equality holds provided n is sufficiently large. The tail bounds above can be ensured by considering the family of nonparametric least squares estimators or regularized nonparametric least squares estimators defined in van de Geer (2000). This family of estimators include linear regression as the simplest case, sparse linear regressions, convex regression, Lipschitz and Isotonic regression, kernel ridge regression based on reproducing kernel Hilbert spaces, estimators based on series expansion, sieves and spline methods. A procedure based on Lipschitz regression for the second-stage nonparametric estimation is illustrated in this chapter for a leading case example.

In the third-stage, regressing $\hat{v}_{i0}$ on $\hat{v}_i$ with l_1-regularization to estimate the main-equation coefficients β^*. In particular, for the third-stage estimation, this chapter considers a non-pivotal Lasso procedure whose regularization parameter depends on the unknown variance of η_i, and a pivotal Dantzig selector whose regularization parameter does not involve the unknown variance of η_i. This pivotal Dantzig selector was originally proposed by Gautier and Tsybakov (2011) in the context of instrumental variables regression. A by-product of the pivotal procedure is a set of non-asymptotic confidence intervals (which also do not involve the unknown variance of η_i). Upon the availability of estimates of the high-dimensional linear coefficients, two different estimation strategies for the selection bias function are proposed: one is a closed form estimator and the other is a nonparametric least squares estimator. Despite that the nonparametric least squares estimator of $g(w_i^T\theta^*)$ is computationally more involved relative to the closed-form estimator, its rate of convergence turns out to be faster. In particular, when β^* is approximately sparse in the sense that it belongs to the l_1-ball, the closed-form estimator cannot achieve *MSE*-consistency even if $n \rightarrow \infty$ while the nonparametric least squares estimator is consistent in *MSE* in this special case.

While existing semiparametric estimation techniques for the selection models limit the number of regressors entering the selection equation and the main equation, the multi-stage estimation procedure with l_1-regularization in the first- and third-stage are more flexible and particularly powerful for applications in which the vector of parameters of interests is high-dimensional but sparse and there is lack of information about the relevant explanatory variables. Moreover, this above-mentioned high-dimensional multi-stage estimation procedure is intuitive and can be easily implemented using existing software packages. In particular, it decomposes the joint search of the optimal values for the high-dimensional linear coefficients and the nonparametric selection bias component into several sequential searches with each search defined over a much smaller parameter space. In particular, the second-stage estimation incurs a computational cost linear in p as it involves solving $p+1$ independent subproblems and each subproblem can be in general solved with a polynomial-time algorithm. The computational efficiency of the first-stage and third-stage estimations is guaranteed by existing algorithms developed for solving the Lasso or the Dantzig program. Upon the availability of estimates of the high-dimensional linear coefficients, the estimator for the selection bias function is simply a closed form estimator or a nonparametric least squares estimator. In additional to the computational efficiency, as we will see in Section 3.4.4 that, under some conditions and when β^* is exactly sparse, the proposed procedures for estimating β^* and $g(\cdot)$ are overall statistically efficient up to the $(n,\, d,\, p)-$factors, relative to any procedure constructed based on model (3.2) for estimating model (3.1), regardless of its computational cost.

The main theoretical results of this chapter are finite-sample bounds from which sufficient scaling conditions on the sample size for estimation consistency in l_2-norm and variable-selection consistency (i.e., the multi-stage high-dimensional estimator correctly selects the non-zero coefficients in the main equation with high probability) are established. These results imply that the estimate from performing the Lasso-type procedures in the third-stage estimation is l_2-consistent as long as β^* belongs to the $l_{q_2}-$"balls" with $q_2 \in [0, 1]$ but inconsistent when $q_2 > 1$. A technical issue related to a set of high-level assumptions on the regressors for estimation consistency and selection consistency arises in the multi-stage estimation procedure from allowing the number of regressors in the main equation to exceed n and this chapter provides analysis to verify these conditions. These verifications also provide a finite-sample guarantee of the population identification condition required by the semiparametric selection models. It is worth mentioning that the multi-stage estimator and the general results in this chapter can be applied to other high-dimensional sparse semiparametric models. Section 3.4.5 discusses estimation of a certain type of high-dimensional semilinear models with the proposed multi-stage strategy when the number of parameters in the linear and (additive) nonparametric components are large relative to the sample size. Statistical efficiency of the proposed estimators is studied via lower bounds on minimax risks and the result shows that, for a class of models with exactly sparse β^* that has at most k_2 non-zero coefficients, the overall convergence rate of the estimator of the high-dimensional linear coefficients in the main equation and the nonparametric least squares estimator of the selection bias function matches the theoretical lower bound up to the $(n, d, p)-$factors, and exceeds it at most by a factor of $k_2^{3/2}$. This statistical efficiency result, however, does not apply to the case where β^* is approximately sparse.

Other theoretical contributions of this chapter include establishing the non-asymptotic counterpart of the familiar asymptotic "oracle" type of results from previous literature: the estimator of the coefficients in the main equation behaves as if the unknown nonparametric component were known, provided the nonparametric component is sufficiently smooth. This new "oracle" result holds for a unified framework of nonparametric least squares estimators and regularized nonparametric least squares estimators considered in the second-stage estimation. In general, for a semiparametric model with two additive components one parametric (linear in parameters) and the other nonparametric, if the prediction error or the $\sqrt{MSE}$ (the square root of the mean squared error) of the nonparametric estimation *per se* is $O_p(t_n)$, this chapter shows that the error arising from not knowing the functional form contributes $O_p\left(t_n^2\right)$ in the l_2-error of the estimator of β^*. The driver behind this "oracle" result lies in the projection strategy. An application of this general result to classical low-dimensional semilinear models would imply that the nonparametric component needs to be estimated at a rate *no slower* than $O\left((\frac{1}{n})^{\frac{1}{4}}\right)$ in order for the estimator of the parametric component to achieve the rate of $O\left(\sqrt{\frac{1}{n}}\right)$. In contrast to the semilinear models, the low-dimensional selection models require the rate of the nonparametric component to be at least $O\left((\frac{1}{n})^{\frac{1}{3}}\right)$ because the nonparametric component in the selection model involves the unknown parameters θ^* that also need to be estimated.

The high-dimensional multi-stage procedure is illustrated with an application to the retail gasoline market in the Greater Saint Louis area. Gasoline stations choose to be either a two-product station offering both self-service and full-service gasoline or a single-product station offering only full-service or self-service gasoline. While a single-product station is unable to price discriminate, a two-product station can charge different prices for full- and self-service gasoline and induce consumers with different valuations to choose the products consistent with their preferences. Similar to Iyer and Seetharaman (2003), this chapter models a retailer's incentive to price discriminate by choosing either single-product or multi-product as a function of market and station characteristics and then models the retailer's pricing decision, conditional on the choice of the product type. However, Iyer and Seetharaman (2003) did not account for interactions between the gas stations in their empirical analysis. This chapter uses geographic information and spatial data to introduce, in the main equation related to the retailers' pricing decisions, a set of variables that are high-dimensional to control for interactions between the gas stations and employ a proposed estimator to identify the competitive market structure. In contrast to other heuristic ways of defining competitive markets as typically seen in the retail gasoline industry literature, the proposed method in this chapter is natural and data-driven. The empirical finding highlights the importance of accounting for potential interactions between stations and suggests that competition effects from retailers that are not in the same local market should not be overlooked.

Section 3.2 presents identification assumptions required for model (3.1) in high-dimensional settings. The estimation procedures are introduced in Section 3.3. Theoretical results are established in Section 3.4. Small-sample performance of the proposed multi-step high-dimensional estimator is evaluated with Monte-Carlo simulations in Section 3.5 and applied to the retail gasoline market in Section 3.6. Section 3.7 concludes this chapter. Proofs of the main results are collected in Section 3.8, with the remaining proofs of technical lemmas contained in Section 3.9.

3.2 Identification assumptions

Notation. The l_q norm of a vector $v \in p \times 1$ is denoted by $|v|_q$, $1 \leq q \leq \infty$ where $|v|_q := (\sum_{i=1}^{p} |v_i|^q)^{1/q}$ when $1 \leq q < \infty$ and $|v|_q := \max_{i=1,...,p} |v_i|$ when $q = \infty$. For a matrix $A \in \mathbb{R}^{p \times p}$, write $|A|_\infty := \max_{i,j} |a_{ij}|$ to be the elementwise $l_\infty-$norm of A. The l_2-operator norm, or spectral norm of the matrix A corresponds to its maximum singular value; i.e., it is defined as $\|A\|_2 := \sup_{v \in S} |Av|_2$, where $S = \{v \in \mathbb{R}^p \mid |v|_2 = 1\}$. The l_∞ matrix norm (maximum absolute row sum) of A is denoted by $\|A\|_\infty := \max_i \sum_j |a_{ij}|$ (note the difference between $|A|_\infty$ and $\|A\|_\infty$). For a square matrix A, denote its minimum eigenvalue and maximum eigenvalue by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$, respectively. The $\mathcal{L}_2(\mathbb{P})-$error of a vector $\Delta(x)$, denoted by $|\Delta|_{\mathcal{L}_2(\mathbb{P})}$, is given by $[\mathbb{E}_{\mathbb{X}}(\Delta(x))^2]^{\frac{1}{2}}$. Define $\mathbb{P}_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$ that places a weight $\frac{1}{n}$ on each observation x_i for $i = 1, ..., n$, and the associated $\mathcal{L}_2(\mathbb{P}_n)-$norm of the vector $\Delta := \{\Delta(x_i)\}_{i=1}^{n}$, denoted by $|\Delta|_n$, is given by $\left[\frac{1}{n}\sum_{i=1}^{n} (\Delta(x_i))^2\right]^{\frac{1}{2}}$. For a vector $v \in \mathbb{R}^p$, let $J(v) = \{j \in \{1, ..., p\} \mid v_j \neq 0\}$ be its support, i.e., the set of indices corresponding to its non-zero components v_j. The cardinality of a set $J \subseteq \{1, ..., p\}$ is denoted by $|J|$. For functions $f(n)$ and $g(n)$, write $f(n) \gtrsim g(n)$ to mean that $f(n) \geq cg(n)$ for a universal constant $c \in (0, \infty)$ and similarly, $f(n) \lesssim g(n)$ to mean that $f(n) \leq c^{'}g(n)$ for a universal constant $c^{'} \in (0, \infty)$, and $f(n) \asymp g(n)$ when $f(n) \gtrsim g(n)$ and $f(n) \lesssim g(n)$ hold simultaneously. Also denote $\max\{a, b\}$ by $a \vee b$ and $\min\{a, b\}$ by $a \wedge b$.

The following assumptions are imposed on model (3.1).

Assumption 3.2.1 (Sampling): The data $\{y_{1i}, y_{2i}, w_i, x_i\}$ are *i.i.d.* with finite second moments.

Remark. The identicalness of $\{y_{1i}, y_{2i}, w_i, x_i\}$ in Assumption 3.2.1 can be relaxed with the condition that $\{\epsilon_{1i}, \eta_i\}$ are identically distributed but $\{w_i, x_i\}$ are not.

Assumption 3.2.2 (Sparsity): The coefficient vector $\beta^* \in \mathbb{R}^p$ belongs to the $l_{q_2}-$"balls" $\mathcal{B}_{q_2}^p(R_{q_2})$ for a "radius" of R_{q_2} and some $q_2 \in [0, 1]$, where the l_q-"balls" of "radius" R

for $q \in [0,\, 1]$ are defined by

$$\begin{aligned} \mathcal{B}_q^p(R) &:= \left\{ \beta \in \mathbb{R}^p \mid |\beta|_q^q = \sum_{j=1}^{p} |\beta_j|^q \leq R \right\} \text{ for } q \in (0,\, 1] \\ \mathcal{B}_0^p(R) &:= \left\{ \beta \in \mathbb{R}^p \mid |\beta|_0 = \sum_{j=1}^{p} \mathbb{I}\left[\beta_j \neq 0\right] \leq R \right\} \text{ for } q = 0. \end{aligned}$$

Remark. Assumption 3.2.2 requires the coefficient vector to be "sparse". As one might expect, if the high-dimensional model lacks any additional structure, then there is no hope of obtaining consistent estimators of β^* when the ratio $\frac{p}{n}$ stays bounded away from 0. For this reason, when working in settings in which $p > n$, it is necessary to impose some type of sparsity assumptions on the unknown coefficient vector $\beta^* \in \mathbb{R}^p$. Assumption 3.2.2 formalizes the sparsity condition by considering the l_q-"balls" $\mathcal{B}_q^p(R_q)$ of "radius" R_q where $q \in [0,\, 1]$. The exact sparsity on β^* corresponds to the case of $q = q_2 = 0$ with $R_{q_2} = k_2$ (in this chapter, the subscript "2" is generally reserved for the main-equation related parameters and the subscript "1" for the selection-equation related parameters), which says that β^* has at most k_2 non-zero components, where the sparsity parameter k_2 is also allowed to increase to infinity with n but slowly compared to n. In the more general setting $q_2 \in (0,\, 1]$, membership in $\mathcal{B}_{q_2}^p(R_{q_2})$ has various interpretations and one of them involves how quickly the ordered coefficients decay. When $q_2 \in [0,\, 1)$, the set $\mathcal{B}_{q_2}^p(R_{q_2})$ is non-convex and the l_1-ball is the *closest convex* approximation of these non-convex sets. In terms of algorithm design, the idea of approximating non-convex problems with their closest convex member (so called "convex relaxation") provides a tremendous computational advantage. This is one of the reasons for favoring the l_1-penalization techniques such as the Lasso (in solving high-dimensional regression problems with sparsity described by the l_q-constraint where $q \in [0,\, 1]$) over estimators based on the l_q-penalty with $q \in [0,\, 1)$ which are computationally more difficult (see the Bridge estimator in Huang, Horowitz, and Ma, 2008 as an example of these nonconvex penalization procedures) and estimators based on l_q-penalty with $q > 1$ (such as the ridge-penalty) which are not the closest convex approximations. On the other hand, if the coefficient vector belongs to an Euclidean ball (the l_2-ball), then it would make more sense to apply a ridge penalty. The focus of this chapter is on high-dimensional sparse β^* that belongs to $\mathcal{B}_{q_2}^p(R_{q_2})$ for $q_2 \in [0,\, 1]$.

Assumption 3.2.3 (Restricted Identifiability): For a subset $S \subseteq \{1,\, 2, ..., p\}$ and all non-zero $\Delta \in \mathbb{C}(S;\, q_2,\, \varphi) \cap \mathbb{S}_\delta$ where

$\mathbb{C}(S;\, q_2,\, \varphi) := \{\Delta \in \mathbb{R}^p \,:\, |\Delta_{S^c}|_1 \leq \varphi|\Delta_S|_1 + (\varphi + 1)|\beta^*_{S^c}|_1\}$ for some constant $\varphi \geq 1$, (with Δ_S denoting the vector in $\mathbb{R}^p$ that has the same coordinates as Δ on S and zero coordinates on the complement S^c of S) and

$$\mathbb{S}_\delta := \{\Delta \in \mathbb{R}^p \,:\, |\Delta|_2 \geq \delta\},$$

the matrix $\mathbb{E}\left[y_{1i}v_i^T v_i\right]$ satisfies

$$\frac{\Delta^T \mathbb{E}\left[y_{1i}v_i^T v_i\right]\Delta}{|\Delta|_2^2} \geq \kappa_L > 0,$$

where

$$v_i = \left(x_{i1} - \mathbb{E}\left(x_{i1}|w_i^T\theta^*\right), \ldots, x_{ip} - \mathbb{E}\left(x_{ip}|w_i^T\theta^*\right)\right).$$

Remark. Assumption 3.2.3 is the high-dimensional counterpart of the familiar identification assumption in the low-dimensional selection model literature (e.g., Powell 1989; Newey, 1991; Ahn and Powell, 1993), which assumes the matrix $\mathbb{E}\left[y_{1i}v_i^T v_i\right]$ is positive definite uniformly over all $\Delta \in \mathbb{R}^p \backslash \{\mathbf{0}\}$. When v_i is a zero-mean Gaussian matrix with covariance $\mathbb{E}\left[y_{1i}v_i^T v_i\right] = \sigma^2 I_{p\times p}$, the smallest eigenvalue of $\mathbb{E}\left[y_{1i}v_i^T v_i\right]$ is σ^2, so the traditional identification condition in the low-dimensional case naturally carries to the high-dimensional case. However, for more general structures on $\mathbb{E}\left[y_{1i}v_i^T v_i\right]$, while this traditional identification condition is plausible for small p, it may become harder to be satisfied when p is large. Assumption 3.2.3 relaxes the uniform positive definiteness but only requires it to hold over a restricted set $\mathbb{C}(S;\, q_2,\, \varphi) \cap \mathbb{S}_\delta$ so that the special case of $x_i \subset w_i$ is allowed even in the high-dimensional settings (the choices of δ and S will be made clear in Section 3.4 when the theoretical results are presented.). If $x_i \subset w_i$, Assumption 3.2.3 says that for any non-zero vector $\lambda \in \mathbb{C}(S;\, q_2,\, \varphi) \cap \mathbb{S}_\delta$, there is no measurable function $f(w_i^T\theta^*)$ such that $x_i^T\lambda = f(w_i^T\theta^*)$ when $y_{1i} = 1$. Consequently, there is at least one component w_{ij} with θ_j^* in the support set of θ^* (namely, the set of non-zero components in θ^*) such that w_{ij} is excluded from x_i. This necessary condition is the high-dimensional extension of the familiar "exclusion restriction" condition in the low-dimensional selection model literature.

When β^* is exactly sparse (namely, $q_2 = 0$), we can take $\delta = 0$ and choose $S = J(\beta^*)$ (where $J(\beta^*)$ denotes the support of β^*), which reduces the set $\mathbb{C}(S;\, q_2,\, \varphi) \cap \mathbb{S}_\delta$ to the following cone:

$$\mathbb{C}(J(\beta^*);\, 0,\, \varphi) := \left\{\Delta \in \mathbb{R}^p \,:\, |\Delta_{J(\beta^*)^c}|_1 \leq \varphi |\Delta_{J(\beta^*)}|_1\right\}.$$

The sample analog of Assumption 3.2.3 over the cone $\mathbb{C}(J(\beta^*);\, 0,\, \varphi)$ is the so-called *restricted eigenvalue* condition on the Gram matrix $\frac{v^T v}{n}$, studied in Bickel, et. al. (2009), Meinshausen and Yu (2009), Raskutti, et al. (2010), Bühlmann and van de Geer (2011), Loh and Wainwright (2012), Negahban, et. al. (2012), etc. Note that in the low-dimensional setting where $p < n$, as long as $\text{rank}(v) = p$, we are guaranteed that the Gram matrix $\frac{v^T v}{n}$ is positive definite. In the high-dimensional setting with $p > n$, the matrix $\frac{v^T v}{n}$ is a $p \times p$ matrix with rank at most n, so it is impossible to have the uniform positive definiteness. It is well-known that the *restricted eigenvalue* assumption, defined more precisely below, is a sufficient condition for the l_2- consistency of the Lasso estimator for the sparse linear regression models in high-dimensional settings. To motivate the restricted set $\mathbb{C}(J(\beta^*); 0, \varphi)$, note that the vectors Δ in this cone have a substantial part of their "mass" concentrated on a set of the cardinality of $J(\beta^*)$. The vectors Δ of interests often concern the error $\hat{\beta} - \beta^*$ where $\hat{\beta}$ is some estimate of β^*. When the high-dimensional sparse linear regression models are estimated by the l_1-penalized techniques, an appropriate choice of the regularization parameter would generally ensure the error $\hat{\beta} - \beta^*$ to be in this restricted set.

In the high-dimensional setting, a sufficient condition for the l_2- consistency of the Lasso estimator $\hat{\beta}_{Las}$ is the *restricted eigenvalue* (RE) condition related to the positive definiteness of the Gram matrix $\frac{X^T X}{n}$ over a restricted set (see, e.g., Bickel, et. al., 2009; Meinshausen and Yu, 2009; Raskutti, et al., 2010; Bühlmann and van de Geer, 2011; Loh and Wainwright 2012; Negahban, et. al., 2012; etc.). Consider the following definition of the RE condition given by Negahban, et. al. (2012) and Wainwright (2015).

Definition 3.2.1 (RE condition). For $q \in [0, 1]$, the matrix $X \in \mathbb{R}^{n \times p}$ satisfies the RE condition over a subset $S \subseteq \{1, 2, ..., p\}$ with parameters $(q, \delta, \kappa, \varphi)$ if

$$\frac{\frac{1}{n}|X\Delta|_2^2}{|\Delta|_2^2} \geq \kappa > 0 \qquad \text{for all nonzero } \Delta \in \mathbb{C}(S;\, q,\, \varphi) \cap \mathbb{S}_\delta, \tag{3.4}$$

where $\mathbb{C}(S; q, \varphi) \cap \mathbb{S}_\delta$ is defined in Assumption 3.2.3.

As discussed previously, when the unknown vector $\beta^* \in \mathbb{R}^p$ is exactly sparse, the set $\mathbb{C}(S; q, 3) \cap \mathbb{S}_\delta$ is reduced to the cone $\mathbb{C}(J(\beta^*); 0, 3)$. When β^* is approximately sparse (namely, $q \in (0, 1]$), in sharp contrast to the case of exact sparsity, the set $\mathbb{C}(S; q, 3)$ is no longer a cone but rather contains a ball centered at the origin. As a consequence, it is never possible to ensure that $\frac{|X\Delta|_2^2}{n}$ is bounded from below for all vectors Δ in the set $\mathbb{C}(S; q, 3)$ (see Negahban, et. al., 2012 for a geometric illustration of this issue). For this reason, in order to obtain a general applicable theory, it is crucial to further restrict the set $\mathbb{C}(S; q, 3)$ for $q \in (0, 1]$ by introducing the set

$$\mathbb{S}_\delta := \{\Delta \in \mathbb{R}^p \,:\, |\Delta|_2 \geq \delta\},$$

where $\delta > 0$ is some parameter depending on the choice of the regularization parameter λ_n in the Lasso program (3.6). Provided the parameter δ and the set S are suitably chosen, the intersection $\mathbb{C}(S;\, q,\, 3) \cap \mathbb{S}_\delta$ excludes many "flat" directions (with eigenvalues of 0) in the space for the case of $q \in (0,\, 1]$. To the best of my knowledge, the necessity of this additional set $\mathbb{S}_\delta$, essential for the approximately sparse case of $q \in (0,\, 1]$, is first recognized explicitly in Negahban, et. al. (2012).

Raskutti et al. (2010) shows that the RE condition (3.4) is satisfied by the design matrix $X \in \mathbb{R}^{n \times p}$ formed by independently sampling each row $X_i \sim N(0,\, \Sigma)$. Rudelson and Zhou (2011) as well as Loh and Wainwright (2012) extend the verification of the RE condition from the case of Gaussian designs to the case of sub-Gaussian designs. The sub-Gaussian assumption says that the explanatory variables need to be drawn from distributions with well-behaved tails like Gaussian. In contrast to the Gaussian assumption, sub-Gaussian variables constitute a more general family of distributions.

When applying the proposed multi-stage procedure in this chapter to estimate the high-dimensional selection models, there is no guarantee that the random matrix $\frac{\hat{v}^T \hat{v}}{n}$ (where $\hat{v}_i$ are the estimates of $v_i = x_i - \mathbb{E}\left(x_i | w_i^T \theta^*\right)$ for $i = 1,\, ...,\, n$) would automatically satisfy these previously established conditions for estimation consistency. For a broad class of sub-Gaussian matrices formed by the true residuals $v_i = x_i - \mathbb{E}\left(x_i | w_i^T \theta^*\right)$ for $i = 1,\, ...,\, n$ whenever $y_{1i} = 1$, this chapter provides results that imply the RE condition (3.4) holds for $\hat{v}^T \hat{v}$ with high probability provided Assumption 3.2.3 is satisfied. Verifications of the RE condition provide a finite-sample guarantee of Assumption 3.2.3 when the unknown residuals v are replaced with their estimate $\hat{v}$ and the expectation is replaced with a sample average.

While the RE assumption is a natural sufficient condition for analyzing l_2-consistency of the Lasso estimator $\hat{\beta}_{Las}$, l_2-consistency of the Dantzig selector $\hat{\beta}_{Dan}$ can be related to a different sufficient condition, the *sensitivity characteristics*, on the term $|X^T X v|_\infty$. These sensitivity characteristics were originally introduced in Ye and Zhang (2010) as the cone invertibility factors and used in Gautier and Tsybakov (2011) for high-dimensional instrumental variable regressions. Gautier and Tsybakov (2011) shows that the sensitivity characteristics can be larger than the usual RE condition of Bickel, et. al (2009) and therefore the Dantzig-type estimators may lead to better results in certain cases[1]. The analysis of a pivotal Dantzig selector in this chapter for estimating the high-dimensional linear coefficients relies on the following definition based on Gautier and Tsybakov (2011):

[1]Recently, another weaker version of the RE condition tailored to the square-root Lasso is developed in Belloni, Chernozhukov, and Wang (2014).

Definition 3.2.2 (l_2-*sensitivity*). The matrix $X \in \mathbb{R}^{n\times p}$ satisfies the l_2-*sensitivity* condition over a subset $S \subseteq \{1, 2, ..., p\}$ with parameters $(q,\, \delta,\, \kappa^{'},\, \varphi)$ if

$$\frac{\frac{1}{n}|X^T X\Delta|_\infty^2}{|v|_2^2} \geq \kappa^{'} > 0 \qquad \text{for all nonzero } \Delta \in \mathbb{C}(S;\, q,\, \varphi) \cap \mathbb{S}_\delta \tag{3.5}$$

where $\mathbb{C}(S;\, q,\, \varphi) \cap \mathbb{S}_\delta$ is defined in Assumption 3.2.3.

When the response variable is a latent variable with only an observable sign, other models such as the high-dimensional binary response models may be considered. In analyzing these models, the RE condition can be replaced with a similar notion, the *restricted strong convexity* (RSC) condition, originally formalized by Negahban, et. al. (2012) in the context of the regularized M-estimation with a general, convex and differentiable loss function. The following definition from Negahban, et. al. (2012) is adopted in this chapter to analyze the theoretical properties of an estimator for the high-dimensional logit and probit model:

Definition 3.2.3 (RSC condition). A convex and differentiable loss function $\mathcal{L}(\theta;\, z_1^n)$ satisfies the RSC condition over a subset $S \subseteq \{1, 2, ..., p\}$ with parameters $(q,\, \delta,\, \kappa^{''},\, \varphi)$ where $\kappa^{''} > 0$ if

$$\mathcal{L}(\theta^* + \Delta;\, z_1^n) - \mathcal{L}(\theta^*;\, z_1^n) - \langle \nabla\mathcal{L}(\theta^*;\, z_1^n),\, \Delta\rangle \geq \kappa^{''}\, |\Delta|_2^2$$

for all nonzero $\Delta \in \mathbb{C}(S;\, q,\, \varphi) \cap \mathbb{S}_\delta$, where $\nabla\mathcal{L}(\theta^*;\, z_1^n)$ denotes the derivative of $\mathcal{L}(\theta;\, z_1^n)$ evaluated at $\theta = \theta^*$, and $\mathbb{C}(S;\, q,\, \varphi) \cap \mathbb{S}_\delta$ is defined in Assumption 3.2.3.

3.3 Estimation procedures

This section presents a 3-stage estimation procedure for the high-dimensional linear coefficients in the main equation and two estimators of the selection bias function. In terms of applicability, the proposed estimators enjoy many computational advantages and can be easily implemented using existing software packages.

3.3.1 The multi-stage estimator of the high-dimensional linear coefficients

To facilitate the presentation of the multi-stage estimator, we reverse the order of the three stages when discussing the estimation procedure; in particular, we will introduce the third-stage estimator and then followed by the second-stage and the first-stage estimators. Note that the second- and third-stage estimations concern only the selected sample (observations with $y_{1i} = 1$) and the first-stage estimation concerns the entire sample. For the third-stage estimation, this chapter considers a non-pivotal Lasso procedure whose regularization parameter depends on the unknown variance of η_i, and a pivotal Dantzig selector (Gautier and Tsybakov, 2011) whose regularization parameter does not involve the unknown variance of η_i.

Non-pivotal third-stage estimation
Revisiting equation (3.3) in Section 3.1 suggests that if an estimate of

$$\left(\mathbb{E}\left(x_{i1}|w_i^T\theta^*\right), \ldots, \mathbb{E}\left(x_{ip}|w_i^T\theta^*\right)\right)$$

is available to us, then we can form estimates

$$\begin{aligned}\hat{v}_i &= \left(x_{i1} - \hat{\mathbb{E}}\left(x_{i1}|w_i^T\hat{\theta}\right), \ldots, x_{ip} - \hat{\mathbb{E}}\left(x_{ip}|w_i^T\hat{\theta}\right)\right), \\ \hat{v}_{i0} &= y_{2i} - \hat{\mathbb{E}}\left(y_{2i}|w_i^T\hat{\theta}\right).\end{aligned}$$

of the nonparametric residuals

$$\begin{aligned}v_i &= \left(x_{i1} - \mathbb{E}\left(x_{i1}|w_i^T\theta^*\right), \ldots, x_{ip} - \mathbb{E}\left(x_{ip}|w_i^T\theta^*\right)\right), \\ v_{i0} &= y_{2i} - \mathbb{E}\left(y_{2i}|w_i^T\theta^*\right).\end{aligned}$$

Then, an estimator of the high-dimensional linear coefficients in the main equation (the **third-stage** estimator) can be obtained by performing the following Lasso program:

$$\hat{\beta}_{HSEL} \in \text{argmin}_{\beta\in\mathbb{R}^p} : \frac{1}{2n}|\hat{v}_0 - \hat{v}\beta|_2^2 + \lambda_{n,3}|\beta|_1, \tag{3.6}$$

where $\lambda_{n,3} > 0$ is some regularization parameter whose choice is to be discussed in Section 3.4. In general, the choice of $\lambda_{n,3}$ depends on $\mathbb{E}(v_{ij}^2)$ and $\mathbb{E}(\eta_i^2)$. To make $\lambda_{n,3}$ and the estimate $\hat{\beta}_{HSEL}$ independent of the effect from $\mathbb{E}(v_{ij}^2)$, we can impose weights on the penalty term as follows

$$\min_{\beta\in\mathbb{R}^p} : \frac{1}{2n}|\hat{v}_0 - \hat{v}\beta|_2^2 + \lambda_{n,3}\sum_{j=1}^{p}\hat{\sigma}_{v_j}|\beta_j|, \tag{3.7}$$

where $\hat{\sigma}_{v_j} := \sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{v}_{ij}^2}$. To make $\lambda_{n,3}$ not depend on the unknown variance of η_i, we can consider the pivotal version of the Dantzig selector as in Gautier and Tsybakov

(2011).

Pivotal third-stage estimation
Set $v_{j*} := \max_{i\in\{1,\dots,n\}}\left(\max\left\{|2x_{ij}|\,,\,|\hat{v}_{ij}|\right\}\right)$ for $j = 1, \dots, p$ and denote D the diagonal $p \times p$ matrix with diagonal entries v_{j*}^{-1}. Consider the following optimization problem:

$$\min_{(\beta,\,\sigma)\in A}\,:\,\left(\left|D^{-1}\beta\right|_1 + C\sigma\right) \tag{3.8}$$

where

$$A = \left\{(\beta,\,\sigma)\,:\,\beta\in\mathbb{R}^p,\,\sigma>0,\,\frac{1}{n}\left|D\hat{v}^T(\hat{v}_0-\hat{v}\beta)\right|_\infty \le \sigma\xi,\,\frac{1}{n}\left|\hat{v}_0-\hat{v}\beta\right|_2^2\le\sigma^2\right\}$$

for some tuning parameter $\xi > 0$ (to be specified in Section 3.4). The computational aspect of this pivotal estimator is detailed in Gautier and Tsybakov (2011).

Remark. The third-stage estimation needs not to be restricted to the Lasso or the Dantzig selector. Other methods with different loss functions (such as the square-root Lasso in Belloni, et. al 2011, 2014) or with different penalty functions (such as the SCAD in Fan and Li, 2001, or the MCP in Zhang, 2010) can be used. This chapter focuses on the analysis of the non-pivotal Lasso and the pivotal Dantzig selector laid out above for the third-stage estimation.

Second-stage estimation
To simplify the notations in the following, write $\mathbb{E}\left(x_{ij}|w_i^T\theta\right) := m_j(w_i^T\theta)$, $\hat{\mathbb{E}}\left(x_{ij}|w_i^T\theta\right) := \hat{m}_j(w_i^T\theta)$, $\mathbb{E}\left(y_{2i}|w_i^T\theta\right) := m_0(w_i^T\theta)$, and $\hat{\mathbb{E}}\left(y_{2i}|w_i^T\theta\right) := \hat{m}_0(w_i^T\theta)$. To estimate $m_j(w_i^T\theta^*)$ for each $j = 0, \dots, p$, we first need some estimate $\hat{\theta}$ of θ^*in the selection equation. Supposing such an estimate is available, to obtain a (**second-stage**) estimator of $m_j(w_i^T\theta^*)$, we consider the following least squares estimator

$$\hat{m}_j \in \arg\min_{\tilde{m}_j\in\mathcal{F}_j}\left\{\frac{1}{n}\sum_{i=1}^n\left(z_{ij}-\tilde{m}_j(w_i\hat{\theta})\right)^2\right\}, \tag{3.9}$$

or the regularized least-squares estimator

$$\hat{m}_j \in \arg\min_{\tilde{m}_j\in\mathcal{F}_j}\left\{\frac{1}{n}\sum_{i=1}^n\left(z_{ij}-\tilde{m}_j(w_i\hat{\theta})\right)^2 + \lambda_{nj,2}\left|\tilde{m}_j\right|_{\mathcal{F}_j}^2\right\}, \tag{3.10}$$

where $|\cdot|_{\mathcal{F}_j}$ is a norm associated with the function class $\mathcal{F}_j$ and $\lambda_{nj,2} \ge 0$ is a regularization parameter and $z_{i0} = y_{2i}$ and $z_{ij} = x_{ij}$ for each $j = 1, \dots, p$. The choice of $\lambda_{nj,2}$ is specified in Section 3.4. A nonparametric regression problem based on (3.9) or (3.10) is a standard setup in many modern statistics books (e.g., van der Vaart and Wellner, 1996; van de Geer, 2000; Wainwright, 2015, etc).

In words, the solutions to program (3.9) are least-squares estimators based on imposing explicit constraints on the function class $\mathcal{F}_j$. The function $\hat{m}_j$ is chosen such that the vector

$$\left(\hat{m}_j(w_1\hat{\theta}),\, ...,\, \hat{m}_j(w_i\hat{\theta}),\, ...,\, \hat{m}_j(w_n\hat{\theta})\right)$$

is closest in l_2-norm to the observation $(z_{1j},\, ...,\, z_{ij},\, ...,\, z_{nj})$ for $j = 0, ..., p$ in terms of the "selected" sample. Examples of (3.9) include the linear regression as the simplest case, sparse linear regressions, convex regression where $\mathcal{F}_j$ is the class of convex functions (e.g., Guntuboyina and Sen, 2013), Lipschitz and Isotonic regression where $\mathcal{F}_j$ is the class of monotone Lipschitz functions (e.g., Kakade, Kalai, Kanade, and Shamir, 2011), etc. In general, this optimization problem defining the non-parametric least squares estimator $\hat{m}_j$ is infinite-dimensional in nature, since $\hat{m}_j$ ranges over the function class $\mathcal{F}_j$. If the function class is "too large", the solution may not exist, in which case $\mathcal{F}_j$ is chosen to be a compact subset of some larger function class by introducing a ball radius in some norm. From the computational point of view, it is sometimes more convenient to implement estimators based on explicit penalization or regularization terms as in (3.10). Examples of (3.10) include kernel ridge regression where $|\cdot|_{\mathcal{F}_j}$ is the norm associated with a reproducing kernel Hilbert space (see e.g., Gu, 2002; Berlinet and Thomas-Agnan, 2004; Wainwright, 2015), estimators based on series expansion (e.g., Cencov, 1962; Andrews, 1991; Newey, 1994, 1997), as well as sieves (e.g., van de Geer, 2000; Chen, 2008) and spline methods (e.g., Wahba, 1980, 1990). A procedure based on Lipschitz regression for the second-stage nonparametric estimation is illustrated in Section 3.4 for a leading case.

It is worth mentioning that although the theoretical guarantees of the multi-stage procedure provided by this chapter requires the second-stage estimation to fit into either (3.9) or (3.10), other nonparametric methods including kernel density estimators, local polynomials, etc., could also be a valid second-stage estimator for the multi-stage procedure in the context of high-dimensional semiparametric selection models and verifying those methods both theoretically and empirically is an open question for future research.

First-stage estimation

Note that in the second-stage estimation of $m_j(w_i^T\theta^*)$ for each $j = 0, ..., p$, the coefficient vector θ^* is unknown and needs to be replaced by some "consistent" **first-stage** estimate $\hat{\theta}$. Parametric and semiparametric estimation of θ^* in the classical low-dimensional settings when the dimension of θ^* is small relative to the sample size n is well-studied (see, e.g., Powell 1994; Pagan and Ullah, 1999). In the high-dimensional settings where the dimension of θ^* grows with and exceeds n, estimation of θ^* in recent development of high-dimensional statistics has been focused on the case where θ^* is either exactly sparse or approximately sparse, and a distributional assumption is imposed on the error term in the linear latent utility models in (3.1). Theoretical guarantees have been established for the high-dimensional binary logit models in the context of Generalized Linear Models (GLM) and M-estimation (e.g., van de Geer, 2008; Bühlmann and van de Geer, 2011; Negahban, et. al, 2012; Loh and Wainwright, 2013). While the main theoretical results of this chapter concern estimators of the high-dimensional linear coefficients β^* in the main equation and estimators of the selection bias function $g(\cdot)$, we illustrate here and also in later sections the high-dimensional parametric estimation procedure for the binary logit and probit models as they are considered the work-horse of many empirical literatures and probit models are widely applied to study selection problems.

As for the high-dimensional sparse linear models, it is natural to consider the estimator based on the l_1-*regularized maximum likelihood* for the binary logit and probit models, namely,

$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \left\{ -\frac{1}{n} \sum_{i=1}^{n} y_{1i} \phi_1(w_i^T \theta) + \frac{1}{n} \sum_{i=1}^{n} \phi_2(w_i^T \theta) + \lambda_{n,1} \left| \theta \right|_1 \right\} \tag{3.11}$$

where n is the sample size of all observations. One can easily verify that when $\phi_1(w_i^T\theta) = w_i^T\theta$ and $\phi_2(w_i^T\theta) = \log(1 + \exp(w_i^T\theta))$, the loss function in the above program corresponds to a binary logit model; when

$$\begin{aligned} \phi_1(w_i^T\theta) &= \log \frac{\Phi(w_i^T\theta)}{1 - \Phi(w_i^T\theta)}, \\ \phi_2(w_i^T\theta) &= -\log \left[1 - \Phi(w_i^T\theta) \right] \end{aligned}$$

($\Phi(\cdot)$ is the standard normal c.d.f.), the loss function corresponds to a binary probit model. The loss function in (3.11) is written in terms of the negative of the likelihood and hence the optimization program is a convex minimization problem. This chapter extends the analysis of the theoretical properties of these estimators from the high-dimensional binary logit models to the high-dimensional probit models, and focuses on the semiparametric estimation of β^* instead of θ^*. Developing semiparametric estimation techniques for the high-dimensional sparse discrete choice models based upon weak restrictions on the error distribution is left for future research.

Remark. Upon solving (3.11), strategies such as the thresholded-Lasso or the post-Lasso may be used before the second-stage estimation, which might boost the performance of the multi-stage estimator in certain situations.

3.3.2 Estimators of the selection bias function

Given the estimates $\hat{\theta}$ and $\hat{\beta}$ of θ^* and β^*, there are two ways to estimate the selection bias function $g(w_i^T\theta^*)$. Recalling (3.2) from Section 3.1,

$$y_{2i} = x_i^T\beta^* + g(w_i^T\theta^*) + \eta_i.$$

where by construction $\mathbb{E}\left[\eta_i | w_i,\, x_i,\, y_{1i} = 1\right] = 0$. Taking the conditional expectation of the above leads to

$$\mathbb{E}\left(y_{2i}|w_i^T\theta^*\right) = \mathbb{E}\left(x_i \mid w_i^T\theta^*\right)\beta^* + g(w_i^T\theta^*),$$

and as a result,

$$g(w_i^T\theta^*) = \mathbb{E}\left(y_{2i}|w_i^T\theta^*\right) - \mathbb{E}\left(x_i \mid w_i^T\theta^*\right)\beta^*, \tag{3.12}$$

where

$$\mathbb{E}\left(x_i \mid w_i^T\theta^*\right) := \left(\mathbb{E}\left(x_{i1}|w_i^T\theta^*\right), \ldots, \mathbb{E}\left(x_{ip}|w_i^T\theta^*\right)\right).$$

Replacing $\mathbb{E}\left(y_{2i}|w_i^T\theta^*\right)$, $\mathbb{E}\left(x_i \mid w_i^T\theta^*\right)$, and β^* with their estimates from Section 3.3.1 yields the estimator $\hat{g}(w_i^T\hat{\theta})$ of $g(w_i^T\theta^*)$:

$$\hat{g}(w_i^T\hat{\theta}) := \hat{\mathbb{E}}\left(y_{2i} \mid w_i^T\hat{\theta}\right) - \hat{\mathbb{E}}\left(x_i \mid w_i^T\hat{\theta}\right)\hat{\beta}, \tag{3.13}$$

where $\hat{\mathbb{E}}\left(x_i \mid w_i^T\hat{\theta}\right) := \left(\hat{\mathbb{E}}\left(x_{i1}|w_i^T\hat{\theta}\right), \ldots, \hat{\mathbb{E}}\left(x_{ip}|w_i^T\hat{\theta}\right)\right)$ is the second-stage estimate of $\mathbb{E}\left(x_i \mid w_i^T\theta^*\right) := \left(\mathbb{E}\left(x_{i1}|w_i^T\theta^*\right), \ldots, \mathbb{E}\left(x_{ip}|w_i^T\theta^*\right)\right)$.

Alternatively, like how we obtain the second-stage estimates in Section 3.3.1, one can estimate $g(w_i^T\theta^*)$ by solving the following least-squares estimator

$$\tilde{g} \in \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\left(y_{2i} - x_i^T\hat{\beta} - f(w_i^T\hat{\theta})\right)^2, \tag{3.14}$$

or the regularized least-squares estimator

$$\tilde{g} \in \arg\min_{f\in\mathcal{F}} \left\{\frac{1}{n}\sum_{i=1}^{n}\left(y_{2i} - x_i^T\hat{\beta} - f(w_i^T\hat{\theta})\right)^2 + \lambda_n^* \left|f\right|_{\mathcal{F}}^2\right\},$$

where $\lambda_n^* \geq 0$ is a regularization parameter. Although this alternative estimator $\tilde{g}(w_i^T\hat{\theta})$ of $g(w_i^T\theta^*)$ is computationally more involved relative to the closed-form estimator $\hat{g}(w_i^T\hat{\theta})$, its rate of convergence turns out to be faster in a leading case as we will see in Section 3.4.

3.4 Main theoretical results

For notational simplicity, in the main theoretical results presented below, we assume the regime of interest is $p \geq n$ and $d \geq n$ (i.e., the number of regressors grows with and exceed the sample size n). The modification to allow $p < n$ or $d < n$ is trivial. Also, as a general rule for this chapter, all the b constants denote positive constants that are independent of n, p, d, R_{q_1} and R_{q_2} but possibly depending on model specific parameters; all the c constants denote universal positive constants that are independent of n, p, d, R_{q_1} and R_{q_2} as well as model specific parameters. The specific values of these constants may change from place to place.

Recall from programs (3.9) and (3.10), $\tilde{m}_j(\cdot) \in \mathcal{F}_j$. Suppose $m_j(\cdot) \in \mathcal{F}_j^*$, which may be different from $\mathcal{F}_j$. Define the shifted version of the function class $\mathcal{F}_j$

$$\bar{\mathcal{F}}_j := \left\{ f = f^{'} - f^{''} \, : \, f^{'}, \, f^{''} \in \mathcal{F}_j \right\}.$$

The following assumptions are imposed to obtain the theoretical results in this section.

Assumption 3.4.1: For any $j = 0, ..., p$, $\bar{\mathcal{F}}_j$ is a *star-shaped* function class; i.e., for any $f \in \bar{\mathcal{F}}_j$, the entire line $\{\alpha f, \, \alpha \in [0, \, 1]\}$ is also contained within $\bar{\mathcal{F}}_j$.

Remark. The star-shaped condition is often seen in literature of nonparametric statistics (see e.g., van der Vaart and Wellner, 1996; Wainwright, 2015; and other textbooks on mathematical statistics). It is relatively mild; for instance, it is satisfied whenever the set $\bar{\mathcal{F}}_j$ is convex and contains the function $f = 0$. It is also satisfied by various non-convex sets of functions, such as in the case of sparse linear regression.

Assumption 3.4.2: The random vector v_j for $j = 0, ..., p$ is sub-Gaussian with parameter at most σ_{v_j}. The matrix $v \in \mathbb{R}^{n \times p}$ is sub-Gaussian with parameters $(\Sigma_v, \, \sigma_v^2)$ where the jth column of v is v_j and $\sigma_v := \max_{j=0,...,p} \sigma_{v_j}$.

Assumption 3.4.3: The random vector η is sub-Gaussian with parameter at most σ_η.

Remark. In the literature of nonparametric estimation, common measures of function complexities associated with sub-Gaussian variables can be controlled with *standard* maximal inequalities as in van der Vaart and Wellner (1996) and van de Geer (2000), etc. There are some special cases of Assumptions 3.4.2 are 3.4.3 where other concentration results (e.g., Maurey, 1991; Ledoux, 1996; Bobkov, 1999; Bobkov and Ledoux, 2000) may provide sharper constants in the tail probability when we relax the identicalness of $\{w_i, x_i\}$ in Assumption 3.2.1. These special cases include: v_j for $j = 0, ..., p$ and η are (i) sub-Gaussian with *strongly log-concave* distribution (defined below) for some $\gamma_{v_j} > 0$ and $\gamma_\eta > 0$, respectively; or, (ii) a bounded vector[2] such that for every $i = 1, ..., n$, v_{ij} and η are supported on the interval (a'_{v_j}, a''_{v_j}) with $B_{v_j} := a''_{v_j} - a'_{v_j}$, and on (a'_η, a''_η) with $B_\eta := a''_\eta - a'_\eta$; or, (iii) a mixture of (i) and (ii) in terms of its probability measure.

Definition 3.4.1 (Strongly log-concave distributions). A distribution $\mathbb{P}$ with density $\mathtt{p}$ (with respect to the Lebesgue measure) is a *strongly log-concave distribution* if the function $\log \mathtt{p}$ is strongly concave. Equivalently stated, the density can be written in the form $\mathtt{p}(x) = \exp(-\psi(x))$, where the function $\psi : \mathbb{R}^n \to \mathbb{R}$ is strongly convex, meaning that there is some $\gamma > 0$ such that

$$\lambda\psi(x) + (1-\lambda)\psi(y) - \psi(\lambda x + (1-\lambda)y) \geq \frac{\gamma}{2}\lambda(1-\lambda)\left|x - y\right|_2^2$$

for all $\lambda \in [0, 1]$, and $x, y \in \mathbb{R}^n$.

Remark. It is easy to verify that the distribution of a standard Gaussian vector in n dimensions is strongly log-concave with parameter $\gamma = 1$. More generally, any Gaussian distribution with covariance matrix $\Sigma \succ 0$ is strongly log-concave with parameter $\gamma = \lambda_{\min}(\Sigma^{-1})$. In addition, there are a variety of non-Gaussian distributions that are also strongly log-concave.

3.4.1 Properties of the non-pivotal Lasso estimator of the high-dimensional linear coefficients

General upper bounds and l_2−consistency

The following theorem (Theorem 3.4.1) provides a general upper bound on the error $\left|\hat{\beta}_{HSEL} - \beta^*\right|_2$ when the second-stage estimation concerns a program as in (3.9). This

[2] A random vector with bounded elements is sub-Gaussian.

result is an "oracle-inequality" type which does not assume the unknown function $m(\cdot)$ belongs to the function class over which the nonparametric estimator from (3.9) is defined. In such settings, the performance of the estimator involves both the estimation error and an approximation error, arising from the fact that $m_j \notin \mathcal{F}_j$.

To state Theorem 3.4.1, we need to introduce a set of definitions. First, we define a quantity that measures the complexity of the function class $\mathcal{F}_j$ (a notion often used in nonparametric literature; e.g., van der Vaart and Wellner, 1996; van de Geer, 2000; Barlett and Mendelson, 2002; Koltchinski, 2006; Wainwright, 2015, etc.). For any radius $r_j > 0$, define the conditional *local complexity*

$$\mathcal{G}_n(r_j;\, \mathcal{F}_j) := \mathbb{E}_{v_j}\left[\sup_{f\in\Omega(r_j;\,\mathcal{F}_j)}\left|\frac{1}{n}\sum_{i=1}^{n} v_{ij} f(w_i^T\theta^*)\right| |w_i^T\theta^*\right],$$

where variables $\{v_{ij}\}_{i=1}^n$ for $j = 0, ..., p$ are *i.i.d.* variates that satisfy Assumption 3.4.2, and

$$\Omega(r_j;\, \mathcal{F}_j) = \left\{f \,:\, f \in \bar{\mathcal{F}}_j \; |f_{\theta^*}|_n \le r_j\right\},$$

where $|f_{\theta^*}|_n := \sqrt{\frac{1}{n}\sum_{i=1}^n \left[f(w_i^T\theta^*)\right]^2}$. For any star-shaped shifted function class $\bar{\mathcal{F}}_j$, the function $t \mapsto \frac{\mathcal{G}_n(t;\,\mathcal{F}_j)}{t}$ is non-decreasing on the interval $(0, \infty)$. Second, let $T_j^* := \sup_{f\in\mathcal{F}_j^*}\frac{1}{n}\sum_{i=1}^n\left[f(w_i^T\hat{\theta}) - f(w_i^T\theta^*)\right]^2$, $T_j := \sup_{f\in\mathcal{F}_j}\frac{1}{n}\sum_{i=1}^n\left[f(w_i^T\hat{\theta}) - f(w_i^T\theta^*)\right]^2$, $T_j^{'} := T_j^* \vee T_j$, and

$$\begin{aligned}
\mathcal{T}_1 &= \max_{j\in\{0,\dots,p\}}\left(T_j^{'} \vee \sqrt{T_j^{'}}\right)\\
\mathcal{T}_2 &= \max_{j\in\{0,\dots,p\}} t_{nj}^2\\
\mathcal{T}_3 &= \max_{j\in\{0,\dots,p\}}\inf_{\tilde{m}_j\in\mathcal{F}_j}\Big(\frac{1}{n}\sum_{i=1}^n\left[\tilde{m}_j(w_i^T\hat{\theta}) - m_j(w_i^T\hat{\theta})\right]^2\\
&\quad +\sqrt{\frac{1}{n}\sum_{i=1}^n\left[\tilde{m}_j(w_i^T\theta^*) - m_j(w_i^T\theta^*)\right]^2}\Big)\\
\mathcal{T}_4 &= \sigma_v\sigma_\eta\sqrt{\frac{\log p}{n}}.
\end{aligned}$$

Third, recall in Section 3.2 the set we introduced:

$$\mathbb{C}(S;\, q_2,\, 3) := \left\{\Delta\in\mathbb{R}^p \,:\, |\Delta_{S^c}|_1 \le 3|\Delta_S|_1 + 4|\beta^*_{S^c}|_1\right\},$$

and the spherical set

$$\mathbb{S}_\delta := \left\{\Delta\in\mathbb{R}^p \,:\, |\Delta|_2 \ge \delta\right\},$$

and the intersection of these two sets $\mathbb{C}(S;\, q_2,\, 3) \cap \mathbb{S}_\delta$. When β^* is approximately sparse (namely, $q_2 \in (0,\, 1]$), we choose S in $\mathbb{C}(S;\, q_2,\, 3)$ to be the following thresholded subset

$$S_{\underline{\tau}} := \left\{ j \in \{1,\, 2,\, ...,\, p\} \,:\, \left|\beta_j^*\right| > \underline{\tau} \right\}$$

with the threshold parameter $\underline{\tau} = \frac{\lambda_{n,3}}{\kappa_L}$ (recall $\lambda_{n,3}$ is the third-stage regularization parameter whose choice is specified in the theorems and the parameter κ_L is defined in Assumption 3.2.3, Section 3.2). When β^* is exactly sparse (namely, $q_2 = 0$), we set $\delta = \underline{\tau} = 0$ and choose $S = J(\beta^*)$, which reduces the set $\mathbb{C}(S;\, q_2,\, 3) \cap \mathbb{S}_\delta$ to the following cone:

$$\mathbb{C}(J(\beta^*);\, 0,\, 3) := \left\{ \Delta \in \mathbb{R}^p \,:\, |\Delta_{J(\beta^*)^c}|_1 \leq 3|\Delta_{J(\beta^*)}|_1 \right\}.$$

Theorem 3.4.1: Let the *critical radius* $r_{nj} > 0$ be the smallest positive quantity satisfying the *critical inequality*

$$\mathcal{G}_n\left(r_{nj};\, \mathcal{F}_j\right) \leq \frac{r_{nj}^2}{\sigma_{v_j}}.$$

Suppose the second-stage estimator solves program (3.9) and Assumptions 3.2.1, 3.2.2, 3.4.1-3.4.3 hold. Additionally, let Assumption 3.2.3 hold over $\mathbb{C}(J(\beta^*);\, 0,\, 3)$ for the exact sparsity case ($q_2 = 0$ with $R_{q_2} = k_2$), and over $\mathbb{C}(S_{\underline{\tau}};\, q_2,\, 3) \cap \mathbb{S}_\delta$ where $\delta \asymp R_{q_2}^{\frac{1}{2}}\left(\lambda_{n,3}\right)^{1-\frac{q_2}{2}}$ and $\underline{\tau} = \frac{\lambda_{n,3}}{\kappa_L}$ for the approximate sparsity case ($q_2 \in (0,\, 1]$), respectively. For any $t_{nj} \geq r_{nj}$, if the third-stage regularization parameter $\lambda_{n,3}$ satisfies

$$\lambda_{n,3} \geq b(\sigma_v,\, \sigma_\eta)\left|\beta^*\right|_1\left(\mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_3\right) + \mathcal{T}_4 := \bar{M}, \tag{3.15}$$

where $b(\sigma_v,\, \sigma_\eta)$ is a known function that only depend on the parameters σ_v and σ_η (and independent of n, d, p, R_{q_2}), and the condition

$$R_{q_2}\underline{\tau}^{-q_2}\left(\frac{\log p}{n} + \mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_3\right) = O(\kappa_L), \tag{3.16}$$

holds, then,

$$|\hat{\beta}_{HSEL} - \beta^*|_2 \leq \frac{c'' R_{q_2}^{\frac{1}{2}}}{\kappa_L^{1-\frac{q_2}{2}}}\left[\bar{M} \vee \lambda_{n,3}\right]^{1-\frac{q_2}{2}} \tag{3.17}$$

with probability at least $1 - c_1 \exp\left(-c_2 \log p\right) - c_3 \sum_{j=0}^p \exp\left(-nC_j^* t_{nj}^2\right)$ for some C_j^* independent of n, d, p, R_{q_2}.

The following theorem (Theorem 3.4.2) provides a general upper bound on the error $\left|\hat{\beta}_{HSEL}-\beta^{*}\right|_{2}$ when the second-stage estimation concerns a regularized program as in (3.10). As in Theorem 3.4.1, this result is an "oracle-inequality" type which does not assume the unknown function $m(\cdot)$ belongs to the function class over which the nonparametric estimator from (3.10) is defined. For Theorem 3.4.2, let the *local complexity* measure $\mathcal{G}_n\left(r_j;\,\mathcal{F}_j\right)$ be defined over the set

$$\Omega(r_j;\,\mathcal{F}_j)=\left\{f\,:\,f\in\bar{\mathcal{F}}_j\;|f_{\theta^*}|_n\leq r_j,\;|f|_{\mathcal{F}_j}\leq 1\right\}$$

where $|f_{\theta^*}|_n:=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[f(w_i^T\theta^*)\right]^2}$ and $j=0,...,p$. Also define the following quantities:

$$\begin{aligned}
\mathcal{T}_1 &= \max_{j\in\{0,\ldots,p\}}\left(T_j^{'}\vee\sqrt{T_j^{'}}\right)\\
\mathcal{T}_2 &= \max_{j\in\{0,\ldots,p\}}\bar{R}_j^2t_{nj}^2\\
\mathcal{T}_3 &= \max_{j\in\{0,\ldots,p\}}\inf_{\tilde{m}_j\in\mathcal{F}_j,\,|\tilde{m}_j|_{\mathcal{F}_j}\leq\bar{R}_j}\Big(\frac{1}{n}\sum_{i=1}^{n}\left[\tilde{m}_j(w_i^T\hat{\theta})-m_j(w_i^T\hat{\theta})\right]^2\\
&\quad+\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\tilde{m}_j(w_i^T\theta^*)-m_j(w_i^T\theta^*)\right]^2}\Big)\\
\mathcal{T}_4 &= \sigma_v\sigma_\eta\sqrt{\frac{\log p}{n}},
\end{aligned}$$

where $T_j^{'}$ is defined prior to the presentation of Theorem 3.4.1.

