Triangular Simultaneous Equations Models in High-Dimensional Settings with Possibly Many Endogenous Regressors

Ying Zhu*

(Preliminary and incomplete - Please do not distribute)

Abstract

This paper explores the validity of the two-stage estimation procedures for triangular simultaneous linear equations models when the number(s) of the first and/or second-stage regressors grow with and exceed the sample size n. In particular, the number of endogenous regressors in the main equation can also grow with and exceed n. The analysis concerns the sparsity case, i.e., $k_1(= k_{1n})$, the maximum number of non-zero components in the vectors of parameters in the first-stage equations, and $k_2(= k_{2n})$, the number of non-zero components in the vector of parameters in the second-stage equation, are allowed to grow with n but small compared to n. I consider the high-dimensional version of the two-stage least square estimator where one obtains the fitted regressors from the first-stage regression by a least square estimator with l_1- regularization (Lasso or Dantzig selector) when the first-stage regression concerns a large number of regressors relative to n, and then apply a Lasso technique with these fitted regressors in the second-stage regression. I establish sufficient conditions for estimation consistency in l_2-norm and variable-selection consistency (i.e., the two-stage high-dimensional estimators correctly select the non-zero coefficients in the main equation with high probability). Depending on the underlying sufficient conditions that are imposed, the rates of convergence in terms of the l_2-error and the smallest sample size required to obtain these consistency results differ by factors involving k_1 and/or k_2. Simulations are conducted to gain insight on the finite sample performance of the two-stage high-dimensional estimator.

*Haas School of Business, UC Berkeley, CA, 94720

1 Introduction

Endogeneity is a very important issue in empirical economic research. Consider the linear model

$$y_i = \mathbf{x}_i^T \beta^* + \epsilon_i, \quad i = 1, ..., n \tag{1}$$

where ϵ_i is a zero-mean random error possibly correlated with x_i. A component in the regressors $\mathbf{x}_i$ is called *endogenous* if it is correlated with ϵ_i (i.e., $\mathbb{E}(x_i\epsilon_i) \neq 0$) and is called *exogenous* (i.e., $\mathbb{E}(x_i\epsilon_i) = 0$) otherwise. Without loss of generality, I will assume all regressors are endogenous throughout the rest of this paper for notational convenience. The modification to allow mix of endogenous and exogenous regressors is trivial. When endogenous regressors present, the classical least squares estimator will be inconsistent for β^* (i.e., $\hat{\beta}_{OLS} \xrightarrow{p} \beta^*$). The classical solution to this problem of endogenous regressors supposes that there is some L-dimensional vector of instrumental variables, denoted $\mathbf{z}_i$, which is observable and satisfies $\mathbb{E}(\mathbf{z}_i\epsilon_i) = \mathbf{0}$ for all i. Among the instrumental variable estimation literature in econometrics, there is a popular class of models called triangular simultaneous equation models that play an important role in accounting for endogeneity that comes from individual choice or market equilibrium. Based on this class of models, the two-step estimation procedures including the two-stage least square (2SLS) estimation and the control function approach deserve the most attention. Consider the following model

$$\begin{aligned} y_i &= \mathbf{x}_i^T \beta^* + \epsilon_i, \qquad i = 1, ..., n, \\ x_{ij} &= \mathbf{z}_{ij}^T \pi_j^* + \eta_{ij}, \qquad i = 1, ..., n,\ j = 1, ..., p. \end{aligned} \tag{2}$$

$\mathbf{x}_i$ are explanatory variables of dimension $p \times 1$ and x_{ij} denotes the j^{th} component of $\mathbf{x}_i$. ϵ_i is a zero-mean random error correlated with x_{ij} for $j = 1, ..., p$. β^* is an unknown vector of parameters of our main interests. Denote the components of β^* by β_j^*. 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 the first equation in (2) as the main equation (or second-stage equation) and the second equations in (2) as the first-stage equations. Throughout the rest of this paper, I will impose the following assumption.

Assumption 1.1: The data $\{y_i, \mathbf{x}_i, \mathbf{z}_i\}_{i=1}^n$ are *i.i.d.*; $\mathbb{E}(\mathbf{z}_{ij}\epsilon_i) = \mathbb{E}(\mathbf{z}_{ij}\eta_{ij}) = \mathbf{0}$ for all $j = 1, ..., p$ and $\mathbb{E}(\mathbf{z}_{ij}\eta_{ij'}) = \mathbf{0}$ for all $j \neq j'$.