Theorem 3.4.2: Let the *critical radius* $r_{nj}>0$ be the smallest positive quantity satisfying the *critical inequality*

$$\mathcal{G}_n\left(r_{nj};\,\mathcal{F}_j\right)\leq\frac{\bar{R}_jr_{nj}^2}{\sigma_{v_j}},$$

where $\bar{R}_j>0$ is a user-defined radius. Suppose the second-stage estimator solves the regularized program (3.10) and Assumptions 3.2.1, 3.2.2, 3.4.1-3.4.3 hold. Additionally, let Assumption 3.2.3 hold over the restricted set $\mathbb{C}(J(\beta^*);\,0,\,3)$ for the exact sparsity case ($q_2=0$ with $R_{q_2}=k_2$), and over $\mathbb{C}(S_{\underline{\tau}};\,q_2,\,3)\cap\mathbb{S}_\delta$ where $\delta\asymp R_{q_2}^{\frac{1}{2}}\left(\lambda_{n,3}\right)^{1-\frac{q_2}{2}}$ and $\underline{\tau}=\frac{\lambda_{n,3}}{\kappa_L}$ for the approximate sparsity case ($q_2\in(0,\,1]$), respectively. For any $t_{nj}\geq r_{nj}$, if the second-stage regularization parameter $\lambda_{nj,2}=2t_{nj}^2+\varsigma$ for any small positive constant $\varsigma > 0$ and the third-stage regularization parameter $\lambda_{n,3}$ satisfies (3.15), and condition (3.16) holds, then, the upper bound (3.17) holds (where the terms $\mathcal{T}_k$, $k = 1, ..., 4$ correspond to the ones defined for Theorem 3.4.2) with probability at least

$$1 - c_1 \exp\left(-c_2 \log p\right) - c_3 \sum_{j=0}^{p} \exp\left(-nC_j^* \bar{R}_j^2 t_{nj}^2\right)$$

for some C_j^* independent of n, d, p, R_{q_2}.

Comments:
(a) For the probability guarantees in Theorems 3.4.1 and 3.4.2, the constant

$$C_j^* = c\frac{\gamma_{v_j} \wedge (B_{v_j}^2 \vee B_\eta^2)^{-1}}{\sigma_{v_j}^2 \vee \sigma_\eta^2}$$

when v_j for $j = 0, ..., p$ and η are (i) sub-Gaussian with *strongly log-concave* distribution for some $\gamma_{v_j} > 0$ and $\gamma_\eta > 0$, respectively; or, (ii) a bounded vector such that for every $i = 1, ..., n$, v_{ij} and η are supported on the interval $(a_{v_j}^{'}, a_{v_j}^{''})$ with $B_{v_j} := a_{v_j}^{''} - a_{v_j}^{'}$, and on $(a_\eta^{'}, a_\eta^{''})$ with $B_\eta := a_\eta^{''} - a_\eta^{'}$; or, (iii) a mixture of (i) and (ii) in terms of its probability measure.

(b) Condition (3.16) in Theorems 3.4.1 and 3.4.2 ensures that with high probability, $\frac{\hat{v}^T \hat{v}}{n}$ satisfies the RE condition (3.4) over $\mathbb{C}(J(\beta^*); 0, 3)$ for the exact sparsity case ($q_2 = 0$ with $R_{q_2} = k_2$), and over $\mathbb{C}(S_{\underline{\tau}}; q_2, 3) \cap \mathbb{S}_\delta$ where $\delta \asymp R_{q_2}^{\frac{1}{2}} (\lambda_{n,3})^{1-\frac{q_2}{2}}$ and $\underline{\tau} = \frac{\lambda_{n,3}}{\kappa_L}$ for the approximate sparsity case ($q_2 \in (0, 1]$), respectively. An implication of this scaling condition is that it provides a finite-sample guarantee of the population identification condition (Assumption 3.2.3) subject to the underlying restricted sets. This result is formalized in the following corollary.

Corollary 3.4.3: Under the assumptions in Theorem 3.4.1 (respectively, the assumptions in Theorem 3.4.2), we have, with the same probability guarantees in Theorem 3.4.1 (respectively, in Theorem 3.4.2),

$$\frac{1}{n} \sum_{i=1}^{n} y_{1i} \left(x_i - \hat{\mathbb{E}}\left[x_i \,|\, w_i^T \hat{\theta},\, y_{1i} = 1\right]\right) \left(x_i - \hat{\mathbb{E}}\left[x_i \,|\, w_i^T \hat{\theta},\, y_{1i} = 1\right]\right)^T$$

is nonsingular on the restricted sets subject to those in Theorem 3.4.1 (respectively, Theorem 3.4.2).

Remarks on Theorems 3.4.1 and 3.4.2

The main proofs for Theorem 3.4.1, Theorem 3.4.2, and Corollary 3.4.3 are provided in Sections 3.8.1-3.8.4.

These theorems imply that if $\lambda_{n,3} \asymp \bar{M}$ and

$$\frac{c^{''}R_{q_2}^{\frac{1}{2}}}{\kappa_L^{1-\frac{q_2}{2}}}\left[b(\sigma_v,\,\sigma_\eta)\left|\beta^*\right|_1(\mathcal{T}_1+\mathcal{T}_2+\mathcal{T}_3)+\mathcal{T}_4\right]^{1-\frac{q_2}{2}}\rightarrow 0,$$

as $n\rightarrow\infty$, then the two-stage estimator $\hat{\beta}_{HSEL}$ is l_2-consistent for β^*. From Theorems 3.4.1 and 4.2, it can be seen that the general upper bounds on $|\hat{\beta}_{HSEL}-\beta^*|_2$ depend on four sources of errors, $\mathcal{T}_k$, $k=1,...,4$. The terms $\mathcal{T}_1$, $\mathcal{T}_2$, $\mathcal{T}_3$, and $\mathcal{T}_4$ are related to the statistical error of the first-stage estimation, the statistical error of the second-stage nonparametric regression, the approximation error arising from the fact that $m_j\notin\mathcal{F}_j$, and the statistical error of the third-stage estimation, respectively. Inspecting the error term $\mathcal{T}_1$ suggests that, given appropriate identification assumptions, the upper bounds on $|\hat{\beta}_{HSEL}-\beta^*|_2$ in Theorems 3.4.1 and 3.4.2 also hold for the more general structure where $\mathbb{E}\left(\epsilon_{2i}|w_i,\,y_{1i}=1\right)=g(h(w_i^T,\,\theta^*))$ and $h(w_i^T,\,\theta^*)$ is a scalar unobservable index.

The extra factor $\left|\beta^*\right|_1$ (in the case of exact sparsity,$\left|\beta^*\right|_1\asymp k_2$) in front of $\mathcal{T}_1$, $\mathcal{T}_2$, and $\mathcal{T}_3$ in the upper bounds on $|\hat{\beta}_{HSEL}-\beta^*|_2$ as well as in the choice of $\lambda_{n,3}$ is unimprovable and arising from the fact that the estimator is a sequential multi-stage procedure[3] based on plugging the first-stage estimator $\hat{\theta}$ in the place of θ^*. When $q_2=1$, the extra factor $\left|\beta^*\right|_1$ in front of $\mathcal{T}_1$, $\mathcal{T}_2$, and $\mathcal{T}_3$ in the upper bounds on $|\hat{\beta}_{HSEL}-\beta^*|_2$ as well as in the in the choice of $\lambda_{n,3}$ is crucial in order for the argument in our analysis to go through. To see this, suppose $\sqrt{\frac{\log p}{n}}$ is small relative to $\mathcal{T}_1$, $\mathcal{T}_2$, and $\mathcal{T}_3$, in which case, condition (3.16) can be reduced to

$$R_{q_2}\left(\left|\beta^*\right|_1\right)^{-q_2}\max_k\left\{\mathcal{T}_k^{1-q_2}\,:\,k=1,\,2,\,3\right\}=O\left(\kappa_L^{1-q_2}\right).$$

When $q_2=1$, if we set $R_{q_2}=\left|\beta^*\right|_1$, then $R_{q_2}\left(\left|\beta^*\right|_1\right)^{-q_2}=1$ so the above condition holds. On the other hand, when $q_2\in[0,\,1)$, condition (3.16) is easier to be satisfied.

When $\mathcal{T}_1$, $\mathcal{T}_2$, and $\mathcal{T}_3$ are small relative to $\mathcal{T}_4$ and β^* is exactly sparse with at most k_2 non-zero coefficients, if we set $\kappa_L=\lambda_{\min}(\Sigma_v)$, the upper bounds in Theorems 3.4.1 and 3.4.2 reduce to $|\hat{\beta}_{HSEL}-\beta^*|_2\precsim\frac{\sigma_v\sigma_\eta}{\lambda_{\min}(\Sigma_v)}\sqrt{\frac{k_2\log p}{n}}$. Note that the scaling $\sqrt{\frac{k_2\log p}{n}}$ is the

[3] Other plug-in type Lasso estimators for the exactly sparse case such as the ones in Rosenbaum and Tsybakov (2011) and the high-dimensional two-stage least-squares estimator in Zhu (2013), also involve the extra factor $\left|\beta^*\right|_1$.

optimal rate of the Lasso for the usual high-dimensional linear regression model with exact sparsity, and the factor $\frac{\sigma_v\sigma_\eta}{\lambda_{\min}(\Sigma_v)}$ has a natural interpretation of an inverse signal-to-noise ratio when v_i is a zero-mean Gaussian matrix with covariance $\Sigma_v = \sigma_v^2 I_{p\times p}$: one has $\lambda_{\min}(\Sigma_v) = \sigma_v^2$, so $\frac{\sigma_v\sigma_\eta}{\lambda_{\min}(\Sigma_v)} = \frac{\sigma_\eta}{\sigma_v}$, which measures the inverse signal-to-noise ratio of the regressors.

For the case of approximately sparse β^* with $q_1,\, q_2 \in (0,\, 1]$, the rate

$$\frac{c^{''} R_{q_2}^{\frac{1}{2}}}{\kappa_L^{1-\frac{q_2}{2}}} \left[b(\sigma_v,\, \sigma_\eta) \left|\beta^*\right|_1 (\mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_3) + \mathcal{T}_4 \right]^{1-\frac{q_2}{2}} \tag{3.18}$$

can be interpreted with the following heuristic. Suppose we choose to the top s_2 coefficients of β^* in absolute values to estimate, then the fast decay imposed by the $l_{q_2}-$balls condition on β^* would mean that the remaining $p - s_2$ coefficients would have relatively little impact. With this intuition, the rate for $q_2 > 0$ can be viewed as the rate that would be achieved by choosing

$$s_2 = \frac{c^{''} R_{q_2}}{\kappa_L^{-q_2}} \left[b(\sigma_v,\, \sigma_\eta) \left|\beta^*\right|_1 (\mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_3) + \mathcal{T}_4 \right]^{-q_2}$$

and then proceeding as if the problem were an instance of an exactly sparse problem $q_2 = 0$ with $k_2 = s_2$. For such a problem, we would expect to obtain the rate

$$\frac{c^{''} \sqrt{s_2}}{\kappa_L} \left[b(\sigma_v,\, \sigma_\eta) \left|\beta^*\right|_1 (\mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_3) + \mathcal{T}_4 \right],$$

which is exactly equal to (3.18).

Notice that the choice of t_{nj} incurs a trade-off between $\mathcal{T}_2$ and $O\left(\sum_{j=0}^p \exp\left(-nt_{nj}^2\right)\right)$ in the probability guarantees in Theorems 3.4.1 and 3.4.2. This is a general phenomenon for these tail bounds. For the problems considered in this chapter, t_{nj} may be chosen in the way that $\mathcal{T}_2$ is dominated by $\mathcal{T}_1$, $\mathcal{T}_3$, and $\mathcal{T}_4$ while the probability guarantees are maximized to allow for the least restrictive requirement on the sample size for l_2-consistency. Section 3.4.1.2 provides a specific example in terms of the choice of t_{nj}. When we set $t_{nj} = r_{nj}$, note that the statistical error related to the second-stage nonparametric regression, $\mathcal{T}_2$, is on the order of $O\left(\max_j r_{nj}^2\right)$ instead of the optimal rate $O\left(\max_j r_{nj}\right)$ that one would expect from a nonparametric regression as (3.9) or (3.10). As long as $\max_j r_{nj} < 1$, we have: $\max_j r_{nj}^2 < \max_j r_{nj}$, and provided $\max_j r_{nj}^2$ is small relative to $\mathcal{T}_1$, $\mathcal{T}_3$ and $\mathcal{T}_4$, the convergence rate of the estimator of the high-dimensional linear coefficients in the main equation behaves as if the unknown nonparametric selection bias were known.

This result establishes the non-asymptotic counterpart of the familiar asymptotic "oracle" type of results from previous literature. One of the drivers behind this oracle result lies on carefully controlling for the term $\left|\frac{1}{n}\sum_{i=1}^{n}\hat{v}_{ij}\left[\hat{m}_j(w_i^T\hat{\theta})-\tilde{m}_j(w_i^T\hat{\theta})\right]\right|$ utilizing the fact that $\hat{v}_{ij}$ estimates the true residual v_{ij} which is obtained by projecting x_{ij} or y_{2i} onto $w_i^T\theta^*$, namely, $v_{ij}=x_{ij}-\mathbb{E}\left(x_{ij}|w_i^T\theta^*\right)$ or $v_{i0}=y_{2i}-\mathbb{E}\left(y_{2i}|w_i^T\theta^*\right)$. When this projection procedure is applied to classical low-dimensional semilinear models with fixed p and d (in which case, there is no first-stage related error $\mathcal{T}_1$), our general upper bounds would imply that the nonparametric component needs to be estimated at a rate *no slower* than $O\left((\frac{1}{n})^{\frac{1}{4}}\right)$ in order for the estimator of the parametric component to achieve the rate of $O\left(\sqrt{\frac{1}{n}}\right)$. In contrast to the semilinear models, low-dimensional selection models require the rate of the nonparametric component to be at least $O\left((\frac{1}{n})^{\frac{1}{3}}\right)$ because the nonparametric component in the selection model involves an unknown single index that also needs to be estimated.

Upper bounds and l_2-consistency for a leading case example

An important consequence of Theorems 3.4.1 and 3.4.2 is when $m_j(\cdot)\in\mathcal{F}_j$ for every $j=0,...,p$ and $\mathcal{F}_j$ in

$$T_j^{'}:=\sup_{f\in\mathcal{F}_j}\frac{1}{n}\sum_{i=1}^{n}\left[f(w_i^T\hat{\theta})-f(w_i^T\theta^*)\right]^2$$

can be restricted to the class of Lipschitz functions, as a result, $\mathcal{T}_3=0$ and $T_j^{'}=\frac{1}{n}\sum_{i=1}^{n}L^2\left[w_i^T\hat{\theta}-w_i^T\theta^*\right]^2:=L^2B^{'}$. Results regarding this leading case are provided in the following corollaries (Corollaries 3.4.4 and 3.4.5). Before stating these results, a procedure based on Lipschitz regression for the second-stage estimation is presented and its theoretical guarantees are provided in Corollaries 3.4.4 and 3.4.5.

We say that a function $f:\mathbb{R}\to\mathbb{R}$ is *L-Lipschitz* if

$$\left|f(t)-f(t^{'})\right|\leq L\left|t-t^{'}\right| \tag{3.19}$$

for all t, $t^{'}\in\mathbb{R}$. When $\mathcal{F}_j$ satisfies the Lipschitz assumption, we restrict $\mathcal{F}_j$ in (3.9) to be the class of Lipschitz functions and consider $\tilde{m}_j$ in this class only, namely,

$$\hat{m}_j\in\arg\min_{\substack{\tilde{m}_j:\,\mathbb{R}\to\mathbb{R}\\ \tilde{m}_j\text{ is }L\text{-Lipschitz}}}\left\{\frac{1}{n}\sum_{i=1}^{n}\left(z_{ij}-\tilde{m}_j(w_i\hat{\theta})\right)^2\right\}\qquad\text{for } j=0,...,p.$$

It can be easily verified that $\bar{\mathcal{F}}_j$, the shifted class of Lipschitz functions is also Lipschitz and satisfies Assumption 3.4.1; i.e., it is star-shaped. By exploiting the structure of Lipschitz functions, the program above can be converted to an equivalent finite-dimensional problem by applying the constraint (3.19) to each of the sampled points $w_i\hat{\theta}$ so that there must exist a real-valued vector $(\tilde{z}_{1j}, ..., \tilde{z}_{ij}, ..., \tilde{z}_{nj})$ which satisfies the constraints in the following convex program

$$\begin{aligned} (\hat{z}_{1j}, ..., \hat{z}_{ij}, ..., \hat{z}_{nj}) \quad &\in \quad \arg \min_{(\tilde{z}_{1j}, ..., \tilde{z}_{ij}, ..., \tilde{z}_{nj})} \left\{ \frac{1}{n} \sum_{i=1}^{n} (z_{ij} - \tilde{z}_{ij})^2 \right\} \\ \text{s.t.} \quad \tilde{z}_{ij} - \tilde{z}_{i'j} \quad &\leq \quad L\left(w_i - w_{i'}\right)^T \hat{\theta} \text{ for all } i,\, i' = 1, ..., n. \end{aligned} \tag{3.20}$$

Given an optimal solution $(\hat{z}_{1j}, ..., \hat{z}_{ij}, ..., \hat{z}_{nj})$, a Lipschitz function $\hat{m}_j$ can be constructed by interpolating linearly between $\hat{z}_{ij}$s and the resulting function $\hat{m}_j$ is an estimate of m_j (namely, the second-stage estimator). Moreover, one can easily see that $\hat{m}_j(w_i^T\hat{\theta}) = \hat{z}_{ij}$. Note that the optimization problem above is a convex program with a quadratic cost function and a total of $\begin{pmatrix} n \\ 2 \end{pmatrix}$ linear constraints and n variables (n here denotes the sample size of the observations for the main equation). There are many computationally efficient algorithms for solving programs like this (e.g., the interior point method). When $m_j(\cdot)$ is a monotonic Lipschitz function, we can impose additional monotonicity constraints together with the Lipschitz constraints in the above convex program. Kakade, et. al (2011) provides an algorithm with provable guarantees for this type of minimization problems.

In the case where the Lipschitz constant L is unknown, cross-validation methods can be used to determine L. For example, we can first solve the optimization problem (3.20) on a subsample of observations by imposing an additional constraint $0 \leq L \leq L^{(0)}$ for a chosen constant $L^{(0)}$ and obtain $(\hat{z}_{1j}, ..., \hat{z}_{ij}, ..., \hat{z}_{nj},\, L) := \varpi^0$. We then test for the prediction quality of this optimal solution ϖ^0 by comparing its predicted values (from interpolating linearly between $\hat{z}_{ij}$s) for the remaining subsample with the actual observed values. If the optimal solution ϖ^0 returns $L \approx L^{(0)}$, we can iterate the process by imposing $0 \leq L \leq L^{(1)} = 2L^{(0)}$ in (3.20) and comparing the new optimal solution ϖ^1 with the previous one ϖ^0 and also testing for the prediction quality of ϖ^1.

Assumption 3.4.4: The matrix w consists of bounded elements[4].

The following proposition (Proposition 3.4.1) regarding the *critical radius* r_{nj} in Theorems 3.4.1 and 3.4.2 is based on results from van der Vaart and Wellner (1996), van de

[4]A random matrix with bounded elements is sub-Gaussian.

Geer (2000), and Wainwright (2015).

Proposition 3.4.1: Let Assumptions 3.2.1 and 3.4.4 hold and $m_j(\cdot) \in \mathcal{F}_j$ for $j = 0, ..., p$. Suppose $\mathcal{F}_j$ belongs to the class of $L-$Lipschitz functions and the Lipschitz regression procedure (3.20) is applied. Then, for every $j = 0, ..., p$, $\mathcal{T}_3 = 0$ and $T_j^{'} = \frac{1}{n}\sum_{i=1}^{n} L^2 \left[w_i^T\hat{\theta} - w_i^T\theta^*\right]^2 := L^2 B^{'}$, and the *critical radius* $r_{nj} = O\left((\frac{|\theta^*|_1}{n})^{\frac{1}{3}}\right)$, in Theorems 3.4.1 and 3.4.2.

The following corollaries (Corollaries 3.4.4 and 3.4.5) provide results regarding the leading case where for every $j = 0, ..., p$, $T_j^{'} = \frac{1}{n}\sum_{i=1}^{n} L^2 \left[w_i^T\hat{\theta} - w_i^T\theta^*\right]^2 := L^2 B^{'}$,$\mathcal{T}_3 = 0$, and the *critical radius* $r_{nj} = O\left((\frac{|\theta^*|_1}{n})^{\frac{1}{3}}\right)$, in Theorems 3.4.1 and 3.4.2. These conditions are ensured by Proposition 3.4.1. The two corollaries differ by the upper bounds on the quantity $B^{'}$. Justifications of these upper bounds on $B^{'}$ are given by Propositions 3.4.2 and 3.4.3. Let Υ_{w,θ^*} be a known function depending only on w and θ^*. The quantity Υ_{w,θ^*} changes according to the assumptions on w, which is to be made clear by Propositions 3.4.2 and 3.4.3. To facilitate the discussion and a later comparison with the minimax lower bounds in Section 3.4.4, the results in Corollaries 3.4.4 and 3.4.5 are presented for the case of exact sparsity on β^* and θ^* ($q_1 = q_2 = 0$). The case of general sparsity on θ^* and β^* (q_1, $q_2 \in [0, 1]$) is presented in Corollary 3.4.6 (which contains Corollary 3.4.4 as a special case).

Corollary 3.4.4 ($q_1 = q_2 = 0$): Suppose θ^* is exactly sparse with at most k_1 non-zero coefficients. Suppose for every $j = 0, ..., p$, $T_j^{'} = \frac{1}{n}\sum_{i=1}^{n} L^2 \left[w_i^T\hat{\theta} - w_i^T\theta^*\right]^2 := L^2 B^{'}$,$\mathcal{T}_3 = 0$, and the *critical radius* $r_{nj} = O\left((\frac{k_1}{n})^{\frac{1}{3}}\right)$, and

$$B^{'} = \frac{1}{n}\sum_{i=1}^{n} \left[w_i^T\hat{\theta} - w_i^T\theta^*\right]^2 \leq c\Upsilon_{w,\theta^*}\frac{k_1 \log d}{n}$$

with probability at least $1 - O\left(\frac{1}{d}\right)$. Assume t_{nj}^2 in $\mathcal{T}_2$ is chosen such that $|\beta^*|_1 \mathcal{T}_2$ is at most

$$O\left(\sqrt{\frac{\log p}{n}} \vee \left(|\beta^*|_1 \sqrt{\frac{k_1 \log d}{n}}\right)\right)$$

and $nt_{nj}^2 \gtrsim \log p$. Suppose Assumptions 3.2.1, 3.4.2-3.4.4 hold. Additionally, let β^* satisfy the exact sparsity in Assumption 3.2.2 ($q_2 = 0$ with $R_{q_2} = k_2$) and Assumption

3.2.3 hold over the restricted set $\mathbb{C}(J(\beta^*);\,0,\,3)$. Assume

$$\kappa_2 \frac{k_2 \log p}{n} + k_2 \sqrt{\frac{k_1 \log d}{n}} = O(\kappa_1),$$

for some strictly positive constants $(\kappa_1,\,\kappa_2)$ depending only on κ_L, σ_v, Υ_{w,θ^*}, and L. If the third-stage regularization parameter $\lambda_{n,3}$ satisfies

$$\lambda_{n,3} \geq c \left(\sigma_v \sigma_\eta \sqrt{\frac{\log p}{n}} \right) \vee \left(Lb\left(\sigma_v,\,\sigma_\eta\right) |\beta^*|_1 \sqrt{\Upsilon_{w,\theta^*}} \sqrt{\frac{k_1 \log d}{n}} \right) := \bar{M}$$

then, with probability at least $1 - O\left(\frac{1}{p \wedge d}\right)$, we have

$$|\hat{\beta}_{HSEL} - \beta^*|_2 \leq \frac{c_1 \sqrt{k_2}}{\kappa_L} \left[\bar{M} \vee \lambda_{n,3} \right]$$

where $b\left(\sigma_v,\,\sigma_\eta\right)$ is some known function depending only on σ_v and σ_η (and independent of n, d, p, k_1, and k_2).

The following assumptions and proposition provide an example in which the upper bound on $B^{'}$ in Corollary 3.4.4 is achieved. In particular, it requires the eigenvalues of Σ_w to be well-behaved over some restricted set.

Assumption 3.4.5: In program (3.11), we have: either (a) $\phi_1(w_i^T\theta) = w_i^T\theta$ and $\phi_2(w_i^T\theta) = \log(1 + \exp(w_i^T\theta))$; namely, the loss function corresponds to a binary logit model. Or, (b) $\phi_1(w_i^T\theta) = \log \frac{\Phi(w_i^T\theta)}{1-\Phi(w_i^T\theta)}$ and $\phi_2(w_i^T\theta) = -\log\left[1 - \Phi(w_i^T\theta)\right]$; namely, the loss function corresponds to a binary probit model.

Assumption 3.4.6: The random matrix w is sub-Gaussian with parameters $(\Sigma_w,\,\sigma_w^2)$. For all $\Delta \in \mathbb{C}(J(\theta^*);\,0,\,3)\backslash\{\mathbf{0}\}$, the matrix Σ_w satisfies

$$0 < \kappa_L^w \leq \frac{\Delta^T \Sigma_w \Delta}{|\Delta|_2^2} \leq \kappa_U^w < \infty$$

Proposition 3.4.2: Suppose the number of regressors $d(= d_n)$ can grow with and exceed the sample size n and the number of non-zero components in θ^* is at most $k_1(= k_{1n})$ and k_1 can increase to infinity with n but slowly compared to n. Let Assumptions

3.2.1, 3.4.5-3.4.6 hold. If $\hat{\theta}$ solves program (3.11) with $\lambda_{n,1} \geq c\sigma_w\sqrt{\alpha_u}\sqrt{\frac{\log d}{n}}$ and $n \gtrsim k_1 \log d$, then, with probability at least $1 - O\left(\frac{1}{d}\right)$,

$$\frac{1}{n}\sum_{i=1}^{n}\left[w_i^T(\hat{\theta}-\theta^*)\right]^2 \leq c^{'}\frac{\kappa_U^w}{\left(\kappa_L^w\right)^2}k_1\left((\lambda_{n,1})^2 \vee \left(\sigma_w^2\alpha_u\frac{\log d}{n}\right)\right),$$

where $\alpha_u > 0$ is a scalar such that $\phi_2^{''}(u) \leq \alpha_u$ for all $u \in \mathbb{R}$.

Remark. From Proposition 3.4.2, we can set $\Upsilon_{w,\theta^*} := \frac{\kappa_U^w\sigma_w^2\alpha_u}{\left(\kappa_L^w\right)^2}$ in Corollary 3.4.4. The boundedness on $\phi_2^{''}(u)$ holds automatically for the binary logit model and binary probit model. For the logit model, we have $\phi_2^{''}(u_i) = \frac{\exp(u_i)}{1+\exp(u_i)}\left(1 - \frac{\exp(u_i)}{1+\exp(u_i)}\right)$. For the probit model, note that $\phi_2^{''}(u_i)$ is $1-Var\left(\epsilon_{1i} \,|\, \epsilon_{1i} \leq u_i\right)$ when $y_{1i} = 1$ and $1-Var\left(\epsilon_{1i} \,|\, \epsilon_{1i} \geq -u_i\right)$ when $y_{1i} = 0$ and the unconditional variance is normalized to 1. Since truncation always reduces variances (Greene, 2003), $\phi_2^{''}(u)$ is bounded from above. If $\lambda_{n,1} \asymp \sigma_w\sqrt{\alpha_u}\sqrt{\frac{\log d}{n}}$, then

$$\frac{1}{n}\sum_{i=1}^{n}\left[w_i^T(\hat{\theta}-\theta^*)\right]^2 \leq c^{'}\frac{\kappa_U^w}{\left(\kappa_L^w\right)^2}\sigma_w^2\alpha_u\frac{k_1\log d}{n}.$$

Corollary 3.4.5: Suppose θ^* is exactly sparse with at most k_1 non-zero coefficients. Suppose for every $j = 0, ..., p$, $T_j^{'} = \frac{1}{n}\sum_{i=1}^{n} L^2\left[w_i^T\hat{\theta} - w_i^T\theta^*\right]^2 := L^2B^{'}$,$\mathcal{T}_3 = 0$, the *critical radius* $r_{nj} = O\left((\frac{k_1}{n})^{\frac{1}{3}}\right)$, and

$$B^{'} = \frac{1}{n}\sum_{i=1}^{n}\left[w_i^T\hat{\theta} - w_i^T\theta^*\right]^2 \leq c\Upsilon_{w,\theta^*}\left|\theta^*\right|_1\sqrt{\frac{\log d}{n}}$$

with probability at least $1 - O\left(\frac{1}{d}\right)$. Assume t_{nj}^2 in $\mathcal{T}_2$ is chosen such that $\left|\beta^*\right|_1\mathcal{T}_2$ is at most

$$O\left(\sqrt{\frac{\log p}{n}} \vee \left(\left|\beta^*\right|_1\left(\frac{k_1^2\log d}{n}\right)^{\frac{1}{4}}\right)\right)$$

and $nt_{nj}^2 \gtrsim \log p$. Suppose Assumptions 3.2.1, 3.4.2-3.4.4 hold. Additionally, let β^* satisfy the exact sparsity in Assumption 3.2.2 ($q_2 = 0$ with $R_{q_2} = k_2$) and Assumption 3.2.3 hold over the restricted set $\mathbb{C}(J(\beta^*);\, 0,\, 3)$. Assume

$$\kappa_2\frac{k_2\log p}{n} + k_2\left(\frac{k_1^2\log d}{n}\right)^{\frac{1}{4}} = O(\kappa_1),$$

for some strictly positive constants $(\kappa_1,\, \kappa_2)$ depending only on κ_L, σ_v, Υ_{w,θ^*}, and L, if the third-stage regularization parameter $\lambda_{n,3}$ satisfies

$$\lambda_{n,3} \geq c^{'} \left(\sigma_v \sigma_\eta \sqrt{\frac{\log p}{n}} \right) \vee \left(Lb\left(\sigma_v,\, \sigma_\eta\right) |\beta^*|_1 \sqrt{\Upsilon_{w,\theta^*}} \left(\frac{|\theta^*|_1^2 \log d}{n} \right)^{\frac{1}{4}} \right) := \bar{M}$$

then, with probability at least $1 - O\left(\frac{1}{p \wedge d}\right)$, we have

$$|\hat{\beta}_{HSEL} - \beta^*|_2 \leq \frac{c_2 \sqrt{k_2}}{\kappa_L} \left[\bar{M} \vee \lambda_{n,3} \right]$$

where $b\left(\sigma_v,\, \sigma_\eta\right)$ is some known function depending only on σ_v and σ_η (and independent of n, d, p, k_1, and k_2).

The following proposition provides an example in which the upper bound on $B^{'}$ in Corollary 3.4.5 is achieved. Let $\rho_{i,\theta} := -y_{1i}\phi_1(w_i^T\theta) + \phi_2(w_i^T\theta)$ and $\rho_{i,\bar{\theta}}^{''}$ be the second derivative of $\rho_{i,\theta}$, evaluated at $\theta = \bar{\theta}$, where $\bar{\theta}$ is some intermediate value between θ^* and $\hat{\theta}$, the solution to program (3.11). Assumption 3.4.4 implies that there is some $\alpha_l > 0$ such that $\rho_{i,\bar{\theta}}^{''} \geq \alpha_l$ for all $i = 1, ..., n$.

Proposition 3.4.3: Let Assumptions 3.2.1, 3.4.4 and 3.4.5 hold. Suppose the number of regressors $d(= d_n)$ can grow with and exceed the sample size n and the number of non-zero components in θ^* is at most $k_1(= k_{1n})$ and k_1 can increase to infinity with n but slowly compared to n. If $\hat{\theta}$ solves program (3.11) with the regularization parameter[5] $\lambda_{n,1} \geq c\sqrt{\frac{\log d}{n}}$, then,

$$\frac{1}{n}\sum_{i=1}^{n} \left[w_i^T(\hat{\theta} - \theta^*) \right]^2 \leq c_1 \Upsilon_{w,\theta^*} |\theta^*|_1 \left(\sqrt{\frac{\log d}{n}} \vee \lambda_{n,1} \right)$$

with probability at least $1 - O\left(\frac{1}{d}\right)$, where $\Upsilon_{w,\theta^*} := \alpha_l^{-1}$.

Remarks on Corollary 3.4.4-3.4.5

The proofs for Corollaries 3.4.4-3.4.5 and Propositions 3.4.1-3.4.3 are provided in Sections 3.8.5-3.8.7.

Corollaries 3.4.4 and 3.4.5 imply that if $\lambda_{n,3} \asymp \bar{M}$ and the upper bounds on $|\hat{\beta}_{HSEL} - \beta^*|_2$ tend to 0 as $n \to \infty$, then the two-stage estimator $\hat{\beta}_{HSEL}$ is l_2-consistent

[5]The choice of $\lambda_{n,1}$ is detailed in Theorems 2.1 or 2.2 in Van de Geer (2008).

for β^*. The difference between Corollary 3.4.4 and Corollary 3.4.5 lies in that the statistical error from the first-stage estimation is smaller in Corollary 3.4.4 relative to Corollary 3.4.5 and as a result, the estimator $\hat{\beta}_{HSEL}$ has a faster rate of convergence in Corollary 3.4.4. The smaller first-stage statistical error in Corollary 3.4.4 is at the expense of imposing conditions on the eigenvalues of Σ_w, as shown in Proposition 3.4.2. Consistency of $\hat{\beta}_{HSEL}$ *per se* does not require restrictions on the eigenvalues of Σ_w, which could be useful in certain applications. Proposition 3.4.3 provides an example where a slower rate of convergence is obtained by the first-stage estimator upon relaxing the assumptions on the eigenvalues of Σ_w.

By Proposition 3.4.1, $\max_j r_{nj}^2 = O\left(\left(\frac{k_1}{n}\right)^{\frac{2}{3}}\right)$. Let us examine various choices of $t_{nj}^2 \geq r_{nj}^2$ in Corollary 3.4.4 (the analysis for Corollary 3.4.5 is similar). Setting $t_{nj}^2 \asymp \sqrt{\frac{\log p}{n|\beta^*|_1^2}} \vee \sqrt{\frac{k_1 \log d}{n}} \gtrsim r_{nj}^2$ makes the second-stage error $|\beta^*|_1 \mathcal{T}_2$ on the same order of $\sqrt{\frac{\log p}{n}} \vee \left(|\beta^*|_1 \sqrt{\frac{k_1 \log d}{n}}\right)$. Under this choice of t_{nj}^2, we require $\sqrt{\frac{n \log p}{|\beta^*|_1^2}} \vee \sqrt{nk_1 \log d} \gtrsim \log p$ in order for the upper bound on $|\hat{\beta}_{HSEL} - \beta^*|_2$ to hold with probability at least $1 - O\left(\frac{1}{p \wedge d}\right)$. Setting $t_{nj}^2 \asymp \left(\frac{\log p \vee (k_1 \log d)}{n}\right)^{\frac{2}{3}} \gtrsim r_{nj}^2$ requires $n^{\frac{1}{3}} \left(\log p \vee (k_1 \log d)\right)^{\frac{2}{3}} \gtrsim \log p$ for the upper bound on $|\hat{\beta}_{HSEL} - \beta^*|_2$ to hold with probability at least $1 - O\left(\frac{1}{p \wedge d}\right)$. If instead, we set $t_{nj}^2 = r_{nj}^2$, then the probability guarantee of $1 - O\left(\frac{1}{p \wedge d}\right)$ would require $k_1^{\frac{2}{3}} n^{\frac{1}{3}} \gtrsim \log p$. Given the exact sparsity of β^* (so$|\beta^*|_1^2 \asymp k_2^2$), if k_2^2 is sufficiently small relative to $n \log p$, the first choice of t_{nj}^2 would provide the least restrictive requirement on the sample size. A later result that concerns with the selection consistency of $\hat{\beta}_{HSEL}$ assumes this choice for t_{nj}^2 and the scaling condition $\sqrt{\frac{n \log p}{|\beta^*|_1^2}} \vee \sqrt{nk_1 \log d} \gtrsim \log p$ on the sample size. When p and d are fixed and small relative to n, the analysis above generalizes existing asymptotic "oracle" results in semiparametric estimation of low-dimensional selection models from specific estimators (such as a series estimator) to a unified framework of nonparametric least squares estimators and regularized nonparametric least squares estimators.

More generally, when $\mathcal{F}_j$ belongs to a Hölder class of order $\nu > 0$, we have $\max_j r_{nj}^2 = O\left(\left(\frac{k_1}{n}\right)^{\frac{2\nu}{2\nu+1}}\right)$. When $v \geq 1$, $\mathcal{T}_1 \asymp \sqrt{\frac{k_1 \log d}{n}}$ and as long as we choose $t_{nj}^2 \asymp \sqrt{\frac{\log p}{n|\beta^*|_1^2}} \vee \sqrt{\frac{k_1 \log d}{n}}$, the second-stage error $|\beta^*|_1 \mathcal{T}_2$ and consequently the upper bound on $|\hat{\beta}_{HSEL} - \beta^*|_2$ would be on the same order of $\sqrt{\frac{\log p}{n}} \vee \left(|\beta^*|_1 \sqrt{\frac{k_1 \log d}{n}}\right)$. On the other hand, when $v \in (0,\, 1)$, we have $\frac{2\nu}{2\nu+1} > \frac{\nu}{2}$ and $\mathcal{T}_1 \asymp \left(\sqrt{B'}\right)^{\nu}$. Provided $B' \geq O(\frac{1}{n})$ (which is indeed the case for Corollaries 3.4.4 and 3.4.5) and the choice of $t_{nj}^2 \asymp \frac{\mathcal{T}_4}{|\beta^*|_1} \vee \left(\sqrt{B'}\right)^{\nu}$, then $|\hat{\beta}_{HSEL} - \beta^*|_2$ is bounded above by $\mathcal{T}_4 \vee \left(|\beta^*|_1 \left(\sqrt{B'}\right)^{\nu}\right)$. However, note in the simple example where $B' = \frac{1}{n}$ and $\mathcal{T}_4 = \sqrt{\frac{1}{n}}$, we have $\mathcal{T}_1 = \left(\frac{1}{n}\right)^{\frac{\nu}{2}} > \sqrt{\frac{1}{n}}$ for any $\nu \in (0,\, 1)$ and therefore $|\hat{\beta}_{HSEL} - \beta^*|_2$ is bounded above by $\left(\frac{1}{n}\right)^{\frac{\nu}{2}}$. Consequently, the minimum requirement for the "oracle" result to hold in the low-dimensional semiparametric selection models with fixed p and d is to have $v = 1$. For the high-dimensional selection models considered in Corollary 3.4.4, the minimum requirement is to have $O\left(\frac{k_1 \log d}{n}\right)^{\frac{\nu}{2}} = O\left(\sqrt{\frac{\log p}{n}}\right)$. In sharp contrast to the low-dimensional semilinear model, the fact that the nonparametric component in the selection model involves an unknown single index that also needed to be estimated increases the requirement on the rate of nonparametric estimation *per se.*

Note that the regularization parameter $\lambda_{n,3}$ and the upper bounds on $|\hat{\beta}_{HSEL} - \beta^*|_2$ depend on σ_v and σ_η, which is intuitive. It is possible to "remove" the dependence on σ_v from the choice of $\lambda_{n,3}$ by imposing weights $\hat{\sigma}_{v_j} := \sqrt{\frac{1}{n}\sum_{i=1}^{n} \hat{v}_{ij}^2}$, $j = 1, ..., p$ on the penalty term as in (3.7). An application of Lemmas A.11 and A.12 yields $\max_{j=1,...,p} \hat{\sigma}_{v_j} \leq 2\sigma_v$ with probability at least $1 - O\left(\frac{1}{p \wedge d}\right)$. The first-stage estimator $\hat{\theta}$ in Corollaries 3.4.4 and 3.4.5 may be replaced with a post-Lasso estimator where a usual low-dimensional estimation procedure is performed on the regressors selected by $\hat{\theta}$ (in a spirit similar to Belloni and Chernozhukov, 2011b, for example); and upon perfect selection or near-perfect selection[6] of $\hat{\theta}$, the term $\sqrt{\frac{k_1 \log d}{n}}$ from Corollary 3.4.4 and the term $\left(\frac{k_1^2 \log d}{n}\right)^{\frac{1}{4}}$ from Corollary 3.4.5 in the upper bounds on $|\hat{\beta}_{HSEL} - \beta^*|_2$ can be reduced to $\sqrt{\frac{k_1}{n}}$ and $\left(\frac{k_1^2}{n}\right)^{\frac{1}{4}}$, respectively.

We now present a result for the general sparsity case where θ^* and β^* belong to the general $l_{q_1}-$ and $l_{q_2}-$ "balls" with $q_1,\, q_2 \in [0,\, 1]$.

Corollary 3.4.6 ($q_1,\, q_2 \in [0,\, 1]$): Suppose for every $j = 0, ..., p$,

$$T_j^{'} = \frac{1}{n}\sum_{i=1}^{n} L^2 \left[w_i^T \hat{\theta} - w_i^T \theta^*\right]^2 := L^2 B^{'},$$

$\mathcal{T}_3 = 0$, and the *critical radius* $r_{nj} = O\left((\frac{|\theta^*|_1}{n})^{\frac{1}{3}}\right)$. Also, assume $\theta^* \in \mathcal{B}_{q_1}^d(R_{q_1})$ for

[6] Ravikumar, Wainwright, and Lafferty (2010) studies selection of a l_1-regularized logistic regression in the high-dimensional setting.

$q_1 \in [0,\, 1]$ with "radius" R_{q_1}, and

$$B^{'} = \frac{1}{n}\sum_{i=1}^{n}\left[w_i^T\hat{\theta} - w_i^T\theta^*\right]^2 \leq c\Upsilon_{w,\theta^*}R_{q_1}\left(\frac{\log d}{n}\right)^{1-\frac{q_1}{2}}$$

with probability at least $1 - O\left(\frac{1}{d}\right)$. Assume t_{nj}^2 in $\mathcal{T}_2$ is chosen such that $|\beta^*|_1 \mathcal{T}_2$ is at most $O(\bar{M})$, where

$$\bar{M} := \max\left\{\sqrt{\frac{\log p}{n}},\, |\beta^*|_1\left(\frac{|\theta^*|_1}{n}\right)^{\frac{2}{3}},\, |\beta^*|_1 R_{q_1}^{\frac{1}{2}}\left(\sqrt{\frac{\log d}{n}}\right)^{1-\frac{q_1}{2}}\right\}$$

and $nt_{nj}^2 \gtrsim \log p$. Moreover, condition (3.16) and Assumptions 3.2.1, 3.2.2, 3.4.2-3.4.4 hold. Additionally, let Assumption 3.2.3 hold over the restricted set $\mathbb{C}(J(\beta^*);\, 0,\, 3)$ for the exact sparsity case ($q_2 = 0$ with $R_{q_2} \asymp k_2$), and over $\mathbb{C}(S_{\underline{\tau}};\, q_2,\, 3) \cap \mathbb{S}_\delta$ where $\delta \asymp R_{q_2}^{\frac{1}{2}}(\lambda_{n,3})^{1-\frac{q_2}{2}}$ and $\underline{\tau} = \frac{\lambda_{n,3}}{\kappa_L}$ for the approximate sparsity case ($q_2 \in (0,\, 1]$), respectively. If the third-stage regularization parameter $\lambda_{n,3} \geq b_0\bar{M}$, then, with probability at least $1 - O\left(\frac{1}{p \wedge d}\right)$, we have

$$|\hat{\beta}_{HSEL} - \beta^*|_2 \leq \frac{b_1\sqrt{R_{q_2}}}{\kappa_L^{1-\frac{q_2}{2}}}\left(\bar{M} \vee \lambda_{n,3}\right)^{1-\frac{q_2}{2}}$$

where b_0 and b_1 are some known functions depending only on σ_v, σ_η, Υ_{w,θ^*}, and L (and independent of n, d, p, R_{q_1}, and R_{q_2}).

Comment on Corollary 3.4.6. Corollary 3.4.6 contains Corollary 3.4.4 as a special case with $q_2 = 0$, $R_{q_2} = k_2$ and $q_1 = 0$, $R_{q_1} = k_1$. When $q_1 = 0$ so that $R_{q_1} = k_1$ and $|\theta^*|_1 \asymp k_1$, the second term in $\bar{M}$ is on the order of $O\left(|\beta^*|_1\left(\frac{k_1}{n}\right)^{\frac{2}{3}}\right)$ and therefore dominated by $\sqrt{\frac{\log p}{n}} \vee |\beta^*|_1\sqrt{\frac{k_1 \log d}{n}}$, as we have seen previously. For more general sparsity of θ^* ($q_1 \in (0,\, 1]$), the second term in $\bar{M}$ may still be small relative to the first and third terms and therefore the "oracle" result is likely to hold for a range of scaling conditions on n, p, d, R_{q_1}, and $|\theta^*|_1$.

Variable-selection consistency of a leading case example with exact sparsity

The following theorem (Theorem 3.4.7) addresses the question: given $\hat{\beta}_{HSEL}$, when does $\hat{\beta}_{HSEL}$ correctly select the non-zero coefficients in the main equation with high probability? This property is referred to as **variable-selection consistency**, which is relevant to case of exactly sparse β^* (and therefore this section assumes β^* is exactly sparse with at most k_2 non-zero coefficients). We say $\hat{\beta}_{HSEL}$ achieves *perfect selection* if $\mathbb{P}[J(\hat{\beta}_{HSEL}) = J(\beta^*)] \to 1$ and *near-perfect selection* if $\mathbb{P}[J(\hat{\beta}_{HSEL}) \supseteq J(\beta^*)] \to 1$ and the number of wrong components selected is on the order of $o_p(k_2)$. Upon perfect selection or near-perfect selection of the regressors, we can then apply low-dimensional techniques to estimate and conduct inference on the important coefficients.

In order for the number of wrong components selected by the standard Lasso to be on the order of $o_p(k_2)$ in the context of standard linear regression models, it is known that the so-called "neighborhood stability condition" (Meinshausen and Bühlmann, 2006) on the design matrix, re-formulated in a nicer form as the "irrepresentable condition" by Zhao and Yu, 2006, or the "mutual incoherence condition" by Wainwright (2009), is sufficient and necessary. Furthermore, it can be shown that the "irrepresentable condition" implies the RE condition (see, e.g., Bühlmann and van de Geer, 2011).

Assumption 3.4.7: $\left\| \mathbb{E}\left[v^T_{1,J(\beta^*)^c} v_{1,J(\beta^*)}\right] \left[\mathbb{E}(v^T_{1,J(\beta^*)} v_{1,J(\beta^*)})\right]^{-1} \right\|_\infty \leq 1 - \phi$ for some constant $\phi \in (0,\, 1]$.

Assumption 3.4.7, the so-called "mutual incoherence condition" originally formalized by Wainwright (2009), captures the intuition that the large number of irrelevant covariates cannot exert an overly strong effect on the subset of relevant covariates. In the most desirable case, the columns indexed by $j \in J(\beta^*)^c$ would all be orthogonal to the columns indexed by $j \in J(\beta^*)$ and then we would have $\phi = 1$. In the high-dimensional setting, this perfect orthogonality is hard to achieve, but one can still hope for a type of "near orthogonality" to hold.

Assumptions 3.2.1 and 3.4.2 ensure that the left-hand-side of the inequality in Assumption 3.4.7 always falls in $[0,\, 1)$. To see this, note that under Assumptions 3.2.1 and 3.4.2, each column v_j, $j = 1, ..., p$ is consisted of *i.i.d.* sub-Gaussian variables. Without loss of generality, we can assume $\mathbb{E}(v_{1j}) = 0$ for all $j = 1, ..., p$. Consequently, the normalization $\max_{j=1,...,p} \frac{|v_j|_2}{\sqrt{n}} \leq \kappa_c$ where $0 < \kappa_c < \infty$ follows from a standard bound for the norms of zero-mean sub-Gaussian vectors and a union bound

$$\mathbb{P}\left[\max_{j=1,...,p} \frac{|v_j|_2}{\sqrt{n}} \leq \kappa_c\right] \geq 1 - 2\exp(-cn + \log p) \geq 1 - 2\exp(-c^{'}n),$$

where the last inequality follows from $n > \log p$. For example, if v_j has a Gaussian design, then we have

$$\max_{j=1,\ldots,p}\frac{|v_j|_2}{\sqrt{n}}\leq\max_{j=1,\ldots,p}\Sigma_{jj}\left(1+\sqrt{\frac{32\log p}{n}}\right),$$

where $\max_{j=1,..,p}\Sigma_{jj}$ corresponds to the maximal variance of any element of v (see Raskutti, et. al, 2011).

Theorem 3.4.7: Under the assumptions in Corollary 3.4.4 and Assumption 3.4.7, if $n\gtrsim(k_2^3\log p)\vee(k_2^2k_1\log d)$, $\sqrt{\frac{n\log p}{|\beta^*|_1^2}}\vee\sqrt{nk_1\log d}\gtrsim\log p$, $\sqrt{\frac{k_1\log d}{n}}=o(1)$, and $\lambda_{n,3}$ satisfies

$$\lambda_{n,3}\geq c\frac{8(2-\frac{\phi}{4})}{\phi}\left[\left(\sigma_v\sigma_\eta\sqrt{\frac{\log p}{n}}\right)\vee\left(L\left|\beta^*\right|_1\sqrt{\Upsilon_{w,\theta^*}}b\left(\sigma_v,\,\sigma_\eta\right)\sqrt{\frac{k_1\log d}{n}}\right)\right],$$

then, we have: (a) the support $J(\hat{\beta}_{HSEL})\subseteq J(\beta^*)$; (b) if $\min_{j\in J(\beta^*)}|\beta_j^*|>\bar{B}$, where

$$\bar{B}:=\frac{c\sqrt{k_2}}{\lambda_{\min}\left(\mathbb{E}\left[v_{1,J(\beta^*)}^Tv_{1,J(\beta^*)}\right]\right)}\left[\left(\sigma_v\sigma_\eta\sqrt{\frac{\log p}{n}}\right)\vee\left(L\left|\beta^*\right|_1\sqrt{\Upsilon_{w,\theta^*}}b\left(\sigma_v,\,\sigma_\eta\right)\sqrt{\frac{k_1\log d}{n}}\right)\right]$$

then $J(\hat{\beta}_{HSEL})\supseteq J(\beta^*)$ and $\hat{\beta}_{HSEL}$ is variable-selection consistent, i.e., $J(\hat{\beta}_{HSEL})=J(\beta^*)$, with probability at least $1-O\left(\frac{1}{p\wedge d}\right)$.

Remark. The proof for Theorem 3.4.7 is provided in Section 3.8.8. Part (a) of Theorem 3.4.7 guarantees that the Lasso does not falsely include elements that are not in the support of β^*. This result hinges on Assumption 3.4.7, namely, the mutual incoherence condition. Part (b) implies that as long as the minimum value of $|\beta_j^*|$ over $j\in J(\beta^*)$ is not too small, then the two-stage Lasso does not falsely exclude elements that are in the support of β^* with high probability. Combining the claims from (a) and (b), the multi-stage estimator is variable-selection consistent with high probability.

Inference with perfect or near perfect selection

When the mutual incoherence condition and the assumption that the true parameters β_j^* over $j\in J(\beta^*)$ is well separated from 0 are plausible for the empirical problem of interest, conditioning on the perfect selection or near-perfect selection result from Theorem 3.4.7, we can then apply low-dimensional techniques to conduct inference on the important coefficients. In the following discussion, we consider the simple case where k_1 and k_2 are fixed. Then, for example, one can apply the estimator

$$\tilde{\beta} := \left(\hat{v}_{\hat{J}}^T \hat{v}_{\hat{J}}\right)^{-1} \left(\hat{v}_{\hat{J}}^T \hat{v}_0\right)$$

where $\hat{J} := J(\hat{\beta}_{HSEL})$. In the multi-stage procedure proposed by this chapter, if the second-stage nonparametric estimation uses the series estimator from Newey (1991), then the post-selection estimator $\tilde{\beta}$ can be shown to be algebraically equivalent to the two-stage estimator of Newey (1991) for the semiparametric selection models when the linear coefficients in the main equation is low-dimensional. In deriving the $\sqrt{n}$-consistency and the asymptotic normality of the two-stage estimator, Newey requires $\sqrt{n}$-consistency on the first-stage estimator of the coefficients in the selection equation. This suggests that in order for the results from Newey (1991) to be applied on the estimator $\tilde{\beta}$, perfect selection or near-perfect selection of $\hat{\theta}$ defined in (3.11) may be required. We may consider a variant of $\hat{\beta}_{HSEL}$. This variant differs from $\hat{\beta}_{HSEL}$ in that, before the second-stage estimation, a post-Lasso procedure is performed on the regressors selected by the first-stage estimator $\hat{\theta}$ to obtain $\tilde{\theta}$, which is then used to form the single index. Rather than imposing perfect selection or near-perfect selection of $\hat{\theta}$, another option may be to use the procedure proposed by Ahn and Powell (1993), which does not require $\sqrt{n}-$consistency on the first-stage estimator and may allow imperfect selection of $\hat{\theta}$. For all these post-selection estimators discussed here, the asymptotic covariance matrix is rather complicated as it involves the derivative of the unobservable selection function. Ahn and Powell (1993) proposes a plug-in estimator for the asymptotic covariance matrix. Alternatively, bootstrap variance estimation can be used to obtain the standard errors of these post-selection estimators.

It is worth noting that while selection-consistency is a desirable property of the Lasso that allows us to conduct post-selection inference, it requires assumptions such as the mutual incoherence condition or the irrepresentable condition which might not hold in economic problems where the design matrices exhibit strong (empirical) correlations. When selection consistency is not achieved by the Lasso procedure, other inference procedures may be useful. While it is possible to construct confidence intervals for individual coefficients (e.g., Belloni, Chernozhukov, and Wei, 2013; Belloni, Chernozhukov, and Hansen, 2014) and linear combinations of several of them in certain high-dimensional models using a low dimensional projection approach (e.g., Zhang and Zhang, 2013; Javanmard and Montanari, 2014), general inference theory with high-dimensional data is still underexplored owing to the complexity of the sampling distributions of existing estimators (see e.g., Efron, 2010). Rather than relying on distributional theory to conduct inference, The following section (Section 3.4.2) provides an alternative way of constructing confidence sets based on the pivotal Dantzig selector (3.8) from Section 3.3. Although developing inference and asymptotic theory for low-dimensional parameters in the high-dimensional selection models is not the focus of this chapter, it makes an interesting topic for future research.

3.4.2 The pivotal Dantzig selector of the high-dimensional linear coefficients and confidence sets

The pivotal Dantzig selector (3.8) was originally proposed by Gautier and Tsybakov (2011) in the context of high-dimensional IV regression. For the particular case of this chapter where the instruments are the fitted regressors $\hat{v}$ themselves, this pivotal estimator is an extension of the Dantzig selector to accommodate for the fact that the variance of the noise η is unknown. It can be related to the square-root Lasso of Belloni, Chernozhukov, and Wang (2010) and Belloni, Chernozhukov, and Wang (2014). The non-asymptotic bounds derived in this section only apply to the case of exactly sparse β^*. However, all these results can be extended to the case of approximately sparse β^* by applying analysis similar to those from previous sections. The confidence sets are the by-products of the non-asymptotic bounds on the pivotal estimator. Construction of confidence sets is based on the following theorem (Theorem 3.4.8), which uses a bound for moderate deviations of self-normalized sums of random variables established by Jing, Shao and Wang (2003). This tool was first applied by Belloni, Chen, and Chernozhukov (2010) and later by Gautier and Tsybakov (2011) as well as Belloni, Chernozhukov, and Wang (2014). The following assumption is needed for this deviation bound to be applied in obtaining Theorem 3.4.8.

Assumption 3.4.8: For all $i = 1, ..., n$, $j = 1, ..., p$ and some constant $\delta' > 0$, $\mathbb{E}\left[|v_{ij}\eta_i|^{2+\delta'}\right] < \infty$ and neither of $v_{ij}\eta_i$ is almost surely equal to 0.
Define

$$b_{n,\delta'} := \min_{j=1,\ldots,p} \frac{\sqrt{\sum_{i=1}^{n} \mathbb{E}\left[v_{ij}^2\eta_i^2\right]}}{\left(\sum_{i=1}^{n} \mathbb{E}\left[|v_{ij}\eta_i|^{2+\delta'}\right]\right)^{1/(2+\delta')}}.$$

Given, for $j = 1, ..., p$, the variables $v_{ij}\eta_i$ are *i.i.d.*, we have

$$b_{n,\delta'} := n^{\frac{\delta'}{4+2\delta'}} \min_{j=1,\ldots,p} \frac{\sqrt{\mathbb{E}\left[v_{ij}^2\eta_i^2\right]}}{\left(\mathbb{E}\left[|v_{ij}\eta_i|^{2+\delta'}\right]\right)^{1/(2+\delta')}}. \tag{3.21}$$

For $a \geq 1$, set

$$\alpha = 2L\left(1 - \Phi\left(a\sqrt{2\log p}\right)\right) + 2a_0 \frac{\left(1 + a\sqrt{2\log p}\right)^{1+\delta'}}{p^{a^2-1} b_{n,\,\delta'}^{2+\delta'}}, \tag{3.22}$$

where $a_0 > 0$ is the absolute constant from the formula (2.11) in Jing, Shao and Wang (2003), and $\Phi(\cdot)$ is the standard normal c.d.f.

Notation. For Theorem 3.4.8, define the quantities $\hat{Q}(\beta) := \frac{1}{n}\left|\hat{v}_0 - \hat{v}\beta\right|_2^2$, and the l_2-sensitivity

$$\kappa^*_{J(\beta^*)} = \inf_{\Delta \in \mathbb{C}(J(\beta^*);0,\,\varphi)} \frac{\frac{1}{n}|\hat{v}^T\hat{v}\Delta|_\infty}{|\Delta|_2}$$

for some $\varphi > 1$. Recall from Section 3.4.1 the notation

$$B' := \frac{1}{n}\sum_{i=1}^{n}\left[w_i^T\hat{\theta} - w_i^T\theta^*\right]^2 \leq c\Upsilon_{w,\theta^*}\frac{k_1 \log d}{n}$$

where Υ_{w,θ^*} is a known function depending only on w and θ^*, and from Section 3.3 the notations

$$v_{j*} := \max_{i\in\{1,\ldots,n\}}\left\{|2x_{ij}| \vee |\hat{v}_{ij}|\right\}$$

for $j = 1, \ldots, p$, and D the diagonal $p \times p$ matrix with diagonal entries v_{j*}^{-1}, $j = 1, \ldots, p$.

Remark. Under Assumptions 3.4.2 and 3.4.3, the condition $\mathbb{E}\left[\left|v_{ij}\eta_i\right|^{2+\delta'}\right] < \infty$ is implied by the fact that v_{ij} (for all $j = 1, ..., p$) and η are sub-Gaussian. To see this, note that the random variable $v_{ij}\eta_i$ is sub-Exponential (using the fact that the product of two sub-Gaussian variables is sub-Exponential) and one of the characterizations of sub-Exponential variables says a zero-mean random variable X is sub-Exponential if and only if the quantity $\sup_{k\geq 2}\left[\frac{\mathbb{E}(X^k)}{k!}\right]^{1/k}$ is finite (see, e.g., Wainwright, 2015).

Theorem 3.4.8: Suppose the assumptions in Corollary 3.4.4 and Assumption 3.4.8 hold. For $a \geq 1$, choose α as in (3.22) and set the tuning parameter

$$\xi \geq a \max\left\{c_0\sqrt{\frac{\log p}{n}},\ \left(\hat{Q}(\beta^*)\right)^{-\frac{1}{2}}|\beta^*|_1 \frac{Lb(\sigma_v)\sqrt{B'}}{\min_{j=1,\ldots,p} v_{j*}}\right\} \tag{3.23}$$

where $c_0 > 1$ and $b(\sigma_v)$ is some known function depending only on σ_v. If $p \leq \exp\left(\frac{b^2_{n,\delta'}}{2a^2}\right)$, then with probability at least $1-\alpha-O\left(\frac{1}{p\wedge d}\right)$, for any solution $(\hat{\beta},\,\hat{\sigma})$ of program (3.8), we have

$$\left|D^{-1}(\hat{\beta}-\beta^*)\right|_2 \leq \frac{1}{\kappa^*_{J(\beta^*)}}\left[\frac{Lb(\sigma_v)\sqrt{B'}}{\min_{j=1,...,p} v_{j*}}\left|\hat{\beta}\right|_1 + 2\xi\hat{\sigma}\right]\left[1-\frac{\xi^2}{\kappa^*_{J(\beta^*)}}\right]^{-1} \cdot\left[1-\frac{1}{\kappa^*_{J(\beta^*)}}\left[\frac{Lb(\sigma_v)\sqrt{k_2B'}}{(\min_{j=1,...,p} v_{j*})^2}\right]\left[1-\frac{\xi^2}{\kappa^*_{J(\beta^*)}}\right]^{-1}\right]^{-1}. \quad (3.24)$$

and, for all $j = 1, ..., p$,

$$\left|\hat{\beta}_j-\beta^*_j\right| \leq \frac{1}{v_{j*}\kappa^*_{J(\beta^*)}}\left[\frac{Lb(\sigma_v)\sqrt{B'}}{\min_{j=1,...,p} v_{j*}}\left|\hat{\beta}\right|_1 + 2\xi\hat{\sigma}\right]\left[1-\frac{\xi^2}{\kappa^*_{J(\beta^*)}}\right]^{-1} \cdot\left[1-\frac{1}{\kappa^*_{J(\beta^*)}}\left[\frac{Lb(\sigma_v)\sqrt{k_2B'}}{(\min_{j=1,...,p} v_{j*})^2}\right]\left[1-\frac{\xi^2}{\kappa^*_{J(\beta^*)}}\right]^{-1}\right]^{-1}. \quad (3.25)$$

Furthermore,

$$C\hat{\sigma} \leq \left|\Delta_{J(\beta^*)}\right|_1 + C\sqrt{\hat{Q}(\beta^*)} \leq \frac{|\Psi_n\Delta|_\infty}{\kappa^*_{J(\beta^*),J(\beta^*)}} + C\sqrt{\hat{Q}(\beta^*)}. \quad (3.26)$$

The proof for Theorem 3.4.8 is provided in Section 3.8.9.

To construct confidence sets based on Theorem 3.4.8, notice that the bounds in (3.24)-(3.26) are meaningful if $\kappa^*_{J(\beta^*)} \geq \bar{\kappa} > 0$ (i.e., the l_2-sensitivity is strictly positive and bounded away from 0). In spite of the appearance, bound (3.24) has the same scaling as the bound in Corollary 3.4.4. This can be verified by Proposition 9.3 in Gautier and Tsybakov (2011) which shows that apart from some positive universal constant, the l_2-sensitivity is no smaller than the restricted eigenvalue multiplied by $k_2^{-\frac{1}{2}}$. However, in cases where the l_2-sensitivity is strictly larger, bound (3.24) would be sharper than the bound in Corollary 3.4.4. Gautier and Tsybakov (2011) provides a data-driven approach of computing $\kappa^*_{J(\beta^*)}$ without knowing $J(\beta^*)$. As long as the tuning parameter ξ is sufficiently small, i.e., $\left\{\sqrt{\frac{\log p}{n}},\,|\beta^*|_1\sqrt{\frac{k_1\log d}{n}}\right\} \longrightarrow 0$, then the term $1 - \frac{\xi^2}{\kappa^*_{J(\beta^*)}}$ in (3.24)-(3.25) is close to 1 provided $\kappa^*_{J(\beta^*)} \geq \bar{\kappa} > 0$. The choice of ξ specified by (3.23) has the same scaling as the choice of $\lambda_{n,3}$ for the non-pivotal Lasso estimator in Corollary 3.4.4 (in either case, the scaling of the tuning parameter needs to match the scaling of the maximum of the first-stage related error and the third-stage related error) except that the choice of ξ does not involve the unknown variance of η_i (and hence *pivotal*). In addition, notice that the upper bounds (3.24)-(3.25) are also *pivotal* to the unknown variance of η_i . The only terms that can involve unknown parameters in the choice of ξ and therefore the upper bounds (3.24)-(3.25) are: $Lb(\sigma_v)\sqrt{B'}$ and $\left(\hat{Q}(\beta^*)\right)^{-\frac{1}{2}} |\beta^*|_1$.

The term $Lb(\sigma_v)\sqrt{B'}$ is relatively easy to deal with: $b(\sigma_v)$ can be replaced with

$$b(\hat{\sigma}_v) := b\left(\max_j \sqrt{\frac{1}{n}\sum_{i=1}^n \hat{v}_{ij}^2}\right)$$

and an application of Lemma A.11 yields $\hat{\sigma}_v \leq 2\sigma_v$ with probability at least $1-O\left(\frac{1}{p\wedge d}\right)$; construction of confidence intervals (that do not contain any unknown parameters) for B' has been considered in the context of several Generalized Linear models (see, e.g., Van de Geer, 2008) and we will assume in this discussion that these confidence sets $\hat{B}'$ for B' are available. Consequently, whenever the term $Lb(\sigma_v)\sqrt{B'}$ shows up in the bounds (3.24)-(3.25), we will replace it with $Lb(\hat{\sigma}_v)\sqrt{\hat{B}'}$. In the case where the constant L is unknown, Section 3.4.1.2 discusses methods to determine this constant.

The term $\left(\hat{Q}(\beta^*)\right)^{-\frac{1}{2}} |\beta^*|_1$ is the harder one here as β^* is unknown and in fact the parameters we want to estimate. One possibility is to consider the following heuristic:

(i) In Step $k = 0$ (initialization), solve program (3.8) with $\xi^k = c_0\sqrt{\frac{\log p}{n}}$ to obtain $\hat{\beta}^k$ for some $c_0 > 1$; update ξ^k with

$$\xi^{k+1} \geq a \max\left\{ c_0\sqrt{\frac{\log p}{n}},\ \left(\hat{Q}(\hat{\beta}^k)\right)^{-\frac{1}{2}} \left|\hat{\beta}^k\right|_1 \frac{\hat{L}b(\hat{\sigma}_v)\sqrt{\hat{B}'}}{\min_{j=1,\ldots,p} v_{j*}} \right\}. \tag{3.27}$$

(ii) In Step $k+1$, solve program (3.8) with ξ^{k+1} to obtain $\hat{\beta}^{k+1}$ and update ξ^{k+1} with ξ^{k+2} as in (3.27). Repeat this step till a pre-specified tolerance level on $\left|\hat{\beta}^{k+1} - \hat{\beta}^k\right|_2$ is reached.

Establishing theoretical guarantees for the heuristic provided above is pursued in a separate ongoing project. In special cases, we may be able to circumvent the fact that

β^* is unknown. For example, when p is large relative to d so that $|\beta^*|_1 \sqrt{\frac{k_1 \log d}{n}} \ll \sqrt{\frac{\log p}{n}}$, then the result in Theorem 3.4.8 is essentially reduced to the case where the pivotal Dantzig selector is applied to the standard high-dimensional linear models with exact sparsity. In a related scenario where a post-Lasso procedure is performed on the regressors selected by the first-stage estimator $\hat{\theta}$ defined in (3.11), upon perfect selection or near-perfect selection of $\hat{\theta}$, the factor $\sqrt{\frac{k_1 \log d}{n}}$ is reduced to $\sqrt{\frac{k_1}{n}}$ which may be smaller relative to $\sqrt{\frac{\log p}{n}}$.

3.4.3 Properties of the estimators of the selection bias function

Given the availability of estimates $\hat{\theta}$ and $\hat{\beta}$ of the high-dimensional linear coefficients from either the non-pivotal procedure or the pivotal procedure, two different estimation strategies for the nonparametric selection bias are considered: one is the closed form estimator (3.13) and the other is the plug-in nonparametric least squares estimator (3.14) which can be obtained from the Lipschitz regression described in Section 3.4.1.2 if we assume $g(\cdot)$ belongs to the class $\mathcal{F}$ of Lipschitz functions. Despite the fact that (3.14) is computationally more involved relative to (3.13), its rate of convergence turns out to be faster as shown in the following. To facilitate the discussion and a later comparison with the minimax lower bounds in Section 3.4.4, we break down the presentations of the results into the case of exact sparsity on β^* and θ^* ($q_1 = q_2 = 0$) in Theorems 3.4.9 and 4.10, and the case of general sparsity on β^* and θ^* (q_1, $q_2 \in [0, 1]$) in Theorems 3.4.11 and 3.4.12 (which contain Theorems 3.4.9 and 3.4.10 as special cases, respectively).

Theorem 3.4.9 ($q_1 = q_2 = 0$): Let the assumptions in Corollary 3.4.4 hold. Suppose $g(\cdot)$ belongs to the class $\mathcal{F}$ of Lipschitz functions. For the estimator $\hat{g}(\cdot)$ of $g(\cdot)$ obtained by (3.13),

$$\left(\mathbb{E}\left[\hat{g}(w_i^T\hat{\theta}) - g(w_i^T\theta^*)\right]^2\right)^{\frac{1}{2}} \leq cb \max\left\{k_2\sqrt{\frac{\log p}{n}},\ |\beta^*|_1\left(\frac{k_1 \log d}{n}\right)^{\frac{1}{4}},\ k_2\,|\beta^*|_1\sqrt{\frac{k_1 \log d}{n}}\right\}$$

where b is some constant depending only on the model-specific structure (and independent of n, d, p, k_1, and k_2).

Theorem 3.4.10 (q_1, $q_2 \in [0, 1]$): Let the assumptions in Corollary 3.4.6 hold. Suppose $g(\cdot)$ belongs to the class $\mathcal{F}$ of Lipschitz functions. For the estimator $\hat{g}(\cdot)$ of $g(\cdot)$ obtained by (3.13),

$$\left(\mathbb{E}\left[\hat{g}(w_i^T\hat{\theta}) - g(w_i^T\theta^*)\right]^2\right)^{\frac{1}{2}} \leq cb\max\left\{R_{q_2}\bar{M}^{1-q_2},\ |\beta^*|_1 R_{q_1}^{\frac{1}{4}}\left(\sqrt{\frac{\log d}{n}}\right)^{\frac{1}{2}-\frac{q_2}{4}}\ |\beta^*|_1\left(\frac{|\theta^*|_1}{n}\right)^{\frac{1}{3}}\right\}$$

where $\bar{M}$ is defined in Corollary 3.4.6 and b is some constant depending only on the model-specific structure (and independent of n, d, p, R_{q_1}, and R_{q_2}).

Theorem 3.4.11 ($q_1 = q_2 = 0$): Let the assumptions in Corollary 3.4.4 hold. Suppose $g(\cdot)$ belongs to the class $\mathcal{F}$ of Lipschitz functions and the random matrix x is sub-Gaussian with parameters $(\Sigma_x,\, \sigma_x^2)$. For all $\Delta \in \mathbb{C}(J(\beta^*);\, 0,\, 3)\backslash\{\mathbf{0}\}$, the matrix Σ_x satisfies $\frac{\Delta^T\Sigma_x\Delta}{|\Delta|_2^2} \leq \kappa_U^x < \infty$. For the estimator $\tilde{g}(\cdot)$ of $g(\cdot)$ obtained by (3.14),

$$\left(\mathbb{E}\left[\tilde{g}(w_i^T\hat{\theta}) - g(w_i^T\theta^*)\right]^2\right)^{\frac{1}{2}} \leq c^{'}b^{'}\max\left\{\sqrt{\frac{k_2\log p}{n}},\ |\beta^*|_1\sqrt{\frac{k_1k_2\log d}{n}},\ \left(\frac{k_1}{n}\right)^{\frac{1}{3}}\right\}$$

where $b^{'}$ is some constant depending only on the model-specific structure (and independent of n, d, p, k_1, and k_2).

Theorem 3.4.12 ($q_1,\, q_2 \in [0,\, 1]$): Let the assumptions in Corollary 3.4.6. Suppose $g(\cdot)$ belongs to the class $\mathcal{F}$ of Lipschitz functions and the random matrix x is sub-Gaussian with parameters $(\Sigma_x,\, \sigma_x^2)$. For all non-zero $\Delta \in \mathbb{C}(S_{\underline{\tau}};\, q_2,\, 3)\cap\mathbb{S}_\delta$ where $\mathbb{C}(S_{\underline{\tau}};\, q_2,\, 3)\cap\mathbb{S}_\delta$ is defined in Corollary 3.4.6, the matrix Σ_x satisfies $\frac{\Delta^T\Sigma_x\Delta}{|\Delta|_2^2} \leq \kappa_U^x < \infty$. For the estimator $\tilde{g}(\cdot)$ of $g(\cdot)$ obtained by (3.14),

$$\left(\mathbb{E}\left[\tilde{g}(w_i^T\hat{\theta}) - g(w_i^T\theta^*)\right]^2\right)^{\frac{1}{2}} \leq c^{'}b^{'}\max\left\{\sqrt{R_{q_2}}\bar{M}^{1-\frac{q_2}{2}},\ \left(\frac{|\theta^*|_1}{n}\right)^{\frac{1}{3}}\right\},$$

where $\bar{M}$ is defined in Corollary 3.4.6 and $b^{'}$ is some constant depending only on the model-specific structure (and independent of n, d, p, R_{q_1}, and R_{q_2}).

Remark. The proofs for Theorems 3.4.9-3.4.12 are provided in Sections 3.8.10 and 3.8.11. First let us look at the case of exactly sparse β^* and θ^* ($q_1 = q_2 = 0$). From Theorem 3.4.11, notice that the terms $\sqrt{\frac{k_2\log p}{n}}$ and $|\beta^*|_1\sqrt{\frac{k_1k_2\log d}{n}}$ are expected from the statistical error of $\hat{\beta}$ that we plug into the nonparametric regression (3.14); and the term $\left(\frac{k_1}{n}\right)^{\frac{1}{3}}$ is expected from the fact that $g(\cdot)$ belongs to the class $\mathcal{F}$ of Lipschitz functions[7]. On the other hand, the term $\left(\frac{k_1}{n}\right)^{\frac{1}{3}}$ is suppressed by $\left(\frac{k_1\log d}{n}\right)^{\frac{1}{4}}$ in Theorem 3.4.9

[7]Note that when ϵ_{1i} and ϵ_{2i} in (3.1) are bivariate normal, the selection bias characterized by the Inverse Mills Ratio is a 1−Lipschitz function (see, e.g., Ruud, 2000). Furthermore, if $m_j(\cdot) \in \mathcal{F}_j$ and

for the closed-form estimator (3.13). When β^* is approximately sparse with $q_2 = 1$, Theorem 3.4.10 implies that the $\sqrt{MSE}$ of the closed-form estimator (3.13) is bounded above by $R_{q_2}\bar{M}^{1-q_2} = |\beta^*|_1$. This upper bound is unimprovable and as a result, it is not possible for (3.13) to achieve *MSE*-consistency even if $n \to \infty$ when $q_2 = 1$. In contrast to (3.13), the nonparametric least squares estimator (3.14) is consistent in *MSE* as $n \to \infty$ when $q_2 = 1$. The key behind the sharp rate achieved by the plug-in nonparametric least squares estimator (3.14) in Theorems 3.4.11 and 3.4.12 lies on the random variables

$$\frac{1}{n}\sum_{i=1}^{n}\eta_i\left[\tilde{g}(w_i^T\theta^*) - g(w_i^T\theta^*)\right],$$

and

$$U_n := \sup_{\delta\in\mathcal{S}(r_1,\, r_2)}\frac{1}{n}\left|\eta^T w\delta\right|,$$

where

$$\mathcal{S}(r_1,\, r_2) := \left\{\delta \in \mathbb{R}^d \mid |\delta|_1 \leq r_1,\ |\delta|_2 \leq r_2\right\}.$$

The analysis for controlling the first term uses a "local function complexity" argument similar to what is done in the proofs for Theorems 3.4.1 and 3.4.2. To upper bound the second term U_n, we can apply a discretization argument over the set $\mathcal{S}(r_1,\, r_2)$ together with results on metric entropy and the fact $\mathbb{E}\left[\eta_i|w_i\right] = 0$. "Small" values of r_1 and r_2 are guaranteed by the upper bounds on $|\hat{\theta} - \theta^*|_2$ from Lemma A.7 and as a result we only need to work with a "small" $\mathcal{S}(r_1,\, r_2)$. The sharp rates provided by these types of analysis seem to be driven by the projection nature of the underlying nonparametric least-squares estimators[8]. As we will see in the following section, the overall convergence rate of the estimator $\hat{\beta}$ (obtained by either the non-pivotal procedure or the pivotal procedure) and the plug-in nonparametric least squares estimator (3.14) is *minimax optimal* in terms of the $(n,\, d,\, p)-$scaling for the case of exactly sparse β^*. However, we will also see that this minimax optimality result does not apply to the case of approximately sparse β^* because of the first-stage related estimation error.

3.4.4 Statistical efficiency via lower bounds on minimax risks

This section studies efficiency of the proposed estimators by deriving lower bounds on minimax rates for the case of l_2-loss. Minimax lower bounds, applicable to any procedure regardless of its computational cost, are complementary to the understanding

$\mathcal{F}_j$ is the class of $L-$Lipschitz functions for every $j = 1, ..., p$, then $g(\cdot)$ is a Lipschitz function.

[8] In fact, a recent chapter by Chatterjee (2014) shows that the least squares estimators are always *admissible* up to a universal constant in many modern statistics problems.

of computationally efficient methods. First, they can reveal gaps between the performance of an optimal procedure in theory and known computationally efficient procedures. Moreover, they can demonstrate regimes in which practical procedures achieve these lower bounds. While one way of interpreting minimax lower bounds is to view the choice of unknown parameters in an adversarial manner, and to compare the estimators based on their worst-case performance, many techniques for deriving minimax lower bounds can be motivated by the Bayesian approach which views the unknown parameters as random variables (e.g., Guntuboyina 2011).

The minimax lower bounds in this section are derived for model (3.2), implied by the original selection model (3.1). As a consequence, these lower bounds provide information-theoretic limits for any procedure constructed based on model (3.2) for estimating model (3.1), regardless of its computational cost. For $q_1,\, q_2 \in [0,\, 1]$, define $\mathcal{H} = \mathcal{B}^p_{q_2}(R_{q_2}) \times \mathcal{F} \circ \mathcal{B}^d_{q_1}(R_{q_1})$, where the l_q- "ball" is defined in Section 3.2 and $\mathcal{F}$ is the class of functions such that $g \in \mathcal{F} : \mathbb{R} \to \mathbb{R}$. When $\beta^* \in \mathcal{B}^p_0(k_2)$ and $\theta^* \in \mathcal{B}^d_0(k_1)$, model (3.2) corresponds to the case of exact sparsity on β^* and θ^*. When $\beta^* \in \mathcal{B}^p_{q_2}(R_{q_2})$ and $\theta^* \in \mathcal{B}^d_{q_1}(R_{q_1})$ for $q \in (0,\, 1]$, model (3.2) corresponds to the case of approximate sparsity based upon imposing a certain decay rate on the ordered entries of β^* and θ^*. Theorem 3.4.13 (Theorem 3.4.14) presents a minimax lower bound for the case of exact sparsity $q_1 = q_2 = 0$ (respectively, the case of approximate sparsity $q_1,\, q_2 \in (0,\, 1]$).

Assumption 3.4.9: There exists a constant $\underline{\kappa}_x > 0$ and a function $f_l(R_{q_2},\, q_2,\, n,\, p)$ such that

$$\frac{1}{\sqrt{n}} \left|x\beta\right|_2 \geq \underline{\kappa}_x \left|\beta\right|_2 - f_l(R_{q_2},\, q_2,\, n,\, p) \quad \text{for all } \beta \in \mathcal{B}^p_{q_2}(R_{q_2}).$$

Assumption 3.4.10: There is no measurable function $f(w_i^T\theta)$ such that $x_i^T\lambda = f(w_i^T\theta)$ when $y_{1i} = 1$ for $\lambda \in \mathcal{B}^p_{q_2}(R_{q_2}) \backslash \{\mathbf{0}\}$.

Remark. Assumptions 3.4.9 and 3.4.10 ensure the identifiability of model (3.2), without which, lower bounds for high-dimensional linear models usually involve a maximum of two quantities: a term involving the diameter of the null-space restricted to the l_q-ball, measuring the degree of non-identifiability of the model, and a term arising from the metric entropy structure for l_q-balls (see Raskutti, et. al, 2011). Assumption 3.4.9 together with Assumption 3.4.10 incurs an upper bound on the $\mathcal{B}^p_q(R)-$kernel diameter in l_2-norm (this result is formalized in Lemma A.10 and proved in Section 3.8.12), and consequently the identifiability of model (3.2).

Theorem 3.4.13 ($q_1 = q_2 = 0$): Let $\mathcal{F}$ be the class of $L-$Lipschitz functions and Assumptions 3.4.9-3.4.10 hold with $f_l(R_{q_2},\, q_2,\, n,\, p) = 0$ and $\underline{\kappa}_x > 0$. Let the parameter space Θ be the set of elements θ in $\mathcal{B}_0^d(k_1)$ such that for any $\lambda \in \mathcal{B}_0^p(k_2)\backslash\{\mathbf{0}\}$, $\nexists$ a measurable $g(\cdot) \in \mathcal{F}$, $x_i^T\lambda = g(w_i^T\theta)$ when $y_{1i} = 1$. Moreover, $\frac{|w\theta|_2}{\sqrt{n}|\theta|_2} \leq \kappa_u$ for all $\theta \in \Theta$ and $\frac{|x\beta|_2}{\sqrt{n}|\beta|_2} \leq \kappa_u^{'}$ for all $\beta \in \mathcal{B}_0^p(k_2)$. If the vector $\eta \sim N(0,\, \sigma_\eta I_{n\times n})$, then, for some constant b depending only on the model-specific structure (and independent of n, d, p, k_1, and k_2),

$$\begin{aligned}
&\min_{\tilde{\beta},\, \tilde{f},\, \tilde{\theta}} \max_{\substack{\theta \in \Theta \\ f(\cdot) \in \mathcal{F} \\ \beta \in \mathcal{B}_0^p(k_2)}} \left(\mathbb{E}\left|\tilde{\beta} - \beta\right|_2^2\right)^{\frac{1}{2}} + \left(\mathbb{E}\left[\tilde{f}(w_i^T\tilde{\theta}) - f(w_i^T\theta)\right]^2\right)^{\frac{1}{2}} \\
\geq\;& b\max\left\{\sqrt{\frac{k_1\log d}{n}},\, \left(\frac{k_1}{n}\right)^{\frac{1}{3}},\, \sqrt{\frac{k_2\log p}{n}}\right\}.
\end{aligned}$$

Theorem 3.4.14 ($q_1,\, q_2 \in (0,\, 1]$): Let $\mathcal{F}$ be the class of $L-$Lipschitz functions and Assumptions 3.4.9-3.4.10 hold with $f_l(R_{q_2},\, q_2,\, n,\, p) = o\left(R_{q_2}^{\frac{1}{2}}\left(\frac{\log p}{n}\right)^{\frac{1}{2}-\frac{q_2}{4}}\right)$ and $\underline{\kappa}_x > 0$. Moreover, $\frac{1}{\sqrt{n}}\max_{j=1,\ldots,d}\left|w_j\right|_2 \leq \kappa_w < \infty$ and $\frac{1}{\sqrt{n}}\max_{j=1,\ldots,p}\left|x_j\right|_2 \leq \kappa_x < \infty$. If the vector $\eta \sim N(0,\, \sigma_\eta I_{n\times n})$, then, for some constant $b^{'}$ depending only on the model-specific structure (and independent of n, d, p, R_{q_1}, and R_{q_2}),

$$\begin{aligned}
&\min_{\tilde{\beta},\, \tilde{f},\, \tilde{\theta}} \max_{\substack{\theta \in \Theta \\ f(\cdot) \in \mathcal{F} \\ \beta \in \mathcal{B}_{q_2}^p(R_{q_2})}} \left(\mathbb{E}\left|\tilde{\beta} - \beta\right|_2^2\right)^{\frac{1}{2}} + \left(\mathbb{E}\left[\tilde{f}(w_i^T\tilde{\theta}) - f(w_i^T\theta)\right]^2\right)^{\frac{1}{2}} \\
\geq\;& b^{'}\max\left\{R_{q_1}^{\frac{1}{2}}\left(\frac{\log d}{n}\right)^{\frac{2-q_1}{4}},\, \left(\frac{R^*}{n}\right)^{\frac{1}{3}},\, R_{q_2}^{\frac{1}{2}}\left(\frac{\log p}{n}\right)^{\frac{2-q_2}{4}}\right\},
\end{aligned}$$

where the parameter space Θ is defined in Theorem 3.4.13 with $\mathcal{B}_0^d(k_1)$ replaced by $\mathcal{B}_{q_1}^d(R_{q_1})$ and $\mathcal{B}_0^p(k_2)$ replaced by $\mathcal{B}_{q_2}^p(R_{q_2})$, and R^* is the "radius" R_{q_1} when $q_1 = 1$.

Remark. The proofs for Theorem 3.4.13 and Theorem 3.4.14 are provided in Sections 3.8.12 and 3.8.13, respectively. These proofs are information-theoretic in nature and based on Fano's inequality (see, e.g., Guntuboyina, 2011; Wainwright, 2015) and results on the metric entropy of the l_q-balls. By Lemma A.10, the conditions on $f_l(R_{q_2},\, q_2,\, n,\, p)$ in Theorems 3.4.13 and 3.4.14 together with Assumption 3.4.10 ensure that the kernel diameter for the nullspace of $\mathcal{B}^p_{q_2}(R_{q_2})$ is dominated by the term related to the metric entropy of $\mathcal{B}^p_{q_2}(R_{q_2})$. In Theorem 3.4.13, we require Assumption 3.4.9 to hold with $f_l(R_{q_2},\, q_2,\, n,\, p) = 0$ and $\underline{\kappa}_x > 0$, which is closely related to the restricted eigenvalue condition on the matrix $\frac{x^T x}{n}$ over the set $\mathcal{B}^p_0(k_2)$. When x is a sub-Gaussian matrix with parameters $(\Sigma_x,\, \sigma^2_x)$ and for all $\Delta \in \mathcal{B}^p_0(k_2)\backslash\{\mathbf{0}\}$, the matrix Σ_x satisfies $\frac{\Delta^T \Sigma_x \Delta}{|\Delta|_2^2} \geq \underline{\kappa}_x > 0$, then Lemma B.2 guarantees Assumption 3.4.9 to hold for $f_l(R_{q_2},\, q_2,\, n,\, p) = 0$ with high probability. Additionally, if w is a sub-Gaussian matrix with parameters $(\Sigma_w,\, \sigma^2_w)$, $\frac{\Delta^T \Sigma_w \Delta}{|\Delta|_2^2} \leq \bar{\kappa}^w < \infty$ for all $\theta \in \Theta$, and $\frac{\Delta^T \Sigma_x \Delta}{|\Delta|_2^2} \leq \bar{\kappa}^x < \infty$ for all $\beta \in \mathcal{B}^p_0(k_2)$, Lemma B.2 also guarantees that $\frac{|w\theta|_2^2}{n|\theta|_2^2} \leq c\bar{\kappa}^w$ for all $\theta \in \Theta$ and $\frac{|x\beta|_2^2}{n|\beta|_2^2} \leq c\bar{\kappa}^x$ for all $\beta \in \mathcal{B}^p_0(k_2)$ hold with high probability (in Theorem 3.4.13). Similarly, the conditions $\frac{1}{\sqrt{n}} \max_{j=1,\ldots,d} |w_j|_2 \leq \kappa_w < \infty$ and $\frac{1}{\sqrt{n}} \max_{j=1,\ldots,p} |x_j|_2 \leq \kappa_x < \infty$ (in Theorem 3.4.14) are also implied by Lemma B.2 with high probability given w_j $(j = 1, ..., d)$ and x_j $(j = 1, ..., p)$ are sub-Gaussian.

Compare the scaling of the lower bound in Theorem 3.4.13 with the upper bounds in Corollary 3.4.4 and Theorem 3.4.11. In particular, from the previous upper bounds, we have

$$\left(\mathbb{E}\left|\hat{\beta}_{HSEL} - \beta^*\right|_2^2\right)^{\frac{1}{2}} + \left(\mathbb{E}\left[\tilde{g}(w_i^T\hat{\theta}) - g(w_i^T\theta^*)\right]^2\right)^{\frac{1}{2}}$$
$$\precsim \max\left\{|\beta^*|_1\sqrt{\tfrac{k_2 k_1 \log d}{n}},\ \left(\tfrac{k_1}{n}\right)^{\frac{1}{3}},\ \sqrt{\tfrac{k_2 \log p}{n}}\right\}$$

(The upper bound on $\mathbb{E}\left|\hat{\beta}_{HSEL} - \beta^*\right|_2^2$ is obtained by converting $\left|\hat{\beta}_{HSEL} - \beta^*\right|_2^2$ with a standard integration over the tail probability in the exponential form). Notice that the scaling in the upper bound above matches the lower bound in Theorem 3.4.13 in terms of $(n,\, d,\, p)-$factors. The only difference between these bounds is that the upper bound exceeds the lower bound by a factor of $|\beta^*|_1\sqrt{k_2} \asymp k_2^{\frac{3}{2}}$ in the term related to the complexity of the set Θ, which is likely due to the fact that the estimator $\hat{\beta}_{HSEL}$ is a sequential multi-stage procedure based on plugging in the first-stage estimator $\hat{\theta}$ in the place of the unknown coefficient θ^* in the selection equation. In a different but somewhat related context which concerns with the high-dimensional sparse linear regression models with many endogenous regressors and instruments (see Zhu 2014), it is found that the upper bound on the $\sqrt{MSE}$ of the l_1-regularized two-stage estimator exceeds the minimax lower bound in Zhu (2014) by a factor of k_2 (where k_2 is the sparsity parameter for the second-stage model).

On the other hand, this minimax optimality result does not apply to the case of approximately sparse β^* when we compare the scaling of the lower bound in Theorem 3.4.14 with the upper bounds in Corollary 3.4.6 and Theorem 3.4.12 for the case $q_2 \in (0, 1]$. In particular, from the previous upper bounds, we have,

$$\left(\mathbb{E}\left|\hat{\beta}_{HSEL} - \beta^*\right|_2^2\right)^{\frac{1}{2}} + \left(\mathbb{E}\left[\tilde{g}(w_i^T\hat{\theta}) - g(w_i^T\theta^*)\right]^2\right)^{\frac{1}{2}}$$

$$\precsim \left\{\sqrt{R_{q_2}}\left[\max\left\{\sqrt{\tfrac{\log p}{n}},\ |\beta^*|_1\left(\tfrac{|\theta^*|_1}{n}\right)^{\frac{2}{3}},\ |\beta^*|_1 R_{q_1}^{\frac{1}{2}}\left(\sqrt{\tfrac{\log d}{n}}\right)^{1-\frac{q_1}{2}}\right\}\right]^{1-\frac{q_2}{2}}\right\} \vee \left\{\left(\tfrac{|\theta^*|_1}{n}\right)^{\frac{1}{3}}\right\} \quad .$$

As in the case of exactly sparse β^*, the terms $\sqrt{R_{q_2}}\left(\sqrt{\frac{\log p}{n}}\right)^{1-\frac{q_2}{2}}$ and $\left(\frac{|\theta^*|_1}{n}\right)^{\frac{1}{3}}$ in the above upper bound match the scalings of the term related to the complexity of the set $\mathcal{B}_{q_2}^p(R_{q_2})$ and the term related to the complexity of the set $\mathcal{F}$, respectively. In sharp contrast to the case of exactly sparse β^* where our sequential multi-stage procedure based on plugging in the first-stage estimator $\hat{\theta}$ only exceeds the minimax optimal result by a factor of $k_2^{\frac{3}{2}}$ in the term related to the complexity of the set Θ, the term $\sqrt{R_{q_2}}\left[|\beta^*|_1 R_{q_1}^{\frac{1}{2}}\left(\sqrt{\frac{\log d}{n}}\right)^{1-\frac{q_1}{2}}\right]^{1-\frac{q_2}{2}}$ in the upper bound above is now worsened by an exponent of $1-\frac{q_2}{2}$ and a factor of $\sqrt{R_{q_2}}\left(|\beta^*|_1\right)^{1-\frac{q_2}{2}}$ when compared to the term related to Θ, $R_{q_2}^{\frac{1}{2}}\left(\sqrt{\frac{\log p}{n}}\right)^{1-\frac{q_1}{2}}$, in the lower bound of Theorem 3.4.14. When $q_2 \in [0, 1]$, note that

$$\left[R_{q_1}^{\frac{1}{2}}\left(\sqrt{\frac{\log d}{n}}\right)^{1-\frac{q_1}{2}}\right]^{1-\frac{q_2}{2}} \geq R_{q_1}^{\frac{1}{2}}\left(\sqrt{\frac{\log d}{n}}\right)^{1-\frac{q_1}{2}}$$

with "=" holds only if $q_2 = 0$ (the case of exactly sparse β^*).

The lower bound in either Theorem 3.4.13 or Theorem 3.4.14 is a "point" result. Even if the main equation in the original selection model (3.1) has a normal error, the normality of η is plausible in model (3.2) only if $g(w_i^T\theta^*) = 0$, i.e., when there is no selection activity. Nevertheless, these "point" results provided by Theorems 3.4.13 and 3.4.14 are still useful because whether $g(w_i^T\theta^*)$ equals 0 or not would be unknown in general and the error from having to estimate $g(\cdot)$ still appears in the lower bounds. Moreover, even if the "point" result does not hold "globally", given that the lower bounds are derived for the minimax risks of the high-dimensional linear coefficients together with the nonparametric selection bias function, at least the second and third terms in the lower bounds of Theorems 3.4.13 and 4.14 should be unimprovable in any "global" result. It is possible to impose distributional assumptions other than normality on η but the derivation of the lower bounds in the proofs may involve more difficult computations related to the Kullback-Leibler divergence or the more general $f-$divergence where f is a convex function with $f(1) = 0$ (see Guntuboyina 2011 for a unified treatment of existing techniques for obtaining lower bounds). For this reason, existing literature on minimax lower bounds almost exclusively focuses on the case of normal errors and lower bounds with less restrictive distributional assumptions other than normality (e.g., sub-Gaussianity) on a random vector are in general impossible to obtain. Recent work of efficiency bounds (e.g., Hansen B., 2014) that proposes a shrinking neighborhood analysis may provide a promising direction for extending these "point" results to the case where $g(w_i^T\theta^*)$ is in a shrinking neighborhood of 0.

3.4.5 Estimation of high-dimensional semilinear models with a two-stage projection strategy

In this section, we discuss how the theory developed in this chapter can be applied to the semilinear models in high-dimensional settings. The multi-stage estimator proposed in this chapter is also useful for estimating the linear coefficients of the following semilinear model:

$$y_i = x_i^T\beta^* + g(w_i) + \eta_i \tag{3.28}$$

where x_i is a $p-$dimensional vector of regressors (and p can grow with and exceed the sample size n). Furthermore, $g(\cdot) : \mathbb{R}^d \to \mathbb{R}$ is an unknown function and w_i is a $d-$dimensional vector of regressors. Our multi-stage projection strategy is now reduced to a two-stage procedure. Based on the analysis in this chapter, it is straightforward to see that Theorems 3.4.1 and 3.4.2 remain valid except that there is no first-stage related error $\mathcal{T}_1$ in the upper bounds on the estimator of β^*. As we have mentioned before in Section 3.4.1, when p and d are fixed and small relative to n, as long as $\mathcal{F}_j$ $(j = 0, ..., p)$ in Theorems 3.4.1 and 3.4.2 are sufficiently smooth so that $\mathbb{E}(x_{ij} \,|\, w_i)$ and $\mathbb{E}(y_i \,|\, w_i)$ can be estimated at a rate *no slower* than $O\left(n^{-\frac{1}{4}}\right)$, the "oracle" property will be achieved.

The case where the dimensions p and d are both large relative to n (namely, $p \geq n$ and $d \geq n$) generalizes the semilinear model considered in Belloni, et. al (2014) in which $d \geq n$ and p remains finite. When $p \geq n$, it is unclear whether the procedure proposed by Belloni, et. al (2014)[9] can be easily extended because the effect from imperfect

[9]When $d \geq n$ and p remains finite, the procedure from Belloni, et. al (2014) includes three steps:

selection in the second step of Belloni, et. al (2014) may not be negligible anymore when the number of components in x_i is also large relative to n. Instead, the projection strategy proposed in this chapter can be used to estimate β^* in the semilinear model (3.28) when $p \geq n$ and $d \geq n$. One way to reduce the curse of dimensionality in the joint multivariate nonparametric component $\mathbb{E}(z_{ij} \,|\, w_i)$ (recall $z_j = x_j$ for $j = 1, ..., p$ and $z_0 = y$) is to consider the class of additive models of the form (Hastie and Tibshirani, 1999):

$$\mathbb{E}(z_{ij} \,|\, w_i) := f_j(w_i) = \sum_{l=1}^{d} f_{jl}(w_{il})$$

where $f_{jl}(\cdot) \in \mathcal{F}_{jl}$ for $l = 1, ..., d$ and $j = 0, ..., p$.

Let us consider the simplest case of $f_{jl}(w_{il}) = w_{il}\theta_l^*$ where θ_l^* is a scalar and $\theta^* \in \mathcal{B}_{q_1}^d(R_{q_1})$ and $\beta^* \in \mathcal{B}_{q_2}^p(R_{q_2})$ for $q_1,\, q_2 \in [0,\, 1]$. Theorem 3.4.1 implies that the l_2-error of the two-stage estimator is bounded above by

$$O\left(\left[R_{q_2}^{\frac{1}{2}}\left(\sqrt{\frac{\log p}{n}}\right)^{1-\frac{q_2}{2}}\right] \vee \left[R_{q_2}^{\frac{1}{2}}\left(|\beta^*|_1\, R_{q_1}\left(\frac{\log d}{n}\right)^{1-\frac{q_1}{2}}\right)^{1-\frac{q_2}{2}}\right]\right).$$

For a more general structure on f_j, suppose $J(f_j) := \{l \,:\, f_{jl} \neq 0\}$ and $k_{1j} = |J(f_j)|$, the cardinality of $J(f_j)$, can increase to infinity with n but slowly compared to n (i.e., f_j is exactly sparse) and

$$\mathbb{E}(z_{ij} \,|\, w_{il}) = \sum_{k=1}^{\infty} \vartheta_{jlk}\phi_{jlk}(w_{il})$$

where $B_{jl} = (\phi_{jlk})_{k=1}^{\infty}$ is an orthonormal basis for $\mathcal{F}_{jl}$. For a truncation parameter M, also define

$$\mathbb{E}^M(z_{ij} \,|\, w_{il}) = \sum_{k=1}^{M} \vartheta_{jlk}\phi_{jlk}(w_{il}).$$

Let Ψ_{jl} denote the $n \times M$ matrix with $\Psi_{jl}(i,\, k) = \phi_{jlk}(w_{il})$. For the first-stage estimation in our two-stage procedure, consider the following minimization problem:

$$\min_{\vartheta_{jl} \in \mathbb{R}^M} \,:\, \frac{1}{2n}\left|z_j - \sum_{l=1}^{d} \Psi_{jl}\vartheta_{jl}\right|_2^2 + \lambda_n \sum_{l=1}^{d} \sqrt{\frac{1}{n}\vartheta_{jl}^T \Psi_{jl}^T \Psi_{jl}\vartheta_{jl}} \qquad (3.29)$$

First, apply the Lasso to the regression of y_i on w_i; second, apply the Lasso to the regression of x_{ij} on w_i for every $j = 1, ..., p$, respectively; and third, regress y_i on x_i and the components of w_i selected by the first and second step.

for some regularization parameter $\lambda_n > 0$. Program (3.29) is the sample version (with truncation) of the following:

$$\min_{f_{jl}\in\mathcal{F}_{jl}} : \frac{1}{2}\mathbb{E}\left(z_{ij} - \sum_{l=1}^{d} f_{jl}(w_{il})\right)^2 + \lambda_n \sum_{l=1}^{d} \sqrt{\mathbb{E}\left(f_{jl}^2(w_{il})\right)}.$$

The optimization program (3.29) is considered in Ravikumar, Lafferty, and Wasserman (2009) and can be viewed as a functional version of the grouped Lasso (Yuan and Lin, 2006). It can be solved with a coordinate descent algorithm proposed by Ravikumar, et. al (2009).

3.5 Monte-Carlo simulation

In this section, simulations are conducted to gain preliminary understanding of the small-sample performance of the non-pivotal multi-stage estimator $\hat{\beta}_{HSEL}$; ongoing work involves implementation of the pivotal procedure described in Section 3.4.2. We consider model (3.1) where $w \in \mathbb{R}^{n\times d}$ is a matrix consisted of independent uniform zero-mean random variables on $[-2, 2]$ with variance $\sigma_w \approx 1.33$ and x takes on the first p columns of w. The *i.i.d.* errors $\epsilon_{1i} \sim \mathcal{N}(0, 1)$ for $i = 1, ..., n$ where n denotes the number of observations generated for the selection equation. We consider two scenarios where $n = 88$ and $n = 200$. Given the setup here, on average 44 (when $n = 88$) and 100 (when $n = 200$) observations, respectively, will be used for estimating the main equation. Conditional on the observations is with $y_{1i} = 1$, the *i.i.d.* errors $(\epsilon_{1i}, \epsilon_{2i})$ have the following joint normal distribution

$$(\epsilon_{1i}, \epsilon_{2i}) \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho\sigma_2 \\ \rho\sigma_2 & \sigma_2 \end{pmatrix}\right),$$

where $\rho \in \{0, 0.9\}$ and $\sigma_2 \in \{0.3, 1, 2\}$. We set $d = 90$, $p = 45$, $k_1 = 4$, and $k_2 = 2$. When $n = 88$, this setup of dimensionality represents a selection model where the number of regressors in the selection equation and the main equation, respectively, exceeds the number of observations used to estimate the corresponding equation, while the number of relevant regressors (ones with nonzero coefficients) is small relative to the sample size. We set $\theta_j^* = 0.5$ for $j = 1, 2, 3, 46$ and the rest of components in θ^* take on values of 0; set $\beta_1^* = \beta_{45}^* = 1$ and the rest of components in β^* take on values of 0. This set up ensures that there is at least one component w_{ij} with θ_j^* in the support set of θ^* such that w_{ij} is excluded from x_i.

We consider four sets of experiments. The first experiment (Experiment 1) concerns the multi-stage estimator $\hat{\beta}_{HSEL}$. As a benchmark for Experiment 1, Experiment 2 applies a one-step Lasso procedure (without correcting selection bias) to the same main equation. Experiments 3 and 4 are benchmarks concerning classical low-dimensional settings. Experiment 3 applies the Heckman's 2-step procedure to model (3.1) where the selection equation and the main equation are in the low-dimensional setting and the supports of the true parameters in both equations are known *a priori*; Experiment 4 applies the OLS to the same low-dimensional model as Experiment 3. We simulate 100 sets of data following the process described above. For each set $t = 1, ..., 100$, we compute the estimates $\hat{\beta}^t$ of the main-equation parameters β^*, l_2-errors of these estimates, $|\hat{\beta}^t - \beta^*|_2$, and selection percentages of $\hat{\beta}^t$ (computed by the number of the elements in $\hat{\beta}^t$ sharing the same sign as their corresponding elements in β^*, divided by the total number of elements in β^*). Results reported in this section include:

(**a**) the mean of the relevant estimates $\frac{1}{100}\sum_{t=1}^{100}\hat{\beta}_1^t$;
(**b**) the mean of the relevant estimates $\frac{1}{100}\sum_{t=1}^{100}\hat{\beta}_{45}^t$;
(**c**) the mean of the averaged irrelevant estimates $\frac{1}{43}\sum_{j\neq 1,45}\frac{1}{100}\sum_{t=1}^{100}\hat{\beta}_j^t$;
(**d**) the mean of the l_2-errors of the estimates $\hat{\beta}^t$ computed as $\frac{1}{100}\sum_{t=1}^{100}|\hat{\beta}^t - \beta^*|_2$;
(**e**) the mean of the selection percentages (computed in a similar fashion as the mean of the l_2-errors of the estimates);
(**f**) the mean of the squared l_2-errors (i.e., the *sample mean squared error*, SMSE, computed as $\frac{1}{100}\sum_{t=1}^{100}|\hat{\beta}^t - \beta^*|_2^2$);
(**g**) the sample squared bias $\sum_{j=1}^{45}(\bar{\hat{\beta}}_j - \beta_j^*)^2$ (where $\bar{\hat{\beta}}_j = \frac{1}{100}\sum_{t=1}^{100}\hat{\beta}_j^t$ for $j = 1, ..., 45$).

The results in this section regarding Experiment 1 are based on the choices of the regularization parameter $\lambda_{n,1} = 0.5\sqrt{\frac{\log d}{n}}$ for the first-stage estimation problem (3.11) and the regularization parameter $\lambda_{n,3} = 0.2k_2\sqrt{\frac{k_1 \log d}{n_s}}$ for the third-stage estimation problem (3.6), where n_s denotes the number of observations with $y_{1i} = 1$. The scalings of $\lambda_{n,1}$ and $\lambda_{n,3}$ are chosen according to Proposition 3.4.2 and Corollary 3.4.4, respectively. The choice of $0.2k_2\sqrt{\frac{k_1 \log d}{n_s}}$ is also used in Experiment 2 for comparing the performance of the proposed procedure and the Lasso without corrective measures. Note that for $\sigma_2 = 1$ and $\sigma_x = \sigma_w \approx 1.33$, $0.2k_2\sqrt{\frac{k_1 \log d}{n_s}}$ is slightly greater than $2\sigma_2 \cdot \sigma_x\sqrt{\frac{\log p}{n_s}}$, the smallest value required for the Lasso estimation of the standard sparse high-dimensional linear models (e.g., Bickel, et. al, 2009). The second-stage estimation in Experiment 1 is based on solving (3.20) with $L = 1$. Ongoing work involves implementing the cross validation procedure described in Section 3.4.1.2 to determine L when ϵ_{2i} has a non-normal distribution.

From Table 5.1, we see that the direct Lasso estimator without correcting selection bias outperforms the multi-stage estimator $\hat{\beta}_{HSEL}$ when $\rho = 0$, and *vice versa* when $\rho = 0.9$. For the design considered here, in the presence of substantial selection activity ($\rho = 0.9$), the mean of the l_2-errors (**row d**) and the sample squared bias of the estimates (**row g**) by the direct Lasso procedure without corrective measures are exacerbated in the high-dimensional setting and this exacerbation mainly comes from the poorer estimates of the relevant regressors as the mean of the averaged irrelevant estimates varies little from the case $\rho = 0$ to the case $\rho = 0.9$. Other simulation results (not included here due to space limit) show that when σ_2 is increased (decreased) from 1 to 2 (respectively, from 1 to 0.3), $\hat{\beta}_{HSEL}$ performs worse (respectively, better) relative to the case $\sigma_2 = 1$, and similar patterns are observed when w_{ij}s are drawn from independent uniform zero-mean random variables on $[-1, 1]$ (respectively, on $[-4, 4]$). Also, as n increases from 88 to 200, $\hat{\beta}_{HSEL}$ performs substantially better. These findings are intuitive and expected. It is worth noting that for the design considered here, in terms of the mean of the selection percentages (**row e**), the direct Lasso procedure without corrective measures is comparable to $\hat{\beta}_{HSEL}$ even in the case $\rho = 0.9$. Ongoing work is exploring situations where variable selection by $\hat{\beta}_{HSEL}$ substantially outperforms variable selection by the direct Lasso procedure.

Table 5.1: Monte-Carlo simulation results for $n = 88$, $\sigma_2 = 1$

	$\rho = 0$				$\rho = 0.9$			
	HSEL	**LASSO**	**HECK**	**OLS**	**HSEL**	**LASSO**	**HECK**	**OLS**
a	0.703	0.730	1.005	1.010	0.627	0.605	1.007	1.002
b	0.742	0.736	0.996	0.994	0.762	0.757	1.006	1.020
c	-0.001	-0.001	NA	NA	-0.001	-0.001	NA	NA
d	0.446	0.430	0.159	0.161	0.474	0.495	0.145	0.164
e	0.969	0.965	NA	NA	0.981	0.979	NA	NA
f	0.227	0.209	0.032	0.033	0.249	0.262	0.025	0.034
g	0.155	0.143	4×10^{-5}	9×10^{-5}	0.197	0.217	9×10^{-5}	4×10^{-4}

3.6 An empirical application to the retail gasoline market

Having established the theoretical properties of the 3-step estimators, we now apply one of these estimators to an empirical example of price-discrimination in the retail gasoline market. When consumers have different valuations for a good and a firm knows the distribution of consumer valuations in the market but not the exact valuation of any specific consumer prior to the sale, the firm can offer a menu of different prices, appropriately bundled with other aspects of the product (such as product quality), and force consumers to choose bundles consistent with their preferences. When differences in costs incurred to produce various bundles in the menu are small compared to the differences in prices, this menu-based offering is a price discrimination mechanism. Retail gasoline markets present a good context to study price discrimination as different gasoline stations in a market typically face similar costs of procuring gasoline (e.g., Shepard, 1991; Iyer and Seetharaman, 2003, etc.). Therefore, any price differences across gasoline stations are likely due to reasons unrelated to the cost of procuring gasoline. Gasoline retailers can choose to be either a two-product station offering both self-service and full-service gasoline or a single-product station offering only full-service or self-service gasoline. While a single-product station is unable to price discriminate, a two-product station can charge different prices for full- and self-service gasoline and induce consumers with different valuations to choose the products consistent with their preferences.

Shepard (1991) estimates pricing decisions of gasoline stations without endogenizing their decisions to price discriminate, i.e., their choice to be single versus multi-product. Iyer and Seetharaman (2003) explicitly examines a firm's incentive to price-discriminate. In doing so, they highlight the importance of accounting for self selectivity considerations in empirical analysis of price discrimination based on market data. Specifically, Iyer and Seetharaman employ a binary probit framework to model a gasoline station's decision to be single-product or multi-product as a function of market and station characteristics, and then model the prices chosen by the gasoline station for its product(s) by estimating linear regressions with Heckman's self-selectivity correction conditional on the station's decision to offer a single- or multi-product. They show that incorrect inferences about the incentive to price discriminate and about the differences in the prices charged between single-product and multi-product stations would result if the endogeneity in the choice of the station-type were ignored in the estimation. Their empirical analysis also shows that a larger income spread in the market implies a greater likelihood of the gasoline station being multi-product. However, Iyer and Seetharaman (2003) did not account for interactions between the gas stations in their empirical analysis. Studies show that pricing decisions of retail gasoline stations may depend on the degree of competitive intensity in the "market" (e.g., Slade, 1992). In the empirical literature on competitive gasoline markets, there have been various ways of defining a "market" (see, e.g., Slade, 1986; Pinkse, Slade and Brett, 2002; Iyer and Seetharaman, 2008). For example, Iyer and Seetharaman (2008) defines mutually exclusive census tracts as local markets, and treat each market as the unit of observation in their empirical analysis. In previous research, markets have been defined based on stations that fall within a circle of half a mile or one mile radius.

One common feature of the previous definitions of competitive markets is that they are subjective heuristics. It would be ideal if one can control for the interactions between different stations without requiring *a priori* knowledge of the structure of the competitive market. Recent work including Manresa (2014) and Bonaldi, Hortacsu, Kastl (2014) develop econometric models to recover the underlying networks in different applications. Both papers hinge on the availability of panel data for each observation in the cross section. In particular, Manresa considers settings where outcomes depend on an agent's own characteristics and on the characteristics of other agents in the data. She applies a Lasso type estimator to identify individuals generating spillovers and their strength using panel data on outcomes and characteristics. Bonaldi, et. al proposes a new measure of systemic risk based on estimating spillovers between funding costs of individual banks with a Lasso type procedure, which is applied to the panel of each individual bank to recover the financial network. However, for the empirical application considered in this section, panel data of each gas station is not available and as a consequence, the econometric model by either Manresa or Bonaldi is not suitable. Instead, we use geographic information and spatial data to create a set of measures that are high-dimensional to control for the interactions between the gas stations and employ one of our proposed estimators to identify the competitive market structure. The following subsection describes the data followed by the empirical model.

Data and the empirical model

This chapter uses the data set from Iyer and Seetharaman (2003). It was collected during July 1998 from a cross-section of 249 gasoline stations in the Greater Saint Louis metropolitan area. Among the 249 stations, 65 are multi-product stations and 172 are single-product self-service stations. In addition, there were 12 single-product full-service stations. In the United States, the low incidence of full-service single product stations is typical and in certain regions full service is required by law. As in Iyer and Seetharaman (2003), we exclude them from the empirical analysis. The survey data include the prices of three grades - 87, 89 and 93 octane levels - of gasoline, along with station-specific characteristics, i.e., number of gasoline pumps, special advertising for cigarettes and soda, presence of convenience store, pay-at-pump facility, car wash, service station, and the number of stations with prices that are visible to a given station. This data set also contains demographic information including income, population density, age distribution, home value, and education levels. This information comes from 1990 U.S. census data, which contain demographic information at the level of each census tract.