If both equations in (2) are in the low-dimensional settings (i.e., $p \ll n$ and $d_j \ll n$ for all $j = 1, ..., p$), the 2SLS estimation and control function approach are algebraically equivalent but have different interpretations. The 2SLS can be motivated by the *generalized instrumental variable* estimator defined as

$$\hat{\beta}_{GIV} = \hat{\beta}(\hat{\Pi}) = (\hat{\Pi}^T Z^T X)^{-1} (\hat{\Pi}^T Z^T Y),$$

where $\hat{\Pi}$ is some $L \times p$ random matrix that can, in general, be estimated. With the choice of $\hat{\Pi}$ being $(Z^T Z)^{-1}(Z^T X)$, the generalized instrumental variable estimator yields the 2SLS estimator, which takes on the following algebraic form

$$\hat{\beta}_{2SLS} = (\hat{X}^T \hat{X})^{-1}(\hat{X}^T Y).$$

Notice in this procedure, one first regresses $\mathbf{x}_i$ on $\mathbf{z}_i$ in the first stage (i.e., $\mathbf{x}_i = \Pi^T \mathbf{z}_i + \boldsymbol{\eta}_i$) to obtain the fitted values $\hat{\mathbf{x}}_i = \mathbf{z}_i \hat{\Pi}$ and then regresses y_i on the fitted values $\hat{\mathbf{x}}_i$ in the second-stage regression. Another way to interpret the 2SLS estimator is through the control function approach. By projecting the error term in the original regression equation, ϵ_i, onto $\boldsymbol{\eta}_i$, one has $\epsilon_i = \alpha \boldsymbol{\eta}_i + \xi_i$ with $\mathbb{E}(\boldsymbol{\eta}_i \xi_i) = \mathbb{E}(\mathbf{x}_i \xi_i) = \mathbf{0}$. Substituting the expression for ϵ_i into the original regression equation, one has $y_i = \mathbf{x}_i^T \beta + \alpha \boldsymbol{\eta}_i + \xi_i$. Based on the second equations in (2), we can consistently estimate $\boldsymbol{\eta}$ by $\hat{\boldsymbol{\eta}} = (I - Z(Z^T Z)^{-1} Z^T)X$. It is easy to show that regressing y_i on $\mathbf{x}_i$ and $\hat{\boldsymbol{\eta}}_i$ leads to the 2SLS estimator.

High dimensionality arises in the triangular simultaneous equations model (2) when either $p \gg n$ or $d_j \gg n$ for at least one j. In this paper, I consider the scenario where only a few coefficients β_j^* are non-zero (i.e., β^* is *sparse*). The case where $d_j \gg n$ for at least one j but $p \ll n$ is similar to the model considered by Belloni and Chernozhukov (2011b), where they showed 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. 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 the triangular simultaneous equation models has not been studied in the literature. In both cases, one can still use the ideas of the 2SLS estimation or control function approach together with the Lasso techniques. 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 square estimation from the first-stage regression as usual and then apply a Lasso technique with these fitted regressors in the second-stage regression. Similarly, in the case where $p \gg n$ and $d_j \gg n$ for at least one j, one can obtain the fitted regressors by a Lasso estimator from the first-stage regression and then apply another Lasso estimator with these fitted regressors in the second-stage regression.

Compared to the 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 regularization in both stages are more flexible and particularly powerful for applications in which the researchers lack of information about the relevant explanatory variables and instruments. In terms of practical implementations, these above-mentioned high-dimensional two-stage estimation procedures enjoy similar computational complexity as the standard Lasso 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 triangular simultaneous equation models is not obvious. This paper 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.

Statistical estimation when the dimension is larger than the sample size is now an active and challenging field. 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; Negahban, Ravikumar, Wainwright, and Yu, 2011; Belloni and Chernozhukov, 2011a; Loh and Wainwright, 2012); more references can be found in the recent books by Bühlmann and van de Geer (2011), as well as the lecture notes by Koltchinskii (2011), Belloni and Chernozhukov (2011b). In recent years, these techniques have received popularity in several areas, such as biostatistics and imaging. Some first applications are now available in economics. Rosenbaum and Tsybakov (2010) deal with the high-dimensional errors-in-variables problem where the non-random regressors are observed with error and discuss an application to hedge fund portfolio replication. Belloni and Chernozhukov (2011a) study the l_1-penalized quantile regression and give an application to cross-country growth analysis. Belloni and Chernozhukov (2010) present various applications of the Lasso to economics including wage regressions, in particular, the selection of instruments in such models. Belloni, Chernozhukov and Hansen (2010) use the Lasso to estimate the optimal instruments with an application to the impact of eminent domain on economic outcomes.