The data also records addresses of each station, from which "Bing Maps REST Services" is used to obtain geographic information including longitude and latitude, travel distance in driving mode between any pair of stations, etc. This information can be used to create variables for partially controlling for the interactions between stations. In particular, given any station, we can count the number of stations and/or stations under one of the three national brands (namely, Amoco, Shell and Mobil), that fall within 1km, 1km and 2km, and so on, from this station. Each of the numbers is then divided by the area (in km^2) of the corresponding layer. The number of stations with prices that are visible to a given station is another useful measure of interaction between stations and this information is available in the data. Using the number of competitors as a measure of interaction between firms has been seen in previous literature (e.g., Bresnahan and Reiss 1991; Iyer and Seetharaman, 2008). The novelty introduced by this section lies in the data-driven nature of the approach: rather than assuming *a priori* knowledge of the structure of the competitive network, it relies on the data to determine the geographic pattern of interaction between stations. If panel data on prices and time-varying instrumental variables for prices are available, we can include prices of other stations in the main equation. Some variants of the 3-step estimators in this chapter combined with the high-dimensional IV estimator in Gautier and Tsybakov (2011) or the high-dimensional 2SLS estimator in Zhu (2013) may be considered as an alternative to identify the sets of competitive markets engaged in pricing. However, in the retail gasoline market, it may be difficult to obtain valid time-varying instrumental variables for prices.

As in Iyer and Seetharaman (2003), we use a binary probit model for the selection of service types where $y_{1i} = 1$ in (3.1) indicates that station i offers multi-service and $y_{1i} = 0$ indicates that station i offers single-self-service. The same set of explanatory variables included in the binary probit model of Iyer and Seetharaman is used here: average income (*AVG*), income spread (*SPREAD*), brand (*BRAND*), pay-at-pump facility (*PAP*), presence of convenience store (*CONV*), car wash (*WASH*), and service station (*SERV*). My empirical model differs from Iyer and Seetharaman mainly in terms of the specifications of the linear pricing model (the main equation): First, while Iyer and Seetharaman assume the selection bias takes on the functional form of the Inverse Mills Ratio, we assume the selection bias function to obey the more general nonparametric single index restriction in (3.1); second, we add a set of measures that are high-dimensional to partially control for the competition effects from other stations. In particular, the following explanatory variables are included in the pricing model: *AVG*, *BRAND*, special advertising for cigarettes and soda (*ADSCC*), the number of stations with visible prices (*VISP*), the total number of stations and the number of stations under one of the three national brands within 1km (*TOT_1* and *BRND_1*), 1km and

2km (*TOT_2* and *BRND_2*), $\cdots$, 34km and 35km (*TOT_35* and *BRND_35*) from a given station. In summary, for the pricing equation, we have $n = 172$, $p = 74$ for stations that serve single-self-service grade-87 gasoline, $n = 168$, $p = 74$ for stations that serve single-self-service grade-93 gasoline, $n = 65$, $p = 74$ for stations that serve multi-service grade-87 gasoline, and $n = 65$, $p = 74$ for stations that serve multi-service grade-93 gasoline. While Iyer and Seetharaman include only average income and brand in their pricing model, they suggest that special advertising for cigarettes and soda might be correlated with the retail gasoline prices and hence we include this information in our pricing model. The last group of variables are measures added to partially control for the competition effects from other stations. Iyer and Seetharaman found the indicators of the presence of pay-at-pump facilities and service stations statistically significant and therefore, the exclusion restriction required by the selection model considered in this application is likely to be satisfied given the setup. Moreover, they justify the exclusion restriction by arguing that a station's decision on the configuration of its station characteristics such as pay-at-pump, convenience store, car-wash, and service station involves costly investments that the station owner has made along with the station-type decision when setting up the retail facility. In contrast, the pricing decisions may vary on a daily basis.

The following summarizes the estimation procedure and empirical findings. We briefly discuss the results pertaining to the effects of the service-type decisions and focus mainly on the empirical findings from the pricing regression because the main difference between Iyer and Seetharaman and the empirical analysis in this section lies in the latter.

Estimation and empirical findings

A standard maximum likelihood procedure for estimating low-dimensional binary probit models is used to obtain estimates of the selection equation of service-type decisions. The estimation is performed for grade-87 stations and grade-93 stations, respectively, and the results are reported in Table 6.1.

Table 6.1: Results of the binary probit model

	Intercept	*AVG*	*SPREAD*	*BRAND*	*PAP*	*CONV*	*WASH*	*SERV*
87	-1.525^{***}	0.008^{*}	-2.591^{**}	0.962^{**}	-0.851^{**}	-0.341	-0.012	2.341^{***}
	(0.453)	(0.005)	(1.227)	(0.411)	(0.414)	(0.321)	(0.346)	(0.292)
93	-1.522^{***}	0.008^{*}	-2.583^{**}	0.958^{**}	-0.848^{**}	-0.338	-0.004	2.335^{***}
	(0.452)	(0.005)	(1.229)	(0.410)	(0.413)	(0.321)	(0.347)	(0.292)

Because individual-level income is not available for this data set, it is not possible to compute the sample standard deviation in income for each tract. Instead, we use two measures to approximate income spread: one is the absolute difference between the percentages of median-level income group and the low-level income group for each tract; the other is the absolute difference between the percentages of median-level income group and the high-level income group for each tract. A smaller value in the first (second) absolute difference indicates a more evenly distributed population in the low-income (respectively, high-income) group and the median income group. It turns out that the second measure is not statistically significant and hence we drop this measure from the probit model. As a consequence, the negative sign of the estimate for $SPREAD$ suggests that more heterogeneous income levels below the $50^{th}-$percentile in the market implies a greater likelihood of the station being multi-product.

For the linear pricing regression model conditional on the service type, the non-pivotal estimator $\hat{\beta}_{HSEL}$ based on (3.6) is used to select the variables with non-zero coefficients and then $\tilde{\beta} := \left(\hat{v}_{\hat{J}}^T \hat{v}_{\hat{J}}\right)^{-1} \left(\hat{v}_{\hat{J}}^T \hat{v}_0\right)$ is computed with $\hat{J} := J(\hat{\beta}_{HSEL})$ (this is the Post-Lasso procedure discussed in Section 3.4.1.3). For the second-stage estimation, program (3.20) is solved where the Lipschitz constant L is determined by the cross-validation procedure described in Section 3.4.1.2 and the choice of $L = 1$ turns out to be robust. For the third-stage estimation, given the setup of our empirical model, the first-stage related estimation error is likely to be dominated by the third-stage related error in the choice of $\lambda_{n,3}$ from Corollary 3.4.4 and hence we choose $\lambda_{n,3}$ based on the third-stage related error. Program (3.6) is first solved with the choice of $\lambda_{n,3}(t) = 2.001 \cdot \hat{\sigma}_v \hat{\sigma}_\eta^t \sqrt{\frac{\log p}{n_s}}$ for $t = 0$ (initialization), where $\hat{\sigma}_v := \max_{j=1,...,p} \sqrt{\frac{1}{n}\sum_{i=1}^n \hat{v}_{ij}^2}$, $\hat{\sigma}_\eta^0 = 1$, and n_s denotes the number of observations used for the pricing regression. Let $\hat{\beta}_{HSEL}^t$ denote the resulting estimate based on $\lambda_{n,3}(t)$ and $\hat{\sigma}_\eta^{t+1}$ denote the updated sample standard deviation of the fitted residuals $\hat{\eta}_i^t := \hat{v}_{i0} - \hat{v}_i \hat{\beta}_{HSEL}^t$ for $i = 1, ..., n_s$. Program (3.6) is then solved with the updated $\lambda_{n,3}(t+1) = 2.001 \cdot \hat{\sigma}_v \hat{\sigma}_\eta^{t+1} \sqrt{\frac{\log p}{n_s}}$. Repeat this process until a pre-specified tolerance level on $\left|\hat{\sigma}_\eta^{t+1} - \hat{\sigma}_\eta^t\right|$ is reached. The result shows that the choice of $\hat{\sigma}_\eta^t \approx 0.04$ is robust. After experimenting with a range of values around the final choice of $\lambda_{n,3}$ determined according to the described procedure, the set of variables selected by $\hat{\beta}_{HSEL}$ for the range of $\lambda_{n,3}$ and the post Lasso estimates $\tilde{\beta}$ based on these selected variables are reported in Table 6.2 for the following groups:

SSL: single-self-service grade-87 gasoline;
SSH: single-self-service grade-93 gasoline;

MSL: multi-self-service grade-87 gasoline;
MSH: multi-self-service grade-93 gasoline.

The variables with blanks in Table 6.2 correspond to those that are not selected by $\hat{\beta}_{HSEL}$ in a particular group. The numerical values within parentheses are bootstrapped standard errors for $\tilde{\beta}$. Estimates with three asterisks, two asterisks, and a single asterisk are statistically significant at level $\alpha = 0.01$, $\alpha = 0.05$, and $\alpha = 0.1$, respectively. Note that our second-stage estimation and third-stage estimation use the demeaned explanatory variables and demeaned prices, so the intercept term is excluded from the pricing regression model.

Table 6.2: Results of pricing regression

	AVG	*BRAND*	*TOT_2*	*TOT_4*
SSL		0.030^{***} (0.005)		-0.052^{***} (0.016)
SSH		0.049^{***} (0.007)	-0.030^{**} (0.015)	-0.030^{*} (0.021)
MSL	$2 \times 10^{-4**}$ (1×10^{-4})			-0.070^{***} (0.029)
MSH				-0.096^{***} (0.032)

Regarding the results of the pricing regression in Table 6.2, the estimate of *BRAND* has a positive sign in SSL and SSH. Moreover, *AVG* (average income) has a positive effect on the pricing decisions in MSL. In Iyer and Seetharaman (2003) which estimated the low-dimensional linear regression counterpart by pooling observations of the single-self-service and multi-self-service each with Heckman's selectivity correction (for grade-87 and grade-93, respectively), *BRAND* and *AVG* are the only two variables included in their pricing model and found to be statistically significant. Based on our empirical results which remove selection bias and partially control for potential interactions between the stations simultaneously, we see that *TOT_4* is selected by $\hat{\beta}_{HSEL}$ in *all* groups and statistically significant at level 0.01 in SSL, MSL, and MSH and at level 0.1 in SSH. *TOT_2* is selected by $\hat{\beta}_{HSEL}$ in SSH and statistically significant at level 0.05. The negative sign of the estimate for *TOT_4* in all groups (*TOT_2* in SSH) suggests that the total number of stations within 3km-4km (respectively, 1km-2km) of a given station has a negative effect on its price. On the other hand, the variable *VISP* (the number of stations with visible prices) is not selected by $\hat{\beta}_{HSEL}$, which is less intuitive.

However, it is possible that in the presence of several clusters of gas stations, there is less competition from adjacent clusters relative to ones that are somewhat further apart. For example, Iyer and Seetharaman (2008) analyzed a similar but richer data set on prices and station characteristics gathered across stations in the Saint Louis metropolitan area and found that closely located retailers who face sufficient heterogeneity in preferences across consumers in a local market may differentiate on product design and pricing strategies (also see, e.g., Png and Reitman, 1994); in contrast, retailers that are farther apart from each other may adopt similar product design and pricing strategies if the market is relatively homogeneous (also see, e.g., Slade 1992). Another explanation for the finding where more competition comes from somewhat intermediate retailers instead of closest ones is that consumers of retail gasoline may travel from their suburban homes located in a neighborhood of one cluster to their work places or central shopping areas located in another cluster that may be somewhat further away; the further located clusters of stations may be linked by routes that are more convenient for commuting (these routes may be more direct or less congested, etc.). This explanation may suggest that retailers consider commuting behavior of their customers when setting the retail price. Investigating this factor requires more substantial empirical analysis and a data set that includes more detailed information on the demographics and business environment, which will be pursued in future research. Nevertheless, the main finding on *TOT_2* and *TOT_4* suggests that in modeling the pricing decisions of retail gas stations, not only it is useful to account for the self selectivity of service-type but also to take into considerations of potential interactions between stations; in particular, competition effects from retailers that are not in the same local market (e.g., the same census tract or neighborhood within a circle of half a mile or one mile radius, etc.) should not be overlooked.

3.7 Conclusion

This chapter provides estimation tools together with their theoretical guarantees for the semiparametric sample selection model in high-dimensional settings under a weak nonparametric restriction on the form of the selection correction. In particular, the number of regressors in the main equation, p, and the number of regressors in the selection equation, d, can grow with and exceed the sample size n. The main theoretical results of this paper are finite-sample bounds from which sufficient scaling conditions on the sample size for estimation consistency and variable-selection consistency (i.e., the multi-stage high-dimensional estimation procedure correctly selects the non-zero coefficients in the main equation with high probability) are established. Statistical efficiency of the proposed estimators is studied via lower bounds on minimax risks. Inference procedures for the coefficients of the main equation, one based on a pivotal Dantzig selector to construct non-asymptotic confidence sets and one based on a post-selection strategy (when perfect or near-perfect selection of the high-dimensional coefficients is achieved), are discussed.

Small-sample performance of one of the proposed procedures is evaluated by Monte-Carlo simulations and illustrated with an empirical application to the retail gasoline market in the Greater Saint Louis area. The preliminary simulation results show that the "bias" from not performing the selection correction is exacerbated in high-dimensional settings. For the empirical application, this paper models a firm's choice of either a single-product or multi-product service as a function of market and station characteristics and then models the station's pricing decision, conditional on the choice of the station type. Using geographic information and spatial data, a set of variables that are high-dimensional is introduced to control for interactions between the gas stations. The empirical finding suggests that competition effects from retailers that are not in the same local market should not be overlooked.

3.8 Main proofs for results in Chapter 3

3.8.1 Lemmas A.1-A.2

Lemma A.1: Let $\hat{\Gamma} = \frac{1}{n}\hat{v}^T\hat{v}$ and

$$\begin{aligned} e &= \left[\hat{\mathbb{E}}\left(x|w\hat{\theta}\right) - \mathbb{E}\left(x|w\hat{\theta}\right) + \mathbb{E}\left(x|w\hat{\theta}\right) - \mathbb{E}\left(x|w\theta^*\right)\right]\beta^* \\ &- \left[\hat{\mathbb{E}}\left(y_2|w\hat{\theta}\right) - \mathbb{E}\left(y_2|w\hat{\theta}\right) + \mathbb{E}\left(y_2|w\hat{\theta}\right) - \mathbb{E}\left(y_2|w\theta^*\right)\right] + \eta. \end{aligned}$$

Suppose the assumptions in Theorems 3.4.1 or 3.4.2 hold. If $\lambda_{n,3}$ in program (3.6) satisfies

$$\lambda_{n,3} \geq 2|\frac{1}{n}\hat{v}^T e|_\infty > 0,$$

and

$$R_{q_2}\underline{\tau}^{-q_2}\left(\frac{\log p}{n} + \mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_3\right) = O(\kappa_L),$$

then there is a constant $c > 0$ such that under Assumption 3.2.2 ($q_2 \in [0,\, 1]$ and when $q_2 = 0$, $R_{q_2} := k_2$),

$$|\hat{\beta}_{HSEL} - \beta^*|_2 \leq \frac{c}{\kappa_L^{1-\frac{q_2}{2}}} R_{q_2}^{\frac{1}{2}}\left(\lambda_{n,3} \vee |\frac{2}{n}\hat{v}^T e|_\infty\right)^{1-\frac{q_2}{2}}.$$

Proof. First, write

$$\begin{aligned} v_0 &= \hat{v}_0 + \hat{\mathbb{E}}\left(y_2|w\hat{\theta}\right) - \mathbb{E}\left(y_2|w\hat{\theta}\right) + \mathbb{E}\left(y_2|w\hat{\theta}\right) - \mathbb{E}\left(y_2|w\theta^*\right) \\ &= v\beta^* + \eta = \left[\hat{v} + \hat{\mathbb{E}}\left(x|w\hat{\theta}\right) - \mathbb{E}\left(x|w\hat{\theta}\right) + \mathbb{E}\left(x|w\hat{\theta}\right) - \mathbb{E}\left(x|w\theta^*\right)\right]\beta^* + \eta, \end{aligned}$$

thus we have

$$\begin{aligned} \hat{v}_0 &= \hat{v}\beta^* + \left[\hat{\mathbb{E}}\left(x|w\hat{\theta}\right) - \mathbb{E}\left(x|w\hat{\theta}\right) + \mathbb{E}\left(x|w\hat{\theta}\right) - \mathbb{E}\left(x|w\theta^*\right)\right]\beta^* \\ &\quad - \left[\hat{\mathbb{E}}\left(y_2|w\hat{\theta}\right) - \mathbb{E}\left(y_2|w\hat{\theta}\right) + \mathbb{E}\left(y_2|w\hat{\theta}\right) - \mathbb{E}\left(y_2|w\theta^*\right)\right] + \eta \\ &:= \hat{v}\beta^* + e \end{aligned}$$

where

$$\begin{aligned} e &= \left[\hat{\mathbb{E}}\left(x|w\hat{\theta}\right) - \mathbb{E}\left(x|w\hat{\theta}\right) + \mathbb{E}\left(x|w\hat{\theta}\right) - \mathbb{E}\left(x|w\theta^*\right)\right]\beta^* \\ &- \left[\hat{\mathbb{E}}\left(y_2|w\hat{\theta}\right) - \mathbb{E}\left(y_2|w\hat{\theta}\right) + \mathbb{E}\left(y_2|w\hat{\theta}\right) - \mathbb{E}\left(y_2|w\theta^*\right)\right] + \eta. \end{aligned}$$

Define the thresholded subset

$$S_{\underline{\tau}} := \left\{j \in \{1,\, 2,\, ...,\, p\} \,:\, \left|\beta_j^*\right| > \underline{\tau}\right\}$$

where $\underline{\tau} = \frac{\lambda_{n,3}}{\kappa_L}$ is the threshold parameter. Define $\hat{\Delta} = \hat{\beta}_{HSEL} - \beta^*$ and the Lagrangian $L(\beta;\, \lambda_{n,3}) = \frac{1}{2n}|\hat{u} - \hat{v}\beta|_2^2 + \lambda_{n,3}|\beta|_1$. Since $\hat{\beta}_{HSEL}$ is optimal, we have

$$L(\hat{\beta}_{HSEL};\, \lambda_{n,3}) \leq L(\beta^*;\, \lambda_{n,3}) = \frac{1}{2n}|e|_2^2 + \lambda_{n,3}|\beta^*|_1,$$

Some algebraic manipulation of the *basic inequality* above yields

$$\begin{aligned} 0 &\leq \frac{1}{2n}|\hat{v}\hat{\Delta}|_2^2 \leq \frac{1}{n}e^T\hat{v}\hat{\Delta} + \lambda_{n,3}\left\{|\beta^*_{S_{\underline{\tau}}}|_1 + |\beta^*_{S^c_{\underline{\tau}}}|_1 - |(\beta^*_{S_{\underline{\tau}}} + \hat{\Delta}_{S_{\underline{\tau}}},\, \beta^*_{S^c_{\underline{\tau}}} + \hat{\Delta}_{S^c_{\underline{\tau}}})|_1\right\} \\ &\leq |\hat{\Delta}|_1|\frac{1}{n}\hat{v}^Te|_\infty + \lambda_{n,3}\left\{|\hat{\Delta}_{S_{\underline{\tau}}}|_1 - |\hat{\Delta}_{S^c_{\underline{\tau}}}|_1 + 2|\beta^*_{S^c_{\underline{\tau}}}|_1\right\} \\ &\leq \frac{\lambda_{n,3}}{2}\left\{3|\hat{\Delta}_{S_{\underline{\tau}}}|_1 - |\hat{\Delta}_{S^c_{\underline{\tau}}}|_1 + 4|\beta^*_{S^c_{\underline{\tau}}}|_1\right\}, \end{aligned} \tag{3.30}$$

where the last inequality holds as long as $\lambda_{n,3} \geq 2|\frac{1}{n}\hat{v}^Te|_\infty > 0$. Consequently,

$$|\hat{\Delta}|_1 \leq 4|\hat{\Delta}_{S_{\underline{\tau}}}|_1 + 4|\beta^*_{S^c_{\underline{\tau}}}|_1 \leq 4\sqrt{\left|S_{\underline{\tau}}\right|}|\hat{\Delta}|_2 + 4|\beta^*_{S^c_{\underline{\tau}}}|_1.$$

We now upper bound the cardinality of $S_{\underline{\tau}}$ in terms of the threshold $\underline{\tau}$ and l_q-ball radius of R_{q_2}. Note that we have

$$R_{q_2} \geq \sum_{j=1}^{p} \left|\beta_j^*\right|^{q_2} \geq \sum_{j\in S_{\underline{\tau}}} \left|\beta_j^*\right|^{q_2} \geq \underline{\tau}^{q_2} \left|S_{\underline{\tau}}\right|$$

and therefore $\left|S_{\underline{\tau}}\right| \leq \underline{\tau}^{-q_2} R_{q_2}$. To upper bound the approximation error $|\beta^*_{S^c_{\underline{\tau}}}|_1$, we use the fact that $\beta^* \in \mathcal{B}^p_{q_2}(R_{q_2})$ and have

$$|\beta^*_{S^c_{\underline{\tau}}}|_1 = \sum_{j\in S^c_{\underline{\tau}}} \left|\beta_j^*\right| = \sum_{j\in S^c_{\underline{\tau}}} \left|\beta_j^*\right|^{q_2} \left|\beta_j^*\right|^{1-q_2} \leq R_{q_2}\underline{\tau}^{1-q_2}.$$

Putting the pieces together yields

$$|\hat{\Delta}|_1 \leq 4\sqrt{\underline{\tau}^{-q_2} R_{q_2}}|\hat{\Delta}|_2 + 4R_{q_2}\underline{\tau}^{1-q_2}. \tag{3.31}$$

Let us first prove the case of $q_2 \in (0,\, 1]$. Note that we also have

$$\begin{aligned}
\frac{1}{2n}|\hat{v}\hat{\Delta}|_2^2 &\leq |\hat{\Delta}|_1|\frac{1}{n}\hat{v}^T e|_\infty + \lambda_{n,3}\left\{|\hat{\Delta}_{S(\underline{\tau})}|_1 - |\hat{\Delta}_{S(\underline{\tau})^c}|_1 + |\beta^*_{S(\underline{\tau})^c}|_1\right\} \\
&\leq \left(8\sqrt{\underline{\tau}^{-q_2} R_{q_2}}|\hat{\Delta}|_2 + 4R_{q_2}\underline{\tau}^{1-q_2}\right)\lambda_{n,3} \\
&\leq c_1\sqrt{\underline{\tau}^{-q_2} R_{q_2}}|\hat{\Delta}|_2\lambda_{n,3} + c_2\underline{\delta} \\
&\leq \max\left\{c_1 R_{q_2}^{\frac{1}{2}}\kappa_L^{\frac{q_2}{2}}\left(\lambda_{n,3}\right)^{1-\frac{q_2}{2}}|\hat{\Delta}|_2,\ c_2\underline{\delta}\right\}
\end{aligned}$$

where the third and fourth inequalities follow from our choices of $\underline{\tau} = \frac{\lambda_{n,3}}{\kappa_L}$ and $\underline{\delta} \asymp R_{q_2}\lambda_{n,3}\tau^{1-q_2}$. Now we proceed by cases. If

$$\max\left\{c_1 R_{q_2}^{\frac{1}{2}}\kappa_L^{\frac{q_2}{2}}\left(\lambda_{n,3}\right)^{1-\frac{q_2}{2}}|\hat{\Delta}|_2,\ c_2\underline{\delta}\right\} = c_1 R_{q_2}^{\frac{1}{2}}\kappa_L^{\frac{q_2}{2}}\left(\lambda_{n,3}\right)^{1-\frac{q_2}{2}}|\hat{\Delta}|_2,$$

so that under the condition

$$R_{q_2}\underline{\tau}^{-q_2}\left(\frac{\log p}{n} + \mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_3\right) = O(\kappa_L),$$

and provided $c > 0$ is sufficiently large, we have

$$|\hat{\Delta}|_2 \geq c\kappa_L^{-1+\frac{q_2}{2}} R_{q_2}^{\frac{1}{2}}\left(\lambda_{n,3}\right)^{1-\frac{q_2}{2}} \geq \delta^* \tag{3.32}$$

where

$$\delta^* = O\left(\max\left\{\mathcal{T}_1^{\frac{3}{2}-q_2},\, \mathcal{T}_2^{\frac{3}{2}-q_2},\, \mathcal{T}_3^{\frac{3}{2}-q_2},\, \mathcal{T}_4^{2-q_2}\right\} R_{q_2}\right).$$

Consequently, (3.30) and (3.32) together imply that

$$\hat{\Delta} \in \mathbb{K}(\underline{\delta},\, S_{\underline{\tau}}) := \mathbb{C}(S_{\underline{\tau}}; q_2,\, 3) \cap \{\Delta \in \mathbb{R}^p \,:\, |\Delta|_2 \geq \delta^*\} \tag{3.33}$$

where $\mathbb{C}(S_{\underline{\tau}};\, 3) := \left\{\Delta \in \mathbb{R}^p \,:\, |\Delta_{S_{\underline{\tau}}^c}|_1 \leq 3|\Delta_{S_{\underline{\tau}}}|_1 + 4|\beta^*_{S_{\underline{\tau}}^c}|_1\right\}$. By Lemma A.2 and Lemma A.3 (or Lemma A.5), the random matrix $\hat{\Gamma} = \hat{v}^T\hat{v}$ satisfies the RE condition (2.1) over

$$\mathbb{C}(S_{\underline{\tau}};\, 3) \cap \{\Delta \in \mathbb{R}^p \,:\, |\Delta|_2 \geq \delta^*\},$$

therefore, we have

$$\kappa_L|\hat{\Delta}|_2^2 \leq \frac{1}{2n}|\hat{v}\hat{\Delta}|_2^2 \leq c_1 R_{q_2}^{\frac{1}{2}} \kappa_L^{\frac{q_2}{2}} \left(\lambda_{n,3}\right)^{1-\frac{q_2}{2}} |\hat{\Delta}|_2$$

so the claim follows. On the other hand, if

$$\max\left\{c_1 R_{q_2}^{\frac{1}{2}} \kappa_L^{\frac{q_2}{2}} \left(\lambda_{n,3}\right)^{1-\frac{q_2}{2}} |\hat{\Delta}|_2,\; c_2\underline{\delta}\right\} = \underline{\delta},$$

then

$$|\hat{\Delta}|_2 \leq c\kappa_L^{-1+\frac{q_2}{2}} R_{q_2}^{\frac{1}{2}} \left(\lambda_{n,3}\right)^{1-\frac{q_2}{2}}$$

so again the claim follows.

To prove the case of $q_2 = 0$, simply choose $\underline{\delta} = 0$ and $S_{\underline{\tau}} = J(\beta^*)$ and the claim follows trivially from the above argument.□

Remark. Inequality (3.31) implies that $|\hat{\Delta}|_1 \precsim R_{q_2}\left(\lambda_{n,3}\right)^{1-q_2}$.

Lemma A.2: Define the thresholded subset

$$S_{\underline{\tau}} := \left\{j \in \{1,\, 2,\, ...,\, p\} \,:\, \left|\beta_j^*\right| > \underline{\tau}\right\}.$$

Under the assumptions in Theorem 3.4.1 (or Theorem 3.4.2) and the choice $\underline{\tau} = \frac{\lambda_{n,3}}{\kappa_L}$, if

$$R_{q_2}\underline{\tau}^{-q_2}\left(\frac{\log p}{n} + \mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_3\right) = O(\kappa_L),$$

the RE condition (2.1) of $\frac{\hat{v}^T\hat{v}}{n}$ holds over the set

$$\mathbb{C}(S_{\underline{\tau}}; q_2,\, 3) \cap \{\Delta \in \mathbb{R}^p \,:\, |\Delta|_2 \geq \delta^*\}$$

where

$$\delta^* = O\left(\max\left\{\mathcal{T}_1^{\frac{3}{2}-q_2},\, \mathcal{T}_2^{\frac{3}{2}-q_2},\, \mathcal{T}_3^{\frac{3}{2}-q_2},\, \mathcal{T}_4^{2-q_2}\right\} R_{q_2}\right).$$

Proof. The argument is similar to what is used in the proof of Lemma 2 from Negahban, et. al (2010). For any $\Delta \in \mathbb{C}(S_{\underline{\mathcal{T}}}; q_2, 3)$, we have

$$|\Delta|_1 \leq 4|\Delta_{S_{\underline{\mathcal{T}}}}|_1 + 4|\beta^*_{S^c_{\underline{\mathcal{T}}}}|_1 \leq 4\sqrt{\left|S_{\underline{\mathcal{T}}}\right|}|\Delta|_2 + 4R_{q_2}\underline{\mathcal{T}}^{1-q_2} \leq 4\sqrt{R_{q_2}}\underline{\mathcal{T}}^{-\frac{q_2}{2}}|\Delta|_2 + 4R_{q_2}\underline{\mathcal{T}}^{1-q_2},$$

where we have used the bound in (3.31) from the proof of Lemma A.1. Therefore, for any vector $\Delta \in \mathbb{C}(S_{\underline{\mathcal{T}}}; q_2, 3)$ and the choice $\underline{\mathcal{T}} = \frac{\lambda_{n,3}}{\kappa_L}$, substituting the upper bound $4\sqrt{R_{q_2}}\underline{\mathcal{T}}^{-\frac{q_2}{2}}|\Delta|_2 + 4R_{q_2}\underline{\mathcal{T}}^{1-q_2}$ on $|\Delta|_1$ into condition (3.34) from Lemma A.3 and condition (3.38) from Lemma A.5 respectively yields

$$\begin{aligned}\left|\Delta^T \frac{\hat{v}^T\hat{v}}{n}\Delta\right| \;\geq\; & |\Delta|_2^2\left\{c\kappa_L - b_0 R_{q_2}\underline{\mathcal{T}}^{-q_2}\left(\frac{\log p}{n} + \mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_3\right)\right\} \\ & - b_1 R_{q_2}^2 \underline{\mathcal{T}}^{2-2q_2}\left(\frac{\log p}{n} + \mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_3\right),\end{aligned}$$

for some positive constants b_0 and b_1. With the choice of

$$\delta^* = O\left(\max\left\{\mathcal{T}_1^{\frac{3}{2}-q_2},\, \mathcal{T}_2^{\frac{3}{2}-q_2},\, \mathcal{T}_3^{\frac{3}{2}-q_2},\, \mathcal{T}_4^{2-q_2}\right\} R_{q_2}\right),$$

if

$$R_{q_2}\underline{\mathcal{T}}^{-q_2}\left(\frac{\log p}{n} + \mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_3\right) = O(\kappa_L),$$

we have

$$\left|\Delta^T \frac{\hat{v}^T\hat{v}}{n}\Delta\right| \geq c^{'}\kappa_L\left\{|\Delta|_2^2 - \frac{|\Delta|_2^2}{2}\right\} = c^{''}\kappa_L|\Delta|_2^2$$

for any Δ such that $|\Delta|_2 \geq \delta^*$. $\square$

3.8.2 Theorem 3.4.1

Lemma A.1 implies that the l_2-consistency of $\hat{\beta}_{HSEL}$ requires verifications of two conditions: (i) $\hat{\Gamma} = \hat{v}^T\hat{v}$ satisfies the RE condition (2.1) with $\gamma = 3$, and (ii) the term $|\frac{1}{n}\hat{v}^T e|_\infty \precsim \phi(k_1,\, k_2,\, d,\, p,\, n)$ with high probability. This is done via Lemmas A.3 and A.4. For the proofs for Lemmas A.3 and A.4, let the *local complexity* measure $\mathcal{G}_n\left(r_j;\, \mathcal{F}_j\right)$ be defined over the set

$$\Omega(r_j;\, \mathcal{F}_j) = \left\{f \;:\; f \in \bar{\mathcal{F}}_j \;|f_{\theta^*}|_n \leq r_j\right\}$$

where $|f_{\theta^*}|_n := \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[f(w_i^T\theta^*)\right]^2}$ and $j = 0, ..., p$. Recall the notations $\mathbb{E}\left(x_{ij}|w_i^T\theta\right) := m_j(w_i^T\theta)$, $\hat{\mathbb{E}}\left(x_{ij}|w_i^T\theta\right) := \hat{m}_j(w_i^T\theta)$, $\mathbb{E}\left(y_{2i}|w_i^T\theta\right) := m_0(w_i^T\theta)$, $\hat{\mathbb{E}}\left(y_{2i}|w_i^T\theta\right) := \hat{m}_0(w_i^T\theta)$. To avoid clustering, write $\left\{\hat{m}_j(w_i^T\theta)\right\}_{i=1}^n := \hat{m}_j(\theta)$, $\left\{\tilde{m}_j(w_i^T\theta)\right\}_{i=1}^n := \tilde{m}_j(\theta)$, and $\left\{m_j(w_i^T\theta)\right\}_{i=1}^n := m_j(\theta)$; these definitions are somewhat abuse of notation but keep in mind that the individual component $\hat{m}_j(\cdot)$ of $\hat{m}_j(\theta)$, $\tilde{m}_j(\cdot)$ of $\tilde{m}_j(\theta)$, and $m_j(\cdot)$ of $m_j(\theta)$ are functions from $\mathbb{R}$ to $\mathbb{R}$. Also recall $T_j^{'} := T_j^* \vee T_j$,

$$\begin{aligned} T_j^* &:= \sup_{f\in\mathcal{F}_j^*}\frac{1}{n}\sum_{i=1}^{n}\left[f(w_i^T\hat{\theta}) - f(w_i^T\theta^*)\right]^2, \\ T_j &:= \sup_{f\in\mathcal{F}_j}\frac{1}{n}\sum_{i=1}^{n}\left[f(w_i^T\hat{\theta}) - f(w_i^T\theta^*)\right]^2, \end{aligned}$$

and the following definitions for Theorem 3.4.1:

$$\begin{aligned} \mathcal{T}_1 &= \max_{j\in\{0,...,p\}}\left(T_j^{'} \vee \sqrt{T_j^{'}}\right) \\ \mathcal{T}_2 &= \max_{j\in\{0,...,p\}} t_{nj}^2 \\ \mathcal{T}_3 &= \max_{j\in\{0,...,p\}}\inf_{\tilde{m}_j\in\mathcal{F}_j}\left(\left|\tilde{m}_j(\hat{\theta}) - m_j(\hat{\theta})\right|_n^2 + |\tilde{m}_j(\theta^*) - m_j(\theta^*)|_n\right) \\ \mathcal{T}_4 &= \sigma_v\sigma_\eta\sqrt{\frac{\log p}{n}}. \end{aligned}$$

To facilitate the presentation, the proofs for Lemmas A.3 and A.4 work through the case $t_{nj} = r_{nj}$. Inspecting these proofs suggests that with minor notation changes, the case where $t_{nj} \geq r_{nj}$ can be analyzed with almost the same argument because for any star-shaped function class $\bar{\mathcal{F}}_j := \left\{f = f^{'} - f^{''} \,:\, f^{'},\, f^{''} \in \mathcal{F}_j\right\}$, the function $t \mapsto \frac{\mathcal{G}_n(t;\mathcal{F}_j)}{t}$ is non-decreasing on $(0, \infty)$.

Lemma A.3 (RE condition): Define $\hat{\Delta} = \hat{\beta}_{HSEL} - \beta^*$. Let $r_{nj} > 0$ be the smallest positive quantity satisfying the *critical inequality*

$$\mathcal{G}_n\left(r_{nj};\, \mathcal{F}_j\right) \leq \frac{r_{nj}^2}{\sigma_{v_j}}.$$

Under Assumptions 3.2.1, 3.2.3, 3.4.1-3.4.3, we have

$$\left|\Delta^T\frac{\hat{v}^T\hat{v}}{n}\Delta\right| \geq \kappa_1|\Delta|_2^2 - \kappa_2\frac{\log p}{n}|\Delta|_1^2 - c\left(\mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_3\right)|\Delta|_1^2 \tag{3.34}$$

with probability at least $1 - c^{'} \sum_{j=1}^{p} \exp\left(-nC_j^* r_{nj}^2\right)$, $\kappa_1 = \frac{\kappa_L}{2}$, $\kappa_2 = c_0 \kappa_L \max\left\{\frac{\sigma_v^4}{\kappa_L^2},\, 1\right\}$, $\sigma_v = \max_j \sigma_{v_j}$.

Proof. We have

$$\left|\Delta^T \frac{\hat{v}^T \hat{v}}{n} \Delta\right| + \left|\Delta^T \left(\frac{v^T v - \hat{v}^T \hat{v}}{n}\right) \Delta\right| \geq \left|\Delta^T \frac{v^T v}{n} \Delta\right|,$$

which implies

$$\begin{aligned}
\left|\Delta^T \frac{\hat{v}^T \hat{v}}{n} \Delta\right| &\geq \left|\Delta^T \frac{v^T v}{n} \Delta\right| - \left|\Delta^T \left(\frac{v^T v - \hat{v}^T \hat{v}}{n}\right) \Delta\right| \\
&\geq \left|\Delta^T \frac{v^T v}{n} \Delta\right| - \left|\frac{v^T v - \hat{v}^T \hat{v}}{n}\right|_\infty |\Delta|_1^2 \\
&\geq \left|\Delta^T \frac{v^T v}{n} \Delta\right| - \left(\left|\frac{v^T (\hat{v} - v)}{n}\right|_\infty + \left|\frac{(\hat{v} - v)^T \hat{v}}{n}\right|_\infty\right) |\Delta|_1^2 \\
&\geq \left|\Delta^T \frac{v^T v}{n} \Delta\right| - \left|\frac{v^T (\hat{v} - v)}{n}\right|_\infty |\Delta|_1^2 \\
&\quad - \left|\frac{(\hat{v} - v)^T v}{n}\right|_\infty |\Delta|_1^2 - \left|\frac{(\hat{v} - v)^T (\hat{v} - v)}{n}\right|_\infty |\Delta|_1^2 .
\end{aligned}$$

To bound the term $\left|\frac{v^T(\hat{v}-v)}{n}\right|_\infty$, let us first fix $(j,\, j^{'})$ and bound the $(j,\, j^{'})$ element of the matrix $\frac{v^T(\hat{v}-v)}{n}$. Notice that

$$\begin{aligned}
\left|\frac{1}{n} v_j^T (\hat{v}_{j'} - v_{j'})\right| &= \left|\frac{1}{n} \sum_{i=1}^{n} v_{ij} (\hat{v}_{ij'} - v_{ij'})\right| \\
&= \left|\frac{1}{n} \sum_{i=1}^{n} v_{ij} \left(\hat{m}_{j'}(w_i^T \hat{\theta}) - m_{j'}(w_i^T \hat{\theta}) + m_{j'}(w_i^T \hat{\theta}) - m_{j'}(w_i^T \theta^*)\right)\right| \\
&\leq \left|\frac{1}{n} \sum_{i=1}^{n} v_{ij} \left[\hat{m}_{j'}(w_i^T \hat{\theta}) - m_{j'}(w_i^T \hat{\theta})\right]\right| + \sqrt{\frac{1}{n} \sum_{i=1}^{n} v_{ij}^2} \sqrt{T_{j'}^*},
\end{aligned}$$

where $T_{j'}^* := \sup_{f \in \mathcal{F}_{j'}^*} \frac{1}{n} \sum_{i=1}^{n} \left[f(w_i^T \hat{\theta}) - f(w_i^T \theta^*)\right]^2$.

For the term $\sqrt{\frac{1}{n} \sum_{i=1}^{n} v_{ij}^2}$, by Lemma B.1, we have

$$\mathbb{P}\left[\max_{j=1,\ldots,p} \left\{\frac{1}{n} \sum_{i=1}^{n} v_{ij}^2\right\} \geq \sigma_{v_j}^2 + t\right] \leq 2\exp(-c\min(\frac{nt^2}{4\sigma_{v_j}^4},\, \frac{nt}{2\sigma_{v_j}^2}) + \log p).$$

Hence, setting $t = \sigma_v^2 := \max_j \sigma_{v_j}^2$ and under the condition $n \gtrsim \log p$, we have

$$\mathbb{P}\left[\max_{j'}\left\{\frac{1}{n}\sum_{i=1}^n v_{ij}^2\right\} \geq 2\sigma_v^2\right] \leq 2\exp(-c' n).$$

Let $\hat{\Delta}_{ij'}(\theta) := \hat{m}_{j'}(w_i^T\theta) - m_{j'}(w_i^T\theta)$, $\bar{\Delta}_{ij'}(\theta) := \hat{m}_{j'}(w_i^T\theta) - \tilde{m}_{j'}(w_i^T\theta)$, and $\tilde{\Delta}_{ij'}(\theta) := \tilde{m}_{j'}(w_i^T\theta) - m_{j'}(w_i^T\theta)$. Note that $\hat{m}_{j'} \in \mathcal{F}_{j'}^*$ and we choose $\hat{m}_{j'}$, $\tilde{m}_{j'} \in \mathcal{F}_{j'}$. For the term

$$\left|\frac{1}{n}\sum_{i=1}^n v_{ij}\left[\hat{m}_{j'}(w_i^T\hat{\theta}) - m_{j'}(w_i^T\hat{\theta})\right]\right|,$$

we have

$$\begin{aligned}
\left|\frac{1}{n}\sum_{i=1}^n v_{ij}\hat{\Delta}_{ij'}(\hat{\theta})\right| &\leq \left|\frac{1}{n}\sum_{i=1}^n v_{ij}\hat{\Delta}_{ij'}(\theta^*)\right| + \sqrt{\frac{1}{n}\sum_{i=1}^n v_{ij}^2}\sqrt{\frac{1}{n}\sum_{i=1}^n\left[\hat{\Delta}_{ij'}(\hat{\theta}) - \hat{\Delta}_{ij'}(\theta^*)\right]^2} \\
&\leq \left|\frac{1}{n}\sum_{i=1}^n v_{ij}\bar{\Delta}_{ij'}(\theta^*)\right| + \left|\frac{1}{n}\sum_{i=1}^n v_{ij}\tilde{\Delta}_{ij'}(\theta^*)\right| + 2\sqrt{\frac{1}{n}\sum_{i=1}^n v_{ij}^2}\sqrt{T_{j'}^* + T_{j'}} \\
&\leq \left|\frac{1}{n}\sum_{i=1}^n v_{ij}\bar{\Delta}_{ij'}(\theta^*)\right| + \sqrt{\frac{1}{n}\sum_{i=1}^n v_{ij}^2}\left|\tilde{m}_{j'}(\theta^*) - m_{j'}(\theta^*)\right|_n \\
&\quad + \sqrt{\frac{1}{n}\sum_{i=1}^n v_{ij}^2}\sqrt{8T_{j'}'}.
\end{aligned}$$

To upper bound the term $\frac{1}{n}\sum_{i=1}^n v_{ij}\bar{\Delta}_{ij'}(\theta^*)$, we argue similarly as in Lemma A.11. First, note that by Lemma A.11 and the triangle inequality, we have

$$\begin{aligned}
\left|\bar{\Delta}_{j'}(\theta^*)\right|_n &\leq \left|\hat{m}_{j'}(\hat{\theta}) - m_{j'}(\hat{\theta})\right|_n + \left|\tilde{m}_{j'}(\hat{\theta}) - m_{j'}(\hat{\theta})\right|_n + 2\sqrt{T_{j'}} \\
&\leq c'\left\{\sqrt{T_{j'}'} + \left(\sigma_v^2 T_{j'}'\right)^{\frac{1}{4}} + \left|\tilde{m}_{j'}(\hat{\theta}) - m_{j'}(\hat{\theta})\right|_n + r_{nj'}\right\} := \bar{r}_{nj'}.
\end{aligned}$$

Setting $u = \bar{r}_{nj'}$ in $\Omega(u; \mathcal{F}_j)$ from Lemma B.3 and following the argument in the proof for Lemma B.3, we obtain

$$\max_{j'}\frac{1}{n}\sum_{i=1}^n v_{ij}\bar{\Delta}_{ij'}(\theta^*) \leq 2\max_{j'}\bar{r}_{nj'}^2$$

with probability at least $1 - c\sum_{j=1}^p \exp\left(-nC_j^*\bar{r}_{nj}^2\right)$. Hence, we obtain

$$\max_{j,j'}\left|\frac{1}{n}\sum_{i=1}^n v_{ij}\left[\hat{m}_{j'}(w_i^T\hat{\theta}) - m_{j'}(w_i^T\hat{\theta})\right]\right|$$

$$\leq c \max_{j'} \left\{ \bar{r}_{nj'}^2 + \sigma_v \left| \tilde{m}_{j'}(\theta^*) - m_{j'}(\theta^*) \right|_n + \sqrt{\sigma_v^2 T_{j'}^{'}} \right\}.$$

Consequently, we have

$$\left| \frac{v^T(\hat{v} - v)}{n} \right|_\infty \leq c \max_j \left\{ \bar{r}_{nj}^2 + \sigma_v \left| \tilde{m}_{j'}(\theta^*) - m_{j'}(\theta^*) \right|_n + \sqrt{\sigma_v^2 T_j^{'}} \right\}.$$

To bound the term $\left| \frac{(\hat{v}-v)^T(\hat{v}-v)}{n} \right|_\infty$, note that we have

$$\left| \frac{(\hat{v} - v)^T(\hat{v} - v)}{n} \right|_\infty$$

$$\begin{aligned}
\leq \quad & 2 \max_{j,j'} \sqrt{\frac{1}{n} \sum_{i=1}^n \left[\hat{m}_j(w_i^T \hat{\theta}) - m_j(w_i^T \hat{\theta}) \right]^2} \cdot \sqrt{\frac{1}{n} \sum_{i=1}^n \left[\hat{m}_{j'}(w_i^T \hat{\theta}) - m_{j'}(w_i^T \hat{\theta}) \right]^2} \\
& + 2 \max_{j,j'} \sqrt{\frac{1}{n} \sum_{i=1}^n \left[m_j(w_i^T \hat{\theta}) - m_j(w_i^T \theta^*) \right]^2} \cdot \sqrt{\frac{1}{n} \sum_{i=1}^n \left[m_{j'}(w_i^T \hat{\theta}) - m_{j'}(w_i^T \theta^*) \right]^2} \\
\leq \quad & c \max_j \left[T_j^{'} + \sqrt{\sigma_v^2 T_j^{'}} + \left| \tilde{m}_j(\hat{\theta}) - m_j(\hat{\theta}) \right|_n^2 + r_{nj}^2 \right]
\end{aligned}$$

with probability at least $1 - c^{'} \sum_{j=1}^p \exp\left(-nC_j^* r_{nj}^2\right)$.

Putting everything together and applying Lemma B.2, we have

$$\left| \Delta^T \frac{\hat{v}^T \hat{v}}{n} \Delta \right| \geq \kappa_1 |\Delta|_2^2 - \kappa_2 \frac{\log p}{n} |\Delta|_1^2 -$$

$$b(\sigma_v) \max_j \left\{ r_{nj}^2 + \inf_{\tilde{m}_j \in \mathcal{F}_j} \left(\left| \tilde{m}_j(\hat{\theta}) - m_j(\hat{\theta}) \right|_n^2 + |\tilde{m}_j(\theta^*) - m_j(\theta^*)|_n \right) + \sqrt{T_j^{'} \vee T_j^{'2}} \right\} |\Delta|_1^2$$

where $\kappa_1 = \frac{\kappa_L}{2}$, $\kappa_2 = c_0 \kappa_L \max\left\{ \frac{\sigma_v^4}{\kappa_L^2}, 1 \right\}$, $\sigma_v = \max_j \sigma_{v_j}$, and a known function $b(\sigma_v)$ depending only on σ_v.□

Lemma A.4 (Upper bound on $\left| \frac{1}{n} \hat{v}^T e \right|_\infty$): Let $r_{nj} > 0$ be the smallest positive quantity satisfying the *critical inequality*

$$\mathcal{G}_n\left(r_{nj}; \mathcal{F}_j\right) \leq \frac{r_{nj}^2}{\sigma_{v_j}}.$$

Under Assumptions 3.2.1 and 3.4.1-3.4.3, we have

$$\left|\frac{\hat{v}^T e}{n}\right|_\infty \leq |\beta^*|_1 \, b(\sigma_v,\, \sigma_\eta) \left(\mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_3\right) + \mathcal{T}_4,$$

with probability at least $1 - c_1 \exp\left(-c_2 \log p\right) - c_3 \sum_{j=0}^{p} \exp\left(-nC_j^* r_{nj}^2\right)$.

Proof. Recall that

$$\begin{aligned} e &= \left[\hat{\mathbb{E}}\left(x|w\hat{\theta}\right) - \mathbb{E}\left(x|w\hat{\theta}\right) + \mathbb{E}\left(x|w\hat{\theta}\right) - \mathbb{E}\left(x|w\theta^*\right)\right]\beta^* \\ &- \left[\hat{\mathbb{E}}\left(y_2|w\hat{\theta}\right) - \mathbb{E}\left(y_2|w\hat{\theta}\right) + \mathbb{E}\left(y_2|w\hat{\theta}\right) - \mathbb{E}\left(y_2|w\theta^*\right)\right] + \eta. \end{aligned}$$

and

$$\begin{aligned} \hat{v} &= x - \hat{\mathbb{E}}\left(x|w\hat{\theta}\right) \\ &= x - \mathbb{E}\left(x|w\theta^*\right) + \mathbb{E}\left(x|w\theta^*\right) - \mathbb{E}\left(x|w\hat{\theta}\right) + \mathbb{E}\left(x|w\hat{\theta}\right) - \hat{\mathbb{E}}\left(x|w\hat{\theta}\right). \end{aligned}$$

Recall $v = x - \mathbb{E}\left(x|w\theta^*\right)$. Let us introduce the following notations

$$\begin{aligned} m_j(w_i^T\hat{\theta}) - m_j(w_i^T\theta^*) &:= T_{ij}^{''}, \\ \hat{m}_j(w_i^T\theta) - m_j(w_i^T\theta) &:= \hat{\Delta}_j(w_i^T\theta), \end{aligned}$$

and

$$T_1 = \max_{j,\,j^{'}=1,\ldots,p}\left|\frac{1}{n}\sum_{i=1}^{n}\left[v_{ij} - T_{ij}^{''} - \hat{\Delta}_j(w_i^T\hat{\theta})\right]\left[T_{ij'}^{''} + \hat{\Delta}_{j'}(w_i^T\hat{\theta})\right]\right| |\beta^*|_1 \quad (3.35)$$

$$T_2 = \max_{j=1,\ldots,p}\left|\frac{1}{n}\sum_{i=1}^{n}\left[v_{ij} - T_{ij}^{''} - \hat{\Delta}_j(w_i^T\hat{\theta})\right]\left[T_{i0}^{''} + \hat{\Delta}_0(w_i^T\hat{\theta})\right]\right| \quad (3.36)$$

$$T_3 = \max_{j=1,\ldots,p}\left|\frac{1}{n}\sum_{i=1}^{n}\left[v_{ij} - T_{ij}^{''} - \hat{\Delta}_j(w_i^T\hat{\theta})\right]\eta_i\right|. \quad (3.37)$$

Expanding the products in (3.35)-(3.37), applying the Cauchy-Schwarz inequality on some of the terms in the expansion, we obtain the following inequalities

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} v_{ij}T^{''}_{ij'} &\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n} v_{ij}^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n} T^{''2}_{ij'}} \\
\frac{1}{n}\sum_{i=1}^{n} v_{ij}\hat{\Delta}_{ij'}(\hat{\theta}) &\leq \left|\frac{1}{n}\sum_{i=1}^{n} v_{ij}\hat{\Delta}_{j'}(w_i^T\theta^*)\right| \\
&\quad + \sqrt{\frac{1}{n}\sum_{i=1}^{n} v_{ij}^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\Delta}_{j'}(w_i^T\hat{\theta}) - \hat{\Delta}_{j'}(w_i^T\theta^*)\right]^2} \\
\frac{1}{n}\sum_{i=1}^{n} T^{''}_{ij}T^{''}_{ij'} &\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n} T^{''2}_{ij}}\sqrt{\frac{1}{n}\sum_{i=1}^{n} T^{''2}_{ij'}} \\
\frac{1}{n}\sum_{i=1}^{n} T^{''}_{ij}\hat{\Delta}_{ij'}(\hat{\theta}) &\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n} T^{''2}_{ij}}\sqrt{\frac{1}{n}\sum_{i=1}^{n} \hat{\Delta}^2_{j'}(w_i^T\hat{\theta})} \\
\frac{1}{n}\sum_{i=1}^{n} \hat{\Delta}_{ij}(\hat{\theta})\hat{\Delta}_{ij'}(\hat{\theta}) &\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n} \hat{\Delta}^2_{j}(w_i^T\hat{\theta})}\sqrt{\frac{1}{n}\sum_{i=1}^{n} \hat{\Delta}^2_{j'}(w_i^T\hat{\theta})} \\
\frac{1}{n}\sum_{i=1}^{n} T^{''}_{ij}\eta_i &\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n} \eta_i^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n} T^{''2}_{ij'}} \\
\frac{1}{n}\sum_{i=1}^{n} \hat{\Delta}_{ij}(\hat{\theta})\eta_i &\leq \left|\frac{1}{n}\sum_{i=1}^{n} \eta_i\hat{\Delta}_{j'}(w_i^T\theta^*)\right| \\
&\quad + \sqrt{\frac{1}{n}\sum_{i=1}^{n} \eta_i^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\Delta}_{j'}(w_i^T\hat{\theta}) - \hat{\Delta}_{j'}(w_i^T\theta^*)\right]^2}.
\end{aligned}$$

To upper bound $\left|\frac{\hat{v}^T e}{n}\right|_\infty$, we need to control the RHS inequalities listed above as well as the additional term $\max_j\left\{\left|\frac{1}{n}\sum_{i=1}^{n} v_{ij}\eta_i\right|\right\}$.

Under Assumptions 3.4.2 and 3.4.3, and the condition $n \gtrsim \log p$, applying Lemma B.1 yields

$$\begin{aligned}
\mathbb{P}\left[\max_j\left\{\frac{1}{n}\sum_{i=1}^{n} v_{ij}^2\right\} \geq 2\sigma_v^2\right] &\leq 2\exp(-c^{'}n), \\
\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} \eta_i^2 \geq 2\sigma_\eta^2\right] &\leq 2\exp(-c^{''}n), \\
\mathbb{P}\left[\max_j\left\{\left|\frac{1}{n}\sum_{i=1}^{n} v_{ij}\eta_i\right|\right\} \geq c_0\sigma_v\sigma_\eta\sqrt{\frac{\log p}{n}}\right] &\leq 2\exp(-c^{'''}\log p).
\end{aligned}$$

An upper bound on the term $\sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{\Delta}_j^2(w_i^T\hat{\theta})}$ is derived in Lemma A.11, which yields

$$\max_j\left\{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{\Delta}_j^2(w_i^T\hat{\theta})}\right\}\leq c\max_j\left\{\sqrt{T_j^{'}}+\left(\sigma_v^2T_j^{'}\right)^{\frac{1}{4}}+\left|\tilde{m}_j(\hat{\theta})-m_j(\hat{\theta})\right|_n+r_{nj}\right\},$$

with probability at least $1-c^{'}\sum_{j=0}^{p}\exp\left(-nC_j^*r_{nj}^2\right)$. An upper bound on the term $\sqrt{\frac{1}{n}\sum_{i=1}^{n}T_{ij}^{''2}}$ is simply $\max_j\sqrt{T_j^{'}}$ by recalling the definition of $T_j^{'}$.

To upper bound the term $\left|\frac{1}{n}\sum_{i=1}^{n}v_{ij}\hat{\Delta}_{j^{'}}(w_i^T\theta^*)\right|$, recall from the proof for Lemma A.3, we have

$$\max_{j,\,j^{'}}\left|\frac{1}{n}\sum_{i=1}^{n}v_{ij}\hat{\Delta}_{j^{'}}(w_i^T\theta^*)\right|\leq c\max_{j^{'}}\left\{\bar{r}_{nj^{'}}^2+\sigma_v\left|\tilde{m}_{j^{'}}(\theta^*)-m_{j^{'}}(\theta^*)\right|_n+\sqrt{\sigma_v^2T_{j^{'}}^{'}}\right\}$$

with probability at least $1-c^{'}\sum_{j=1}^{p}\exp\left(-nC_j^*\bar{r}_{nj}^2\right)$. Applying the same argument from above to control the term $\left|\frac{1}{n}\sum_{i=1}^{n}\eta_i\hat{\Delta}_{ij^{'}}(\theta^*)\right|$ yields

$$\max_{j^{'}}\left|\frac{1}{n}\sum_{i=1}^{n}\eta_i\hat{\Delta}_{j^{'}}(w_i^T\theta^*)\right|\leq c^{'}\max_{j^{'}}\left\{\bar{r}_{nj}^2+\sigma_\eta\left|\tilde{m}_{j^{'}}(\theta^*)-m_{j^{'}}(\theta^*)\right|_n+\sqrt{\sigma_\eta^2T_j^{'}}\right\}$$

with probability at least $1-c^{''}\sum_{j=1}^{p}\exp\left(-nC_j^*\bar{r}_{nj}^2\right)$.

Putting all the pieces together, we have

$$\left|\frac{\hat{v}^Te}{n}\right|_\infty\leq|\beta^*|_1\,b(\sigma_v,\,\sigma_\eta)\left(\mathcal{T}_1+\mathcal{T}_2+\mathcal{T}_3\right)+\mathcal{T}_4$$

with probability at least $1-c_1\exp\left(-c_2\log p\right)-c_3\sum_{j=1}^{p}\exp\left(-nC_j^*r_{nj}^2\right)$. $\square$

Now, by applying Lemma A.1, setting $\lambda_{n,3}\geq|\beta^*|_1\,b(\sigma_v,\,\sigma_\eta)\left(\mathcal{T}_1+\mathcal{T}_2+\mathcal{T}_3\right)+\mathcal{T}_4$ and combining with Lemma A.3, we obtain

$$|\hat{\beta}_{HSEL}-\beta^*|_2\leq\frac{c^{''}R_{q_2}^{\frac{1}{2}}}{\kappa_L^{1-\frac{q_2}{2}}}\left[|\beta^*|_1\,b(\sigma_v,\,\sigma_\eta)\left(\mathcal{T}_1+\mathcal{T}_2+\mathcal{T}_3\right)+\mathcal{T}_4\right]^{1-\frac{q_2}{2}},\quad q_2\in[0,\,1]$$

with probability at least $1-c_1\exp\left(-c_2\log p\right)-c_3\sum_{j=1}^{p}\exp\left(-nC_j^*r_{nj}^2\right)$.

3.8.3 Theorem 3.4.2

As in the proofs for Theorem 3.4.1, the proofs for Lemmas A.5 and A.6 work through the case $t_{nj} = r_{nj}$. The case where $t_{nj} \geq r_{nj}$ can be analyzed with almost the same argument because for any star-shaped function class $\bar{\mathcal{F}}_j := \left\{ f = f' - f'' \,:\, f',\, f'' \in \mathcal{F}_j \right\}$, the function $t \mapsto \frac{\mathcal{G}_n(t;\,\mathcal{F}_j)}{t}$ is non-decreasing on $(0,\,\infty)$. For the proofs for Lemmas A.5 and A.6, let the *local complexity* measure $\mathcal{G}_n\left(r_j;\,\mathcal{F}_j\right)$ be defined over the set

$$\Omega(r_j;\,\mathcal{F}_j) = \left\{ f \,:\, f \in \bar{\mathcal{F}}_j \ \left|f_{\theta^*}\right|_n \leq r_j,\, |f|_{\mathcal{F}_j} \leq 1 \right\}$$

where $\left|f_{\theta^*}\right|_n := \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[f(w_i^T\theta^*)\right]^2}$ and $j = 0, ..., p$. Recall the following definitions for Theorem 3.4.2:

$$\begin{aligned}
\mathcal{T}_1 &= \max_{j\in\{0,\ldots,p\}} \left(T_j^{'} \vee \sqrt{T_j^{'}}\right) \\
\mathcal{T}_2 &= \max_{j\in\{0,\ldots,p\}} \bar{R}_j^2 t_{nj}^2 \\
\mathcal{T}_3 &= \max_{j\in\{0,\ldots,p\}} \inf_{\tilde{m}_j\in\mathcal{F}_j,|\tilde{m}_j|_{\mathcal{F}_j}\leq\bar{R}_j} \left(\left|\tilde{m}_j(\hat{\theta}) - m_j(\hat{\theta})\right|_n^2 + \left|\tilde{m}_j(\theta^*) - m_j(\theta^*)\right|_n\right) \\
\mathcal{T}_4 &= \sigma_v\sigma_\eta\sqrt{\frac{\log p}{n}}.
\end{aligned}$$

Lemma A.5 (RE condition): Let $r_{nj} > 0$ be the smallest positive quantity satisfying the *critical inequality*

$$\mathcal{G}_n\left(r_{nj};\,\mathcal{F}_j\right) \leq \frac{\bar{R}_j r_{nj}^2}{\sigma_{v_j}},$$

where $\bar{R}_j > 0$ is a user-defined radius. Under Assumptions 3.2.1, 3.2.3, 3.4.1-3.4.3, if we choose the second-stage regularization parameter $\lambda_{nj,2} = (2+\varsigma)r_{nj}^2$ for $j = 1, ..., p$ where ς is a small positive constant, then

$$\left|\Delta^T\frac{\hat{v}^T\hat{v}}{n}\Delta\right| \geq \kappa_1|\Delta|_2^2 - \kappa_2\frac{\log p}{n}|\Delta|_1^2 - c\left(\mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_3\right)|\Delta|_1^2 \tag{3.38}$$

with probability at least $1 - c'\sum_{j=1}^{p}\exp\left(-nC_j^*\bar{R}_j^2 r_{nj}^2\right)$

Proof. The proof is almost identical to that of Lemma A.3 except that in upper bounding the term $\frac{1}{n}\sum_{i=1}^{n}\left[\hat{m}_j(w_i^T\hat{\theta}) - m_j(w_i^T\hat{\theta})\right]^2$, we apply Lemma A.12 with the second-stage regularization parameter $\lambda_{nj,2} = (2+\varsigma)r_{nj}^2$ where ς is a small positive constant for $j = 1, ..., p$, and obtain

$$\frac{1}{n}\sum_{i=1}^{n}\left[\hat{m}_j(w_i^T\hat{\theta}) - m_j(w_i^T\hat{\theta})\right]^2 \leq c_0\left|\tilde{m}_j(\hat{\theta}) - m_j(\hat{\theta})\right|_n^2 + c_1\bar{R}_j^2 r_{nj}^2 + c_2 T_j^{'} + c_3\sqrt{\sigma_v^2 T_j^{'}}$$

with probability at least $1 - c\exp\left(-nC_j^*\bar{R}_j^2 r_{nj}^2\right)$. The second-stage regularization parameter $\lambda_{nj,2}$ is chosen such that $\lambda_{nj,2} \geq 2r_{nj}^2$ for $j = 1, ..., p$. Using similar argument as in the proof for Lemma A.3 yields

$$\left|\frac{v^T(\hat{v}-v)}{n}\right|_\infty \leq$$

$$c\max_j\left(\inf_{\tilde{m}_j\in\mathcal{F}_j, |\tilde{m}_j|_{\mathcal{F}_j}\leq\bar{R}_j}\left(\left|\tilde{m}_j(\hat{\theta}) - m_j(\hat{\theta})\right|_n^2 + |\tilde{m}_j(\theta^*) - m_j(\theta^*)|_n\right)\right.$$

$$\left. + \bar{R}_j^2 r_{nj}^2 + T_j^{'} + \sqrt{\left(\sigma_v^2 T_j^{'}\right)}\right)$$

with probability at least $1 - c_1\exp(-c_2\log p) - c_3\sum_{j=1}^p \exp\left(-nC_j^*\bar{R}_j^2 r_{nj}^2\right)$.

For the term $\left|\frac{(\hat{v}-v)^T(\hat{v}-v)}{n}\right|_\infty$, we have

$$\left|\frac{(\hat{v}-v)^T(\hat{v}-v)}{n}\right|_\infty \leq c^{'}\max_j\left[\left|\tilde{m}_j(\hat{\theta}) - m_j(\hat{\theta})\right|_n^2 + \bar{R}_j^2 r_{nj}^2 + T_j^{'} + \sqrt{\sigma_v^2 T_j^{'}}\right]$$

with probability at least $1 - c\sum_{j=1}^p \exp\left(-nC_j^*\bar{R}_j^2 r_{nj}^2\right)$.

Putting everything together and applying Lemma B.2, we have

$$\left|\Delta^T\frac{\hat{v}^T\hat{v}}{n}\Delta\right| \geq \kappa_1|\Delta|_2^2 - \kappa_2\frac{\log p}{n}|\Delta|_1^2 -$$

$$b(\sigma_v)\max_j\left\{\bar{R}_j^2 r_{nj}^2 + \inf_{\tilde{m}_j\in\mathcal{F}_j}\left(\left|\tilde{m}_j(\hat{\theta}) - m_j(\hat{\theta})\right|_n^2 + |\tilde{m}_j(\theta^*) - m_j(\theta^*)|_n\right) + \sqrt{T_j^{'}\vee T_j^{'2}}\right\}|\Delta|_1^2$$

where $\kappa_1 = \frac{\kappa_L}{2}$, $\kappa_2 = c_0\kappa_L\max\left\{\frac{\sigma_v^4}{\kappa_L^2}, 1\right\}$, and $\sigma_v = \max_j\sigma_{v_j}$, and a known function $b(\sigma_v)$ depending only on σ_v. $\square$

Lemma A.6 (Upper bound on $\left|\frac{1}{n}\hat{v}^T e\right|_\infty$): Let $r_{nj} > 0$ be the smallest positive quantity satisfying the *critical inequality*

$$\mathcal{G}_n\left(r_{nj};\,\mathcal{F}_j\right) \le \frac{\bar{R}_j r_{nj}^2}{\sigma_{v_j}},$$

and $\bar{R}_j > 0$ is a user-defined radius. Under Assumptions 3.2.1 and 3.4.1-3.4.3 and if we choose the second-stage regularization parameter $\lambda_{nj,2} = (2+\varsigma)r_{nj}^2$ where ς is a small positive constant for $j = 1, ..., p$, we have

$$\left|\frac{\hat{v}^T e}{n}\right|_\infty \le |\beta^*|_1\, b(\sigma_v,\,\sigma_\eta)\left(\mathcal{T}_1+\mathcal{T}_2+\mathcal{T}_3\right)+\mathcal{T}_4,$$

with probability at least $1 - c_1 \exp\left(-c_2 \log p\right) - c_3 \sum_{j=1}^{p} \exp\left(-nC_j^* \bar{R}_j^2 r_{nj}^2\right)$.

Proof. The proof is almost identical to that of Lemma A.4 except that in upper bounding the term $\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\hat{m}_j(w_i^T\hat{\theta}) - m_j(w_i^T\hat{\theta})\right]^2}$, we apply Lemma A.12 just as what we have done in the proof for Lemma A.5. □

Consequently, by applying Lemma A.1, setting $\lambda_{n,3} \ge |\beta^*|_1\, b(\sigma_v,\,\sigma_\eta)\left(\mathcal{T}_1+\mathcal{T}_2+\mathcal{T}_3\right)+\mathcal{T}_4$, and combining with Lemma A.5, we obtain

$$|\hat{\beta}_{HSEL} - \beta^*|_2 \le \frac{c^{''} R_{q_2}^{\frac{1}{2}}}{\kappa_L^{1-\frac{q_2}{2}}}\left[|\beta^*|_1\, b(\sigma_v,\,\sigma_\eta)\left(\mathcal{T}_1+\mathcal{T}_2+\mathcal{T}_3\right)+\mathcal{T}_4\right]^{1-\frac{q_2}{2}}, \quad q_2 \in (0,\,1]$$

with probability at least $1 - c_1^{'} \exp\left(-c_2^{'} \log p\right) - c_3^{'} \sum_{j=1}^{p} \exp\left(-nC_j^* \bar{R}_j^2 r_{nj}^2\right)$.

3.8.4 Corollary 3.4.3

Proof. This follows from the previous proofs for the RE condition and Lemma A.1. □

3.8.5 Proposition 3.4.1 and Corollaries 3.4.4-3.4.6

Proof of Proposition 3.4.1. Note that under the assumption $m_j(\cdot) \in \mathcal{F}_j$ for $j = 0, ..., p$, we have $\inf \left|\tilde{m}_j(\hat{\theta}) - m_j(\hat{\theta})\right|_n^2 = 0$. When $\mathcal{F}_j$ is $L-$Lipschitz, under the Lipschitz regression (22),

$$T_j^{'} = \frac{1}{n}\sum_{i=1}^{n} L^2\left[w_i^T\hat{\theta} - w_i^T\theta^*\right]^2 := L^2 B^{'}$$

for every $j = 0, ..., p$. Now let us compute the critical radius r_{nj} that satisfies

$$\mathcal{G}_n\left(r_{nj};\, \mathcal{F}_j\right) \leq \frac{r_{nj}^2}{\sigma_{v_j}},$$

where $\mathcal{F}_j$ is the space of $L-$Lipschitz functions. The metric entropy of this class (in the sup-norm) scales as $\log N_\infty(t;\, \bar{\mathcal{F}}_j) \asymp \frac{1}{t}$ by Lemma B.11. Consequently, applying Lemma B.9 yields

$$\begin{aligned}
\frac{1}{\sqrt{n}}\int_0^{r_{nj}} \sqrt{\log N_n(t;\, \mathbb{B}_n(r_{nj};\, \mathcal{H}_j)}dt &\leq \frac{1}{\sqrt{n}}\int_0^{r_{nj}} \sqrt{\log N_\infty(t;\, \mathrm{sym}(\mathcal{H}_j)}dt \\
&\leq \frac{c}{\sqrt{n}}\int_0^{r_{nj}} \left(\frac{|\theta^*|_1}{t}\right)^{\frac{1}{2}} dt \\
&= \frac{c}{\sqrt{n}}\sqrt{|\theta^*|_1 r_{nj}}.
\end{aligned}$$

Thus, it suffices to choose $r_{nj} > 0$ such that

$$\begin{aligned}
\frac{\sqrt{r_{nj}}}{\sqrt{n}} &\asymp \frac{r_{nj}^2}{\sigma_{v_j}} \\
\Longrightarrow r_{nj}^2 &\asymp \left(\frac{|\theta^*|_1 \sigma_{v_j}^2}{n}\right)^{\frac{2}{3}}.
\end{aligned}$$

□

Proof of Corollaries 3.4.4-3.4.6. In proving the two corollaries, we follow the proof for Theorem 3.4.1. We prove the case where β^* and θ^* are exactly sparse and the approximate sparse case of β^* and θ^* follows the same argument. We have

$$\max_{j=0,\ldots,p} T_j^{'} \leq \frac{1}{n}\sum_{i=1}^n L^2\left[w_i^T\hat{\theta} - w_i^T\theta^*\right]^2 = L^2 B^{'}$$

Now, depending on the assumptions on w_i, we have

$$\begin{aligned}
B^{'} &= \frac{1}{n}\sum_{i=1}^n \left[w_i^T\hat{\theta} - w_i^T\theta^*\right]^2 \leq c\frac{\kappa_U^w \sigma_w^2 \alpha_u}{(\kappa_L^w)^2}\frac{k_1 \log d}{n} \text{ (Corollary 3.4.4)} \\
B^{'} &= \frac{1}{n}\sum_{i=1}^n \left[w_i^T\hat{\theta} - w_i^T\theta^*\right]^2 \leq c^{'}\frac{|\theta^*|_1}{\alpha_l}\sqrt{\frac{k_1 \log d}{n}} \text{ (Corollary 3.4.5)}
\end{aligned}$$

with probability at least $1-O\left(\frac{1}{d}\right)$. Putting the pieces together and applying Theorem 3.4.1, choosing $t_{nj}^2 \geq r_{nj}^2$ in $\mathcal{T}_2$ such that $|\beta^*|_1 \mathcal{T}_2$ is at most

$$O\left(\sqrt{\frac{\log p}{n}} \vee \left(|\beta^*|_1 \sqrt{\frac{k_1 \log d}{n}}\right)\right).$$

and $\lambda_{n,3}$ in Corollaries 3.4.4 and 3.4.5, if $\sqrt{\frac{n \log p}{|\beta^*|_1^2}} \vee \sqrt{n k_1 \log d} \gtrsim \log p$, and

$$\kappa_2 \frac{k_2 \log p}{n} + c k_2 \sigma_v \left[\left(\frac{k_1 \sigma_v^2}{n}\right)^{\frac{2}{3}}\right] + k_2 B^{'} + k_2 \sqrt{\sigma_v^2 B^{'}} = O(\kappa_1),$$

then, with probability at least $1-O\left(\frac{1}{d \wedge p}\right)$, the error satisfies the bound

$$\begin{aligned} |\hat{\beta}_{HSEL} - \beta^*|_2 \quad \leq \quad & \frac{c_1 \sqrt{k_2}}{\kappa_L} \left[\left(\sigma_v \sigma_\eta \sqrt{\frac{\log p}{n}}\right) \vee \left(L \sqrt{\frac{\kappa_U^w \sigma_w^2 \alpha_u}{(\kappa_L^w)^2}} b\left(\sigma_v,\, \sigma_\eta\right) |\beta^*|_1 \sqrt{\frac{k_1 \log d}{n}}\right)\right] \\ & \text{(Corollary 3.4.4)} \end{aligned}$$

and

$$\begin{aligned} |\hat{\beta}_{HSEL} - \beta^*|_2 \quad \leq \quad & \frac{c_2 \sqrt{k_2}}{\kappa_L} \left[\left(\sigma_v \sigma_\eta \sqrt{\frac{\log p}{n}}\right) \vee \left(L \sqrt{\frac{|\theta^*|_1}{\alpha_l}} b^{'}\left(\sigma_v,\, \sigma_\eta\right) |\beta^*|_1 \left(\frac{k_1 \log d}{n}\right)^{\frac{1}{4}}\right)\right] \\ & \text{(Corollary 3.4.5).} \end{aligned}$$

□

3.8.6 Proposition 3.4.2

Proof. First, by Corollary 5 in Negahban, et. al (2010) for the logit model and Lemma A.7 for the probit model, we have, for $\lambda_{n,1} \asymp c \sigma_w \sqrt{\alpha_u} \sqrt{\frac{\log d}{n}}$,

$$\left|\hat{\theta} - \theta^*\right|_2^2 \leq c_1 \frac{\sigma_w^2 \alpha_u}{\left(\kappa_L^w\right)^2} \frac{k_1 \log d}{n}$$

with probability at least $1-O\left(\frac{1}{d}\right)$. Let $\Delta = \hat{\theta} - \theta^*$. Under Assumption 3.4.6, Lemma B.2 implies

$$\frac{|w\Delta|_2^2}{n} \leq \frac{3\kappa_U^w}{2} |\Delta|_2^2 + \alpha^{'} \frac{\log d}{n} |\Delta|_1^2.$$

Combining the inequality above with Lemma 1 in Negahban, et. al which shows that $|\Delta|_1 \leq 4\sqrt{k_1}\,|\Delta|_2$ yields the desired result. □

Lemma A.7: Suppose the number of regressors $d(=d_n)$ can grow with and exceed the sample size n and the number of non-zero components in θ^* is at most $k_1(=k_{1n})$ and k_1 can increase to infinity with n but slowly compared to n. Let Assumptions 3.2.1 and 3.4.6 (the lower bound part) hold. If $\hat{\theta}$ solves program (13) with $\lambda_{n,1} \geq c\sigma_w\sqrt{\alpha_u}\sqrt{\frac{\log d}{n}}$ and $n \gtrsim k_1 \log d$, then,

$$\left|\hat{\theta}-\theta^*\right|_2 \leq c\frac{\sqrt{k_1}}{\kappa_L^w}\left(\sqrt{\alpha_u\sigma_w^2}\sqrt{\frac{\log d}{n}} \vee \lambda_{n,1}\right)$$

with probability at least $1-O\left(\frac{1}{d}\right)$.

Proof. Recall

$$L_n(\theta)=\left\{-\frac{1}{n}\sum_{i=1}^n y_{1i}\phi_1(w_i^T\theta)+\frac{1}{n}\sum_{i=1}^n \phi_2(w_i^T\theta)+\lambda_{n,1}\,|\theta|_1\right\},$$

where

$$\begin{aligned}
\phi_1(w_i^T\theta) &= \log\frac{\Phi(w_i^T\theta)}{1-\Phi(w_i^T\theta)},\\
\phi_2(w_i^T\theta) &= -\log\left[1-\Phi(w_i^T\theta)\right]=-\log\frac{1}{1+\exp\left(\phi_1(w_i^T\theta)\right)}
\end{aligned}$$

for the probit model. Define the following quantity

$$\delta L_n(\theta^*,\,\Delta):=L_n(\theta^*+\Delta)-L_n(\theta^*)-\langle\nabla L_n(\theta^*),\,\Delta\rangle\,.$$

To prove Lemma A.7, by Theorem 1 in Negahban, et. al (2010), it suffices to show that Step 1: $\delta L_n(\theta^*,\,\hat{\Delta})$ satisfies the RSC condition (2.4) where $\hat{\Delta}=\hat{\theta}-\theta^*$; and Step 2: with high probability, we have

$$|\nabla L_n(\theta^*)|_\infty \leq \sqrt{\alpha_u\sigma_w^2\frac{\log d}{n}}.$$

Step 1: We need to show that $\delta L_n(\theta^*,\,\hat{\Delta})$ satisfies the RSC condition where $\hat{\Delta}=\hat{\theta}-\theta^*$. Define $\rho_{i,\theta}=-y_{1i}\phi_1(w_i^T\theta)+\phi_2(w_i^T\theta)$, $\hat{Q}_1(\theta)=\mathbb{E}_n\rho_{i,\theta}$, and $Q_1(\theta)=\mathbb{E}\rho_{i,\theta}$. By the mean value theorem, we have

$$L_n(\theta^*,\,\hat{\Delta})=\hat{Q}_1(\hat{\theta})-\hat{Q}_1(\theta^*)=\frac{1}{n}\sum_{i=1}^n\rho''_{i,\bar{\theta}}\left[w_i^T(\hat{\theta}-\theta^*)\right]^2,$$

where $\rho''_{i,\bar{\theta}}$ denotes the second derivative of $\rho_{i,\bar{\theta}}$ and $\bar{\theta}$ is some intermediate value.

Proposition 1 in Loh and Wainwright (2013) implies that there exists a constant $\underline{\alpha} > 0$, depending on the probit model and $(\sigma_w,\, \Sigma_w)$, such that for all vectors $\theta_2 \in \mathcal{B}_2(3) \cap \mathcal{B}_1(R)$

$$\hat{Q}_1(\theta_2) - \hat{Q}_1(\theta_1) \geq \begin{cases} \frac{\underline{\alpha}}{2}\left|\Delta\right|_2^2 - \frac{c^2\sigma_w^2}{2\underline{\alpha}}\frac{\log p}{n}\left|\Delta\right|_1^2 & \text{for all } \left|\Delta\right|_2 \leq 3, \\ \frac{3\underline{\alpha}}{2}\left|\Delta\right|_2 - 3c\sigma_w\sqrt{\frac{\log p}{n}}\left|\Delta\right|_1 & \text{for all } \left|\Delta\right|_2 \geq 3, \end{cases}$$

with probability at least $1 - c_1\exp(-c_2 n)$ and $\Delta = \theta_2 - \theta_1$. Combining this result with Lemma 1 in Negahban, et. al (2010) which shows that $\left|\hat{\Delta}\right|_1 \leq 4\sqrt{k_1}\left|\hat{\Delta}\right|_2$ yields the RSC of $\delta L_n(\theta^*,\, \hat{\Delta})$.

Step 2: In the following, we show that $\left|\nabla L_n(\theta^*)\right|_\infty \leq \sqrt{\alpha_u\sigma_w^2\frac{\log d}{n}}$ with high probability. For a fixed index $j = 1, ..., d$, we begin by establishing an upper bound on

$$\frac{1}{n}\sum_{i=1}^{n} V_{ij} := \frac{1}{n}\sum_{i=1}^{n}\left[w_{ij}y_{1i}\phi_1'(w_i^T\theta^*) - w_{ij}\phi_2'(w_i^T\theta^*)\right]$$

where $\phi_1'(u) = \frac{\partial\phi_1}{\partial u}$ and $\phi_2'(u) = \frac{\partial\phi_2}{\partial\phi_1}\frac{\partial\phi_1}{\partial u}$. Let us condition on $\{w_i\}_{i=1}^n$, so that y_{1i} is drawn from the exponential family with parameter $\phi_1(w_i^T\theta) = \log(\frac{\Phi(w_i^T\theta)}{1-\Phi(w_i^T\theta)})$ for the probit model. For any $t \in \mathbb{R}$, compute the cumulant function

$$\begin{aligned} & \log\mathbb{E}\left[\exp(tV_{ij})\,|\,w_i\right] \\ = \; & \log\left\{\mathbb{E}\left[\exp\left(tw_{ij}y_{1i}\phi_1'(w_i^T\theta^*)\right)\,|\,w_i\right]\exp\left(-tw_{ij}\phi_2'(w_i^T\theta^*)\right)\right\} \\ = \; & \phi_2\left(\phi_1'(w_i^T\theta^*)tw_{ij} + \phi_1(w_i^T\theta^*)\right) - \phi_2(\phi_1(w_i^T\theta^*)) - \phi_2'(w_i^T\theta^*))(tw_{ij}) \\ = \; & \frac{t^2}{2}w_{ij}^2\left[\phi_2''\left(\phi_1(w_i^T\theta^*) + \upsilon_i\phi_1'(w_i^T\theta^*)tw_{ij}\right)\right] \end{aligned}$$

for some $\upsilon_i \in [0,\, 1]$. Since this upper bound holds for each $i = 1, ..., n$, we obtain

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}\log\mathbb{E}\left[\exp(tV_{ij})\,|\,w_i\right] & \leq \frac{t^2}{2}\left\{\frac{1}{n}\sum_{i=1}^{n}w_{ij}^2\left[\phi_2''\left(\phi_1(w_i^T\theta^*) + \upsilon_i\phi_1'(w_i^T\theta^*)tw_{ij}\right)\right]\right\} \\ & \leq \frac{t^2\alpha_u}{2}\left\{\frac{1}{n}\sum_{i=1}^{n}w_{ij}^2\right\} \end{aligned}$$

where the inequality follows since $\phi_2''(u) \leq \alpha_u$ for some $\alpha_u > 0$ and all $u \in \mathbb{R}$. Note that for each $j = 1, ..., d$, the variables $\left\{w_{ij}^2 - \mathbb{E}\left[w_{ij}^2\right]\right\}_{i=1}^n$ are *i.i.d.* zero-mean and sub-Gaussian with parameter at most σ_w. Consequently, we have $\mathbb{E}\left[w_{ij}^2\right] \leq \sigma_w^2$. Since the squared variables are sub-exponential, by Lemma B.1, we have the tail bound

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} w_{ij}^2 \geq 2\sigma_w^2\right] \leq 2\exp\left(-c_1 n\right).$$

Define the event $\Xi = \left\{\max_{j=1,\ldots,d} \frac{1}{n}\sum_{i=1}^{n} w_{ij}^2 \leq 2\sigma_w^2\right\}$. An application of union bound yields

$$\mathbb{P}\left[\Xi^c\right] \leq 2\exp\left(-c_1 n + \log d\right) \leq 2\exp(-c_1 n)$$

under the condition $n \gtrsim \log d$. Conditioning on the event Ξ, we obtain

$$\frac{1}{n}\sum_{i=1}^{n} \log \mathbb{E}\left[\exp(tV_{ij}) \,|\, w_i\right] \leq t^2 \alpha_u \sigma_w^2, \quad \text{for each } j = 1, ..., d.$$

An application of the Chernoff bound and the union bound yields

$$\mathbb{P}\left[\max_{j=1,\ldots,d}\left|\frac{1}{n}\sum_{i=1}^{n} V_{ij}\right| \geq \varepsilon \,|\, \Xi\right] \leq 2\exp\left(-c_1 n \frac{\varepsilon^2}{\alpha_u \sigma_w^2} + \log d\right).$$

Setting $\varepsilon = \sqrt{\alpha_u \sigma_w^2 \frac{\log d}{n}}$ and combining with the bound on $\mathbb{P}\left[\Xi^c\right]$ yields

$$\begin{aligned}
\mathbb{P}\left[\max_{j=1,\ldots,d}\left|\frac{1}{n}\sum_{i=1}^{n} V_{ij}\right| \geq \sqrt{\alpha_u \sigma_w^2 \frac{\log d}{n}}\right] &\leq \mathbb{P}\left[\Xi^c\right] + \mathbb{P}\left[\max_{j=1,\ldots,d}\left|\frac{1}{n}\sum_{i=1}^{n} V_{ij}\right| \geq \varepsilon \,|\, \Xi\right] \\
&\leq c_2 \exp\left(-c_3 n \frac{\varepsilon^2}{\alpha_u \sigma_w^2} + \log d\right) \\
&\leq c_4 \exp\left(-c_5 \log d\right),
\end{aligned}$$

where the last inequality follows since $p > n$ is the regime of our interest.

Now put the pieces together, we obtain

$$\left|\nabla L_n(\theta^*)\right|_\infty \leq \sqrt{\alpha_u \sigma_w^2 \frac{\log d}{n}}$$

with probability at least $1 - c^{'} \exp\left(-c^{''} \log d\right)$.

Combining Step 1 and Step 2 and applying Theorem 1 in Negahban, et. al (2010) yields the desired result. □

Remark. Theorem 1 in Negahban, et. al (2010) can be applied to obtain the upper bound in Proposition 3.4.2 for the case of approximate sparsity $q_1 \in (0, 1]$.

3.8.7 Proposition 3.4.3

Proof. Recall from the proof for Proposition 3.4.2:

$$\hat{Q}_1(\hat{\theta}) - \hat{Q}_1(\theta^*) = \frac{1}{n}\sum_{i=1}^{n} \rho^{''}_{i,\bar{\theta}} \left[w_i^T(\hat{\theta} - \theta^*)\right]^2,$$

where $\rho^{''}_{i,\bar{\theta}}$ denotes the second derivative of $\rho_{i,\bar{\theta}}$ and $\bar{\theta}$ is some intermediate value. Assumption 3.4.4 implies $\rho^{''}_{i,\bar{\theta}} \geq \alpha_l > 0$, so we further have

$$\frac{1}{n}\sum_{i=1}^{n} \left[w_i^T(\hat{\theta} - \theta^*)\right]^2 \leq \frac{1}{\alpha_l}\left[\hat{Q}_1(\hat{\theta}) - \hat{Q}_1(\theta^*)\right]. \tag{3.39}$$

Hence, it remains to upper bound $\hat{Q}_1(\hat{\theta}) - \hat{Q}_1(\theta^*)$. Write

$$\begin{aligned}\hat{Q}_1(\hat{\theta}) - \hat{Q}_1(\theta^*) &= \hat{Q}_1(\hat{\theta}) - Q_1(\hat{\theta}) - \left[\hat{Q}_1(\theta^*) - Q_1(\theta^*)\right] + \left[Q_1(\hat{\theta}) - Q_1(\theta^*)\right] \\ &:= \Lambda_{\hat{\theta}} - \Lambda_{\theta^*} + E_{\hat{\theta}}, \end{aligned} \tag{3.40}$$

where we define $\hat{Q}_1(\hat{\theta}) - Q_1(\hat{\theta}) - \left[\hat{Q}_1(\theta^*) - Q_1(\theta^*)\right] := \Lambda_{\hat{\theta}} - \Lambda_{\theta^*}$ and the excess risk $Q_1(\hat{\theta}) - Q_1(\theta^*) := E_{\hat{\theta}}$. Furthermore, by the definition of $\hat{\theta}$, we have $\hat{Q}_1(\hat{\theta}) + \lambda_{n,1}\left|\hat{\theta}\right|_1 \geq \hat{Q}_1(\theta^*) + \lambda_{n,1}\left|\theta^*\right|_1$, which yields the following *basic inequality*

$$E_{\hat{\theta}} + \lambda_{n,1}\left|\hat{\theta}\right|_1 \leq -\left[\Lambda_{\hat{\theta}} - \Lambda_{\theta^*}\right] + \lambda_{n,1}\left|\theta^*\right|_1 + E_{\theta^*}$$

Under our model setup, $E_{\theta^*} = 0$. This implies

$$E_{\hat{\theta}} \leq \left|\Lambda_{\hat{\theta}} - \Lambda_{\theta^*}\right| + \lambda_{n,1}\left|\hat{\theta} - \theta^*\right|_1,$$

by the triangle inequality. To control for $\Lambda_{\hat{\theta}} - \Lambda_{\theta^*}$ and $E_{\hat{\theta}}$, we follow Lemma 6.7 in Bühlmann and Van de Geer (2011) (alternatively, see Theorems 2.1 or 2.2 in Van de Geer, 2008) to obtain the desired upper bound on $\hat{Q}_1(\hat{\theta}) - \hat{Q}_1(\theta^*)$ in (3.40).□

3.8.8 Theorem 3.4.7

The proof for Theorem 3.4.7 hinges on an intermediate result that shows the "mutual incoherence" assumption on $\mathbb{E}[v_1^T v_1]$ (the population version of $\frac{1}{n}v^T v$) guarantees that, with high probability, analogous conditions hold for the estimated quantity $\frac{1}{n}\hat{v}^T\hat{v}$. This result is established in Lemma A.13. The main proof for Theorem 3.4.7 is based on a construction called Primal-Dual Witness (PDW) method developed by Wainwright (2009). This method constructs a pair $(\hat{\beta}, \hat{\mu})$. When this procedure succeeds, the constructed pair is primal-dual optimal, and acts as a witness for the fact that the Lasso has a unique optimal solution with the correct signed support. The procedure is described in the following.

(i) Set $\hat{\beta}_{J(\beta^*)^c} = 0$.

(ii) Obtain $(\hat{\beta}_{J(\beta^*)}, \hat{\mu}_{J(\beta^*)})$ by solving the oracle subproblem

$$\hat{\beta}_{J(\beta^*)} \in \arg \min_{\beta_{J(\beta^*)} \in \mathbb{R}^{k_2}} \{ \frac{1}{2n} |\hat{v}_0 - \hat{v}_{J(\beta^*)} \beta_{J(\beta^*)}|_2^2 + \lambda_{n,3} |\beta_{J(\beta^*)}|_1 \},$$

and choose $\hat{\mu}_{J(\beta^*)} \in \partial |\hat{\beta}_{J(\beta^*)}|_1$, where $\partial |\hat{\beta}_{J(\beta^*)}|_1$ denotes the set of subgradients at $\hat{\beta}_{J(\beta^*)}$ for the function $|\cdot|_1 : \mathbb{R}^{k_2} \to \mathbb{R}$.

(iii) Solve for $\hat{\mu}_{J(\beta^*)^c}$ via the zero-subgradient equation

$$\frac{1}{n} \hat{v}^T (\hat{v}_0 - \hat{v}\hat{\beta}) + \lambda_{n,3} \hat{\mu} = 0,$$

and check whether or not the *strict dual feasibility* condition $|\hat{\mu}_{J(\beta^*)^c}|_\infty < 1$ holds.

The proof includes four parts. Lemma A.8 guarantees the uniqueness of the optimal solution of the two-stage Lasso procedure, $\hat{\beta}_{HSEL}$, and shows that $\hat{\beta}_{HSEL} = (\hat{\beta}_{J(\beta^*)}, \mathbf{0})$ where $\hat{\beta}_{J(\beta^*)}$ is the solution obtained in step 2 of the PDW construction. Based on this uniqueness claim, one can then talk unambiguously about the support of the two-stage Lasso estimate. Lemma A.9 proves part (a) of Theorem 3.4.7 by verifying the *strict dual feasibility* condition in step 3 of PDW and ensures that $\hat{\beta}_{HSEL,\, J(\beta^*)}$ is uniformly close to $\beta^*_{J(\beta^*)}$ in the $l_\infty-$norm. Part (b) of Theorem 3.4.7 is a consequence of this uniform norm bound in Lemma A.9: as long as the minimum value of $|\beta^*_j|$ over $j \in J(\beta^*)$ is not too small, then the multi-stage estimator does not falsely exclude elements that are in the support of β^* with high probability. To simplify the notations in the following analysis, let $J(\beta^*) = K$, $J(\beta^*)^c := K^c$, $\Sigma_{K^cK} := \mathbb{E}\left[v_{1,K^c}^T v_{1,K}\right]$, $\hat{\Sigma}_{K^cK} := \frac{1}{n} v_{K^c}^T v_K$, and $\check{\Sigma}_{K^cK} := \frac{1}{n} \hat{v}_{K^c}^T \hat{v}_K$. Similarly, let $\Sigma_{KK} := \mathbb{E}\left[v_{1,K}^T v_{1,K}\right]$, $\hat{\Sigma}_{KK} := \frac{1}{n} v_K^T v_K$, and $\check{\Sigma}_{KK} := \frac{1}{n} \hat{v}_K^T \hat{v}_K$.

Lemma A.8: If the PDW construction succeeds and if $\lambda_{\min}\left(\mathbb{E}\left[v_{1,J(\beta^*)}^T v_{1,J(\beta^*)}\right]\right) \geq C_{\min} > 0$, then the vector $(\hat{\beta}_K, \mathbf{0}) \in \mathbb{R}^p$ is the unique optimal solution of the Lasso.

Proof. The proof for Lemma A.8 adopts the proof for Lemma 1 from Chapter 6.4.2 of Wainwright (2015). If the PDW construction succeeds, then $\hat{\beta} = (\hat{\beta}_K, \mathbf{0})$ is an optimal solution with associated subgradient vector $\hat{\mu} \in \mathbb{R}^p$ satisfying $|\hat{\mu}_{K^c}|_\infty < 1$, and $\left\langle \hat{\mu}, \hat{\beta} \right\rangle = |\hat{\beta}|_1$. Suppose $\tilde{\beta}$ is another optimal solution. Letting $F(\beta) = \frac{1}{2n}|\hat{v}_0 - \hat{v}\beta|_2^2$, then $F(\hat{\beta}) + \lambda_{n,3} \left\langle \hat{\mu}, \hat{\beta} \right\rangle = F(\tilde{\beta}) + \lambda_{n,3}|\tilde{\beta}|_1$, and hence $F(\hat{\beta}) - \lambda_{n,3} \left\langle \hat{\mu}, \tilde{\beta} - \hat{\beta} \right\rangle = F(\tilde{\beta}) + \lambda_{n,3} \left(|\tilde{\beta}|_1 - \left\langle \hat{\mu}, \tilde{\beta} \right\rangle\right)$. However, by the zero-subgradient[10] conditions for optimality, we have $\lambda_{n,3}\hat{\mu} = -\nabla F(\hat{\beta})$, which implies that $F(\hat{\beta}) + \left\langle \nabla F(\hat{\beta}), \tilde{\beta} - \hat{\beta} \right\rangle - F(\tilde{\beta}) = \lambda_{n,3} \left(|\tilde{\beta}|_1 - \left\langle \hat{\mu}, \tilde{\beta} \right\rangle\right)$. By convexity of F, the left-hand side is non-positive, which implies that $|\tilde{\beta}|_1 \leq \left\langle \hat{\mu}, \tilde{\beta} \right\rangle$. But since we also have $\left\langle \hat{\mu}, \tilde{\beta} \right\rangle \leq |\hat{\mu}|_\infty |\tilde{\beta}|_1$, we must have $|\tilde{\beta}|_1 = \left\langle \hat{\mu}, \tilde{\beta} \right\rangle$. Since $|\hat{\mu}_{K^c}|_\infty < 1$, this equality can only occur if $\tilde{\beta}_j = 0$ for all $j \in K^c$. Thus, all optimal solutions are supported only on K, and hence can be obtained by solving the oracle subproblem in the PDW procedure. Given $\lambda_{\min}\left(\mathbb{E}\left[v_{1,J(\beta^*)}^T v_{1,J(\beta^*)}\right]\right) \geq C_{\min} > 0$, this subproblem is strictly convex, and hence it has a unique minimizer. □

Lemma A.9: Suppose the assumptions in Theorem 3.4.7 hold. With the choice of the regularization parameter

$$\lambda_{n,3} \geq c\frac{8(2-\frac{\phi}{4})}{\phi}\left[\left(\sigma_v\sigma_\eta\sqrt{\frac{\log p}{n}}\right) \vee \left(L\left|\beta^*\right|_1\sqrt{\Upsilon_{w,\theta^*}}b\left(\sigma_v,\, \sigma_\eta\right)\sqrt{\frac{k_1\log d}{n}}\right)\right]$$

and under the condition $n \gtrsim (k_2^3 \log p) \vee (k_2^2 k_1 \log d)$, $\sqrt{\frac{n \log p}{|\beta^*|_1^2}} \vee \sqrt{nk_1 \log d} \gtrsim \log p$, and $\sqrt{\frac{k_1 \log d}{n}} = o(1)$, we have $|\hat{\mu}_{K^c}|_\infty \leq 1 - \frac{\phi}{8}$ with probability at least $1 - c_1 \exp(-c_2 \log(p \wedge d))$. Furthermore,

$$|\hat{\beta}_K - \beta_K^*|_\infty \leq \frac{c\sqrt{k_2}}{\lambda_{\min}(\Sigma_{KK})}\left[\left(\sigma_v\sigma_\eta\sqrt{\frac{\log p}{n}}\right) \vee \left(L\left|\beta^*\right|_1\sqrt{\Upsilon_{w,\theta^*}}b\left(\sigma_v,\, \sigma_\eta\right)\sqrt{\frac{k_1\log d}{n}}\right)\right],$$

with probability at least $1 - c_1 \exp(-c_2 \log(p \wedge d))$.

Proof. By construction, the sub-vectors $\hat{\beta}_K$, $\hat{\mu}_K$, and $\hat{\mu}_{K^c}$ satisfy the zero-subgradient

[10] Given a convex function $g : \mathbb{R}^p \mapsto \mathbb{R}$, $\mu \in \mathbb{R}^p$ is a subgradient at β, denoted by $\mu \in \partial g(\beta)$, if $g(\beta + \triangle) \geq g(\beta) + \langle \mu, \triangle \rangle$ for all $\triangle \in \mathbb{R}^p$. When $g(\beta) = |\beta|_1$, notice that $\mu \in \partial|\beta|_1$ if and only if $\mu_j = \text{sign}(\beta_j)$ for all $j = 1, ..., p$, where $\text{sign}(0)$ is allowed to be any number in $[-1, 1]$.

condition in the PDW construction. Recall the definition of e from Lemma A.1. With the fact that $\hat{\beta}_{K^c} = \beta^*_{K^c} = 0$, we have

$$\begin{aligned} \frac{1}{n}\hat{v}_K^T\hat{v}_K\left(\hat{\beta}_K - \beta^*_K\right) + \frac{1}{n}\hat{v}_K^T e + \lambda_{n,3}\hat{\mu}_K &= 0, \\ \frac{1}{n}\hat{v}_{K^c}^T\hat{v}_K\left(\hat{\beta}_K - \beta^*_K\right) + \frac{1}{n}\hat{v}_{K^c}^T e + \lambda_{n,3}\hat{\mu}_{K^c} &= 0. \end{aligned}$$

From the equations above, by solving for the vector $\hat{\mu}_{K^c} \in \mathbb{R}^{p-k_2}$, we obtain

$$\begin{aligned} \hat{\mu}_{K^c} &= -\frac{1}{n\lambda_{n,3}}\hat{v}_{K^c}^T\hat{v}_K\left(\hat{\beta}_K - \beta^*_K\right) - \hat{v}_{K^c}^T\frac{e}{n\lambda_{n,3}}, \\ \hat{\beta}_K - \beta^*_K &= -\left(\frac{1}{n}\hat{v}_K^T\hat{v}_K\right)^{-1}\frac{\hat{v}_K^T e}{n} - \lambda_{n,3}\left(\frac{\hat{v}_K^T\hat{v}_K}{n}\right)^{-1}\hat{\mu}_K, \end{aligned}$$

which yields

$$\hat{\mu}_{K^c} = \left(\check{\Sigma}_{K^cK}\check{\Sigma}_{KK}^{-1}\right)\hat{\mu}_K + \left(\hat{v}_{K^c}^T\frac{e}{n\lambda_{n,3}}\right) - \left(\check{\Sigma}_{K^cK}\check{\Sigma}_{KK}^{-1}\right)\hat{v}_K^T\frac{e}{n\lambda_{n,3}}.$$

By the triangle inequality, we have

$$|\hat{\mu}_{K^c}|_\infty \le \left\|\check{\Sigma}_{K^cK}\check{\Sigma}_{KK}^{-1}\right\|_\infty + \left|\hat{v}_{K^c}^T\frac{e}{n\lambda_{n,3}}\right|_\infty + \left\|\check{\Sigma}_{K^cK}\check{\Sigma}_{KK}^{-1}\right\|_\infty\left|\hat{v}_K^T\frac{e}{n\lambda_{n,3}}\right|_\infty,$$

where the fact that $|\hat{\mu}_K|_\infty \le 1$ is used in the inequality above. By Lemma A.13, we have $\left\|\check{\Sigma}_{K^cK}\check{\Sigma}_{KK}^{-1}\right\|_\infty \le 1 - \frac{\phi}{4}$ with probability at least $1 - c\exp(-\log(p \wedge d))$. Hence,

$$\begin{aligned} |\hat{\mu}_{K^c}|_\infty &\le 1 - \frac{\phi}{4} + \left|\hat{v}_{K^c}^T\frac{e}{n\lambda_{n,3}}\right|_\infty + \left\|\check{\Sigma}_{K^cK}\check{\Sigma}_{KK}^{-1}\right\|_\infty\left|\hat{v}_K^T\frac{e}{n\lambda_{n,3}}\right|_\infty \\ &\le 1 - \frac{\phi}{4} + \left(2 - \frac{\phi}{4}\right)\left|\hat{v}^T\frac{e}{n\lambda_{n,3}}\right|_\infty. \end{aligned}$$

Therefore, it suffices to show that $\left(2 - \frac{\phi}{4}\right)\left|\hat{v}^T\frac{e}{n\lambda_{n,3}}\right|_\infty \le \frac{\phi}{8}$ with high probability. This result is established in Lemma A.16. Thus, we have $|\hat{\mu}_{K^c}|_\infty \le 1 - \frac{\phi}{8}$ with high probability.

It remains to establish a bound on the l_∞−norm of the error $\hat{\beta}_K - \beta^*_K$. By the triangle inequality, we have

$$\begin{aligned} |\hat{\beta}_K - \beta^*_K|_\infty &\le \left|\left(\frac{\hat{v}_K^T\hat{v}_K}{n}\right)^{-1}\frac{\hat{v}_K^T e}{n}\right|_\infty + \lambda_{n,3}\left\|\left(\frac{\hat{v}_K^T\hat{v}_K}{n}\right)^{-1}\right\|_\infty \\ &\le \left\|\left(\frac{\hat{v}_K^T\hat{v}_K}{n}\right)^{-1}\right\|_\infty\left|\frac{\hat{v}_K^T e}{n}\right|_\infty + \lambda_{n,3}\left\|\left(\frac{\hat{v}_K^T\hat{v}_K}{n}\right)^{-1}\right\|_\infty, \end{aligned}$$

Using bound (3.63) from Lemma A.15, we have

$$\left\| \left(\frac{\hat{v}_K^T \hat{v}_K}{n} \right)^{-1} \right\|_\infty \leq \frac{2\sqrt{k_2}}{\lambda_{\min}(\hat{\Sigma}_{KK})} \leq \frac{4\sqrt{k_2}}{\lambda_{\min}(\Sigma_{KK})}.$$

From the proof for Corollary 3.4.4, we have,

$$|\frac{1}{n}\hat{v}^T e|_\infty \precsim \left(\sigma_v \sigma_\eta \sqrt{\frac{\log p}{n}}\right) \vee \left(L\,|\beta^*|_1 \sqrt{\Upsilon_{w,\theta^*}} b\,(\sigma_v,\, \sigma_\eta) \sqrt{\frac{k_1 \log d}{n}}\right),$$

with probability at least $1 - c_1 \exp(-c_2 \log(p \wedge d))$. Putting everything together, with the choice of $\lambda_{n,3}$, we obtain

$$|\hat{\beta}_K - \beta_K^*|_\infty \leq \frac{c\sqrt{k_2}}{\lambda_{\min}(\Sigma_{KK})} \left[\left(\sigma_v \sigma_\eta \sqrt{\frac{\log p}{n}}\right) \vee \left(L\,|\beta^*|_1 \sqrt{\Upsilon_{w,\theta^*}} b\,(\sigma_v,\, \sigma_\eta) \sqrt{\frac{k_1 \log d}{n}}\right) \right],$$

with probability at least $1 - c_1 \exp(-c_2 \log(p \wedge d)))$, as claimed. □

3.8.9 Theorem 3.4.8

Proof. Let $\Delta := D^{-1}(\hat{\beta} - \beta^*)$ and $\Psi_n := \frac{1}{n} D\hat{v}^T \hat{v} D$ and we have

$$|\Psi_n \Delta| = \left| \frac{1}{n} D\hat{v}^T(\hat{v}_0 - \hat{v}\hat{\beta}) - \frac{1}{n} D\hat{v}^T(\hat{v}_0 - \hat{v}\beta^*) \right|.$$

Consequently,

$$\begin{aligned} |\Psi_n \Delta|_\infty &\leq \left| \frac{1}{n} D\hat{v}^T(\hat{v}_0 - \hat{v}\hat{\beta}) \right|_\infty + \left| \frac{1}{n} D\hat{v}^T(\hat{v}_0 - \hat{v}\beta^*) \right|_\infty \\ &\leq \hat{\sigma}\xi + \left| \frac{1}{n} D\hat{v}^T e \right|_\infty, \end{aligned}$$

recalling $e := \hat{v}_0 - \hat{v}\beta^*$ defined in Lemma A.1. As in the nonpivotal case, to upper bound $\left| D \frac{\hat{v}^T e}{n} \right|_\infty$, we control the terms except $\max_j \left\{ \frac{1}{n} \sum_{i=1}^n v_{ij}\eta_i \right\}$ in (3.35)-(3.37) with the argument from the proofs for Lemma A.4 and Corollary 3.4.4. To control for the term $\max_j \left\{ \frac{1}{n} \sum_{i=1}^n v_{ij}\eta_i \right\}$, we use an argument similar to Gautier and Tsybakov (2011). Let $Q(\beta) := \frac{1}{n} |v_0 - v\beta|_2^2$ and define the event

$$\mathcal{E} = \left\{ \left| \frac{1}{n} Dv^T \eta \right|_\infty \leq \sqrt{Q(\beta^*)}\xi \right\}.$$

Since $Q(\beta^*) = \frac{1}{n}\sum_{i=1}^n \eta^2$, we have

$$\begin{aligned}\mathbb{P}(\mathcal{E}^c) &\leq \sum_{j=1}^p \mathbb{P}\left(\frac{1}{n}\left|\frac{\sum_{i=1}^n v_{ij}\eta_i}{v_{j*}\sqrt{\frac{1}{n}\sum_{i=1}^n \eta_i^2}}\right| \geq \xi\right) \\ &\leq \sum_{j=1}^p \mathbb{P}\left(\left|\frac{\sum_{i=1}^n v_{ij}\eta_i}{\sqrt{\sum_{i=1}^n (v_{ij}\eta_i)^2}}\right| \geq \sqrt{n}\xi\right).\end{aligned}$$

By Lemma B.16, for all $j = 1, \dots, p$,

$$\mathbb{P}\left(\left|\frac{\sum_{i=1}^n v_{ij}\eta_i}{\sqrt{\sum_{i=1}^n (v_{ij}\eta_i)^2}}\right| \geq \sqrt{n}\xi\right) \leq 2\left(1 - \Phi\left(\sqrt{n}\xi\right)\right) + 2a_0\frac{(1+\sqrt{n}\xi)^{1+\delta'}}{p^{a^2} b_{n,\delta'}^{2+\delta'}}.$$

Thus, the event $\mathcal{E}$ holds with probability at least $1 - \alpha$ where

$$\alpha = 2L\left(1 - \Phi\left(\sqrt{n}\xi\right)\right) + 2a_0\frac{(1+\sqrt{n}\xi)^{1+\delta'}}{p^{a^2-1} b_{n,\delta'}^{2+\delta'}}.$$

Now, let $\hat{Q}(\beta) := \frac{1}{n}|\hat{v}_0 - \hat{v}\beta|_2^2$. The following shows that, conditioning on the event $\mathcal{E}$ and by appropriately choosing ξ, we have

$$\left|\frac{1}{n}D\hat{v}^T(\hat{v}_0 - \hat{v}\beta^*)\right|_\infty \leq \hat{Q}(\beta^*)\xi,$$

namely, $\left(\beta^*, \hat{Q}(\beta^*)\right) \in A$. First, notice that we have

$$\begin{aligned}\left|\frac{1}{n}D\hat{v}^T(\hat{v}_0 - \hat{v}\beta^*)\right|_\infty &\leq \left|\frac{1}{n}Dv^T(v_0 - v\beta^*)\right|_\infty + \left|\frac{1}{n}D\hat{v}^T(\hat{v}_0 - \hat{v}\beta^*) - \frac{1}{n}Dv^T(v_0 - v\beta^*)\right|_\infty \\ &\leq \left|\frac{1}{n}Dv^T(v_0 - v\beta^*)\right|_\infty + 2\left|\frac{1}{n}Dv^T(\hat{v}_0 - v_0)\right|_\infty \\ &\quad + \left|\frac{1}{n}D(\hat{v} - v)^T(\hat{v}_0 - v_0)\right|_\infty \\ &\quad + 2\left|\frac{1}{n}Dv^T(\hat{v} - v)\beta^*\right|_\infty + \left|\frac{1}{n}D(\hat{v} - v)^T(\hat{v} - v)\beta^*\right|_\infty \\ &\leq \sqrt{Q(\beta^*)}\xi + \frac{Lb(\sigma_v)\sqrt{B'}}{\min_{j=1,\dots,p} v_{j*}}|\beta^*|_1 \\ &\leq \sqrt{\hat{Q}(\beta^*)}\xi + \left|\sqrt{\hat{Q}(\beta^*)} - \sqrt{Q(\beta^*)}\right|\xi + \frac{Lb(\sigma_v)\sqrt{B'}}{\min_{j=\{1,\dots,p\}} v_{j*}}|\beta^*|_1\end{aligned}$$

where the third inequality follows from argument similar to the proofs of the upper bounds for Lemma A.3 and Corollary 3.4.4, and $b(\sigma_v)$ is a positive constant only depending on σ_v. We now control for the term $\left|\sqrt{\hat{Q}(\beta^*)} - \sqrt{Q(\beta^*)}\right|$. Note that

$$\begin{aligned}
\left|\hat{Q}(\beta^*) - Q(\beta^*)\right| \;\leq\; & \left|\frac{1}{n}\sum_{i=1}^{n}\left(\hat{v}_{i0}^2 - v_{i0}^2\right)\right| + \left|\frac{2}{n}\sum_{i=1}^{n}\left(\hat{v}_{i0}\hat{v}_i - v_{i0}v_i\right)\beta^*\right| \\
& + \left|\frac{1}{n}\sum_{i=1}^{n}\beta^{*T}\left(\hat{v}_i^T\hat{v}_i - v_i^T v_i\right)\beta^*\right| \\
\leq\; & \left|\frac{1}{n}\sum_{i=1}^{n}\left(\hat{v}_{i0}^2 - v_{i0}^2\right)\right| + \max_{j=1,\ldots,p}\left|\frac{2}{n}\sum_{i=1}^{n}\left(\hat{v}_{i0}\hat{v}_{ij} - v_{i0}v_{ij}\right)\right| |\beta^*|_1 \\
& + \max_{j,\, j'=1,\ldots,p}\left|\frac{2}{n}\sum_{i=1}^{n}\left(\hat{v}_{ij}^T\hat{v}_{ij'} - v_{ij}^T v_{ij'}\right)\right| |\beta^*|_1^2 .
\end{aligned}$$

Following from argument similar to the proofs of the upper bounds for Lemma A.3 and Corollary 3.4.4, we can upper bound the terms

$$\left|\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\left(\hat{v}_{i0}^2 - v_{i0}^2\right)\right|,$$
$$\max_{j=1,\ldots,p}\left|\tfrac{2}{n}\textstyle\sum_{i=1}^{n}\left(\hat{v}_{i0}\hat{v}_{ij} - v_{i0}v_{ij}\right)\right|,$$
$$\max_{j,\, j'=1,\ldots,p}\left|\tfrac{2}{n}\textstyle\sum_{i=1}^{n}\left(\hat{v}_{ij}^T\hat{v}_{ij'} - v_{ij}^T v_{ij'}\right)\right|,$$

which yields

$$\left|\hat{Q}(\beta^*) - Q(\beta^*)\right| = O\left(\left(\frac{1}{n}\right)^{\frac{2}{3}}|\beta^*|_1 + \left(\frac{\log d}{n}\right)^{\frac{1}{2}}|\beta^*|_1\right).$$