In the presence of endogenous regressors, the direct implementation of the Lasso or Dantzig selector fails as the zero coefficients in equation (1) do not correspond to the zero coefficients in a linear projection type of model. The linear instrumental variable model with a single or a few endogenous regressors and many instruments has been studied among the high dimensional econometrics literature. For example, Belloni, Chernozhukov, and Hansen (2011) consider the following triangular simultaneous equation model:

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

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) in the wage equation such as innate abilities.

In many applications, the number of endogenous regressors is also large 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 $\mathbb{E}(X_i | \epsilon_i) \neq 0$ 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.) sold within a market (say, market i) or by a firm (say, firm i). Ideally, one might want to regress the price on the production costs, which tend to be exogenous. However, since costs are usually unobserved by the researchers, a proxy of costs is needed. Intuitively, the characteristics of firm i's (or, market i's) products can be used as a proxy. There are two major issues with using the characteristics of firm i's (or, market i's) products as the instruments. First, the number of explanatory variables formed by the characteristics (and the transformations of these characteristics) of products such as personal computers, automobiles, and residential houses 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, including 23 processor-type dummies and 9 operating system-type dummies (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, 2010). Second, firm i's 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 actually correlated with unobserved components of price (Ackerberg and Crawford, 2009). An alternative is to use other firms' (other markets') characteristics as the instruments for firm i's (market i's) characteristics. In demand estimation literature, this type of instruments are sometime referred to as BLP instruments, e.g., Berry, et. al., 1995 (respectively, Hausman instruments, e.g., Nevo, 2001).

The case of many endogenous regressors and many instrumental variables has been studied in the Generalized Method of Moments (GMM) context by Fan and Liao (2011), and Gautier and Tsybakov (2011). 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 $\mathbb{E}(\mathbf{z}_i\epsilon_i) = \mathbf{0}$. They discuss without proofs the STIV procedure with estimated linear projection type instruments, akin to the 2SLS procedure, and find it works successfully in simulation. Gautier and Tsybakov also speculate the rate of convergence for this type of two-stage estimation procedures when both stage equations are in the high-dimensional settings. As will be shown in the subsequent section, their speculation partially agrees with the results in this paper.

In the low-dimensional setting, the properties of the 2SLS and GMM estimators are well-understood. However, it is unclear how the regularized 2SLS procedures compare to the regularized GMM procedures in the high-dimensional setting. Consequently, it is important to study these regularized two-stage high-dimensional estimation procedures in depth. Moreover, the regularized 2SLS procedures are more intuitive and easier to be understood relative to the high-dimensional GMM estimators proposed in previous literature. Furthermore, while existing studies have provided

useful tools for analyzing the statistical properties of the GMM type of high-dimensional estimators, an important contribution of this paper is introducing a set of techniques that are particularly suitable for showing estimation consistency and selection consistency of the two-step type of high-dimensional estimators. In summary, the aims of this paper, as mentioned earlier, are to provide a theoretical justification that has not been given in literature for these regularized 2SLS procedures in the high-dimensional setting.

I present the basic definitions and notations in Section 2. Results regarding the statistical properties (including the estimation consistency and selection consistency) of the high-dimensional 2SLS procedure under the sparsity scenario are established in Section 3. Section 4 presents simulation results. Section 5 discusses future work. All the proofs are collected in the appendices (Section 6).

2 Notations and definitions

Notation. For the convenience of the reader, I summarize here notations to be used throughout this paper. The l_q norm of a vector $v \in m \times 1$ is denoted by $|v|_q$, $1 \leq q \leq \infty$ where $|v|_q := (\sum_{i=1}^m |v_i|^q)^{1/q}$ 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_1-operator norm (maximum absolute column sum) of A is denoted by $||A||_1 := \max_i \sum_j |a_{ij}|$. I make use of the bound $||A||_1 \leq \sqrt{m}||A||_2$ for any symmetric matrix $A \in \mathbb{R}^{m \times m}$. For a matrix Σ, denote its minimum eigenvalue and maximum eigenvalue by $\lambda_{\min}(\Sigma)$ and $\lambda_{\max}(\Sigma)$, respectively. 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 $c^{'} \in (0, \infty)$. $f(n) \asymp g(n)$ when $f(n) \gtrsim g(n)$ and $f(n) \lesssim g(n)$ 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 $\mathbb{B}_1^m(r) := \{v \in \mathbb{R}^m \,|\, |v|_2 \leq r\}$. 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 $|J|$.