Now, notice that if $|a_1 - a_2| \leq a_3$, then $\left|\sqrt{a_1} - \sqrt{a_2}\right| \leq \sqrt{a_3}$, we have

$$\left|\sqrt{\hat{Q}(\beta^*)} - \sqrt{Q(\beta^*)}\right| = O\left(\left(\frac{1}{n}\right)^{\frac{1}{3}}|\beta^*|_1^{\frac{1}{2}} + \left(\frac{\log d}{n}\right)^{\frac{1}{4}}|\beta^*|_1^{\frac{1}{2}}\right),$$

and therefore,

$$\left|\sqrt{\hat{Q}(\beta^*)} - \sqrt{Q(\beta^*)}\right|\xi \leq c\sqrt{\hat{Q}(\beta^*)}\xi,$$

for $c > 1$. Consequently, if we choose

$$\xi \geq a\max\left\{c\sqrt{\frac{\log p}{n}},\ \frac{Lb(\sigma_v)\sqrt{B'}}{\min_{j=1,\ldots,p} v_{j*}}\left(\hat{Q}(\beta^*)\right)^{-\frac{1}{2}}|\beta^*|_1\right\}$$

we have

$$\left|\frac{1}{n}D\hat{v}^T(\hat{v}_0 - \hat{v}\beta^*)\right|_\infty < \sqrt{\hat{Q}(\beta^*)}\xi$$

with probability at least $1 - O\left(\frac{1}{p\wedge d}\right) - \alpha$. On the other hand, $(\hat{\beta}, \hat{\sigma})$ minimizes the criterion $|D^{-1}\beta|_1 + C\sigma$ on the same set A. Thus, we have

$$\left|D^{-1}\hat{\beta}\right|_1 + C\hat{\sigma} \leq \left|D^{-1}\beta^*\right|_1 + C\sqrt{\hat{Q}(\beta^*)},$$

which implies

$$\begin{aligned}
\left|\Delta_{J(\beta^*)^c}\right|_1 &= \sum_{j\in J(\beta^*)^c}\left|v_{j*}\hat{\beta}_j\right| \\
&\leq \sum_{j\in J(\beta^*)}\left(\left|v_{j*}\beta_j^*\right| - \left|v_{j*}\hat{\beta}_j\right|\right) + C\left(\sqrt{\hat{Q}(\beta^*)} - \sqrt{\hat{Q}(\hat{\beta})}\right) \\
&\leq \left|\Delta_{J(\beta^*)}\right|_1 + C\left(\sqrt{\hat{Q}(\beta^*)} - \sqrt{\hat{Q}(\hat{\beta})}\right) \\
&\leq \left|\Delta_{J(\beta^*)}\right|_1 + C\left|\frac{\left(\frac{1}{n}\sum_{i=1}^n \hat{v}_{ij}e_i\right)D\Delta}{\sqrt{\frac{1}{n}\sum_{i=1}^n e_i^2}}\right| \\
&\leq \left|\Delta_{J(\beta^*)}\right|_1 + C\left|\frac{\left(\frac{1}{n}\sum_{i=1}^n \hat{v}_{ij}e_i\right)D}{\sqrt{\frac{1}{n}\sum_{i=1}^n e_i^2}}\right|_\infty |\Delta|_1 \\
&\leq \left|\Delta_{J(\beta^*)}\right|_1 + C\,|\Delta|_1\,,
\end{aligned}$$

where the third inequality follows from the convexity of $\beta \mapsto \sqrt{\hat{Q}(\beta)}$ and the last inequality follows from the concavity of $\sqrt{\cdot}$ and the Cauchy-Schwarz inequality. Consequently, we have the following cone condition:

$$\left|\Delta_{J(\beta^*)^c}\right|_1 \leq \frac{1+C}{1-C}\left|\Delta_{J(\beta^*)}\right|_1.$$

We now finish upper bounding $|\Psi_n\Delta|_\infty$. Recall that

$$\begin{aligned}
|\Psi_n\Delta|_\infty &\leq \hat{\sigma}\xi + \left|\frac{1}{n}D\hat{v}^Te\right|_\infty \\
&\leq \hat{\sigma}\xi + \sqrt{Q(\beta^*)}\xi + \frac{Lb(\sigma_v)\sqrt{B'}}{\min_{j=1,\ldots,p} v_{j*}}|\beta^*|_1 \\
&\leq \xi\left(2\hat{\sigma} + \sqrt{Q(\beta^*)} - \sqrt{Q(\hat{\beta})}\right) + \xi\left(\sqrt{Q(\hat{\beta})} - \sqrt{\hat{Q}(\hat{\beta})}\right) \\
&\quad + \frac{Lb(\sigma_v)\sqrt{B'}}{\min_{j=1,\ldots,p} v_{j*}}|\beta^*|_1.
\end{aligned}$$

We first upper bound the term $\sqrt{Q(\beta^*)} - \sqrt{Q(\hat{\beta})}$. Arguing as in the proof for the cone condition, we have

$$\begin{aligned}
\sqrt{Q(\beta^*)} - \sqrt{Q(\hat{\beta})} &\leq \left| \frac{\left(\frac{1}{n}\sum_{i=1}^n v_{ij}\eta_i\right) D\Delta}{\sqrt{\frac{1}{n}\sum_{i=1}^n \eta_i^2}} \right| \\
&\leq \max_{j\in 1,\ldots,p} \left| \frac{\left(\frac{1}{n}\sum_{i=1}^n v_{ij}\eta_i\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^n v_{ij}^2\eta_i^2}} \right| |\Delta|_1 \\
&\leq \xi\, |\Delta|_1 .
\end{aligned}$$

To upper bound the term $\sqrt{Q(\hat{\beta})} - \sqrt{\hat{Q}(\hat{\beta})}$, we argue as in the proof for upper bounding $\left|\sqrt{\hat{Q}(\beta^*)} - \sqrt{Q(\beta^*)}\right|$ from the above and obtain

$$\sqrt{\hat{Q}(\hat{\beta})} - \sqrt{Q(\hat{\beta})} = O\left(\left(\frac{1}{n}\right)^{\frac{1}{3}} \left|\hat{\beta}\right|_1^{\frac{1}{2}} + \left(\frac{\log d}{n}\right)^{\frac{1}{4}} \left|\hat{\beta}\right|_1^{\frac{1}{2}} \right),$$

and therefore,

$$\xi\left(\sqrt{\hat{Q}(\hat{\beta})} - \sqrt{Q(\hat{\beta})}\right) < \xi\hat{\sigma}.$$

Combining the pieces together, we obtain

$$\begin{aligned}
|\Psi_n\Delta|_\infty &\leq 2\xi\hat{\sigma} + \xi^2 |\Delta|_1 + \frac{Lb(\sigma_v)\sqrt{B'}}{\min_{j=1,\ldots,p} v_{j*}} |\beta^*|_1 \\
&\leq \xi\left(2\hat{\sigma} + \xi\frac{|\Psi_n\Delta|_\infty}{\kappa^*_{J(\beta^*)}}\right) + \frac{Lb(\sigma_v)\sqrt{B'}}{\min_{j=1,\ldots,p} v_{j*}} |\beta^*|_1
\end{aligned}$$

where the second inequality uses the definition of the sensitivity $\kappa^*_{J(\beta^*)}$. The above yields

$$|\Psi_n\Delta|_\infty \leq \left[\frac{Lb(\sigma_v)\sqrt{B'}}{\min_{j=1,\ldots,p} v_{j*}} |\beta^*|_1 + 2\xi\hat{\sigma}\right]\left[1 - \frac{\xi^2}{\kappa^*_{J(\beta^*)}}\right]^{-1}.$$

Now, we can apply the cone condition to obtain

$$\left|D^{-1}(\hat{\beta} - \beta^*)\right|_2 \leq \frac{1}{\kappa^*_{J(\beta^*)}} \left[\frac{Lb(\sigma_v)\sqrt{B'}}{\min_{j=1,\ldots,p} v_{j*}} |\beta^*|_1 + 2\xi\hat{\sigma}\right]\left[1 - \frac{\xi^2}{\kappa^*_{J(\beta^*)}}\right]^{-1}.$$

Some further algebraic manipulation shows

$$\left|D^{-1}(\hat{\beta}-\beta^*)\right|_2 \le \frac{1}{\kappa^*_{J(\beta^*)}}\left[\frac{Lb(\sigma_v)\sqrt{B'}}{\min_{j=1,...,p} v_{j*}}\left|\hat{\beta}\right|_1 + 2\xi\hat{\sigma}\right]\left[1-\frac{\xi^2}{\kappa^*_{J(\beta^*)}}\right]^{-1} + \frac{1}{\kappa^*_{J(\beta^*)}}\left[\frac{Lb(\sigma_v)\sqrt{B'}}{(\min_{j=1,...,p} v_{j*})^2}\sqrt{k_2}\left|D^{-1}\left(\hat{\beta}-\beta^*\right)\right|_2\right]\left[1-\frac{\xi^2}{\kappa^*_{J(\beta^*)}}\right]^{-1}$$

which implies

$$\left|D^{-1}(\hat{\beta}-\beta^*)\right|_2 \le \frac{1}{\kappa^*_{J(\beta^*)}}\left[\frac{Lb(\sigma_v)\sqrt{B'}}{\min_{j=1,...,p} v_{j*}}\left|\hat{\beta}\right|_1 + 2\xi\hat{\sigma}\right]\left[1-\frac{\xi^2}{\kappa^*_{J(\beta^*)}}\right]^{-1} \cdot \left[1-\frac{1}{\kappa^*_{J(\beta^*)}}\left[\frac{Lb(\sigma_v)\sqrt{k_2 B'}}{(\min_{j=1,...,p} v_{j*})^2}\right]\left[1-\frac{\xi^2}{\kappa^*_{J(\beta^*)}}\right]^{-1}\right]^{-1}.$$

Simple calculations yield, for all $j = 1, ..., p$,

$$\left|\hat{\beta}_j-\beta^*_j\right| \le \frac{1}{v_{j*}\kappa^*_{J(\beta^*)}}\left[\frac{Lb(\sigma_v)\sqrt{B'}}{\min_{j=1,...,p} v_{j*}}\left|\hat{\beta}\right|_1 + 2\xi\hat{\sigma}\right]\left[1-\frac{\xi^2}{\kappa^*_{J(\beta^*)}}\right]^{-1} \cdot \left[1-\frac{1}{\kappa^*_{J(\beta^*)}}\left[\frac{Lb(\sigma_v)\sqrt{k_2 B'}}{(\min_{j=1,...,p} v_{j*})^2}\right]\left[1-\frac{\xi^2}{\kappa^*_{J(\beta^*)}}\right]^{-1}\right]^{-1}.$$

To establish the upper bound on $\hat{\sigma}$, note that the optimality of $(\hat{\beta}, \hat{\sigma})$ implies

$$\begin{aligned} C\hat{\sigma} &\le \left|\Delta_{J(\beta^*)}\right|_1 + C\sqrt{\hat{Q}(\beta^*)} \\ &\le \frac{|\Psi_n\Delta|_\infty}{\kappa^*_{J(\beta^*),J(\beta^*)}} + C\sqrt{\hat{Q}(\beta^*)}. \end{aligned}$$

□

3.8.10 Theorems 3.4.9 and 3.4.10

Proof. The estimator of $g(w_i^T\theta^*)$ in Theorems 3.4.9 and 3.4.10 takes on the following form

$$\hat{\mathbb{E}}\left(y_{2i}\,|\,w_i^T\hat{\theta}\right) - \hat{\mathbb{E}}\left(x_i\,|\,w_i^T\hat{\theta}\right)^T\hat{\beta} := \hat{g}(w_i^T\hat{\theta}),$$

where $\hat{\beta} = \hat{\beta}_{HSEL}$. We now derive an upper bound for

$$\left|\hat{g}(\hat{\theta}) - g(\theta^*)\right|_2 := \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\hat{g}(w_i^T\hat{\theta}) - g(w_i^T\theta^*)\right]^2}$$

Note that we can write

$$\begin{aligned}
& \hat{g}(w_i^T\hat{\theta}) - g(w_i^T\theta^*) \\
= & \left[\hat{\mathbb{E}}\left(y_{2i} \,|\, w_i^T\hat{\theta}\right) - \mathbb{E}\left(y_{2i} \,|\, w_i^T\theta^*\right)\right] - \left[\hat{\mathbb{E}}\left(x_i \,|\, w_i^T\hat{\theta}\right)^T \hat{\beta} - \mathbb{E}\left(x_i \,|\, w_i^T\theta^*\right)^T \beta^*\right] \\
= & \left[\hat{\mathbb{E}}\left(y_{2i} \,|\, w_i^T\hat{\theta}\right) - \mathbb{E}\left(y_{2i} \,|\, w_i^T\theta^*\right)\right] - \left[\hat{\mathbb{E}}\left(x_i \,|\, w_i^T\hat{\theta}\right) - \mathbb{E}\left(x_i \,|\, w_i^T\hat{\theta}\right)\right]^T \hat{\beta} \\
& - \left[\mathbb{E}\left(x_i \,|\, w_i^T\hat{\theta}\right) - \mathbb{E}\left(x_i \,|\, w_i^T\theta^*\right)\right]^T \hat{\beta} - \mathbb{E}\left(x_i \,|\, w_i^T\theta^*\right)^T (\hat{\beta} - \beta^*).
\end{aligned}$$

By an elementary inequality and Hölder's inequality, we have

$$\begin{aligned}
& \left[\hat{g}(w_i^T\hat{\theta}) - g(w_i^T\theta^*)\right]^2 \\
\leq & \; 4\left[\hat{\mathbb{E}}\left(y_{2i} \,|\, w_i^T\hat{\theta}\right) - \mathbb{E}\left(y_{2i} \,|\, w_i^T\theta^*\right)\right]^2 \\
& +4\left[\left[\hat{\mathbb{E}}\left(x_i \,|\, w_i^T\hat{\theta}\right) - \mathbb{E}\left(x_i \,|\, w_i^T\hat{\theta}\right)\right]^T \hat{\beta}\right]^2 \\
& +4\left[\left[\mathbb{E}\left(x_i \,|\, w_i^T\hat{\theta}\right) - \mathbb{E}\left(x_i \,|\, w_i^T\theta^*\right)\right]^T \hat{\beta}\right]^2 + 4\left[\mathbb{E}\left(x_i \,|\, w_i^T\theta^*\right)^T (\hat{\beta} - \beta^*)\right]^2 \\
\leq & \; 8\left[\hat{\mathbb{E}}\left(y_{2i} \,|\, w_i^T\hat{\theta}\right) - \mathbb{E}\left(y_{2i} \,|\, w_i^T\hat{\theta}\right)\right]^2 + 8\left[\mathbb{E}\left(y_{2i} \,|\, w_i^T\hat{\theta}\right) - \mathbb{E}\left(y_{2i} \,|\, w_i^T\theta^*\right)\right]^2 \\
& +8\max_j\left[\hat{\mathbb{E}}\left(x_{ij} \,|\, w_i^T\hat{\theta}\right) - \mathbb{E}\left(x_{ij} \,|\, w_i^T\hat{\theta}\right)\right]^2 |\beta^*|_1^2 \\
& +8\max_j\left[\hat{\mathbb{E}}\left(x_{ij} \,|\, w_i^T\hat{\theta}\right) - \mathbb{E}\left(x_{ij} \,|\, w_i^T\hat{\theta}\right)\right]^2 \left|\hat{\beta} - \beta^*\right|_1^2 \\
& +8\max_j\left[\mathbb{E}\left(x_{ij} \,|\, w_i^T\hat{\theta}\right) - \mathbb{E}\left(x_{ij} \,|\, w_i^T\theta^*\right)\right]^2 |\beta^*|_1^2 \\
& +8\max_j\left[\mathbb{E}\left(x_{ij} \,|\, w_i^T\hat{\theta}\right) - \mathbb{E}\left(x_{ij} \,|\, w_i^T\theta^*\right)\right]^2 \left|\hat{\beta} - \beta^*\right|_1^2 \\
& +4\max_j\left(\mathbb{E}\left(x_{ij} \,|\, w_i^T\theta^*\right)\right)^2 \left|\hat{\beta} - \beta^*\right|_1^2.
\end{aligned}$$

Applying previous results yields the desired result. By a standard integration over the tail probability (in the exponential form), the bound above with high probability can be converted to a bound in expectation as stated in Theorem 3.4.8.□

3.8.11 Theorems 3.4.11 and 3.4.12

Proof. We prove the case where β^* and θ^* are exactly sparse and the approximate sparse case of β^* and θ^* follows the same argument (see the remark at the end of this section). The estimator of $g(w_i^T\theta^*)$ in Theorem 3.4.11 is obtained by solving the following program

$$\tilde{g} \in \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\left(y_{2i} - x_i^T\hat{\beta} - f(w_i^T\hat{\theta})\right)^2$$

where $\hat{\beta} = \hat{\beta}_{HSEL}$. We derive an upper bound for

$$\left|\tilde{g}(\hat{\theta}) - g(\theta^*)\right|_2 := \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\tilde{g}(w_i^T\hat{\theta}) - g(w_i^T\theta^*)\right]^2}.$$

Since $\tilde{g}$ is optimal and g is feasible, we have the basic inequality

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_{2i} - x_i^T\hat{\beta} - \tilde{g}(w_i^T\hat{\theta})\right)^2 \leq \frac{1}{n}\sum_{i=1}^{n}\left(y_{2i} - x_i^T\hat{\beta} - g(w_i^T\hat{\theta})\right)^2.$$

Some algebra leads to the equivalent expression

$$\begin{aligned}
\frac{1}{2}\left|\tilde{g}(\hat{\theta}) - g(\hat{\theta})\right|_n^2 \quad\leq\quad & \frac{1}{n}\sum_{i=1}^{n}\left[x_i^T(\beta^* - \hat{\beta})\right]\left[\tilde{g}(w_i^T\hat{\theta}) - g(w_i^T\hat{\theta})\right] \\
& + \frac{1}{n}\sum_{i=1}^{n}\left[g(w_i^T\theta^*) - g(w_i^T\hat{\theta})\right]\left[\tilde{g}(w_i^T\hat{\theta}) - g(w_i^T\hat{\theta})\right] \\
& + \frac{1}{n}\sum_{i=1}^{n}\eta_i\left[\tilde{g}(w_i^T\theta^*) - g(w_i^T\theta^*)\right] \\
& + \frac{1}{n}\sum_{i=1}^{n}\eta_i\left[\tilde{g}(w_i^T\hat{\theta}) - \tilde{g}(w_i^T\theta^*)\right] \\
& - \frac{1}{n}\sum_{i=1}^{n}\eta_i\left[g(w_i^T\hat{\theta}) - g(w_i^T\theta^*)\right].
\end{aligned}$$

Now, by the Fenchel-Young inequality, we have

$$\begin{aligned}
&\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\left[\left[x_i^T(\beta^* - \hat{\beta})\right]\left[\tilde{g}(w_i^T\hat{\theta}) - g(w_i^T\hat{\theta})\right] + \left[g(w_i^T\theta^*) - g(w_i^T\hat{\theta})\right]\left[\tilde{g}(w_i^T\hat{\theta}) - g(w_i^T\hat{\theta})\right]\right] \\
&\quad\leq \tfrac{2}{n}\textstyle\sum_{i=1}^{n}\left[x_i^T(\beta^* - \hat{\beta})\right]^2 + \tfrac{1}{8n}\textstyle\sum_{i=1}^{n}\left[\tilde{g}(w_i^T\hat{\theta}) - g(w_i^T\hat{\theta})\right]^2 \\
&\quad + \tfrac{2}{n}\textstyle\sum_{i=1}^{n}\left[g(w_i^T\theta^*) - g(w_i^T\hat{\theta})\right]^2 + \tfrac{1}{8n}\textstyle\sum_{i=1}^{n}\left[\tilde{g}(w_i^T\hat{\theta}) - g(w_i^T\hat{\theta})\right]^2.
\end{aligned}$$

Consequently, we have

$$\begin{aligned}\frac{1}{4}\left|\tilde{g}(\hat{\theta})-g(\hat{\theta})\right|_n^2 \;\leq\; & \frac{2}{n}\sum_{i=1}^{n}\left[x_i^T(\beta^*-\hat{\beta})\right]^2+\frac{2}{n}\sum_{i=1}^{n}\left[g(w_i^T\theta^*)-g(w_i^T\hat{\theta})\right]^2\\ &+\frac{1}{n}\sum_{i=1}^{n}\eta_i\left[\tilde{g}(w_i^T\theta^*)-g(w_i^T\theta^*)\right]\\ &+\frac{1}{n}\sum_{i=1}^{n}\eta_i\left[\tilde{g}(w_i^T\hat{\theta})-\tilde{g}(w_i^T\theta^*)\right]\\ &-\frac{1}{n}\sum_{i=1}^{n}\eta_i\left[g(w_i^T\hat{\theta})-g(w_i^T\theta^*)\right].\end{aligned}$$

First, recall the upper bound on $\left|\hat{\beta}-\beta^*\right|_2^2$ from Corollary 3.4.4. Let $\Delta=\hat{\beta}-\beta^*$. Applying the inequality in Lemma B.2 yields

$$\frac{|x\Delta|_2^2}{n}\leq\frac{3\kappa_U^x}{2}|\Delta|_2^2+\alpha'\frac{\log d}{n}|\Delta|_1^2,$$

and combining with the fact that $|\Delta|_1\leq 4\sqrt{k_2}\,|\Delta|_2$ shown previously yields an upper bound on the term $\frac{1}{n}\sum_{i=1}^{n}\left[x_i^T(\beta^*-\hat{\beta})\right]^2$.

Now, by the $\bar{L}-$Lipschitz condition on $g(\cdot)$, we have

$$\frac{1}{n}\sum_{i=1}^{n}\left[g(w_i^T\theta^*)-g(w_i^T\hat{\theta})\right]^2\leq\frac{\bar{L}^2}{n}\sum_{i=1}^{n}\left[w_i^T\hat{\theta}-w_i^T\theta^*\right]^2$$

which is upper bounded by Proposition 3.4.2.

For the term $\frac{1}{n}\sum_{i=1}^{n}\eta_i\left[\tilde{g}(w_i^T\theta^*)-g(w_i^T\theta^*)\right]$, we follow the argument as in the proof for Corollary 3.4.4. Using the fact that $\mathbb{E}\left[\eta_i|w_i\right]=0$, we obtain

$$\frac{1}{n}\sum_{i=1}^{n}\eta_i\left[\tilde{g}(w_i^T\theta^*)-g(w_i^T\theta^*)\right]\leq c\left(\frac{k_1\sigma_\eta^2}{n}\right)^{\frac{2}{3}}.$$

It remains to upper bound the terms

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\eta_i\left[\tilde{g}(w_i^T\hat{\theta})-\tilde{g}(w_i^T\theta^*)\right],\\ &\frac{1}{n}\sum_{i=1}^{n}\eta_i\left[g(w_i^T\hat{\theta})-g(w_i^T\theta^*)\right].\end{aligned}$$

By the $\bar{L}-$Lipschitz assumption on $g(\cdot)$ and $\tilde{g}(\cdot)$, we have

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}\eta_i\left[g(w_i^T\theta^*)-g(w_i^T\hat{\theta})\right] &\leq \frac{\bar{L}}{n}\sum_{i=1}^{n}\eta_i\left[w_i^T\hat{\theta}-w_i^T\theta^*\right], \\ \frac{1}{n}\sum_{i=1}^{n}\eta_i\left[\tilde{g}(w_i^T\theta^*)-\tilde{g}(w_i^T\hat{\theta})\right] &\leq \frac{\bar{L}}{n}\sum_{i=1}^{n}\eta_i\left[w_i^T\hat{\theta}-w_i^T\theta^*\right]. \end{aligned}$$

We provide upper bounds on the RHS terms in the inequalities above. For $r_1,\, r_2 > 0$, define the set

$$\mathcal{S}(r_1,\, r_2) := \left\{\delta \in \mathbb{R}^d \mid |\delta|_1 \leq r_1,\; |\delta|_2 \leq r_2\right\},$$

and the random variables $U_n = U_n(r_1,\, r_2)$ given by

$$U_n := \sup_{\delta\in\mathcal{S}(r_1,\, r_2)}\frac{1}{n}\left|\eta^T w\delta\right|.$$

For a given $t \in (0,\, 1)$ to be chosen, let us upper bound the minimal cardinality of a set that covers $\mathcal{S}(r_1,\, r_2)$ up to r_2t-accuracy in l_2-norm. By Lemma B.10, we can find such a covering set $\left\{\delta^1,\, \ldots,\, \delta^N\right\} \subset \mathcal{S}(r_1,\, r_2)$ with cardinality $N = N(r_1,\, r_2,\, t)$ that is upper bounded as

$$\log N(r_1,\, r_2,\, t) \leq c_0 r_1^2\left(\frac{1}{t}\right)^2 \log d.$$

Consequently, for each $\delta \in \mathcal{S}(r_1,\, r_2)$, we may find some δ^i such that $|\delta - \delta^i|_2 \leq r_2 t$. By the triangle inequality, we then have

$$\begin{aligned} \frac{1}{n}\left|\eta^T w\delta\right| &\leq \frac{1}{n}\left|\eta^T w\delta^i\right| + \frac{1}{n}\left|\eta^T w(\delta-\delta^i)\right| \\ &\leq \frac{1}{n}\left|\eta^T w\delta^i\right| + \frac{|\eta|_2}{\sqrt{n}}\frac{|w(\delta-\delta^i)|_2}{\sqrt{n}}. \end{aligned}$$

Given the assumptions on w, we have $\frac{\left|w(\delta-\delta^i)\right|_2}{\sqrt{n}} \leq \sqrt{\kappa_U^w}\,|\delta-\delta^i|_2 \leq \sqrt{\kappa_U^w} r_2 t$. Moreover, by Lemma B.1, we have $\frac{|\eta|_2}{\sqrt{n}} \leq 2\sigma_\eta$ with probability $1 - c_1\exp(-c_2 n)$. Putting together the pieces, we obtain

$$\frac{1}{n}\left|\eta^T w\delta\right| \leq \frac{1}{n}\left|\eta^T w\delta^i\right| + 2\sqrt{\kappa_U^w}\sigma_\eta r_2 t$$

with probability $1 - c_1\exp(-c_2 n)$. Taking the supremum over δ on both sides yields

$$U_n \leq \max_{i=1,\ldots,N}\frac{1}{n}\left|\eta^T w\delta^i\right| + 2\sqrt{\kappa_U^w}\sigma_\eta r_2 t.$$

It remains to bound the finite maximum over the covering set. Note that conditioning on $\{w_i\}_{i=1}^n$, each variate $\frac{1}{n}\left|\eta^T w\delta^i\right|$ is zero-mean (by the fact that $\mathbb{E}\left[\eta_i|w_i\right]=0$) sub-Gaussian with parameter $\frac{\sigma_\eta^2\kappa_U^w r_2^2}{n}$. Under the assumptions on δ^i and w, by the property of sub-Gaussian variables (Lemma B.17), we conclude that

$$U_n \leq \sigma_\eta r_2\sqrt{\kappa_U^w}\left\{\sqrt{\frac{3\log N(r_1,\, r_2,\, t)}{n}}+2t\right\}.$$

Now, by Proposition 3.4.2, we can set $r_1=c^{'}k_1\sqrt{\frac{\log d}{n}}$ and $r_2=c^{''}\sqrt{\frac{k_1\log d}{n}}$. By choosing $t\asymp\sqrt{\frac{k_1\log d}{n}}$ so that $\sqrt{\frac{3\log N(r_1,r_2,t)}{n}}=2t$, we obtain

$$U_n \leq c\sigma_\eta\sqrt{\kappa_U^w}\frac{k_1\log d}{n}.$$

Consequently, we conclude that,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n\eta_i\left[g(w_i^T\theta^*)-g(w_i^T\hat{\theta})\right] &\leq c\sigma_\eta\sqrt{\kappa_U^w}\frac{k_1\log d}{n},\\
\frac{1}{n}\sum_{i=1}^n\eta_i\left[\tilde{g}(w_i^T\theta^*)-\tilde{g}(w_i^T\hat{\theta})\right] &\leq c\sigma_\eta\sqrt{\kappa_U^w}\frac{k_1\log d}{n}.
\end{aligned}$$

Putting the pieces together, we have,

$$\begin{aligned}
\left|\tilde{g}(\hat{\theta})-g(\theta^*)\right|_n &\leq \left|\tilde{g}(\hat{\theta})-g(\hat{\theta})\right|_n+\left|g(\hat{\theta})-g(\theta^*)\right|_n\\
&\leq c^{'}b^{'}\max\left\{\sqrt{\frac{k_2\log p}{n}},\ |\beta^*|_1\sqrt{\frac{k_1k_2\log d}{n}},\ \left(\frac{k_1}{n}\right)^{\frac{1}{3}}\right\}.
\end{aligned}$$

where $b^{'}$ is some constant depending only on the model-specific structure and independent of n, d, p, k_1, and k_2. By a standard integration over the tail probability, the bound above with high probability can be converted to a bound in expectation as stated in Theorem 3.4.11.□

Remark. To prove for the general sparsity case on β^* and $\theta^*(q_1,\, q_2\in[0,\,1])$, we replace $r_1=c^{'}k_1\sqrt{\frac{\log d}{n}}$ and $r_2=c^{''}\sqrt{\frac{k_1\log d}{n}}$ with $r_1=c^{'}R_{q_2}\left(\sqrt{\frac{\log d}{n}}\right)^{1-q_2}$ and $r_2=c^{''}R_{q_2}^{\frac{1}{2}}\left(\sqrt{\frac{\log d}{n}}\right)^{1-\frac{q_2}{2}}$.

3.8.12 Theorem 3.4.13

Define $\mathcal{H} = \mathcal{B}^p_{q_2}(R_{q_2}) \times \mathcal{F} \circ \mathcal{B}^d_{q_1}(R_{q_1})$ where $q_1,\, q_2 \in [0,\, 1]$. Let $\tilde{x} := (x,\, \overrightarrow{f_\theta}) \in \mathbb{R}^{n\times(p+1)}$ where $\overrightarrow{f_\theta} := \left(f(w_1^T\theta),\, \ldots,\, f(w_n^T\theta)\right)^T$ with $\theta \in \mathcal{B}^d_{q_1}(R_{q_1})$ and $f(\cdot) \in \mathcal{F}$. The lower bounds derived in the following involve the set defined by intersecting the kernel of $\tilde{x}$ with $\mathcal{B}^p_{q_2}(R_{q_2})$, which we denote $\mathcal{N}_{q_2}(\tilde{x}) := \mathrm{Ker}(\tilde{x}) \cap \mathcal{B}^p_{q_2}(R_{q_2})$ where

$$\mathrm{Ker}(\tilde{x}) := \left\{\beta \,:\, \tilde{x}\begin{pmatrix}\beta\\ 1\end{pmatrix} = 0\right\},$$

and define the $\mathcal{B}^p_{q_2}(R_{q_2})-$kernel diameter in the l_2-norm

$$\mathrm{diam}\left(\mathcal{N}_{q_2}(\tilde{x})\right) := \max_{\beta\in\mathcal{N}_{q_2}(\tilde{x})} |\beta|_2 \,.$$

The following lemma controls for the term $\mathrm{diam}\left(\mathcal{N}_{q_2}(\tilde{x})\right)$.

Lemma A.10: If Assumptions 3.4.9-3.4.10 hold for any $q_2 \in [0,\, 1]$, then the $\mathcal{B}^p_{q_2}(R_{q_2})$-kernel diameter in l_2-norm is upper bounded as

$$\mathrm{diam}\left(\mathcal{N}_{q_2}(\tilde{x})\right) \leq \frac{f_l(R_{q_2},\, q_2,\, n,\, p)}{\kappa_l}.$$

Proof. This proof is similar to the proof for Lemma 1 in Raskutti, et. al (2011). Under Assumption 3.4.10, we are guaranteed that there is no measurable function $f(w_i^T\theta)$ such that $x_i^T\lambda = f(w_i^T\theta)$ when $y_{1i} = 1$ for $\lambda \in \mathcal{B}^p_{q_2}(R_{q_2})$. On the other hand, if

$$\mathrm{diam}\left(\mathcal{N}_{q_2}(\tilde{x})\right) > \frac{f_l(R_{q_2},\, q_2,\, n,\, p)}{\kappa_l},$$

then there must exist some $\beta \in \mathcal{B}^p_{q_2}(R_{q_2})$ with $x\beta = 0$ and $|\beta|_2 > \frac{f_l(R_{q_2},\, q_2,\, n,\, p)}{\kappa_l}$. We then have

$$0 = \frac{1}{\sqrt{n}}\,|x\beta|_2 < \kappa_l\,|\beta|_2 - f_l(R_{q_2},\, q_2,\, n,\, p),$$

which contradicts the lower bound condition on $\frac{1}{\sqrt{n}}\,|x\beta|_2$ in Assumption 3.4.9. □

Proof. Let $M = M_2(\delta_n;\, \mathcal{H})$ be the cardinality of a maximal packing of $\mathcal{H}$ in the $L^2(\mathbb{P})-$metric with elements $\left\{h^1,\, \ldots,\, h^M\right\}$. Also let the random vector $H \in \mathbb{R}^{p+1}$ be uniformly distributed over the packing set $\left\{h^1,\, \ldots,\, h^M\right\}$. From a standard argument in terms of the error for a multi-way hypothesis testing problem, we obtain the following lower bound on the minimax $L^2(\mathbb{P})-$risk:

$$\min_{\hat{h}} \max_{h\in\mathcal{H}} \mathbb{E}\left|\hat{h}-h\right|_2^2 \geq \frac{1}{4}\delta_n^2 \min_{\tilde{h}} \mathbb{P}\left[\tilde{h}\neq H\right] \tag{3.41}$$

where the estimator $\tilde{h}$ takes values in the packing set. Fano's inequality implies the lower bound

$$\mathbb{P}\left[\tilde{h}\neq H\right] \geq 1-\frac{I(H;\, y_2)+\log 2}{\log M_2(\delta_n;\, \mathcal{H})}$$

where $I(H;\, y_2)$ is the mutual information between H and $y_2 \in \mathbb{R}^n$.

Let us first upper bound the mutual information via the procedure of Yang and Barron (1999): by covering $\mathcal{P}(\mathcal{H}) := \{\mathbb{P}_h : h\in\mathcal{H}\}$ under the square-root Kullback-Leibler (KL) divergence, the mutual information is upper bounded as

$$I(H;\, y_2) \leq \log N(\varepsilon_n;\, \mathcal{P}(\mathcal{H})) + \frac{c^2 n}{\sigma_\eta^2}(\kappa_{1u}\vee\kappa_{2u})\varepsilon_n^2$$

where $N(\varepsilon_n;\, \mathcal{P}(\mathcal{H}))$ is the covering number of $\mathcal{P}(\mathcal{H})$ under the KL divergence. This upper bound combined with the Fano lower bound implies

$$\mathbb{P}\left[\tilde{h}\neq H\right] \geq 1-\frac{\log N(\varepsilon_n;\, \mathcal{P}(\mathcal{H}))+\frac{c^2 n}{\sigma_\eta^2}(\kappa_{1u}\vee\kappa_{2u})\varepsilon_n^2+\log 2}{\log M_2(\delta_n;\, \mathcal{H})}. \tag{3.42}$$

An upper bound on $\log N(\varepsilon_n;\, \mathcal{P}(\mathcal{H}))$ can be obtained using the following argument. For the Gaussian models considered here, the KL divergence

$$\begin{aligned}
D(\mathbb{P}_h \quad \| \quad \mathbb{P}_{h'}) &= \frac{1}{2\sigma_\eta^2}\left|x(\beta-\beta')+f_\theta-f'_{\theta'}\right|_2^2 \\
&\leq \frac{1}{\sigma_\eta^2}\left|x(\beta-\beta')\right|_2^2+\frac{n}{\sigma_\eta^2}\left|f_\theta-f'_{\theta'}\right|_n^2 \\
&\leq \frac{1}{\sigma_\eta^2}\left|x(\beta-\beta')\right|_2^2+\frac{2n}{\sigma_\eta^2}\left|f_\theta-f'_{\theta}\right|_n^2+\frac{2n}{\sigma_\eta^2}\left|f'_\theta-f'_{\theta'}\right|_n^2 \\
&\leq \frac{n\kappa_{2u}}{\sigma_\eta^2}\left|\beta-\beta'\right|_2^2+\frac{2n}{\sigma_\eta^2}\left|f_\theta-f'_{\theta}\right|_\infty^2+\frac{2n\bar{L}^2\kappa_{1u}}{\sigma_\eta^2}\left|\theta-\theta'\right|_2^2
\end{aligned}$$

where the last inequality uses the $\bar{L}-$Lipschitz assumption on $f'\in\mathcal{F}$ and the upper-RE condition on w and x as well as the fact that $|f-g|_n \leq |f-g|_\infty$.

Now, applying the upper bound on the covering number of $\mathcal{B}_0^d(k_1)$ in the l_2-metric provided by Lemma B.14, we conclude that there exists a set $\left\{\theta^1, \ldots, \theta^{N_1}\right\}$ such that for all $\theta \in \mathcal{B}_0^d(k_1)$, there exists some index i with$|\theta - \theta^i|_2^2 \leq \varepsilon_n^2$ and $\log N_2\left(\varepsilon_n;\, \mathcal{B}_0^d(k_1)\right) \leq k_1 \log d + k_1 \log \frac{\sqrt{k_1}}{\varepsilon_n}$; similarly, there also exists a set $\left\{\beta^1, \ldots, \beta^{N_2}\right\}$ such that for all $\beta \in \mathcal{B}_0^p(k_2)$, there exists some index j with $|\beta - \beta^j|_2^2 \leq \varepsilon_n^2$ and $\log N_2\left(\varepsilon_n;\, \mathcal{B}_0^p(k_2)\right) \leq k_2 \log p + k_2 \log \frac{\sqrt{k_2}}{\varepsilon_n}$. Applying the upper bound on the covering number of the $\bar{L}-$Lipschitz function class $\mathcal{F}$ in the sup-norm provided by Lemma B.11, we conclude that there exists a set $\left\{f_\theta^1, \ldots, f_\theta^{N_3}\right\}$ such that for all $f \in \mathcal{F}$ and $\theta \in \left\{\theta^1, \ldots, \theta^{N_1}\right\}$, there exists some index l with $\sup_{w_i\theta}\left|f(w_i^T\theta) - f^l(w_i^T\theta)\right| \leq \varepsilon_n$ and $\log N_\infty\left(\varepsilon_n;\, \mathcal{F}_\theta\right) \leq \frac{2k_1}{\varepsilon_n}$ for all $\theta \in \left\{\theta^1, \ldots, \theta^{N_1}\right\}$. It is not hard to see that a covering set of $\mathcal{P}(\mathcal{H})$ under the KL divergence can be formed by

$$\left\{\theta^1, \ldots, \theta^{N_1}\right\} \times \left\{\beta^1, \ldots, \beta^{N_2}\right\} \times \left\{f_\theta^1, \ldots, f_\theta^{N_3} \,:\, \theta \in \left\{\theta^1, \ldots, \theta^{N_1}\right\}\right\}$$

and

$$\begin{aligned}\log N(\varepsilon_n;\, \mathcal{P}(\mathcal{H})) &\leq \log\left[\left(N_2\left(\varepsilon_n;\, \mathcal{B}_0^d(k_1)\right)\right)^2 \cdot N_2\left(\varepsilon_n;\, \mathcal{B}_0^p(k_2)\right) \cdot N_\infty\left(\varepsilon_n;\, \mathcal{F}\right)\right] \\ &\leq c^{'}\left\{k_1 \log d + k_1 \log \frac{\sqrt{k_1}}{\varepsilon_n} + k_2 \log p + k_2 \log \frac{\sqrt{k_2}}{\varepsilon_n} + \frac{k_1}{\varepsilon_n}\right\} \quad (3.43)\end{aligned}$$

It remains to lower bound $\log M_2(\delta_n;\, \mathcal{H})$; we do so by constructing a minimal covering set of $\mathcal{H} = \mathcal{B}_0^p(k_2) \times \mathcal{F} \circ \mathcal{B}_0^d(k_1)$ in the $L^2(\mathbb{P})-$metric namely, $\log N_2(\delta_n;\, \mathcal{H})$. Note that

$$\begin{aligned}\left|h - h^{'}\right|_{L^2(\mathbb{P})}^2 &= \left|\beta - \beta^{'}\right|_2^2 + \mathbb{E}\left[f(w_i\theta) - f^{'}(w_i\theta^{'})\right]^2 \\ &\leq \left|\beta - \beta^{'}\right|_2^2 + 2\mathbb{E}\left[f(w_i\theta) - f^{'}(w_i\theta)\right]^2 + 2\mathbb{E}\left[f^{'}(w_i\theta) - f^{'}(w_i\theta^{'})\right]^2 \\ &\leq \left|\beta - \beta^{'}\right|_2^2 + 2\mathbb{E}\left[f(w_i\theta) - f^{'}(w_i\theta)\right]^2 + 2\kappa_{1u}\bar{L}^2\left|\theta - \theta^{'}\right|_2^2\end{aligned}$$

where the last inequality uses the $\bar{L}-$Lipschitz assumption on $f^{'} \in \mathcal{F}$ and the population upper-RE condition on w. Using the similar argument from above, we conclude from the first inequality of Lemma B.12 and Lemma B.15 that there exists a set $\left\{\theta^1, \ldots, \theta^{N_1}\right\}$ such that for all $\theta \in \mathcal{B}_0^d(k_1)$, there exists some index i with$|\theta - \theta^i|_2^2 \leq \delta_n^2$ and $\log N_2\left(\delta_n;\, \mathcal{B}_0^d(k_1)\right) \geq \log M_2\left(2\delta_n;\, \mathcal{B}_0^d(k_1)\right) \geq \frac{k_1}{2} \log \frac{d-k_1}{k_1/2}$; similarly, there also exists a set $\left\{\beta^1, \ldots, \beta^{N_2}\right\}$ such that for all $\beta \in \mathcal{B}_0^p(k_2)$, there exists some index j with $|\beta - \beta^j|_2^2 \leq \delta_n^2$ and $\log N_2\left(\delta_n;\, \mathcal{B}_0^p(k_2)\right) \geq \log M_2\left(2\delta_n;\, \mathcal{B}_0^p(k_2)\right) \geq \frac{k_2}{2} \log \frac{d-k_2}{k_2/2}$. Furthermore, by Lemma B.11 and the first inequality of Lemma B.12, we conclude that there exists a set $\left\{f_\theta^1, \dots, f_\theta^{N_3}\right\}$ such that for all $f \in \mathcal{F}$ and $\theta \in \left\{\theta^1, \dots, \theta^{N_1}\right\}$, there exists some index l with $\mathbb{E}\left[f(w_i^T\theta) - f^l(w_i^T\theta)\right]^2 \le \delta_n^2$ and $\log N_2(\delta_n;\, \mathcal{F}_\theta) \ge \log M_2(2\delta_n;\, \mathcal{F}_\theta) \ge \frac{k_1}{\delta_n}$ for all $\theta \in \left\{\theta^1, \dots, \theta^{N_1}\right\}$. It is not hard to see that a covering set of $\mathcal{H}$ under the $L^2(\mathbb{P})-$metric can be formed by

$$\left\{\theta^1, \dots, \theta^{N_1}\right\} \times \left\{\beta^1, \dots, \beta^{N_2}\right\} \times \left\{f_\theta^1, \dots, f_\theta^{N_3} \,:\, \theta \in \left\{\theta^1, \dots, \theta^{N_1}\right\}\right\}$$

and

$$\begin{aligned} \log N_2(\delta_n;\, \mathcal{H}) &\ge \log\left[\left(M_2\left(2\delta_n;\, \mathcal{B}_0^d(k_1)\right)\right)^2 \cdot M_2\left(2\delta_n;\, \mathcal{B}_0^p(k_2)\right) \cdot M_2\left(2\delta_n;\, \mathcal{F}\right)\right] \\ &\ge c''\left\{\frac{k_1}{2}\log\frac{d-k_1}{k_1/2} + \frac{k_2}{2}\log\frac{d-k_2}{k_2/2} + \frac{k_1}{\delta_n}\right\}. \end{aligned}$$

Now applying the second inequality of Lemma B.12 yields

$$\begin{aligned} \log M(\delta_n;\, \mathcal{H}) &\ge \log N(\delta_n;\, \mathcal{H}) \\ &\ge c''\left\{\frac{k_1}{2}\log\frac{d-k_1}{k_1/2} + \frac{k_2}{2}\log\frac{d-k_2}{k_2/2} + \frac{k_1}{\delta_n}\right\}. \end{aligned} \tag{3.44}$$

Substituting (3.43) and (3.44) into (3.42) yields

$$\mathbb{P}\left[\tilde{h} \neq H\right] \ge$$

$$1 - \frac{c'\left\{k_1\log d + k_1\log\frac{\sqrt{k_1}}{\varepsilon_n} + k_2\log p + k_2\log\frac{\sqrt{k_2}}{\varepsilon_n} + \frac{k_1}{\varepsilon_n}\right\} + \frac{c^2 n}{\sigma_\eta^2}(\kappa_{1u} \vee \kappa_{2u})\varepsilon_n^2 + \log 2}{c''\left\{\frac{k_1}{2}\log\frac{d-k_1}{k_1/2} + \frac{k_2}{2}\log\frac{d-k_2}{k_2/2} + \frac{k_1}{\delta_n}\right\}}.$$

Under the condition $\varepsilon \ge \frac{1}{d}$, setting

$$\begin{aligned} \frac{c^2 n}{\sigma_\eta^2}(\kappa_{1u} \vee \kappa_{2u})\varepsilon_n^2 &= c'\left\{k_1\log d + k_1\log\frac{\sqrt{k_1}}{\varepsilon_n} + k_2\log p + k_2\log\frac{\sqrt{k_2}}{\varepsilon_n} + \frac{k_1}{\varepsilon_n}\right\}, \\ \text{and } 4\frac{c^2 n}{\sigma_\eta^2}(\kappa_{1u} \vee \kappa_{2u})\varepsilon_n^2 &= c''\left\{\frac{k_1}{2}\log\frac{d-k_1}{k_1/2} + \frac{k_2}{2}\log\frac{d-k_2}{k_2/2} + \frac{k_1}{\delta_n}\right\} \end{aligned}$$

yields the following

$$\begin{aligned} \mathbb{P}\left[\tilde{h} \neq H\right] &\ge \frac{1}{4}, \\ \text{and } \delta_n^2 &\asymp \varepsilon_n^2 \asymp \max\left\{\frac{k_1\log d}{n},\, \frac{k_2\log p}{n},\, \left(\frac{k_1}{n}\right)^{\frac{2}{3}}\right\}. \end{aligned}$$

It can be easily verified that under the scaling of ε_n, we are guaranteed to have $\varepsilon_n \geq \frac{1}{d}$.

Consequently, substituting the pieces above into (3.41) yields

$$\min_{\hat{h}} \max_{h\in\mathcal{H}} \mathbb{E}\left|\hat{h}-h\right|_2^2 \gtrsim \max\left\{\frac{k_1\log d}{n},\ \frac{k_2\log p}{n},\ \left(\frac{k_1}{n}\right)^{\frac{2}{3}}\right\}.$$

□

3.8.13 Theorem 3.4.14

Proof. The proof for Theorem 3.4.14 is almost identical to the proof for Theorem 3.4.13 except that instead of Lemma B.15, we apply Lemma B.10 to lower bound the packing number of $\mathcal{B}^d_{q_1}(R_{q_1})$ and $\mathcal{B}^p_{q_2}(R_{q_2})$ in the l_2-metric where $q_1,\, q_2 \in (0,\, 1]$. These steps give us

$$\mathbb{P}\left[\tilde{h}\neq H\right] \geq 1- \frac{u_{q_1}\left[R_{q_1}^{\frac{2}{2-q_1}}\left(\frac{1}{\varepsilon_n}\right)^{\frac{2q_1}{2-q_1}}\log d\right]+u_{q_2}\left[R_{q_2}^{\frac{2}{2-q_2}}\left(\frac{1}{\varepsilon_n}\right)^{\frac{2q_2}{2-q_2}}\log p\right]+c^{'}\frac{R}{\varepsilon_n}+\frac{c^2n}{\sigma_\eta^2}(\kappa_{1u}\vee\kappa_{2u})\varepsilon_n^2+\log 2}{l_{q_1}\left[R_{q_1}^{\frac{2}{2-q_1}}\left(\frac{1}{\delta_n}\right)^{\frac{2q_1}{2-q_1}}\log d\right]+l_{q_2}\left[R_{q_2}^{\frac{2}{2-q_2}}\left(\frac{1}{\delta_n}\right)^{\frac{2q_2}{2-q_2}}\log p\right]+c^{''}\frac{R}{\delta_n}}.$$

Setting

$$\begin{aligned}
\frac{c^2n}{\sigma_\eta^2}(\kappa_{1u}\vee\kappa_{2u})\varepsilon_n^2 &= u_{q_1}\left[R_{q_1}^{\frac{2}{2-q_1}}\left(\frac{1}{\varepsilon_n}\right)^{\frac{2q_1}{2-q_1}}\log d\right] \\
&\quad +u_{q_2}\left[R_{q_2}^{\frac{2}{2-q_2}}\left(\frac{1}{\varepsilon_n}\right)^{\frac{2q_2}{2-q_2}}\log p\right]+c^{'}\frac{R}{\varepsilon_n}, \\
\text{and}\ \ 4\frac{c^2n}{\sigma_\eta^2}(\kappa_{1u}\vee\kappa_{2u})\varepsilon_n^2 &= \left[l_{q_1}R_{q_1}^{\frac{2}{2-q_1}}\left(\frac{1}{\delta_n}\right)^{\frac{2q_1}{2-q_1}}\log d\right] \\
&\quad +\left[l_{q_2}R_{q_2}^{\frac{2}{2-q_2}}\left(\frac{1}{\delta_n}\right)^{\frac{2q_2}{2-q_2}}\log p\right]+c^{''}\frac{R}{\delta_n}
\end{aligned}$$

yields the following

$$\begin{aligned}
\mathbb{P}\left[\tilde{h}\neq H\right] &\geq \frac{1}{4}, \\
\text{and}\ \ \delta_n^2 &\asymp \varepsilon_n^2 \asymp \max\left\{R_{q_1}\left(\frac{\log d}{n}\right)^{\frac{2-q_1}{2}},\ R_{q_2}\left(\frac{\log p}{n}\right)^{\frac{2-q_1}{2}},\ \left(\frac{R}{n}\right)^{\frac{2}{3}}\right\}.
\end{aligned}$$

Consequently, substituting the pieces above into (3.41) yields

$$\min_{\hat{h}} \max_{h\in\mathcal{H}} \mathbb{E}\left|\hat{h}-h\right|_2^2 \gtrsim \max\left\{R_{q_1}\left(\frac{\log d}{n}\right)^{\frac{2-q_1}{2}}, R_{q_2}\left(\frac{\log p}{n}\right)^{\frac{2-q_1}{2}}, \left(\frac{R}{n}\right)^{\frac{2}{3}}\right\}.$$

□

3.8.14 Upper bounds on the second-stage estimators

The following lemma upper bounds the term

$$\frac{1}{n}\sum_{i=1}^{n}\left[\hat{m}_j(w_i^T\hat{\theta}) - m_j(w_i^T\hat{\theta})\right]^2$$

(for $j = 0, ..., p$). Recall $y_i = z_{i0}$ and $x_{ij} = z_{ij}$ for $j = 1, ..., p$. From Section 3.3 we have

$$z_{ij} = \mathbb{E}\left(z_{ij}|w_i^T\theta^*\right) + v_{ij} := m_j(w_i^T\theta^*) + v_{ij} \qquad \text{for } j = 0, ..., p$$

which can be rewritten as

$$z_{ij} = m_j(w_i^T\hat{\theta}) + v_{ij} - \left(m_j(w_i^T\hat{\theta}) - m_j(w_i^T\theta^*)\right).$$

To avoid notation clustering, write $\left\{\hat{m}_j(w_i^T\theta)\right\}_{i=1}^n := \hat{m}_j(\theta)$. Recall the set

$$\bar{\mathcal{F}}_j := \left\{f = f' - f'' \, : \, f', \, f'' \in \mathcal{F}_j\right\},$$

and

$$\Omega(r_j; \, \mathcal{F}_j) = \left\{f \, : \, f \in \bar{\mathcal{F}}_j, \, |f_{\theta^*}|_n \leq r_j\right\},$$

where $|f_{\theta^*}|_n := \sqrt{\frac{1}{n}\sum_{i=1}^n \left[f(w_i^T\theta^*)\right]^2}$, and the conditional *local complexity*

$$\mathcal{G}_n(r_j; \, \mathcal{F}_j) := \mathbb{E}_{v_j}\left[\sup_{f\in\Omega(r_j;\,\mathcal{F}_j)}\left|\frac{1}{n}\sum_{i=1}^{n} v_{ij} f(w_i^T\theta^*)\right| |w_i^T\theta^*\right], \tag{3.45}$$

where variables $\{v_{ij}\}_{i=1}^n$ for $j = 0, ..., p$ are *i.i.d.* variates that satisfy Assumption 3.4.2. In terms of this complexity measure (3.45), we have the following result.

Lemma A.11: Let $r_{nj} > 0$ be the smallest positive quantity satisfying the *critical inequality*

$$\mathcal{G}_n\left(r_{nj};\, \mathcal{F}_j\right) \leq \frac{r_{nj}^2}{\sigma_{v_j}}. \tag{3.46}$$

Then there are universal positive constants $(c_1,\, c_2)$ such that for any $t \geq r_{nj}$ and $\tilde{m}_j \in \mathcal{F}_j$, the non-parametric least squares estimate $\hat{m}_j$ via program (11) satisfies

$$\left|\hat{m}_j(\hat{\theta}) - m_j(\hat{\theta})\right|_n^2 \leq 7\left|\tilde{m}_j(\hat{\theta}) - m_j(\hat{\theta})\right|_n^2 + 128tr_{nj} + 4\sqrt{32\sigma_{v_j}^2 T_j^{'}} + 16T_j^{'}$$

with probability at least $1 - c_1 \exp\left(-nC_j^* tr_{nj}\right)$.

Remark. The constants in the error bound is not optimal and can be improved.

Proof. Since $\hat{m}_j$ is optimal and $\tilde{m}_j$ is feasible for the program (11), we have

$$\frac{1}{2}\sum_{i=1}^n \left(z_{ij} - \hat{m}_j(w_i^T\hat{\theta})\right)^2 \leq \frac{1}{2}\sum_{i=1}^n \left(z_{ij} - \tilde{m}_j(w_i^T\hat{\theta})\right)^2.$$

Define $\hat{\Delta}_{ij}(\theta) = (\hat{m}_j - m_j) \circ (w_i^T\theta)$, $\left\{\hat{\Delta}_{ij}(\theta)\right\}_{i=1}^n = \hat{\Delta}_j(\theta)$, $\bar{\Delta}_{ij}(\theta) = (\hat{m}_j - \tilde{m}_j) \circ (w_i^T\theta)$, $\left\{\bar{\Delta}_{ij}(\theta)\right\}_{i=1}^n = \bar{\Delta}_j(\theta)$, $\tilde{\Delta}_{ij}(\theta) = (\tilde{m}_{xj} - m_{xj}) \circ (w_i^T\theta)$, and $\left\{\tilde{\Delta}_{ij}(\theta)\right\}_{i=1}^n = \tilde{\Delta}_j(\theta)$. Performing some algebra yields the *basic inequality*

$$\left|\hat{\Delta}_j(\hat{\theta})\right|_n^2 \leq \left|\tilde{\Delta}_j(\hat{\theta})\right|_n^2 + 2\left|\frac{1}{n}\sum_{i=1}^n \left(v_{ij} + T_{ij}^{''}\right)\bar{\Delta}_{ij}(\hat{\theta})\right|$$

Furthermore, we have

$$\begin{aligned}\frac{1}{n}\left|\sum_{i=1}^n T_{ij}^{''}\bar{\Delta}_{ij}(\hat{\theta})\right| &\leq \frac{1}{n}\left|\sum_{i=1}^n T_{ij}^{''}\hat{\Delta}_{ij}(\hat{\theta})\right| + \frac{1}{n}\left|\sum_{i=1}^n T_{ij}^{''}\tilde{\Delta}_{ij}(\hat{\theta})\right| \\ &\leq \frac{2}{n}\left|\sum_{i=1}^n T_{ij}^{''2}\right| + \frac{1}{4n}\left|\sum_{i=1}^n \hat{\Delta}_{ij}^2(\hat{\theta})\right| + \frac{1}{4n}\left|\sum_{i=1}^n \tilde{\Delta}_{ij}^2(\hat{\theta})\right|,\end{aligned}$$

where the first inequality follows from the triangle inequality and the second inequality follows from the Fenchel-Young inequality. Consequently, we have

$$\frac{1}{2}\left|\hat{\Delta}_j(\hat{\theta})\right|_n^2 \leq \frac{3}{2}\left|\tilde{\Delta}_j(\hat{\theta})\right|_n^2 + 2\left|\frac{1}{n}\sum_{i=1}^n v_{ij}\bar{\Delta}_{ij}(\hat{\theta})\right| + 4T_j^{'}, \tag{3.47}$$

recalling the definition of $T_j^{\prime}$.

We need to upper bound the stochastic process $\frac{1}{n}\left|\sum_{i=1}^n v_{ij}\bar{\Delta}_{ij}(\hat{\theta})\right|$. Now, note that by the elementary inequalities, we have

$$\begin{aligned}
\frac{1}{n}\left|\sum_{i=1}^n v_{ij}\bar{\Delta}_{ij}(\hat{\theta})\right| &\leq \frac{1}{n}\left|\sum_{i=1}^n v_{ij}\left[\bar{\Delta}_{ij}(\hat{\theta})-\bar{\Delta}_{ij}(\theta^*)\right]\right| + \frac{1}{n}\left|\sum_{i=1}^n v_{ij}\bar{\Delta}_{ij}(\theta^*)\right| \\
&\leq \sqrt{\frac{1}{n}\sum_{i=1}^n v_{ij}^2}\sqrt{\frac{1}{n}\sum_{i=1}^n\left[\bar{\Delta}_{ij}(\hat{\theta})-\bar{\Delta}_{ij}(\theta^*)\right]^2} + \frac{1}{n}\left|\sum_{i=1}^n v_{ij}\bar{\Delta}_{ij}(\theta^*)\right| \\
&\leq \sqrt{\frac{1}{n}\sum_{i=1}^n v_{ij}^2}\sqrt{4\sup_{f\in\mathcal{F}_j}\frac{1}{n}\sum_{i=1}^n\left[f(w_i^T\hat{\theta})-f(w_i^T\theta^*)\right]^2} + \frac{1}{n}\left|\sum_{i=1}^n v_{ij}\bar{\Delta}_{ij}(\theta^*)\right| \\
&\leq \sqrt{\frac{1}{n}\sum_{i=1}^n v_{ij}^2}\sqrt{4T_j^{\prime}} + \frac{1}{n}\left|\sum_{i=1}^n v_{ij}\bar{\Delta}_{ij}(\theta^*)\right|
\end{aligned}$$

For the term $\sqrt{\frac{1}{n}\sum_{i=1}^n v_{ij^{\prime}}^2}$, by Lemma B.1, we have

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n v_{ij^{\prime}}^2 \geq 2\sigma_{v_{j^{\prime}}}^2\right] \leq 2\exp(-cn).$$

For the term $\frac{1}{n}\left|\sum_{i=1}^n v_{ij}\bar{\Delta}_{ij}(\theta^*)\right|$, notice that the error function $\bar{\Delta}_j$ belongs to the set $\bar{\mathcal{F}}_j$. If $\left|\bar{\Delta}_j(\theta^*)\right|_n \geq u = \sqrt{tr_{nj}}$, an upper bound on the term $\frac{1}{n}\left|\sum_{i=1}^n v_{ij}\bar{\Delta}_{ij}(\theta^*)\right|$ can be obtained by Lemma B.3:

$$\begin{aligned}
\frac{1}{n}\left|\sum_{i=1}^n v_{ij}\bar{\Delta}_{ij}(\theta^*)\right| &\leq \sup_{f(\theta^*)\in\Omega(\left|\bar{\Delta}\right|_n;\mathcal{F}_j)}\left|\frac{1}{n}\sum_{i=1}^n v_{ij}f(w_i^T\theta^*)\right| \\
&\leq 2\left|\bar{\Delta}_j(\theta^*)\right|_n\sqrt{tr_{nj}},
\end{aligned}$$

with probability at least $1-\exp\left(-nC_j^*tr_{nj}\right)$, where $r_{nj}>0$ is the smallest positive quantity satisfying the *critical inequality*

$$\mathcal{G}_n\left(r_{nj};\,\mathcal{F}_j\right) \leq \frac{r_{nj}^2}{\sigma_{v_j}}.$$

Applying the triangle inequality and the Fenchel-Young inequality twice yields

$$\left|\bar{\Delta}_j(\theta^*)\right|_n\sqrt{tr_{nj}} \leq 8tr_{nj} + \frac{1}{16}\left(\left|\hat{\Delta}_j(\theta^*)\right|_n^2 + \left|\tilde{\Delta}_j(\theta^*)\right|_n^2\right).$$

Putting the pieces together, we obtain

$$\frac{1}{4}\left|\hat{\Delta}_j(\hat{\theta})\right|_n^2 \leq \frac{7}{4}\left|\tilde{\Delta}_j(\hat{\theta})\right|_n^2 + 32tr_{nj} + \sqrt{32\sigma_{v_j}^2 T_j^{'}} + 4T_j^{'}$$

where

$$T_j^{'} = \sup_{f\in\mathcal{F}_j} \frac{1}{n}\sum_{i=1}^{n}\left[f(w_i^T\hat{\theta}) - f(w_i^T\theta^*)\right]^2.$$

If $\left|\bar{\Delta}_j(\theta^*)\right|_n < \sqrt{tr_{nj}}$, then the claim follows trivially by triangle inequality and the Fenchel-Young inequality.□

Lemma A.12: For a user-defined radius $\bar{R}_j > 0$, let $r_{nj} > 0$ be the smallest positive quantity satisfying the *critical inequality*

$$\mathcal{G}_n\left(r_{nj};\, \mathcal{F}_j\right) \leq \frac{\bar{R}_j r_{nj}^2}{\sigma_{v_j}}, \tag{3.48}$$

where $\Omega(r_j;\, \mathcal{F}_j) = \left\{f \,:\, f \in \bar{\mathcal{F}}_j \; |f_{\theta^*}|_n \leq r_j,\, |f|_{\mathcal{F}_j} \leq 1\right\}$ for $j = 0, ..., p$. Suppose that we solve program (12) with $\lambda_{nj,2} \geq 2r_{nj}^2$. Let $\tilde{m}_j \in \mathcal{F}_j$ with $|\tilde{m}_j|_{\mathcal{F}_j} \leq \bar{R}_j$. Then there are universal positive constants (c_0, c_1, c_2, c_3) such that the non-parametric least squares estimate $\hat{m}_j$ via (12) satisfies

$$\left|\hat{m}_j(\hat{\theta}) - m_j(\hat{\theta})\right|_n^2 \leq c_0\left|\tilde{m}_j(\hat{\theta}) - m_j(\hat{\theta})\right|_n^2 + c_1\bar{R}_j^2\left\{r_{nj}^2 + \lambda_{nj,2}\right\} + c_2T_j^{'} + c_3\sqrt{\sigma_v^2 T_j^{'}}$$

with probability at least $1 - c\exp\left(-nC_j^*\bar{R}_j^2 r_{nj}^2\right)$.

Proof. Introduce the shorthand $\tilde{\sigma}_{v_j} = \frac{\sigma_{v_j}}{\bar{R}_j}$ and we work with an equivalent model with noise variance $\left(\frac{\sigma_{v_j}}{\bar{R}_j}\right)^2$ and the rescaled approximation error $\left|\tilde{m}_j(\hat{\theta}) - m_j(\hat{\theta})\right|_n^2$. Note that the final error then should be multiplied by $\bar{R}_j^2$. Now, since $\hat{m}_j$ is optimal and $\tilde{m}_j$ is feasible for program (12), we have

$$\frac{1}{2}\sum_{i=1}^{n}\left(z_{ij} - \hat{m}_j(w_i^T\hat{\theta})\right)^2 + \lambda_{nj,2}\left|\hat{m}_j\right|_{\mathcal{F}_j}^2 \leq \frac{1}{2}\sum_{i=1}^{n}\left(z_{ij} - \tilde{m}_j(w_i^T\hat{\theta})\right)^2 + \lambda_{nj,2}\left|\tilde{m}_j\right|_{\mathcal{F}_j}^2.$$

Recall the definitions

$$\begin{aligned}
\hat{\Delta}_{ij}(\theta) &:= (\hat{m}_j - m_j) \circ (w_i^T\theta), \\
\tilde{\Delta}_{ij}(\theta) &:= (\tilde{m}_j - m_j) \circ (w_i^T\theta), \\
\bar{\Delta}_{ij}(\theta) &:= (\hat{m}_j - \tilde{m}_j) \circ (w_i^T\theta), \\
\left\{\hat{\Delta}_{ij}(\theta)\right\}_{i=1}^n &:= \hat{\Delta}_j(\theta), \\
\left\{\tilde{\Delta}_{ij}(\theta)\right\}_{i=1}^n &:= \tilde{\Delta}_j(\theta), \\
\left\{\bar{\Delta}_{ij}(\theta)\right\}_{i=1}^n &:= \bar{\Delta}_j(\theta).
\end{aligned}$$

Performing some algebra yields the modified basic inequality

$$\begin{aligned}
\frac{1}{2}\left|\hat{\Delta}_j(\hat{\theta})\right|_n^2 &\leq \frac{1}{2}\left|\tilde{\Delta}_j(\hat{\theta})\right|_n^2 + \frac{1}{n}\left|\sum_{i=1}^n (v_{ij} + T_{ij})\,\bar{\Delta}_{ij}(\hat{\theta})\right| + \lambda_{nj,2}\left\{|\tilde{m}_j|_{\mathcal{F}_j}^2 - |\hat{m}_j|_{\mathcal{F}_j}^2\right\} \\
&\leq \frac{1}{2}\left|\tilde{\Delta}_j(\hat{\theta})\right|_n^2 + \frac{1}{n}\left|\sum_{i=1}^n v_{ij}\bar{\Delta}_{ij}(\hat{\theta})\right| + \frac{1}{n}\left|\sum_{i=1}^n T_{ij}\bar{\Delta}_{ij}(\hat{\theta})\right| \\
&\quad + \lambda_{nj,2}\left\{|\tilde{m}_j|_{\mathcal{F}_j}^2 - |\hat{m}_j|_{\mathcal{F}_j}^2\right\}. \qquad (3.49)
\end{aligned}$$

Furthermore, we have

$$\begin{aligned}
\frac{1}{n}\left|\sum_{i=1}^n T_{ij}\bar{\Delta}_{ij}(\hat{\theta})\right| &\leq \frac{1}{n}\left|\sum_{i=1}^n T_{ij}\hat{\Delta}_{ij}(\hat{\theta})\right| + \frac{1}{n}\left|\sum_{i=1}^n T_{ij}\tilde{\Delta}_{ij}(\hat{\theta})\right| \\
&\leq \frac{2}{n}\left|\sum_{i=1}^n T_{ij}^2\right| + \frac{1}{4n}\left|\sum_{i=1}^n \hat{\Delta}_{ij}^2(\hat{\theta})\right| + \frac{1}{4n}\left|\sum_{i=1}^n \tilde{\Delta}_{ij}^2(\hat{\theta})\right|,
\end{aligned}$$

where the first inequality uses the triangle inequality and the second uses the Fenchel-Young inequality. Substituting the inequality into (3.49), we obtain

$$\frac{1}{4}\left|\hat{\Delta}_j(\hat{\theta})\right|_n^2 \leq \frac{3}{4}\left|\tilde{\Delta}_j(\hat{\theta})\right|_n^2 + \frac{1}{n}\left|\sum_{i=1}^n v_{ij}\bar{\Delta}_{ij}(\hat{\theta})\right| + \frac{2}{n}\left|\sum_{i=1}^n T_{ij}^2\right| + \lambda_{nj,2}\left\{|\tilde{m}_j|_{\mathcal{F}_j}^2 - |\hat{m}_j|_{\mathcal{F}_j}^2\right\}. \quad (3.50)$$

The remainder of the proof can be divided into two cases.
Case (i): $|\hat{m}_j|_{\mathcal{F}_j} \leq 2$. In this case, (3.50) implies that

$$\frac{1}{4}\left|\hat{\Delta}_j\right|_n^2 \leq \frac{3}{4}\left|\tilde{\Delta}_j\right|_n^2 + \frac{1}{n}\left|\sum_{i=1}^n v_{ij}\bar{\Delta}_{ij}(\hat{\theta})\right| + \frac{2}{n}\left|\sum_{i=1}^n T_{ij}^2\right| + 9\lambda_{nj,2},$$

using the fact that $|\tilde{m}_j|_{\mathcal{F}_j} \leq 1$. Since $|\hat{m}_j|_{\mathcal{F}_j} \leq 2$, we have $\left|\bar{\Delta}_j\right|_{\mathcal{F}_j} \leq 3$ and hence

$$|\tilde{m}_j|^2_{\mathcal{F}_j} - |\hat{m}_j|^2_{\mathcal{F}_j} = \left(|\tilde{m}_j|_{\mathcal{F}_j} + |\hat{m}_j|_{\mathcal{F}_j}\right)\left(|\tilde{m}_j|_{\mathcal{F}_j} - |\hat{m}_j|_{\mathcal{F}_j}\right) \leq 9$$

by the triangle inequality.

We need to upper bound the stochastic process $\frac{1}{n}\left|\sum_{i=1}^n v_{ij}\bar{\Delta}_{ij}(\hat{\theta})\right|$.

$$\begin{aligned}
\frac{1}{n}\left|\sum_{i=1}^n v_{ij}\bar{\Delta}_{ij}(\hat{\theta})\right| &\leq \frac{1}{n}\left|\sum_{i=1}^n v_{ij}\left[\bar{\Delta}_{ij}(\hat{\theta}) - \bar{\Delta}_{ij}(\theta^*)\right]\right| + \frac{1}{n}\left|\sum_{i=1}^n v_{ij}\bar{\Delta}_{ij}(\theta^*)\right| \\
&\leq \sqrt{\frac{1}{n}\sum_{i=1}^n v_{ij}^2}\sqrt{\frac{1}{n}\sum_{i=1}^n \left[\bar{\Delta}_{ij}(\hat{\theta}) - \bar{\Delta}_{ij}(\theta^*)\right]^2} + \frac{1}{n}\left|\sum_{i=1}^n v_{ij}\bar{\Delta}_{ij}(\theta^*)\right| \\
&\leq \sqrt{\frac{1}{n}\sum_{i=1}^n v_{ij}^2}\sqrt{4\sup_{f\in\mathcal{F}_j}\frac{1}{n}\sum_{i=1}^n \left[f(w_i^T\hat{\theta}) - f(w_i^T\theta^*)\right]^2} + \frac{1}{n}\left|\sum_{i=1}^n v_{ij}\bar{\Delta}_{ij}(\theta^*)\right| \\
&\leq \sqrt{\frac{1}{n}\sum_{i=1}^n v_{ij}^2}\sqrt{4T_j'} + \frac{1}{n}\left|\sum_{i=1}^n v_{ij}\bar{\Delta}_{ij}(\theta^*)\right|.
\end{aligned}$$

Now, applying Lemma B.4 yields

$$\frac{1}{n}\left|\sum_{i=1}^n v_{ij}\bar{\Delta}_{ij}(\theta^*)\right| \leq 6r_{nj}\left|\bar{\Delta}_j(\theta^*)\right|_n + \frac{1}{32}\left|\bar{\Delta}_j(\theta^*)\right|_n^2 \tag{3.51}$$

with probability at least $1 - c\exp\left(-nC_j^*\bar{R}_j^2 r_{nj}^2\right)$. Consequently,

$$\begin{aligned}
\frac{1}{4}\left|\hat{\Delta}_j(\hat{\theta})\right|_n^2 &\leq \frac{3}{4}\left|\tilde{\Delta}_j(\hat{\theta})\right|_n^2 + 6r_{nj}\left|\bar{\Delta}_j(\hat{\theta})\right|_n + \frac{1}{32}\left|\bar{\Delta}_j(\hat{\theta})\right|_n^2 + \frac{2}{n}\left|\sum_{i=1}^n T_{ij}^2\right| + 9\lambda_{nj,2} \\
&\leq \frac{3}{4}\left|\tilde{\Delta}_j(\hat{\theta})\right|_n^2 + 216r_{nj}^2 + \frac{1}{12}\left|\hat{\Delta}_j(\hat{\theta})\right|_n^2 + \frac{1}{12}\left|\tilde{\Delta}_j(\hat{\theta})\right|_n^2 \\
&\quad + \frac{1}{16}\left|\hat{\Delta}_j(\hat{\theta})\right|_n^2 + \frac{1}{16}\left|\tilde{\Delta}_j(\hat{\theta})\right|_n^2 + \frac{2}{n}\left|\sum_{i=1}^n T_{ij}^2\right| + 9\lambda_{nj,2},
\end{aligned}$$

where the second inequality follows by the triangle inequality, and the Fenchel-Young inequality. This inequality above implies

$$\frac{5}{48}\left|\hat{\Delta}_j(\hat{\theta})\right|_n^2 \leq \frac{43}{48}\left|\tilde{\Delta}_j(\hat{\theta})\right|_n^2 + 216r_{nj}^2 + \frac{2}{n}\left|\sum_{i=1}^n T_{ij}^2\right| + 9\lambda_{nj,2}.$$

Case (ii): $|\hat{m}_j|_{\mathcal{F}_j} > 2 > 1 \geq |\tilde{m}_j|_{\mathcal{F}_j}$. In this case, we have

$$\begin{aligned} |\tilde{m}_j|^2_{\mathcal{F}_j} - |\hat{m}_j|^2_{\mathcal{F}_j} &= \left(|\tilde{m}_j|_{\mathcal{F}_j} + |\hat{m}_j|_{\mathcal{F}_j}\right)\left(|\tilde{m}_j|_{\mathcal{F}_j} - |\hat{m}_j|_{\mathcal{F}_j}\right) \\ &\leq |\tilde{m}_j|_{\mathcal{F}_j} - |\hat{m}_j|_{\mathcal{F}_j}, \end{aligned}$$

since $|\tilde{m}_j|_{\mathcal{F}_j} - |\hat{m}_j|_{\mathcal{F}_j} < 0$. Writing $\hat{m}_j = \tilde{m}_j + \bar{\Delta}_j$ and applying the triangle inequality $|\hat{m}_j|_{\mathcal{F}_j} \geq \left|\bar{\Delta}_j\right|_{\mathcal{F}_j} - |\tilde{m}_j|_{\mathcal{F}_j}$ yields

$$\begin{aligned} \lambda_{nj,2}\left\{|\tilde{m}_j|^2_{\mathcal{F}_j} - |\hat{m}_j|^2_{\mathcal{F}_j}\right\} &\leq \lambda_{nj,2}\left\{2\,|\tilde{m}_j|_{\mathcal{F}_j} - \left|\bar{\Delta}_j\right|_{\mathcal{F}_j}\right\} \\ &\leq 2\lambda_{nj,2} - \lambda_{nj,2}\left|\bar{\Delta}_j\right|_{\mathcal{F}_j}. \end{aligned}$$

Substituting this upper bound into (3.50) yields

$$\begin{aligned} \frac{1}{4}\left|\hat{\Delta}_j(\hat{\theta})\right|^2_n &\leq \frac{3}{4}\left|\tilde{\Delta}_j(\hat{\theta})\right|^2_n + \frac{1}{n}\left|\sum_{i=1}^n v_{ij}\bar{\Delta}_{ij}(\hat{\theta})\right| + \frac{2}{n}\left|\sum_{i=1}^n T^2_{ij}\right| + 2\lambda_{nj,2} - \lambda_{nj,2}\left|\bar{\Delta}_j\right|_{\mathcal{F}_j} \\ &\leq \frac{3}{4}\left|\tilde{\Delta}_j(\hat{\theta})\right|^2_n + \frac{1}{n}\left|\sum_{i=1}^n v_{ij}\bar{\Delta}_{ij}(\theta^*)\right| + \sqrt{\frac{1}{n}\sum_{i=1}^n v^2_{ij}}\sqrt{4T'_j} \qquad (3.52) \\ &\quad + \frac{2}{n}\left|\sum_{i=1}^n T^2_{ij}\right| + 2\lambda_{nj,2} - \lambda_{nj,2}\left|\bar{\Delta}_j\right|_{\mathcal{F}_j} \end{aligned}$$

Lemma B.5 upper bounds the stochastic component $\frac{1}{n}\left|\sum_{i=1}^n v_{ij}\bar{\Delta}_{ij}(\theta^*)\right|$ in inequality (3.52): There are universal positive constants $(c_1,\, c_2)$ such that for all $|\Delta_j|_{\mathcal{F}_j} \geq 1$,

$$\frac{1}{n}\left|\sum_{i=1}^n v_{ij}\Delta_{ij}(\theta^*)\right| \leq 2r_{nj}\,|\Delta_j(\theta^*)|_n + 2r^2_{nj}\,|\Delta_j|_{\mathcal{F}_j} + \frac{1}{16}\,|\Delta_j(\theta^*)|^2_n \qquad (3.53)$$

with probability at least $1 - c\exp\left(-nC^*_j\bar{R}^2_j r^2_{nj}\right)$. We now complete the proof using inequality (3.53). Since $|\hat{m}_j|_{\mathcal{F}_j} > 2 > 1 \geq |\tilde{m}_j|_{\mathcal{F}_j}$, the triangle inequality implies that

$\left|\bar{\Delta}_j\right|_{\mathcal{F}_j} \geq |\hat{m}_j|_{\mathcal{F}_j} - |\tilde{m}_j|_{\mathcal{F}_j} > 1$, so that inequality (3.53) can be applied. Consequently,

$$\begin{aligned}\frac{1}{4}\left|\hat{\Delta}_j(\hat{\theta})\right|_n^2 &\leq \frac{3}{4}\left|\tilde{\Delta}_j(\hat{\theta})\right|_n^2 + 2r_{nj}\left|\bar{\Delta}_j(\hat{\theta})\right|_n + \left(2r_{nj}^2 - \lambda_{nj,2}\right)\left|\bar{\Delta}_j\right|_{\mathcal{F}_j} \\ &\quad + \frac{1}{16}\left|\bar{\Delta}_j(\hat{\theta})\right|_n^2 + \sqrt{\frac{1}{n}\sum_{i=1}^n v_{ij}^2}\sqrt{4T_j'} + \frac{2}{n}\left|\sum_{i=1}^n T_{ij}^2\right| + 2\lambda_{nj,2} \\ &\leq \frac{3}{4}\left|\tilde{\Delta}_j(\hat{\theta})\right|_n^2 + 32r_{nj}^2 + \frac{1}{16}\left|\hat{\Delta}_j(\hat{\theta})\right|_n^2 + \frac{1}{16}\left|\tilde{\Delta}_j(\hat{\theta})\right|_n^2 \\ &\quad + \frac{1}{8}\left|\hat{\Delta}_j(\hat{\theta})\right|_n^2 + \frac{1}{8}\left|\tilde{\Delta}_j(\hat{\theta})\right|_n^2 + \sqrt{\frac{1}{n}\sum_{i=1}^n v_{ij}^2}\sqrt{4T_j'} + \frac{2}{n}\left|\sum_{i=1}^n T_{ij}^2\right| + 2\lambda_{nj,2},\end{aligned}$$

where the second inequality follows by the choice of $\lambda_{nj,2}$, the triangle inequality, and the Fenchel-Young inequality. This inequality above implies

$$\frac{1}{16}\left|\hat{\Delta}_j(\hat{\theta})\right|_n^2 \leq \frac{15}{16}\left|\tilde{\Delta}_j(\hat{\theta})\right|_n^2 + 32r_{nj}^2 + \sqrt{\frac{1}{n}\sum_{i=1}^n v_{ij}^2}\sqrt{4T_j'} + \frac{2}{n}\left|\sum_{i=1}^n T_{ij}^2\right| + 2\lambda_{nj,2}.$$

□

3.8.15 Lemmas for Theorem 3.4.7

Lemma A.13: If the assumptions in Corollary 3.4.4 and Assumption 3.4.7 hold, then under the conditions $n \gtrsim (k_2^3 \log p) \vee (k_2^2 k_1 \log d)$, $\sqrt{\frac{n \log p}{|\beta^*|_1^2}} \vee \sqrt{n k_1 \log d} \gtrsim \log p$, and $\sqrt{\frac{k_1 \log d}{n}} = o(1)$, the sample matrix $\frac{1}{n}\hat{v}^T\hat{v}$ satisfies an analogous version of the "mutual incoherence" assumption with high probability, i.e.,

$$\mathbb{P}\left[\left\|\frac{1}{n}\hat{v}_{K^c}^T\hat{v}_K\left(\frac{1}{n}\hat{v}_K^T\hat{v}_K\right)^{-1}\right\|_\infty \geq 1 - \frac{\phi}{4}\right] \leq O\left(\frac{1}{p \wedge d}\right).$$

Proof. We use the following decomposition similar to the method used in Ravikumar, et. al. (2010)

$$\check{\Sigma}_{K^cK}\check{\Sigma}_{KK}^{-1} - \Sigma_{K^cK}\Sigma_{KK}^{-1} = R_1 + R_2 + R_3 + R_4 + R_5 + R_6,$$

where

$$\begin{aligned} R_1 &= \Sigma_{K^cK}[\hat{\Sigma}_{KK}^{-1} - \Sigma_{KK}^{-1}], \\ R_2 &= [\hat{\Sigma}_{K^cK} - \Sigma_{K^cK}]\Sigma_{KK}^{-1}, \\ R_3 &= [\hat{\Sigma}_{K^cK} - \Sigma_{K^cK}][\hat{\Sigma}_{KK}^{-1} - \Sigma_{KK}^{-1}], \\ R_4 &= \hat{\Sigma}_{K^cK}[\check{\Sigma}_{KK}^{-1} - \hat{\Sigma}_{KK}^{-1}], \\ R_5 &= [\check{\Sigma}_{K^cK} - \hat{\Sigma}_{K^cK}]\hat{\Sigma}_{KK}^{-1}, \\ R_6 &= [\check{\Sigma}_{K^cK} - \hat{\Sigma}_{K^cK}][\check{\Sigma}_{KK}^{-1} - \hat{\Sigma}_{KK}^{-1}]. \end{aligned}$$

By Assumption 3.4.7, we have

$$\left\| \Sigma_{K^cK}\Sigma_{KK}^{-1} \right\|_\infty \le 1 - \phi.$$

It suffices to show that $||R_i||_\infty \le \frac{\phi}{6}$ for $i = 1, ..., 3$ and $||R_i||_\infty \le \frac{\phi}{12}$ for $i = 4, ..., 6$.

For the first term R_1, we have

$$R_1 = -\Sigma_{K^cK}\Sigma_{KK}^{-1}[\hat{\Sigma}_{KK} - \Sigma_{KK}]\hat{\Sigma}_{KK}^{-1},$$

Using the sub-multiplicative property $||AB||_\infty \le ||A||_\infty ||B||_\infty$ and the elementary inequality $||A||_\infty \le \sqrt{m}||A||_2$ for any symmetric matrix $A \in \mathbb{R}^{m\times m}$, we can bound R_1 as follows:

$$\begin{aligned} ||R_1||_\infty &\le \left\| \Sigma_{K^cK}\Sigma_{KK}^{-1} \right\|_\infty \left\| \hat{\Sigma}_{KK} - \Sigma_{KK} \right\|_\infty \left\| \hat{\Sigma}_{KK}^{-1} \right\|_\infty \\ &\le (1-\phi) \left\| \hat{\Sigma}_{KK} - \Sigma_{KK} \right\|_\infty \sqrt{k_2} \left\| \hat{\Sigma}_{KK}^{-1} \right\|_2, \end{aligned}$$

where the last inequality follows from Assumption 3.4.7. Using bound (3.60) from the proof for Lemma A.14, we have

$$\left\| \hat{\Sigma}_{KK}^{-1} \right\|_2 \le \frac{2}{\lambda_{\min}(\Sigma_{KK})}$$

with probability at least $1 - c_1 \exp(-c_2 n)$. Next, applying bound (3.55) from Lemma A.14 with $\varepsilon = \frac{\phi\lambda_{\min}(\Sigma_{KK})}{12(1-\phi)\sqrt{k_2}}$, we have

$$\mathbb{P}\left[\left\| \hat{\Sigma}_{KK} - \Sigma_{KK} \right\|_\infty \ge \frac{\phi\lambda_{\min}(\Sigma_{KK})}{12(1-\phi)\sqrt{k_2}} \right] \le 2\exp(-b\frac{n}{k_2^3} + 2\log k_2).$$

Then, we are guaranteed that

$$\mathbb{P}[||R_1||_\infty \ge \frac{\phi}{6}] \le 2\exp(-b\frac{n}{k_2^3} + 2\log k_2).$$

For the second term R_2, we first write

$$\begin{aligned} ||R_2||_\infty &\leq \sqrt{k_2}\left\|\Sigma_{KK}^{-1}\right\|_2\left\|\hat{\Sigma}_{K^cK}-\Sigma_{K^cK}\right\|_\infty \\ &\leq \frac{\sqrt{k_2}}{\lambda_{\min}(\Sigma_{KK})}\left\|\hat{\Sigma}_{K^cK}-\Sigma_{K^cK}\right\|_\infty. \end{aligned}$$

An application of bound (3.54) from Lemma A.14 with $\varepsilon = \frac{\phi}{6}\frac{\lambda_{\min}(\Sigma_{KK})}{\sqrt{k_2}}$ to bound the term $\left\|\hat{\Sigma}_{K^cK}-\Sigma_{K^cK}\right\|_\infty$ yields

$$\mathbb{P}[||R_2||_\infty \geq \frac{\phi}{6}] \leq 2\exp(-b\frac{n}{k_2^3}+\log(p-k_2)+\log k_2).$$

For the third term R_3, by applying bounds (3.54) from Lemma A.14 with $\varepsilon = \frac{\phi\lambda_{\min}(\Sigma_{KK})}{6}$ to bound the term $\left\|\hat{\Sigma}_{K^cK}-\Sigma_{K^cK}\right\|_\infty$ and (3.56) from Lemma A.14 to bound the term $\left\|\hat{\Sigma}_{KK}^{-1}-\Sigma_{KK}^{-1}\right\|_\infty$, we have

$$\mathbb{P}[||R_3||_\infty \geq \frac{\phi}{6}] \leq 2\exp(-b\frac{n}{k_2^2}+\log(p-k_2)+\log k_2).$$

Putting everything together, we conclude that

$$\mathbb{P}[||\hat{\Sigma}_{K^cK}\hat{\Sigma}_{KK}^{-1}||_\infty \geq 1-\frac{\phi}{2}] \leq c\exp(-b\frac{n}{k_2^3}+2\log p).$$

For the fourth term R_4, we have, with probability at least $1-c\exp(-bn\min\{\frac{1}{k_2^3},\frac{1}{k_2^{3/2}}\}+2\log p)$,

$$\begin{aligned} ||R_4||_\infty &\leq \left\|\hat{\Sigma}_{K^cK}\hat{\Sigma}_{KK}^{-1}\right\|_\infty\left\|\check{\Sigma}_{KK}-\hat{\Sigma}_{KK}\right\|_\infty\left\|\check{\Sigma}_{KK}^{-1}\right\|_\infty \\ &\leq (1-\frac{\phi}{2})\left\|\check{\Sigma}_{KK}-\hat{\Sigma}_{KK}\right\|_\infty\sqrt{k_2}\left\|\check{\Sigma}_{KK}^{-1}\right\|_2, \end{aligned}$$

where the last inequality follows from the bound on $||\hat{\Sigma}_{K^cK}\hat{\Sigma}_{KK}^{-1}||_\infty$ established previously. Using bounds (3.63) from the proof for Lemma A.15, we have

$$\left\|\check{\Sigma}_{KK}^{-1}\right\|_2 \leq \frac{4}{\lambda_{\min}(\Sigma_{KK})}$$

with probability at least $1-c_1\exp(-c_2\log(p\vee d))$. Next, applying Lemma A.15 with $\varepsilon = \frac{\phi\lambda_{\min}(\Sigma_{KK})}{48(1-\frac{\phi}{2})\sqrt{k_2}}$ to bound the term $\left\|\check{\Sigma}_{KK}-\hat{\Sigma}_{KK}\right\|_\infty$ yields,

$$\mathbb{P}[||R_4||_\infty \geq \frac{\phi}{12}] \leq ck_2^2\cdot\exp\left(-\frac{c_1 n}{k_2^{\frac{3}{2}}}\right)+c_2\exp(-c_3\log d)$$

For the fifth term R_5, using bound (3.60) from the proof for Lemma A.14, we have

$$\begin{aligned} ||R_5||_\infty &\leq \sqrt{k_2} \left\|\hat{\Sigma}_{KK}^{-1}\right\|_2 \left\|\check{\Sigma}_{K^cK} - \hat{\Sigma}_{K^cK}\right\|_\infty \\ &\leq \frac{2\sqrt{k_2}}{\lambda_{\min}(\Sigma_{KK})} \left\|\check{\Sigma}_{K^cK} - \hat{\Sigma}_{K^cK}\right\|_\infty . \end{aligned}$$

Applying Lemma A.15 with $\varepsilon = \frac{\phi\lambda_{\min}(\Sigma_{KK})}{24\sqrt{k_2}}$ to bound the term $\left\|\check{\Sigma}_{K^cK} - \hat{\Sigma}_{K^cK}\right\|_\infty$ yields

$$\mathbb{P}[||R_5||_\infty \geq \frac{\phi}{12}] \leq c(p-k_2)k_2 \cdot \exp\left(-\frac{c_1 n}{k_2^{\frac{3}{2}}}\right) + c_2\exp(-c_3 \log d)$$

For the sixth term R_6, by applying Lemma A.15 to bound the terms $\left\|\check{\Sigma}_{K^cK} - \hat{\Sigma}_{K^cK}\right\|_\infty$ and $\left\|\check{\Sigma}_{KK}^{-1} - \hat{\Sigma}_{KK}^{-1}\right\|_\infty$ respectively, with $\varepsilon = \frac{\phi}{12}\frac{\lambda_{\min}(\Sigma_{KK})}{8}$, we are guaranteed that

$$\mathbb{P}[||R_6||_\infty \geq \frac{\phi}{12}] \leq c(p-k_2)k_2 \cdot \exp\left(-\frac{c_1 n}{k_2}\right) + c_2\exp(-c_3 \log d)$$

Under the conditions $n \gtrsim (k_2^3 \log p) \vee (k_2^2 k_1 \log d)$, $\sqrt{\frac{n\log p}{|\beta^*|_1^2}} \vee \sqrt{nk_1 \log d} \gtrsim \log p$, and $\sqrt{\frac{k_1 \log d}{n}} = o(1)$, putting the bounds on $R_1 - R_6$ together, we conclude that

$$\mathbb{P}[||\check{\Sigma}_{K^cK}\check{\Sigma}_{KK}^{-1}||_\infty \geq 1 - \frac{\phi}{4}] \leq O\left(\frac{1}{p \wedge d}\right).$$

□

Lemma A.14: Suppose Assumptions 3.2.1 and 3.4.2 hold. For any $\varepsilon > 0$ and constant c, we have

$$\mathbb{P}\left\{\left\|\hat{\Sigma}_{K^cK} - \Sigma_{K^cK}\right\|_\infty \geq \varepsilon\right\} \leq (p-k_2)k_2 \cdot 2\exp(-cn\min\{\frac{\varepsilon^2}{4k_2^2\sigma_v^4}, \frac{\varepsilon}{2k_2\sigma_v^2}\}), \tag{3.54}$$

$$\mathbb{P}\left\{\left\|\hat{\Sigma}_{KK} - \Sigma_{KK}\right\|_\infty \geq \varepsilon\right\} \leq k_2^2 \cdot 2\exp(-cn\min\{\frac{\varepsilon^2}{4k_2^2\sigma_v^4}, \frac{\varepsilon}{2k_2\sigma_v^2}\}). \tag{3.55}$$

Furthermore, under the scaling $n \gtrsim k_2 \log p$, for constants b_1 and b_2, we have

$$\left\|\hat{\Sigma}_{KK}^{-1} - \Sigma_{KK}^{-1}\right\|_\infty \leq \frac{1}{\lambda_{\min}(\Sigma_{KK})}, \tag{3.56}$$

with probability at least $1 - c_1 \exp(-c_2 n \min\{\frac{\lambda^2_{\min}(\Sigma_{KK})}{4k_2\sigma_v^4}, \frac{\lambda_{\min}(\Sigma_{KK})}{2\sqrt{k_2}\sigma_v^2}\})$.

Proof. Denote the element (j', j) of the matrix difference $\hat{\Sigma}_{K^cK} - \Sigma_{K^cK}$ by $\omega_{j'j}$. By the definition of the l_∞matrix norm, we have

$$\begin{aligned}
\mathbb{P}\left\{\left\|\hat{\Sigma}_{K^cK} - \Sigma_{K^cK}\right\|_\infty \geq \varepsilon\right\} &= \mathbb{P}\left\{\max_{j' \in K^c} \sum_{j \in K} |\omega_{j'j}| \geq \varepsilon\right\} \\
&\leq (p - k_2)\mathbb{P}\left\{\sum_{j \in K} |\omega_{j'j}| \geq \varepsilon\right\} \\
&\leq (p - k_2)\mathbb{P}\left\{\exists j \in K \,|\, |\omega_{j'j}| \geq \frac{\varepsilon}{k_2}\right\} \\
&\leq (p - k_2)k_2\mathbb{P}\left\{|\omega_{j'j}| \geq \frac{\varepsilon}{k_2}\right\} \\
&\leq (p - k_2)k_2 \cdot 2\exp(-cn \min\{\frac{\varepsilon^2}{k_2^2\sigma_v^4}, \frac{\varepsilon}{k_2\sigma_v^2}\}),
\end{aligned}$$

where the last inequality follows the deviation bound for sub-exponential random variables, i.e., Lemma B.1. Bound (3.55) can be obtained in a similar way except that the pre-factor $(p - k_2)$ is replaced by k_2. To prove the last bound (3.56), write

$$\begin{aligned}
\left\|\hat{\Sigma}^{-1}_{KK} - \Sigma^{-1}_{KK}\right\|_\infty &= \left\|\Sigma^{-1}_{KK}\left[\Sigma_{KK} - \hat{\Sigma}_{KK}\right]\hat{\Sigma}^{-1}_{KK}\right\|_\infty \\
&= \sqrt{k_2}\left\|\Sigma^{-1}_{KK}\left[\Sigma_{KK} - \hat{\Sigma}_{KK}\right]\hat{\Sigma}^{-1}_{KK}\right\|_2 \\
&= \sqrt{k_2}\left\|\Sigma^{-1}_{KK}\right\|_2\left\|\Sigma_{KK} - \hat{\Sigma}_{KK}\right\|_2\left\|\hat{\Sigma}^{-1}_{KK}\right\|_2 \\
&\leq \frac{\sqrt{k_2}}{\lambda_{\min}(\Sigma_{KK})}\left\|\Sigma_{KK} - \hat{\Sigma}_{KK}\right\|_2\left\|\hat{\Sigma}^{-1}_{KK}\right\|_2. \qquad (3.57)
\end{aligned}$$

To bound the term $\left\|\Sigma_{KK} - \hat{\Sigma}_{KK}\right\|_2$ in (3.57), applying Lemma B.1 with $v^Tv = \hat{\Sigma}_{KK}$ and $t = \frac{\lambda_{\min}(\Sigma_{KK})}{2\sqrt{k_2}}$ yields

$$\left\|\hat{\Sigma}_{KK} - \Sigma_{KK}\right\|_2 \leq \frac{\lambda_{\min}(\Sigma_{KK})}{2\sqrt{k_2}},$$

with probability at least $1 - c_1 \exp(-c_2 n \min\{\frac{\lambda^2_{\min}(\Sigma_{KK})}{4k_2\sigma_v^4}, \frac{\lambda_{\min}(\Sigma_{KK})}{2\sqrt{k_2}\sigma_v^2}\})$.

To bound the term $\left\|\hat{\Sigma}_{KK}^{-1}\right\|_2$ in (3.57), note that we can write

$$\begin{aligned}\lambda_{\min}(\Sigma_{KK}) &= \min_{\|h'\|_2=1} h'^T \Sigma_{KK} h' \\ &= \min_{\|h'\|_2=1} \left[h'^T \hat{\Sigma}_{KK} h' + h'^T (\Sigma_{KK} - \hat{\Sigma}_{KK}) h'\right] \\ &\leq h^T \hat{\Sigma}_{KK} h + h^T (\Sigma_{KK} - \hat{\Sigma}_{KK}) h \qquad (3.58)\end{aligned}$$

where $h \in \mathbb{R}^{k_2}$ is a unit-norm minimal eigenvector of $\hat{\Sigma}_{KK}$. Applying Lemma B.1 yields

$$\left|h^T \left(\Sigma_{KK} - \hat{\Sigma}_{KK}\right) h\right| \leq \frac{\lambda_{\min}(\Sigma_{KK})}{2}$$

with probability at least $1 - c_1 \exp(-c_2 n)$. Therefore,

$$\begin{aligned}\lambda_{\min}(\Sigma_{KK}) &\leq \lambda_{\min}(\hat{\Sigma}_{KK}) + \frac{\lambda_{\min}(\Sigma_{KK})}{2} \\ \Longrightarrow \lambda_{\min}(\hat{\Sigma}_{KK}) &\geq \frac{\lambda_{\min}(\Sigma_{KK})}{2}, \qquad (3.59)\end{aligned}$$

and consequently,

$$\left\|\hat{\Sigma}_{KK}^{-1}\right\|_2 \leq \frac{2}{\lambda_{\min}(\Sigma_{KK})}. \qquad (3.60)$$

Putting everything together, we have

$$\left\|\hat{\Sigma}_{KK}^{-1} - \Sigma_{KK}^{-1}\right\|_\infty \leq \frac{\sqrt{k_2}}{\lambda_{\min}(\Sigma_{KK})} \frac{\lambda_{\min}(\Sigma_{KK})}{2\sqrt{k_2}} \frac{2}{\lambda_{\min}(\Sigma_{KK})} = \frac{1}{\lambda_{\min}(\Sigma_{KK})}.$$

with probability at least $1 - c_1 \exp(-c_2 n \min\{\frac{\lambda_{\min}^2(\Sigma_{KK})}{4k_2\sigma_v^4}, \frac{\lambda_{\min}(\Sigma_{KK})}{2\sqrt{k_2}\sigma_v^2}\})$. $\square$

Lemma A.15: Suppose the assumptions in Lemma A.13 hold. For any $\varepsilon > 0$, under the condition $n \gtrsim (k_2^3 \log p) \vee (k_2^2 k_1 \log d)$, $\sqrt{\frac{n \log p}{|\beta^*|_2}} \vee \sqrt{n k_1 \log d} \gtrsim \log p$, and $\sqrt{\frac{k_1 \log d}{n}} = o(1)$, we have

$$\mathbb{P}\left\{\left\|\check{\Sigma}_{K^cK} - \hat{\Sigma}_{K^cK}\right\|_\infty \geq \varepsilon\right\} \leq c(p - k_2)k_2 \cdot \exp\left(-\frac{n\varepsilon}{c_1 k_2}\right) + c_2 \exp(-c_3 \log d)$$

$$\mathbb{P}\left\{\left\|\check{\Sigma}_{KK} - \hat{\Sigma}_{KK}\right\|_\infty \geq \varepsilon\right\} \leq c' k_2^2 \cdot \exp\left(-\frac{n\varepsilon}{c_1 k_2}\right) + c_2 \exp(-c_3 \log d)$$

Furthermore, we have

$$\left\|\check{\Sigma}_{KK}^{-1} - \hat{\Sigma}_{KK}^{-1}\right\|_\infty \leq \frac{8}{\lambda_{\min}(\Sigma_{KK})} \quad \text{with probability at least } 1 - c_1 \exp(-c_2 \log(p \wedge d)). \tag{3.61}$$

Proof. Denote the element $(j^{'}, j)$ of the matrix difference $\check{\Sigma}_{K^cK} - \hat{\Sigma}_{K^cK}$ by $\omega_{j^{'}j}$. Using the same argument as in the proof for Lemma A.14, we have

$$\mathbb{P}\left\{\left\|\check{\Sigma}_{K^cK} - \hat{\Sigma}_{K^cK}\right\|_\infty \geq \varepsilon\right\} \leq (p-k_2)k_2\mathbb{P}\left\{|\omega_{j'j}| \geq \frac{\varepsilon}{k_2}\right\}.$$

Following the derivation of the upper bounds on $\left|\frac{(\hat{x}-\tilde{x})^T\tilde{x}}{n}\right|_\infty$ and $\left|\frac{(\hat{x}-\tilde{x})^T(\hat{x}-\tilde{x})}{n}\right|_\infty$ in the proofs for Lemma A.3 and Corollary 3.4.4 and the identity

$$\frac{1}{n}\left(\check{\Sigma}_{K^cK} - \hat{\Sigma}_{K^cK}\right) = \frac{1}{n}v_{K^c}^T(\hat{v}_K - v_K) + \frac{1}{n}(\hat{v}_{K^c} - v_{K^c})^T v_K + \frac{1}{n}(\hat{v}_{K^c} - v_{K^c})^T(\hat{v}_K - v_K),$$

we notice that to upper bound $|\omega_{j'j}|$, it suffices to upper bound $3\cdot\left|\frac{1}{n}v_{j'}^T(\hat{v}_j - v_j)\right|$. From the proofs for Lemma A.3 and Corollary 3.4.4, we have

$$\left|\frac{1}{n}v_{j'}^T(\hat{v}_j - v_j)\right| \leq b\left(u^2 + \sqrt{\frac{k_1 \log d}{n}}\right),$$

with probability at least $1-c_1\exp\left(-nC_j^*u^2\right)-O\left(\frac{1}{d}\right)$. If $n \gtrsim k_2^2 k_1 \log d$, setting $u = \sqrt{\frac{\varepsilon}{bk_2}}$ for any $\varepsilon > 0$ yields

$$\mathbb{P}\left[|\omega_{j'j}| \geq \frac{\varepsilon}{k_2}\right] \leq c\exp\left(-\frac{n\varepsilon}{c_1 k_2}\right) + c_2\exp(-c_3 \log d).$$

Therefore,

$$\mathbb{P}\left\{\left\|\check{\Sigma}_{K^cK} - \hat{\Sigma}_{K^cK}\right\|_\infty \geq \varepsilon\right\} \leq c(p-k_2)k_2\cdot\exp\left(-\frac{n\varepsilon}{c_1 k_2}\right) + c_2\exp(-c_3\log d).$$

The second bound in Lemma A.15 can be obtained in a similar way except that the pre-factor $(p - k_2)$ is replaced by k_2.

To prove the third bound in Lemma A.15, by applying the same argument as in the proof for Lemma A.14, we have

$$\begin{aligned}\left\|\check{\Sigma}_{KK}^{-1} - \hat{\Sigma}_{KK}^{-1}\right\|_\infty &\leq \frac{\sqrt{k_2}}{\lambda_{\min}(\hat{\Sigma}_{KK})}\left\|\hat{\Sigma}_{KK} - \check{\Sigma}_{KK}\right\|_2\left\|\check{\Sigma}_{KK}^{-1}\right\|_2 \\ &\leq \frac{2\sqrt{k_2}}{\lambda_{\min}(\Sigma_{KK})}\left\|\hat{\Sigma}_{KK} - \check{\Sigma}_{KK}\right\|_2\left\|\check{\Sigma}_{KK}^{-1}\right\|_2,\end{aligned}$$

where the last inequality comes from bound (3.59).

For the term $\left\|\hat{\Sigma}_{KK} - \check{\Sigma}_{KK}\right\|_2$, setting $\varepsilon = \frac{\lambda_{\min}(\Sigma_{KK})}{\sqrt{k_2}}$ in the previous upper bound on $\left\|\hat{\Sigma}_{KK} - \check{\Sigma}_{KK}\right\|_\infty$ yields

$$\begin{aligned}\left\|\hat{\Sigma}_{KK} - \check{\Sigma}_{KK}\right\|_2 &\leq \left\|\hat{\Sigma}_{KK} - \check{\Sigma}_{KK}\right\|_\infty \\ &\leq \frac{\lambda_{\min}(\Sigma_{KK})}{\sqrt{k_2}}, \end{aligned} \tag{3.62}$$

with probability at least $1 - ck_2^2 \cdot \exp\left(-\frac{n}{c_1 k_2^{3/2}}\right) - c_2 \exp(-c_3 \log d)$.

To bound the term $\left\|\check{\Sigma}_{KK}^{-1}\right\|_2$, again we have,

$$\begin{aligned}\lambda_{\min}(\hat{\Sigma}_{KK}) &\leq h^T \check{\Sigma}_{KK} h + h^T(\hat{\Sigma}_{KK} - \check{\Sigma}_{KK})h \\ &\leq h^T \check{\Sigma}_{KK} h + k_2 \left|\hat{\Sigma}_{KK} - \check{\Sigma}_{KK}\right|_\infty \\ &\leq h^T \check{\Sigma}_{KK} h + b' \left\{\left(\frac{1}{n}\right)^{\frac{2}{3}} + \sqrt{\frac{k_1 \log d}{n}}\right\},\end{aligned}$$

where $h \in \mathbb{R}^{k_2}$ is a unit-norm minimal eigenvector of $\check{\Sigma}_{KK}$. The last inequality follows from the bounds on $\left|\frac{(\hat{v}-v)^T v}{n}\right|_\infty$ and $\left|\frac{(\hat{v}-v)^T(\hat{v}-v)}{n}\right|_\infty$ from the proofs for Lemma A.3 and Corollary 3.4.4. Therefore, if $\sqrt{\frac{k_1 \log d}{n}} = o(1)$, then we have

$$\lambda_{\min}(\check{\Sigma}_{KK}) \geq \frac{\lambda_{\min}(\hat{\Sigma}_{KK})}{2}$$

$$\begin{aligned}\Longrightarrow \left\|\check{\Sigma}_{KK}^{-1}\right\|_2 &\leq \frac{2}{\lambda_{\min}(\hat{\Sigma}_{KK})} \\ &\leq \frac{4}{\lambda_{\min}(\Sigma_{KK})},\end{aligned} \tag{3.63}$$

where the last inequality follows from bound (3.59) in the proof for Lemma A.14. Putting everything together, we have

$$\left\|\hat{\Sigma}_{KK}^{-1} - \tilde{\Sigma}_{KK}^{-1}\right\|_\infty \leq \frac{2\sqrt{k_2}}{\lambda_{\min}(\Sigma_{KK})} \frac{\lambda_{\min}(\Sigma_{KK})}{\sqrt{k_2}} \frac{4}{\lambda_{\min}(\Sigma_{KK})} = \frac{8}{\lambda_{\min}(\Sigma_{KK})}.$$

with probability at least $1 - O\left(\frac{1}{d \wedge p}\right)$. $\square$

Lemma A.16: Suppose the conditions in Corollary 3.4.4 hold. With the choice of the tuning parameter

$$\lambda_{n,3} \geq c\frac{8(2-\frac{\phi}{4})}{\phi}\left[\left(\sigma_v\sigma_\eta\sqrt{\frac{\log p}{n}}\right) \vee \left(L\left|\beta^*\right|_1\sqrt{\Upsilon_{w,\theta^*}}b\left(\sigma_v,\,\sigma_\eta\right)\sqrt{\frac{k_1\log d}{n}}\right)\right]$$

for some sufficiently large constant $c > 0$, under the conditions

$$\begin{aligned} n &\gtrsim (k_2\log p)\vee\left(k_2^2k_1\log d\right), \\ \sqrt{\frac{n\log p}{|\beta^*|_1^2}}\vee\sqrt{nk_1\log d} &\gtrsim \log p, \end{aligned}$$

then, we have

$$\left(2-\frac{\phi}{4}\right)\left|\hat{v}^T\frac{e}{n\lambda_{n,3}}\right|_\infty \leq \frac{\phi}{8},$$

with probability at least $1-c_1\exp(-c_2\log(d\wedge p))$.

Proof. Recall from the proofs for Lemma A.4 and Corollary 3.4.4 on $|\frac{\hat{v}^Te}{n\lambda_{n,3}}|_\infty$, so as long as

$$\lambda_{n,3} \geq c\frac{8(2-\frac{\phi}{4})}{\phi}\left[\left(\sigma_v\sigma_\eta\sqrt{\frac{\log p}{n}}\right) \vee \left(L\left|\beta^*\right|_1\sqrt{\Upsilon_{w,\theta^*}}b\left(\sigma_v,\,\sigma_\eta\right)\sqrt{\frac{k_1\log d}{n}}\right)\right]$$

for some sufficiently large constant $c > 0$, under the conditions

$$\begin{aligned} n &\gtrsim (k_2\log p)\vee\left(k_2^2k_1\log d\right), \\ \sqrt{\frac{n\log p}{|\beta^*|_1^2}}\vee\sqrt{nk_1\log d} &\gtrsim \log p, \end{aligned}$$

we have

$$\left(2-\frac{\phi}{4}\right)\left|\hat{v}^T\frac{e}{n\lambda_{n,3}}\right|_\infty \leq \frac{\phi}{8},$$

with probability at least $1-c_1\exp(-c_2\log(d\wedge p))$.□

3.9 Technical lemmas and the proofs for Section 3.8

Lemma B.1: Let $X \in \mathbb{R}^{n\times p_1}$ be a zero-mean sub-Gaussian matrix with parameters $(\Sigma_X,\, \sigma_X^2)$. For any fixed (unit) vector $\Delta \in \mathbb{R}^{p_1}$, we have

$$\mathbb{P}\left[\left|\frac{|X\Delta|_2^2}{n} - \mathbb{E}[\frac{|X\Delta|_2^2}{n}]\right| \geq \varepsilon\right] \leq 2\exp\left(-c_0 n \min\left\{\frac{\varepsilon^2}{\sigma_X^4},\, \frac{\varepsilon}{\sigma_X^2}\right\}\right).$$

Moreover, if $Y \in \mathbb{R}^{n\times p_2}$ is a zero-mean sub-Gaussian matrix with parameters $(\Sigma_Y,\, \sigma_Y^2)$, then

$$\mathbb{P}\left[\left|\frac{Y^TX}{n} - \text{cov}(Y_i,\, X_i)\right|_\infty \geq \varepsilon\right] \leq 6\exp\left(-c_1 n \min\left\{\frac{\varepsilon^2}{\sigma_X^2\sigma_Y^2},\, \frac{\varepsilon}{\sigma_X\sigma_Y}\right\} + \log p_1 + \log p_2\right)$$

where X_i and Y_i are the i^{th} rows of X and Y, respectively.

Remark. This lemma is Lemma 14 in Loh and Wainwright (2012). □

Lemma B.2: Let $X \in \mathbb{R}^{n\times p_1}$ be a sub-Gaussian matrix with parameters $(\Sigma_X,\, \sigma_X^2)$. We have

$$\begin{aligned}\frac{|X\Delta|_2^2}{n} &\geq \frac{\underline{\alpha}}{2}|\Delta|_2^2 - \alpha'\frac{\log p_1}{n}|\Delta|_1^2, \qquad \text{for all } \Delta \in \mathbb{R}^{p_1}\\ \frac{|X\Delta|_2^2}{n} &\leq \frac{3\bar{\alpha}}{2}|\Delta|_2^2 + \alpha'\frac{\log p_1}{n}|\Delta|_1^2, \qquad \text{for all } \Delta \in \mathbb{R}^{p_1}\end{aligned}$$

with probability at least $1-c_1\exp(-c_2 n)$, where $\underline{\alpha}$, $\bar{\alpha}$, and α' only depend on Σ_X and σ_X.

Remark. This lemma is Lemma 13 in Loh and Wainwright (2012). The choice of $\underline{\alpha}$, $\bar{\alpha}$, and α' depends on the problems.□

Lemma B.3: For all $u \geq r_{nj}$, we have

$$\mathbb{P}\left[\mathcal{A}_j(u) \mid \left\{w_i^T\theta^*\right\}_{i=1}^n\right] \leq \exp\left(-nC_j^* u^2\right)$$

where

$$\mathcal{A}_j(u) := \left\{\exists f \in \bar{\mathcal{F}}_j \cap \{|f_{\theta^*}|_n \geq u\} \,:\, \frac{1}{n}\left|\sum_{i=1}^n v_{ij} f(w_i^T\theta^*)\right| \geq 2\,|f_{\theta^*}|_n\, u\right\},$$

and

$$C_j^* = c\frac{\gamma_{v_j} \wedge (B_{v_j}^2 \vee B_\eta^2)^{-1}}{\sigma_{v_j}^2 \vee \sigma_\eta^2} \tag{3.64}$$

when v_j is either (i) sub-Gaussian with *strongly log-concave* distribution for some $\gamma_{v_j} > 0$; or, (ii) a bounded vector such that for every $i = 1, ..., n$, v_{ij} is supported on the interval (a'_{v_j}, a''_{v_j}) with $B_{v_j} := a''_{v_j} - a'_{v_j}$; or, (iii) a mixture of (i) and (ii) in terms of its probability measure.

Proof. This proof is a modification of the proofs for Corollary 8.3 in van de Geer (2000) and Lemma 13.2 in Wainwright (2015). Suppose that there exists some $f \in \bar{\mathcal{F}}_j$ with $|f_{\theta^*}|_n \geq u$ such that

$$\frac{1}{n}\left|\sum_{i=1}^{n} v_{ij} f(w_i^T\theta^*)\right| \geq 2\,|f_{\theta^*}|_n\, u. \tag{3.65}$$

Defining the function $\tilde{f}_{\theta^*} := \frac{u}{|f_{\theta^*}|_n} f_{\theta^*}$, observe that $\left|\tilde{f}_{\theta^*}\right|_n = u$. Since $f \in \bar{\mathcal{F}}_j$ and $u \leq |f_{\theta^*}|_n$ by construction, the star-shaped assumption implies that $\tilde{f} \in \bar{\mathcal{F}}_j$. Consequently, we have shown that if there exists a function f satisfying inequality (3.65), which occurs whenever the event $\mathcal{A}_j(u)$ is true, then there exists a function $\tilde{f} \in \bar{\mathcal{F}}_j$ with $\left|\tilde{f}_{\theta^*}\right|_n = u$ such that

$$\frac{1}{n}\left|\sum_{i=1}^{n} v_{ij} \tilde{f}(w_i^T\theta^*)\right| = \frac{u}{|f_{\theta^*}|_n}\frac{1}{n}\left|\sum_{i=1}^{n} v_{ij} f(w_i^T\theta^*)\right| \geq 2u^2.$$

Summarizing then, we have established the inequality

$$\mathbb{P}\left[\mathcal{A}_j(u) \mid \left\{w_i^T\theta^*\right\}_{i=1}^n\right] \leq \mathbb{P}\left[A_{nj}(u) \geq 2u^2 \mid \left\{w_i^T\theta^*\right\}_{i=1}^n\right],$$

where

$$A_{nj}(u) := \sup_{\tilde{f}(\theta^*)\in\Omega(u;\,\mathcal{F}_j)} \frac{1}{n}\left|\sum_{i=1}^{n} v_{ij} \tilde{f}(w_i^T\theta^*)\right|$$

where

$$\Omega(u;\,\mathcal{F}_j) = \left\{f \,:\, f \in \bar{\mathcal{F}}_j,\, |f_{\theta^*}|_n \leq u\right\}.$$

By construction, $\mathbb{E}(v_{ij}|w_i^T\theta^*) = 0$ for all $i = 1, ..., n$. Consequently, conditioning on $\left\{w_i^T\theta^*\right\}_{i=1}^n$, for each fixed $\tilde{f}$, the variable $\frac{1}{n}\left|\sum_{i=1}^n v_{ij}\tilde{f}(w_i^T\theta^*)\right|$ is zero-mean sub-Gaussian, so that the variable $A_{nj}(u)$ is the supremum of a sub-Gaussian process. If we view this supremum as a function of a standardized sub-Gaussian vector, then by standard maximal inequality (e.g., Corollary 8.3 in van de Geer, 2000), we have

$$\mathbb{P}\left[A_{nj}(u) \geq \mathbb{E}_{v_j}\left[A_{nj}(u) \mid \left\{w_i^T\theta^*\right\}_{i=1}^n\right] + b \wedge 2\sigma_{v_j}^2 \mid \left\{w_i^T\theta^*\right\}_{i=1}^n\right] \leq \exp\left(-\frac{nC_j^* b^2}{u^2}\right).$$

Consequently, for any $b = u^2$,

$$\mathbb{P}\left[A_{nj}(u) \geq \mathbb{E}_{v_j}\left[A_{nj}(u) \mid \left\{w_i^T\theta^*\right\}_{i=1}^n\right] + u^2 \mid \left\{w_i^T\theta^*\right\}_{i=1}^n\right] \leq \exp\left(-nC_j^* u^2\right). \tag{3.66}$$

Now let us look at special cases that yield sharp constant C_j^*.

Special case (i): If v_j is sub-Gaussian with *strongly log-concave* distribution for some $\gamma_{v_j} > 0$, then by Cauchy-Schwarz inequality, it can be verified that the associated Lipschitz constant is at most $\frac{\sigma_{v_j} u}{\sqrt{n}}$. Consequently, by Lemma B.6, for any $b > 0$, we have the sub-Gaussian tail bound

$$\mathbb{P}\left[A_{nj}(u) \geq \mathbb{E}_{v_j}\left[A_{nj}(u) \mid \left\{w_i^T\theta^*\right\}_{i=1}^n\right] + b \mid \left\{w_i^T\theta^*\right\}_{i=1}^n\right] \leq \exp\left(-\frac{n\gamma_{v_j} b^2}{2u^2\sigma_{v_j}^2}\right).$$

Setting $b = u^2$ yields

$$\mathbb{P}\left[A_{nj}(u) \geq \mathbb{E}_{v_j}\left[A_{nj}(u) \mid \left\{w_i^T\theta^*\right\}_{i=1}^n\right] + u^2 \mid \left\{w_i^T\theta^*\right\}_{i=1}^n\right] \leq \exp\left(-\frac{n\gamma_{v_j} u^2}{2\sigma_{v_j}^2}\right).$$

Special case (ii): If v_j is a bounded vector such that for every $i = 1, ..., n$, v_{ij} is supported on the interval (a'_{v_j}, a''_{v_j}) with $B_{v_j} := a''_{v_j} - a'_{v_j}$, then by Lemma B.7, following similar argument as above, we have

$$\mathbb{P}\left[A_{nj}(u) \geq \mathbb{E}_{v_j}\left[A_{nj}(u) \mid \left\{w_i^T\theta^*\right\}_{i=1}^n\right] + u^2 \mid \left\{w_i^T\theta^*\right\}_{i=1}^n\right] \leq \exp\left(-\frac{nu^2}{2\sigma_{v_j}^2 B_{v_j}^2}\right).$$

Finally, by definition of $A_{nj}(u)$ and $\mathcal{G}_n(u; \mathcal{F}_j)$, we have $\mathbb{E}_{v_j}\left[A_{nj}(u) \mid \left\{w_i^T\theta^*\right\}_{i=1}^n\right] = \sigma_{v_j}\mathcal{G}_n(u; \mathcal{F}_j)$. By Lemma B.8, the function $t \mapsto \frac{\mathcal{G}_n(t; \mathcal{F}_j)}{t}$ is non-decreasing, and since $u \geq r_{nj}$ by assumption, we have

$$\sigma_{v_j}\frac{\mathcal{G}_n(u; \mathcal{F}_j)}{u} \leq \sigma_{v_j}\frac{\mathcal{G}_n(r_{nj}; \mathcal{F}_j)}{r_{nj}} \leq r_{nj},$$

where the last inequality uses the definition of r_{nj}. Putting everything together, we have shown that $\mathbb{E}_{v_j}\left[A_{nj}(u) \mid \left\{w_i^T\theta^*\right\}_{i=1}^n\right] \leq ur_{nj}$. Combined with the tail bound (3.66), respectively, we obtain

$$\begin{aligned} \mathbb{P}\left[A_{nj}(u) \geq 2u^2 \mid \left\{w_i^T\theta^*\right\}_{i=1}^n\right] &\leq \mathbb{P}\left[A_{nj}(u) \geq ur_{nj} + u^2 \mid \left\{w_i^T\theta^*\right\}_{i=1}^n\right] \\ &\leq \exp\left(-nC_j^* u^2\right) \end{aligned}$$

where the inequality uses the fact that $u^2 \geq ur_{nj}$. □

Lemma B.4: There are universal positive constants (c_1, c_2) such that for all $\Delta_j \in \left\{f \in \bar{\mathcal{F}}_j \,:\, |f|_{\mathcal{F}_j} \leq 3,\, |f_{\theta^*}|_n \geq u\right\}$,

$$\frac{1}{n}\left|\sum_{i=1}^n v_{ij}\Delta_j(w_i^T\theta^*)\right| \leq 6r_{nj}\,|\Delta_j(\theta^*)|_n + \frac{1}{32}\,|\Delta_j(\theta^*)|_n^2 \tag{3.67}$$

with probability at least $1 - c_1\exp\left(-nC_j^* r_{nj}^2\right)$ and C_j^* follows (3.64) under the same special cases in Lemma B.3.

Proof. To establish bound (3.67), we first consider over the ball

$$\left\{\Delta_j \in \bar{\mathcal{F}}_j \,:\, |\Delta_j|_{\mathcal{F}_j} \leq 3\right\} \cap \left\{\Delta_j \in \bar{\mathcal{F}}_j \,:\, |\Delta_j(\theta^*)|_n \leq u\right\}$$

for some fixed radius $u \geq r_{nj}$. Later in the proof we extend the bound to one that is uniform in $|\Delta_j(\theta^*)|_n$ via a peeling argument. Define the random variable

$$A'_{nj}(u) := \sup_{\Delta_j \in \Omega'(u;\,\mathcal{F}_j)} \frac{1}{n}\left|\sum_{i=1}^n v_{ij}\Delta_j(w_i^T\theta^*)\right|.$$

where $\Omega'(u;\,\mathcal{F}_j) = \left\{\Delta_j \in \bar{\mathcal{F}}_j \,:\, |\Delta_j|_{\mathcal{F}_j} \leq 3,\, |\Delta_j(\theta^*)|_n \leq u\right\}$. Following the same argument in Lemma B.3, we obtain

$$\mathbb{P}\left[A'_{nj}(u) \geq \mathbb{E}_{v_j}\left[A'_{nj}(u) \mid \left\{w_i^T\theta^*\right\}_{i=1}^n\right] + b \mid \left\{w_i^T\theta^*\right\}_{i=1}^n\right] \leq \exp\left(-\frac{nC_j^* b^2}{u^2}\right). \tag{3.68}$$

We first derive a bound for $u = r_{nj}$. By definition of $A'_{nj}(u)$ and $\mathcal{G}_n\left(r_{nj};\,\mathcal{F}_j\right)$ and the critical radius r_{nj}, we have

$$\mathbb{E}_{v_j}\left[A'_{nj}(r_{nj}) \mid \left\{w_i^T\theta^*\right\}_{i=1}^n\right] = \tilde{\sigma}_{v_j}\mathcal{G}_n\left(r_{nj};\,\mathcal{F}_j\right) \leq 3r_{nj}^2.$$

Setting $b = 3r_{nj}^2$ in the tail bound (3.68) yields

$$\mathbb{P}\left[A_{nj}'(r_{nj}) \geq 6r_{nj}^2 \mid \left\{w_i^T\theta^*\right\}_{i=1}^n\right] \leq \exp\left(-nC_j^* r_{nj}^2\right).$$

On the other hand, for any $u > r_{nj}$, we have

$$\begin{aligned}\mathbb{E}_{v_j}\left[A_{nj}'(u) \mid \left\{w_i^T\theta^*\right\}_{i=1}^n\right] &= \tilde{\sigma}_{v_j}\mathcal{G}_n\left(u;\, \mathbb{B}_{\mathcal{F}_j}(3)\right) = u\frac{\tilde{\sigma}_{v_j}\mathcal{G}_n\left(u;\, \mathbb{B}_{\mathcal{F}_j}(3)\right)}{u} \\ &\leq u\frac{\tilde{\sigma}_{v_j}\mathcal{G}_n\left(r_{nj};\, \mathbb{B}_{\mathcal{F}_j}(3)\right)}{r_{nj}} \leq 3ur_{nj},\end{aligned}$$

where the first inequality follows from Lemma B.8 and the second inequality follows from the critical radius r_{nj}. Setting $b = \frac{u^2}{128}$ in the tail bound (3.68) yields

$$\mathbb{P}\left[A_{nj}'(u) \geq 3ur_{nj} + \frac{u^2}{128} \mid \left\{w_i^T\theta^*\right\}_{i=1}^n\right] \leq \exp\left(-nC_j^* u^2\right). \tag{3.69}$$

It remains to prove the bound (3.67) via a peeling argument. Let $\mathcal{E}$ denote the event that the bound (3.67) is violated for some function Δ_j with $|\Delta_j|_{\mathcal{F}_j} \leq 3$. For real numbers $0 \leq a_1 < a_2$, let $\mathcal{E}(a_1,\, a_2)$ denote the event that it is violated for some function such that $|\Delta_j(\theta^*)|_n \in [a_1,\, a_2]$, and $|\Delta_j|_{\mathcal{F}_j} \leq 3$. For $m = 0,\, 1,\, 2,\, ...$, define $t_m = 2^m r_{nj}$. We then have the decomposition $\mathcal{E} = \mathcal{E}(0,\, t_0) \cup (\cup_{m=0}^{\infty}\mathcal{E}(t_m,\, t_{m+1}))$ and hence by union bound,

$$\mathbb{P}[\mathcal{E}] \leq \mathbb{P}[\mathcal{E}(0,\, t_0)] + \sum_{m=0}^{\infty}\mathbb{P}[\mathcal{E}(t_m,\, t_{m+1})]. \tag{3.70}$$

The final step is to bound each of the terms in this summation. Since $t_0 = r_{nj}$, we have

$$\mathbb{P}[\mathcal{E}(0,\, t_0)] \leq \mathbb{P}\left[A_{nj}'(r_{nj}) \geq 6r_{nj}^2 \mid \left\{w_i^T\theta^*\right\}_{i=1}^n\right] \leq \exp\left(-nC_j^* r_{nj}^2\right). \tag{3.71}$$

On the other hand, suppose that $\mathcal{E}(t_m,\, t_{m+1})$ holds, meaning that there exists some function Δ_j with $|\Delta_j|_{\mathcal{F}_j} \leq 3$ and $|\Delta_j(\theta^*)|_n \leq t_{m+1}$ such that

$$\begin{aligned}\frac{1}{n}\left|\sum_{i=1}^n v_{ij}\Delta_j(w_i^T\theta^*)\right| &> 6r_{nj}\left|\Delta_j(\theta^*)\right|_n + \frac{1}{32}\left|\Delta_j(\theta^*)\right|_n^2 \\ &> 6r_{nj}t_m + \frac{1}{32}t_m^2 \\ &= 3r_{nj}t_{m+1} + \frac{1}{128}t_{m+1}^2,\end{aligned}$$

where the second inequality follows since $|\Delta_j(\theta^*)|_n \geq t_m$; and the third inequality follows since $t_{m+1} = 2t_m$. This lower bound implies that

$$A'_{nj}(t_{m+1}) \geq 3r_{nj}t_{m+1} + \frac{1}{128}t_{m+1}^2,$$

hence the bound (3.69) implies that

$$\mathbb{P}\left[\mathcal{E}(t_m,\, t_{m+1})\right] \leq \exp\left(-\frac{nt_{m+1}^2}{c_2\tilde{\sigma}_{v_j}}\right) = \exp\left(-\frac{n2^{2(m+1)}r_{nj}^2}{c_2\tilde{\sigma}_{v_j}}\right).$$

Combining this tail bound with our earlier bound (3.71), and substituting into the union bound (3.70) yields

$$\mathbb{P}\left[\mathcal{E}\right] \leq \exp\left(-\frac{n\gamma_{v_j}r_{nj}^2}{c_2\tilde{\sigma}_{v_j}}\right) + \sum_{m=0}^{\infty}\exp\left(-\frac{n\gamma_{v_j}2^{2(m+1)}r_{nj}^2}{c_2\tilde{\sigma}_{v_j}}\right) \leq c_1\exp\left(-\frac{n\gamma_{v_j}r_{nj}^2}{c_2\tilde{\sigma}_{v_j}^2}\right).$$

Lemma B.5: There are universal positive constants $(c_1,\, c_2)$ such that for all $\Delta_j \in \left\{f \in \bar{\mathcal{F}}_j \,:\, |f|_{\mathcal{F}_j} \geq 1\right\}$,

$$\frac{1}{n}\left|\sum_{i=1}^{n} v_{ij}\Delta_{ij}(\theta^*)\right| \leq 2r_{nj}\left|\Delta_j(\theta^*)\right|_n + 2r_{nj}^2\left|\Delta_j\right|_{\mathcal{F}_j} + \frac{1}{16}\left|\Delta_j(\theta^*)\right|_n^2$$

with probability at least $1 - c_1\exp\left(-nC_j^*r_{nj}^2\right)$ and C_j^* follows (3.64) under the same special cases in Lemma B.3.

Proof. This proof is a modification of the proofs for Corollary 8.3 in van de Geer (2000) and Lemma 13.2 in Wainwright (2015). It suffices to show that

$$\frac{1}{n}\left|\sum_{i=1}^{n} v_{ij}f(w_i^T\theta^*)\right| \leq 2r_{nj}\left|f_{\theta^*}\right|_n + 2r_{nj}^2 + \frac{1}{32}\left|f_{\theta^*}\right|_n^2, \quad \text{for all } |f|_{\mathcal{F}_j} = 1. \tag{3.72}$$

To see this, noting that for any function Δ with $|\Delta|_{\mathcal{F}_j} \geq \frac{1}{2}$, we can define $f = \frac{\Delta(\theta^*)}{|\Delta(\theta^*)|_{\mathcal{F}_j}}$. Substituting this definition and then multiplying both sides of inequality (3.72) by $|\Delta(\theta^*)|_{\mathcal{F}_j}$, we obtain

$$\begin{aligned}\frac{1}{n}\left|\sum_{i=1}^{n} v_{ij}\Delta_{ij}(\theta^*)\right| &\leq 2r_{nj}\left|\Delta_j(\theta^*)\right|_n + 2r_{nj}^2\left|\Delta_j(\theta^*)\right|_{\mathcal{F}_j} + \frac{1}{32}\frac{\left|\Delta_j(\theta^*)\right|_n^2}{\left|\Delta_j(\theta^*)\right|_{\mathcal{F}_j}} \\ &\leq 2r_{nj}\left|\Delta_j(\theta^*)\right|_n + 2r_{nj}^2\left|\Delta_j(\theta^*)\right|_{\mathcal{F}_j} + \frac{1}{16}\left|\Delta_j(\theta^*)\right|_n^2,\end{aligned}$$

where the second inequality follows by the fact that $|\Delta|_{\mathcal{F}_j} \geq \frac{1}{2}$.

To establish the bound (3.72), we first consider it over the ball $\{|f_{\theta^*}|_n \leq u\}$ for some fixed radius $u \geq r_{nj}$. Later in the proof we extend bound (3.72) to one that is uniform in $|f_{\theta^*}|_n$ via a peeling argument. Define the random variable

$$A^{''}_{nj}(u) := \sup_{f(\theta)\in\Omega^{''}(u;\,\mathcal{F}_j)} \frac{1}{n}\left|\sum_{i=1}^{n} v_{ij} f(w_i^T\theta^*)\right|.$$

where $\Omega^{''}(u;\,\mathcal{F}_j) = \left\{f \in \bar{\mathcal{F}}_j \,:\, |f|_{\mathcal{F}_j} \leq 1,\, |f_{\theta^*}|_n \leq u\right\}$.

The next part of the proof is almost identical to that in the proof for Lemma B.4 except that we replace $A^{'}_{nj}(u)$ with $A^{''}_{nj}(u)$ and $\Omega^{'}(u;\,\mathcal{F}_j)$ with $\Omega^{''}(u;\,\mathcal{F}_j)$. We first derive a bound for $u = r_{nj}$. By definition of $A^{''}_{nj}(u)$ and $\mathcal{G}_n\left(r_{nj};\,\mathcal{F}_j\right)$ and the critical radius r_{nj}, we have

$$\mathbb{E}_{v_j}\left[A^{''}_{nj}(r_{nj}) \mid \left\{w_i^T\theta^*\right\}_{i=1}^{n}\right] = \tilde{\sigma}_{v_j}\mathcal{G}_n\left(r_{nj};\,\mathbb{B}_{\mathcal{F}_j}(1)\right) \leq r_{nj}^2.$$

Setting $b = r_{nj}^2$ in the tail bound (3.68) with $A^{'}_{nj}(u)$ replaced by $A^{''}_{nj}(u)$ and $\Omega^{'}(u;\,\mathcal{F}_j)$ replaced by $\Omega^{''}(u;\,\mathcal{F}_j)$ yields

$$\mathbb{P}\left[A^{''}_{nj}(r_{nj}) \geq 2r_{nj}^2 \mid \left\{w_i^T\theta^*\right\}_{i=1}^{n}\right] \leq \exp\left(-nC_j^* r_{nj}^2\right). \tag{3.73}$$

On the other hand, for any $u > r_{nj}$, we have

$$\begin{aligned}\mathbb{E}_{v_j}\left[A^{''}_{nj}(u) \mid \left\{w_i^T\theta^*\right\}_{i=1}^{n}\right] &= \tilde{\sigma}_{v_j}\mathcal{G}_n\left(u;\,\mathbb{B}_{\mathcal{F}_j}(1)\right) = u\frac{\tilde{\sigma}_{v_j}\mathcal{G}_n\left(u;\,\mathbb{B}_{\mathcal{F}_j}(1)\right)}{u} \\ &\leq u\frac{\tilde{\sigma}_{v_j}\mathcal{G}_n\left(r_{nj};\,\mathbb{B}_{\mathcal{F}_j}(1)\right)}{r_{nj}} \leq u r_{nj},\end{aligned}$$

where the first inequality follows from Lemma B.8 and the second inequality follows from the critical radius r_{nj}. Setting $b = \frac{u^2}{128}$ in the tail bound (3.68) with $A^{'}_{nj}(u)$ replaced by $A^{''}_{nj}(u)$ and $\Omega^{'}(u;\,\mathcal{F}_j)$ replaced by $\Omega^{''}(u;\,\mathcal{F}_j)$ yields

$$\mathbb{P}\left[A^{''}_{nj}(u) \geq u r_{nj} + \frac{u^2}{128} \mid \left\{w_i^T\theta^*\right\}_{i=1}^{n}\right] \leq \exp\left(-nC_j^* u^2\right). \tag{3.74}$$

It remains to prove the bound (3.72) via a peeling argument. Let $\mathcal{E}^{'}$ denote the event that the bound (3.72) is violated for some function f with $|f|_{\mathcal{F}_j} = 1$. For real numbers

$0 \le a_1 < a_2$, let $\mathcal{E}'(a_1, a_2)$ denote the event that it is violated for some function such that $|f_{\theta^*}|_n \in [a_1, a_2]$, and $|f|_{\mathcal{F}_j} = 1$. For $m = 0, 1, 2, ...,$ define $t_m = 2^m r_{nj}$. We then have the decomposition $\mathcal{E}' = \mathcal{E}'(0, t_0) \cup \left(\cup_{m=0}^{\infty} \mathcal{E}'(t_m, t_{m+1})\right)$ and hence by union bound,

$$\mathbb{P}\left[\mathcal{E}'\right] \le \mathbb{P}\left[\mathcal{E}'(0, t_0)\right] + \sum_{m=0}^{\infty} \mathbb{P}\left[\mathcal{E}'(t_m, t_{m+1})\right]. \tag{3.75}$$

The final step is to bound each of the terms in this summation. Since $t_0 = r_{nj}$, we have

$$\mathbb{P}\left[\mathcal{E}'(0, t_0)\right] \le \mathbb{P}\left[A''_{nj}(r_{nj}) \ge 2r_{nj}^2 \mid \left\{w_i^T \theta^*\right\}_{i=1}^n\right] \le \exp\left(-nC_j^* r_{nj}^2\right). \tag{3.76}$$

On the other hand, suppose that $\mathcal{E}'(t_m, t_{m+1})$ holds, meaning that there exists some function f with $|f|_{\mathcal{F}_j} = 1$ and $|f_{\theta^*}|_n \le t_{m+1}$ such that

$$\begin{aligned} \frac{1}{n}\left|\sum_{i=1}^n v_{ij} f(w_i^T \theta^*)\right| &> 2r_{nj}\,|f_{\theta^*}|_n + 2r_{nj}^2 + \frac{1}{32}\,|f_{\theta^*}|_n^2 \\ &> 2r_{nj} t_m + 2r_{nj}^2 + \frac{1}{32} t_m^2 \\ &= r_{nj} t_{m+1} + 2r_{nj}^2 + \frac{1}{128} t_{m+1}^2, \end{aligned}$$

where the second inequality follows since $|f_{\theta^*}|_n \ge t_m$; and the third inequality follows since $t_{m+1} = 2t_m$. This lower bound implies that

$$A''_{nj}(t_{m+1}) \ge r_{nj} t_{m+1} + \frac{1}{128} t_{m+1}^2,$$

hence the bound (3.74) implies that

$$\mathbb{P}\left[\mathcal{E}'(t_m, t_{m+1})\right] \le \exp\left(-\frac{n t_{m+1}^2}{c_2 \tilde{\sigma}_{v_j}}\right) = \exp\left(-\frac{n 2^{2(m+1)} r_{nj}^2}{c_2 \tilde{\sigma}_{v_j}}\right).$$

Combining this tail bound with our earlier bound (3.76), and substituting into the union bound (3.75) yields

$$\mathbb{P}\left[\mathcal{E}'\right] \le \exp\left(-\frac{n\gamma_{v_j} r_{nj}^2}{2\tilde{\sigma}_{v_j}}\right) + \sum_{m=0}^{\infty} \exp\left(-\frac{n\gamma_{v_j} 2^{2(m+1)} r_{nj}^2}{c_2 \tilde{\sigma}_{v_j}}\right) \le c_1 \exp\left(-\frac{n\gamma_{v_j} r_{nj}^2}{c_2 \tilde{\sigma}_{v_j}^2}\right).$$

□

Lemma B.6: Suppose the variable X has *strongly log-concave* distribution with parameter $\gamma > 0$. Then for any function $f : \mathbb{R}^n \to \mathbb{R}$ that is L-Lipschtiz with respect to Euclidean norm, we have

$$\mathbb{P}\left[|f(X) - \mathbb{E}\left[f(X)\right]| \geq t\right] \leq 2\exp\left(-\frac{\gamma t^2}{2L^2}\right).$$

Remark: This result was initiated by Maurey (1991) and further developed by Bobkov and Ledoux (2000).

Definition B.1 (Separate convex functions). A function $f : \mathbb{R}^n \to \mathbb{R}$ is *separately convex* if, for each $j \in \{1, 2, ..., n\}$, the co-ordinate function $f_j : \mathbb{R} \to \mathbb{R}$ defined by varying only the j^{th} co-ordinate, is a convex function of x_j.

Lemma B.7: Let $\{X_i\}_{i=1}^n$ be independent random variables, each supported on the interval $[a, b]$, and let $f : \mathbb{R}^n \to \mathbb{R}$ be separately convex, and $L-$Lipschitz with respect to the Euclidean norm. Then for all $t > 0$, we have

$$\mathbb{P}\left[f(X) - \mathbb{E}\left[f(X)\right] \geq t\right] \leq 2\exp\left(-\frac{t^2}{2L^2(b-a)^2}\right).$$

Remark: The proof can be found in Wainwright (2015), which is based on Ledoux (1996).

Lemma B.8: For any star-shaped function class $\mathcal{F}$, the function $t \mapsto \frac{\mathcal{G}_n(t;\mathcal{F})}{t}$ is non-decreasing on the interval $(0, \infty)$.

Proof. This is Lemma 13.1 from Wainwright (2015).

Definition B.2 (Covering and packing numbers). For a metric space consisting of a set $\mathcal{X}$ and a metric $\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$:

(i) An $t-$covering of $\mathcal{X}$ with respect to ρ is a set $\left\{\beta^1, ..., \beta^N\right\} \subset \mathcal{X}$ such that for all $\beta \in \mathcal{X}$, there exists some $i \in \{1, ..., N\}$ with $\rho(\beta, \beta^i) \leq t$. The $t-$covering number $N(t; \mathcal{X}, \rho)$ is the cardinality of the smallest $t-$covering.

(ii) An $t^{'}-$packing of $\mathcal{X}$ with respect to ρ is a set $\left\{\beta^1, ..., \beta^N\right\} \subset \mathcal{X}$ such that $\rho(\beta^i, \beta^j) \geq t$ for all $i \neq j$. The $t^{'}-$packing number $M(t^{'}; \mathcal{X}, \rho)$ is the cardinality of the largest $t^{'}-$covering.

Lemma B.9: Let $N_n(t;\, \mathbb{B}_n(r_n;\, \mathcal{F}))$ denote the $t-$covering number of the set

$$\mathbb{B}_n(r_n;\, \mathcal{F}) = \{f \in 2\mathcal{F} \,:\, |f|_n \leq r_n\}$$

in the empirical $L^2(\mathbb{P}_n)$ norm. Then the critical radius condition (3.46) holds for any $r_n \in (0,\, \sigma]$ such that

$$\frac{32}{\sqrt{n}} \int_{\frac{r_n^2}{2\sigma}}^{r_n} \sqrt{\log N_n(t;\, \mathbb{B}_n(r_n;\, \mathcal{F}))} dt \leq \frac{r_n^2}{\sigma}.$$

Remark. This result is established by van der Vaart and Wellner (1996), van de Geer (2000), Barlett and Mendelson (2002), Koltchinski (2006), Wainwright (2015), etc.

Lemma B.10: For $q \in (0,\, 1]$, let

$$\theta \in \mathcal{B}^d(R_q) := \left\{ \theta^{'} \in \mathbb{R}^d \,:\, \left|\theta^{'}\right|_1^q = \sum_{j=1}^{d} \left|\theta_j^{'}\right|^q \leq R_q \right\}$$

and $N_2(t;\, \mathcal{B}^d(R_q))$ be the $t-$covering number of the set $\mathcal{B}^d(R_q)$ in the l_2-norm. Then there is a universal constant c such that

$$\log N_2(t;\, \mathcal{B}^d(R_q)) \leq c R_q^{\frac{2}{2-q}} \left(\frac{1}{t}\right)^{\frac{2q}{2-q}} \log d \quad \text{for all}\, t \in (0,\, R_q^{\frac{1}{q}})$$

Conversely, assume in addition that $t < 1$ and $t^2 \geq c^{''} \left(R_q^{\frac{2}{2-q}} \frac{\log d}{d^{\kappa}}\right)^{1-\frac{q}{2}}$ for some fixed $\kappa \in (0,\, 1)$. Then there is a constant $c^{'} \leq c$ such that

$$\log N_2(t;\, \mathcal{B}^d(R_q)) \geq c^{'} R_q^{\frac{2}{2-q}} \left(\frac{1}{t}\right)^{\frac{2q}{2-q}} \log d.$$

Remark. These bounds are obtained by inverting known results on (dyadic) entropy numbers (e.g., Schütt, 1984; Guedon and E. Litvak, 2000; Kühn, 2001) of l_q-balls as in the proof for Lemma 2 from Raskutii, Wainwright, and Yu (2011). □

Lemma B.11: Consider the class of smooth functions

$$\mathcal{F}_L := \left\{ f \,:\, [0,\, 1] \to \mathbb{R} \,:\, f(0) = 0,\, \left|f(x) - f(x^{'})\right| \leq L \left|x - x^{'}\right| \quad \forall x,\, x^{'} \in [0,\, 1] \right\}.$$

We have

$$\log N_\infty(t;\, \mathcal{F}_L) \asymp \frac{L}{t},$$
$$\text{and } \log N_2(t;\, \mathcal{F}_L) \asymp \frac{L}{t}$$

where $N_\infty(t;\, \mathcal{F}_L)$ is the $t-$covering number of the set $\mathcal{F}_L$ in the sup-norm and $N_2(t;\, \mathcal{F}_L)$ is the $t-$covering number of the set $\mathcal{F}_L$ in the $L^2\left(\mathbb{P}\right)-$norm.

Remark. The proof for $\log N_\infty(t;\, \mathcal{F}_L)$ is a well-known result due to Kolmogorov and Tikhomirov (1961). To proof for $\log N_2(t;\, \mathcal{F}_L)$ is an application of Varshamov-Gilbert lemma. □

Lemma B.12: For all $\epsilon > 0$, the packing and covering number of a set $\mathbb{T}$ with respect to a metric ρ are related as follows:

$$M(2\epsilon;\, \mathbb{T},\, \rho) \leq N(\epsilon;\, \mathbb{T},\, \rho) \leq M(\epsilon;\, \mathbb{T},\, \rho).$$

Proof. To prove the first inequality, let $\left\{\theta^1,\, ...,\, \theta^{2M}\right\}$ be a maximal $2\epsilon-$packing, so that $\rho(\theta^i,\, \theta^j) > 2\epsilon$ for all $i \neq j$. Given a $\epsilon-$cover, let γ be an arbitrary element. For any distinct pair $i \neq j$, we have

$$2\epsilon < \rho(\theta^i,\, \theta^j) \leq \rho(\theta^i,\, \gamma) + \rho(\gamma,\, \theta^j),$$

so that it is not possible to have $\rho(\gamma,\, \theta^i) \leq \epsilon$ and $\rho(\gamma,\, \theta^j) \leq \epsilon$ simultaneously. Conversely, we must have at least one element in the cover for each element of the packing set, showing that $M(2\epsilon;\, \mathbb{T},\, \rho) \leq N(\epsilon;\, \mathbb{T},\, \rho)$. Turning to the second inequality, let $\left\{\theta^1,\, ...,\, \theta^{2M}\right\}$ be a maximal $\epsilon-$packing. Maximality implies that for any other θ not already in the packing set, we must have $\rho(\theta,\, \theta^i) \leq \epsilon$ for some θ^i. Thus, the given set forms a $\epsilon-$cover, showing that $N(\epsilon;\, \mathbb{T},\, \rho) \leq M(\epsilon;\, \mathbb{T},\, \rho)$. □

Lemma B.13: Let $k \geq 1$. There exists a subset W of $\{0,\, 1\}^k$ with $|W| > e^{\frac{k}{8}}$ such that the Hamming distance,

$$d(\tau,\, \tau^{'}) := \sum_{j=1}^{k} \mathbb{I}\left\{\tau_i \neq \tau_i^{'}\right\} > \frac{k}{4}$$

for all $\tau,\, \tau^{'} \in W$ with $\tau \neq \tau^{'}$.

Remark. This is Varshamov-Gilbert Lemma.

Lemma B.14: For a given radius $r > 0$, define the set

$$\mathcal{S}(s,\, r) := \left\{ \delta \in \mathbb{R}^d \mid |\delta|_0 \le 2s,\ |\delta|_2 \le r \right\}.$$

There is a covering set $\left\{\delta^1,\, ...,\, \delta^N\right\} \in \mathcal{S}(s,\, r)$ with cardinality $N = N(s,\, r,\, \epsilon)$ such that

$$\log N(s,\, r,\, \epsilon) \le \log \begin{pmatrix} d \\ 2s \end{pmatrix} + 2s \log \left(\frac{1}{\epsilon}\right).$$

Remark. This result is established in Matousek (2002).

Lemma B.15: There exists a subset $\tilde{\mathcal{M}} \subset \mathcal{M}$ with cardinality $\left|\tilde{\mathcal{M}}\right| \ge \exp\left(\frac{s}{2}\log\frac{d-s}{s/2}\right)$ such that $\rho_{\mathcal{M}}(z,\, z') \ge \frac{s}{2}$ for all $z,\, z' \in \tilde{\mathcal{M}}$.

Remark. This is Lemma 4 from Raskutii, Wainwright, and Yu (2011).

Lemma B.16: Let $X_1,\, ...,\, X_n$ be independent random variables such that, for every i, $\mathbb{E}(X_i) = 0$ and $0 < \mathbb{E}\left(|X_i|^{2+\delta}\right) < \infty$ for some $0 < \delta < 1$. Set $S_n = \sum_{i=1}^n X_i$, $B_n^2 = \sum_{i=1}^n \mathbb{E}(X_i^2)$, $V_n^2 = \sum_{i=1}^n X_i^2$, $L_{n,\delta} = \sum_{i=1}^n \mathbb{E}\left(|X_i|^{2+\delta}\right)$, $b_{n,\delta} = \frac{B_n}{L_{n,\delta}^{1/(2+\delta)}}$. Then for all $0 \le t \le b_{n,\delta}$ and an absolute constant $a_0 > 0$

$$\left|\mathbb{P}\left[\frac{S_n}{V_n} \ge t\right] - [1 - \Phi(t)]\right| \le a_0(1+t)^{1+\delta}\frac{e^{-\frac{t^2}{2}}}{b_{n,\delta}^{2+\delta}}.$$

Remark. This is formula 2.11 from Jing, Shao, and Wang (2003) on moderate deviations for self-normalized sums.

Lemma B.17: Let $X_k,\ k \ge 1$ be *i.i.d.* sub-Gaussian variables with parameters at most σ and $Z := \max_{k=1,\ldots,N} |X_k|$. For some integer $N \ge 10$, we have $\mathbb{E}\left[Z\right] \le 3\sigma\sqrt{N}$.

Remark. This is a well-known result on sub-Gaussian maxima (e.g., van der Vaart and Wellner, 1996).

Bibliography

Ackerberg, D. A., and G. S. Crawford (2009). "Estimating Price Elasticities in Differentiated Product Demand Models with Endogenous Characteristics". Working Paper.

Ahn, H. and J. Powell (1993). "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism". *Journal of Econometrics*, 58, 3-29.

Amemiya, T. (1974). "The Non-Linear Two-Stage Least Squares Estimator". *Journal of Econometrics*, 2, 105-110.

Andrews, D. W. K. (1991a). "Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Regression Models". *Econometrica*, 59, 307-345.

Angrist, J. D. and A. B. Krueger (1991). "Does Compulsory School Attendance Affect Schooling and Earnings?". *Quarterly Journal of Economics*, 106, 979-1014.

Bartlett, P. and S. Mendelson (2002). "Gaussian and Rademacher Complexities: Risk Bounds and Structural Results". *Journal of Machine Learning Research*, 3, 463-482.

Belloni, A., V. Chernozhukov, and L. Wang (2011). "Square-Root Lasso: Pivotal Recovery of Sparse Signals Via Conic Programming". *Biometrika*, 98, 791-806.

Belloni, A., V. Chernozhukov, and L. Wang (2014). "Pivotal Estimation of Nonparametric Functions via Square-Root Lasso". arXiv:1105.1475.

Belloni, A. and V. Chernozhukov (2011a). "L1-Penalized Quantile Regression in High-Dimensional Sparse Models". *The Annals of Statistics*, 39, 82-130.

Belloni, A. and V. Chernozhukov (2011b). "High Dimensional Sparse Econometric Models: an Introduction", in: Inverse Problems and High Dimensional Estimation, Stats in the Château 2009, Alquier, P., E. Gautier, and G. Stoltz, Eds., *Lecture Notes in Statistics*, 203, 127-162, Springer, Berlin.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). "Sparse Models and Methods for Instrumental Regression, with an Application to Eminent Domain". *Econometrica*, 80, 2369-2429.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). "Inference on Treatment Effects after Selection amongst High-Dimensional Controls". cemmap.

Belloni, A. and V. Chernozhukov (2013). "Least Squares after Model Selection in High-Dimensional Sparse Models". *Bernoulli*, 19, 521-547.

Benkard, C. L., and P. Bajari (2005). "Hedonic Price Indexes with Unobserved Product Characteristics, and Application to Personal Computers". *Journal of Business and Economic Statistics*, 23, 61-75.

Berlinet, A. and C. Thomas-Agnan (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic, Norwell, MA.

Berry, S. T., J. A. Levinsohn, and A. Pakes (1995). "Automobile Prices in Market Equilibrium". *Econometrica*, 63, 841-890.

Bickel, P., J. Y. Ritov, and A. B. Tsybakov (2009). "Simultaneous Analysis of Lasso and Dantzig Selector". *The Annals of Statistics*, 37, 1705-1732.

Bobkov, S. G. (1999). "Isoperimetric and Analytic Inequalities for Log-Concave Probability Measures". *The Annals of Probability*, 27, 1903-1921.

Bobkov, S. G. and M. Ledoux (2000). "From Brunn-Minkowski to Brascamp-Lieb and to Logarithmic Sobolev Inequalities". *Geometric and Functional Analysis*, 10, 1028-1052.

Bonaldi, P., A. Hortacsu, and J. Kastl (2014). "An Empirical Analysis of Systemic Risk in the EURO-Zone". Manuscript. Stanford University.

Bresnahan, T. F. and P. C. Reiss (1991). "Entry and Competition in Concentrated Markets". *Journal of Political Economy*, 99, 5, 977-1009.

Bühlmann, P. and S. A. van de Geer (2011). *Statistics for High-Dimensional Data*. Springer, New-York.

Caner, M. (2009). "LASSO Type GMM Estimator". *Econometric Theory*, 25, 1-23.

Candès, E. and T. Tao (2007). "The Dantzig Selector: Statistical Estimation when p is Much Larger Than n". *The Annals of Statistics*, 35, 2313-2351.

Carrasco, M. and J. P. Florens (2000). "Generalization of GMM to a Continuum of Moment Conditions". *Econometric Theory*, 16, 797-834.

Carrasco, M. (2012). "A Regularization Approach to the Many Instruments Problem". *Journal of Econometrics*, 170, 383-398.

Čencov, N. N. (1962). "Evaluation of an Unknown Distribution Density from Observations". *Soviet Mathematics*, 3, 1559-1562.

Chatterjee, S. (2014). "A New Perspective on Least Squares under Convex Constraint". arXiv:1402.0830.

Chen, X. H. (2007). "Large Sample Sieve Estimation of Semi-Nonparametric Models". *Handbook of Econometrics*, 6, 5549-5632.

Chen, X. H. and M. Reiss (2011). "On Rate Optimality for Ill-Posed Inverse Problems in Econometrics". *Econometric Theory*, 27, 497-521.

Chintagunta, P. K. (1993). "Investigating Purchase Incidence, Brand Choice and Purchase Quantity Decisions of Households". *Marketing Science*, 12, 1993, 184-208.

Cosslett, S. R. (1981). "Distribution-Free Maximum Likelihood Estimator of the Binary Choice Model". *Econometrica*, 51, 765-782.

Cosslett, S. R. (1991). "Semiparametric Estimation of a Regression Model with Sample Selectivity", in: W.A. Barnett, J.L. Powell, and G. Tauchen, eds., *Nonparametric and Semiparametric Methods in Econometrics and Statistics* (Cambridge University Press, Cambridge).

Dalalyan, A., and A. B. Tsybakov (2008). "Aggregation by Exponential Weighting, Sharp PAC-Bayesian Bounds and Sparsity". *Journal of Machine Learning Research*, 72, 39-61.

Donoho, D. L., M. Elad, and V. N. Temlyakov (2006). "Stable Recovery of Sparse Overcomplete Representations in the Presence of Noise". *IEEE Transactions on Information Theory*, 52, 6-18.

Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, vol. 1, Cambridge University Press.

Fan, J. and R. Li (2001). "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties". *Journal of American Statistical Association*, 96, 1348-1360.

Fan, J. and Liao, Y. (2014). "Endogeneity in Ultrahigh Dimension". *The Annals of Statistics*, 42, 872-917.

Fan, J. and J. Lv (2010). "A Selective Overview of Variable Selection in High Dimensional Feature Space". *Statistica Sinica*, 20, 101-148.

Fan, J. and J. Lv (2011). "Non-Concave Penalized Likelihood with NP-Dimensionality". *IEEE Transactions on Information Theory*, 57, 5467-5484.

Fan, J., J. Lv, and L, Qi (2011). "Sparse High Dimensional Models in Economics". *Annual Review of Economics*, 3, 291-317.

Friedman, J., T. Hastie, and R. Tibshirani (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent". *Journal of Statistical Software*, 33, 1–22.

Fox, J. T., K. Kim, and C. Yang (2013). "A Simple Nonparametric Approach to Estimating the Distribution of Random Coefficients in Structural Models". Working paper. University of Michigan.

Gallant, A. R. and D. W. Nychka (1987). "Semi-Nonparametric Maximum Likelihood Estimation". *Econometrica*, 55, 363-391.

Garen, J. (1984). "The Returns to Schooling: A Selectivity Bias Approach with a Continuous Choice Variable". *Econometrica*, 52, 1199-1218.

Gautier, E. and A. B. Tsybakov (2011). "High-dimensional Instrumental Variables Regression and Confidence Sets". Manuscript. CREST (ENSAE).

Greene, W. H. (2003). *Econometric Analysis*. Pearson Education India.

Gronau, R. (1973). "The Effects of Children on the Housewife's Value of Time". *Journal of Political Economy*, 81, S168-S199.

Gu, C. (2002). *Smoothing Spline ANOVA Models*. Springer Series in Statistics. Springer, New York, NY.

Guedon, O. and A. E. Litvak (2000). "Euclidean Projections of $p-$Convex Body". In *Geometric Aspects of Functional Analysis*, 95-108. Springer-Verlag.

Guntuboyina, A. (2011). "Lower Bounds for the Minimax Risk Using f-divergences and Applications". *IEEE Transactions on Information Theory*, 57, 2386-2399.

Guntuboyina, A. (2011). *Minimax Lower Bounds*. Ph.D. thesis, Department of Statistics, Yale University.

Guntuboyina, A. and B. Sen (2013). "Global Risk Bounds and Adaptation in Univariate Convex Regression". arXiv:1305.1648.

Han, A. K. (1987). "Non-parametric Analysis of a Generalized Regression Model: The Maximum Rank Correlation Estimator". *Journal of Econometrics*, 35, 303-316.

Hansen, C., J. Hausman, and W. K. Newey (2008). "Estimation with Many Instrumental Variables". *Journal of Business and Economic Statistics*, 26, 398-422.

Hansen, B. (2014). "Efficient Shrinkage in Parametric Models". Working paper. University of Wisconsin, Madison.

Heckman, J. (1974). "Shadow Prices, Market Wages, and Labor Supply". *Econometrica*, 42, 679-693.

Heckman, J. (1976). "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models". *Annals of Economic and Social Measurement*, 5, 475-492.

Huang, J., J. L. Horowitz, and S, Ma (2008). "Asymptotic Properties of Bridge Estimators in Sparse High-Dimensional Regression Models". *The Annals of Statistics*, 36, 587-613.

Ichimura H. (1987). *Consistent Estimation of Index Model Coefficients*. Ph.D. thesis, Department of Economics, M.I.T.

Ichimura, H. and L. Lee (1991). "Semiparametric Least Squares Estimation of Multiple Index Models: Single Equation Estimation", in: W. A. Barnett, J. L. Powell, G. Tauchen (Eds.), *Nonparametric and Semiparametric Methods in Econometrics and Statistics*. Cambridge University Press, Cambridge.

Iyer, G. and P. B. Seetharaman (2003). "To Price Discriminate or Not: Product Choice and the Selection Bias Problem". *Quantitative Marketing and Economics*, 1, 155-178.

Iyer, G. and P. B. Seetharaman (2008). "Too Close to be Similar: Product and Price

Competition in Retail Gasoline Markets". *Quantitative Marketing and Economics*, 6, 205-234.

Javanmard, A. and A. Montanari (2014). "Confidence Intervals and Hypothesis Testing for High-Dimensional Regression". Manuscript. Stanford University.

Jing, B.-Y., Q. M. Shao, and Q. Wang (2003). "Self-Normalized Cramér-Type Large Deviations for Independent Random Variables". *The Annals of Probability*, 31, 2167-2215.

Kakade, S, A. T. Kalai, V. Kanade, O. Shamir (2011). "Efficient Learning of Generalized Linear and Single Index Models with Isotonic Regression". In *Advances in Neural Information Processing Systems*, 23 (NIPS), 927-935.

Klein, W. and R. S. Spady (1987). "An Efficient Semiparametric Estimator of the Binary Response Model". Manuscript (*Bell Communications Research*).

Kolmogorov, A. and B. Tikhomirov (1959). "ϵ-Entropy and ϵ-Capacity of Sets in Functional Spaces". *Uspekhi Mat. Nauk.*, 86, 3-86. Appeared in English as Amer. Math. Soc. Translations, 17, 277-364, 1961.

Koltchinskii V (2006). "Local Rademacher Complexities and Oracle Inequalities in Risk Minimization". *The Annals of Statistics*, 34, 2593-2656.

Koltchinskii, V. (2009). "The Dantzig Selector and Sparsity Oracle Inequalities". *Bernoulli*, 15, 799-828.

Koltchinskii, V. (2011). "Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems". Forthcoming in *Lecture Notes in Mathematics*, Springer, Berlin.

Krishnamurthi, L. and S. P. Raj (1988). "A Model of Brand Choice and Purchase Quantity Price Sensitivities". *Marketing Science*, 7, 1-20.

Kühn, T. (2001). "A Lower Estimate for Entropy Numbers". *Journal of Approximation Theory*, 110, 120-124.

Ledoux, M. (1996). "On Talagrand's Deviation Inequalities for Product Measures". *ESAIM: Probability and Statistics*, 1, 63-87.

Ledoux, M. (2001). *The Concentration of Measure Phenomenon. Mathematical Surveys and Monographs.* American Mathematical Society, Providence, RI.

Ledoux, M., and M. Talagrand (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, New York, NY.

Lee L. (1982). "Some Approaches to the Correction of Selectivity Bias". *Review of Economic Studies*, 49, 355-372.

Lee B. (1988). *Nonparametric Tests using a Kernel Estimation Method.* Ph.D. thesis, Department of Economics, University of Wisconsin, Madison.

Leeb, H. and B. M. Potscher (2006). "Can One Estimate the Conditional Distribution of Post-Model-Selection Estimators?" *The Annals of Statistics*, 34, 2554-2591.

Lin, Y., and H. H. Zhang (2006). "Component Selection and Smoothing in Multivariate Nonparametric Regression". *The Annals of Statistics*, 34(5): 2272-2297.

Loh, P., and M. Wainwright (2012). "High-Dimensional Regression with Noisy and Missing data: Provable Guarantees with Non-convexity". *The Annals of Statistics*, 40, 1637-1664.

Loh, P., and M. Wainwright (2013). "Regularized M-estimators with Nonconvexity: Statistical and Algorithmic Theory for Local Optima". NIPS, NV.

Lounici, K. (2008). "Sup-Norm Convergence Rate and Sign Concentration Property of the Lasso and Dantzig Selector". *Electronic Journal of Statistics*, 2, 90-102.

Manresa, E. (2014). "Estimating the Structure of Social Interactions Using Panel Data". Working paper. CEMFI.

Manski, C .F. (1975). "Maximum Score Estimation of the Stochastic Utility Models of Choice", *Journal of Econometrics*, 3, 205-228.

Manski, C. F. (1985). "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator", *Journal of Econometrics*, 27, 313-334.

Matousek, J. (2002). *Lectures on Discrete Geometry.* Springer-Verlag, New York.

Maurey, B (1991). "Some Deviation Inequalities". *Geometric and Functional Analysis*, 1, 188-197.

McDiarmid. C. (1989). "On the Method of Bounded Differences". In *Surveys in Combinatorics*, no. 141 in *Long Math. Soc. Lecture Notes*, 148-188. Cambridge University Press.

Meinshausen, N., and B. Yu (2009). "Lasso-type Recovery of Sparse Representations for High-dimensional Data". *The Annals of Statistics*, 37, 246-270.

Mussa, M. and S. Rosen (1978). "Monopoly and Product Quality". *Journal of Economic Theory*, 18, 301-317.

Negahban, S., P. Ravikumar, M. J. Wainwright, and B. Yu (2012). "A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers". *Statistical Science*, 27, 538-557.

Nevo, A. (2001). "Measuring Market Power in the Ready-to-Eat Cereal Industry". *Econometrica*, 69, 307-342.

Newey, W., J. L. Powell, J. R. Walker (1990). "Semiparametric Estimation of Selection Models: Some Empirical Results". *American Economic Review Papers and Proceedings*, 80, 324-328.

Newey, W. K. (1991). "Two-step Series Estimation of Sample Selection Models". Department of Economics, Princeton University, Princeton, NJ Manuscript.

Newey, W. K. (1994). "Series Estimation of Regression Functionals". *Econometric Theory*, 10, 1-28.

Newey, W. K. (1997). "Convergence Rates and Asymptotic Normality of Series Estimators". *Journal of Econometrics*, 29, 147-168.

Ravikumar, P., H. Liu, J. Lafferty, and L. Wasserman (2009). "Sparse Additive Models". *Journal of the Royal Statistical Society*, *Series B*, 71, 1009-1030.

Ravikumar, P., M. J. Wainwright, and J. Lafferty (2010). "High-dimensional Ising Model Selection Using l_1- Regularized Logistic Regression". *The Annals of Statistics*, 38, 1287-1319.

Raskutti, G., M. J. Wainwright, and B. Yu (2010). "Restricted Eigenvalue Conditions for Correlated Gaussian Designs". *Journal of Machine Learning Research*, 11, 2241-2259.

Raskutti, G., M. J. Wainwright, and B. Yu (2011). "Minimax Rates of Estimation for High-dimensional Linear Regression over l_q-Balls". *IEEE Trans. Information Theory*, 57, 6976-6994.

Rigollet, P., and A. B. Tsybakov (2011). "Exponential Screening and Optimal Rates of Sparse Estimation". *The Annals of Statistics*, 35, 731-771.

Roberts, J. and A. Sweeting (2011). "Competition versus Auction Design," Working Paper, Duke University.

Roberts, J. and A. Sweeting (2012). "When Should Sellers Use Auctions?". Manuscript. Duke University.

Robinson, P. M. (1988). "Root-n-Consistent Semiparametric Regression". *Econometrica*, 56, 932-954.

Rosenbaum, M. and A. B. Tsybakov (2010). "Sparse Recovery Under Matrix Uncertainty". *The Annals of Statistics*, 38, 2620-2651.

Rosenbaum, M. and A. B. Tsybakov (2013). "Improved Matrix Uncertainty Selector", in: *From Probability to Statistics and Back: High-Dimensional Models and Processes* - A Festschrift in Honor of Jon A. Wellner, Banerjee, M. et al. Eds, *IMS Collections*, 9, 276-290, Institute of Mathematical Statistics.

Rudelson, M. and S. Zhou (2011). "Reconstruction from Anisotropic Random Measurements". Technical report, University of Michigan.

Ruud, P. A. (2000). *An Introduction to Classical Econometric Theory.* Oxford University Press.

Pagan, A. and A. Ullah (1999). *Nonparametric Econometrics.* Cambridge University Press.

Pinkse, J., M. E. Slade, and C. Brett (2002). "Spatial Price Competition: A Semiparametric Approach". *Econometrica*, 70, 1111-1153.

Png, I. P. L. and D. Reitman (1994). "Service Time Competition". *The RAND Journal of Economics*, 25, 619-634.

Powell, J. L. (1987). "Semiparametric Estimation of Bivariate Latent Variable Models". Working paper no. 8704, Revised April 1989 (Social Systems Research

Institute, University of Wisconsin, Madison, WI).

Powell, J., J. H. Stock, and T. M. Stoker (1989). "Semiparametric Estimation of Weighted Average Derivatives". *Econometrica*, 57, 1403-1430.

Powell, J. (1994). "Estimation of Semiparametric Models". *Handbook of Econometrics*, 4. Eds. R. F. Engle and D. L. McFadden, 2443-2521

Sala-i-Martin, X. (1997). "I Just Ran Two Million Regressions". *The American Economic Review*, 87, 178-183.

Schütt, C. (1984). "Entropy Numbers of Diagonal Operators between Symmetric Banach Spaces". *Journal of Approximation Theory*, 40, 121-128.

Shepard, A. (1991). "Price Discrimination and Retail Configuration". *Journal of Political Economy*, 99, 30-53.

Slade, M. E. (1986). "Exogeneity Tests of Market Boundaries Applied to Petroleum Products". *The Journal of Industrial Economics*, 34, 291-303.

Slade, M. E. (1992). "Vancouver Gasoline Price Wars: An Empirical Exercise in Uncovering Supergame Strategies". *The Review of Economic Studies*, 59, 257-276.

Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso". *Journal of the Royal Statistical Society*, Series B, 58, 267-288.

van de Geer, S (2000). *Empirical Processes in M-Estimation*. Cambridge University Press.

van de Geer, S (2008). "High-Dimensional Generalized Linear Models and the Lasso". *The Annals of Statistics*, 36, 614-645.

van der Vaart, A. W. and J. Wellner (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, NY.

Vershynin, R. (2012). "Introduction to the Non-Asymptotic Analysis of Random Matrices", in Eldar, Y. and G. Kutyniok, Eds, *Compressed Sensing: Theory and Applications*, 210-268, Cambridge.

Wahba, G. (1980). "Spline Bases, Regularization, and Generalized Cross Validation for Solving Approximation Problems with Large Quantities of Noisy Data", in *Approximation Theory III*, W. Cheney, Eds., 905-912, Academic Press.

Wahba, G. (1990). *Spline Models for Observational Data. CBMS-NSF Regional Conference Series in Applied Mathematics*. SIAM, Philadelphia, PN.

Wainwright, J. M. (2009). "Sharp Thresholds for High-dimensional and Noisy Sparsity Recovery Using l_1- Constrained Quadratic Programming (Lasso)". *IEEE Trans. Information Theory*, 55, 2183-2202.

Wainwright, J. M. (2015). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. In preparation. University of California, Berkeley.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge.

Yang, Y., and A. Barron (1999). "Information-Theoretic Determination of Minimax Rates of Convergence". *The Annals of Statistics*, 27, 1564-1599.

Ye, F., and C.-H. Zhang (2010). "Rate Minimaxity of the Lasso and Dantzig Selector for the l_q Loss in l_r Balls". *Journal of Machine Learning Research*, 11, 3519-3540.

Zhang C.-H. (2010). "Nearly Unbiased Variable Selection under Minimax Concave Penalty". *The Annals of Statistics*, 38, 894-942.

Zhang C.-H. and S. S. Zhang (2013). "Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models". *Journal of the Royal Statistical Society*, Series B, 76, 217-242.

Zhao, P., and Yu, B. (2007). "On Model Selection Consistency of Lasso". *Journal of Machine Learning Research*, 7, 2541-2567.

Zhu, Y. (2013). "Sparse Linear Models and Two-Stage Estimation in High-Dimensional Settings with Possibly Many Endogenous Regressors". Working paper, University of California, Berkeley.

Zhu, Y. (2014). "High-Dimensional Linear Models with Endogeneity and Sparsity". *The California Econometrics Conference*. Stanford University.

Zhu, Y. (2014). "High-Dimensional Semiparametric Selection Models: Estimation Theory with an Application to the Retail Gasoline Market". Working paper, University of California, Berkeley.