I will begin with a brief discussion of the case where all components in X in (1) are *exogenous*. Assume the number of regressors p in equation (1) grows with and exceeds the sample size n. Let us focus on the class of models where β^* has at most k non-zero parameters, where k is also allowed to increase to infinity with p and n. Consider the following Lasso program:

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

where $\lambda_n > 0$ is some tuning parameter.

In the high-dimensional setting, it is well-known that a sufficient condition for l_q- consistency of the Lasso estimate $\hat{\beta}_{Las}$ is that the matrix $\frac{1}{n}X^TX$ satisfies some type of restricted eigenvalue (RE) conditions (see, e.g., Bickel, et. al., 2009; Meinshausen and Yu, 2009; Negahban, et. al., 2010; Raskutti et al., 2010; Bühlmann and van de Geer, 2011; Loh and Wainwright, 2012). In this paper, I will use the following definitions, referred to as RE1 and RE2, respectively.

Definition 1 (RE1): The matrix $X \in \mathbb{R}^{n\times p}$ satisfies the RE1 condition over a subset $S \subseteq \{1, 2, ..., p\}$ with parameter (δ, γ) if

$$\frac{\frac{1}{n}|Xv|_2^2}{|v|_2^2} \geq \delta > 0 \qquad \text{for all } v \in \mathbb{C}(S; \gamma)\backslash\{\mathbf{0}\}, \tag{3}$$

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.

The intuition behind RE1 is that in the high-dimensional setting where $p > n$, the Hessian is a $p \times p$ matrix with rank at most n, so it is impossible to guarantee that its eigenvalues will be uniformly bounded away from 0. RE1 condition relaxes the stringency of the uniform eigenvalue condition but only requires it to hold for a suitable subset $\mathbb{C}(S; \gamma)$ of vectors. When the unknown vector $\beta^* \in \mathbb{R}^p$ is sparse, a natural choice of S is the support set of β^*, i.e., $J(\beta^*)$. RE1 is a much milder condition than 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 RE1 condition but not vice versa. Additionally, Raskutti et al., 2010 give examples of matrix families for which the RE1 condition holds, but the restricted isometry constants tend to infinity as $(n, |S|)$ grow. Furthermore, they show that even if a matrix exhibits a high amount of dependency among the covariates, it might still satisfy RE1. To make it 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, and universal constants c_1, c_2 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 $1 - c_1 \exp(-c_2 n)$. The bound above ensures the RE1 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 $n > 32(\frac{\kappa_2}{\kappa_1})^2 k \log p$. Rudelson and Zhou (2011) as well as Loh and Wainwright (2012) extend this type of analysis from the case of Gaussian designs to the case of sub-Gaussian designs. In this paper, I make use of the following definition for a sub-Gaussian matrix.

Definition 2: A random variable X with mean $\mu = \mathbb{E}[X]$ is sub-Gaussian if there is a positive number σ such that

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

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 $u^T A_i$ is sub-Gaussian with parameter at most σ_A^2.

For example, if $A \in \mathbb{R}^{n \times p}$ is formed by independently sampling each row $A_i \sim N(0, \Sigma_A)$, then the resulting matrix $A \in \mathbb{R}^{n \times p}$ is a sub-Gaussian matrix with parameters $(\Sigma_A, ||\Sigma_A||_2)$, recalling $||\Sigma_A||_2$ denotes the spectral norm of Σ_A.

In addition, the following definition from Loh and Wainwright will be useful for analyzing the statistical properties of the two-stage estimators for the triangular simultaneous equations models in this paper. In some sense, Definition 3 (RE2) can be viewed as a sufficient condition for RE1.

Definition 3 (RE2): The random matrix $\hat{\Gamma} \in \mathbb{R}^{p \times p}$ satisfies the RE2 condition with $\alpha > 0$ and tolerance $\tau(n, p) > 0$ if

$$v^T \hat{\Gamma} v \geq \alpha |v|_2^2 - \tau(n, p)|v|_1^2, \qquad \text{for all } v \in \mathbb{R}^p. \tag{4}$$

3 High-dimensional 2SLS estimation

Suppose from our first-stage regression, we obtain estimates $\hat{\pi}_j$ and let $\hat{\mathbf{x}}_j = \mathbf{z}_j \hat{\pi}_j$ 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:

$$\hat{\beta}_{H2SLS} \in \text{argmin}_{\beta \in \mathbb{R}^p} : \frac{1}{2n}|y - \hat{X}\beta|_2^2 + \lambda_n |\beta|_1. \tag{5}$$

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

Lemma 3.1 (General upper bound on the l_2-error). Let $\hat{\Gamma} = \hat{X}^T\hat{X}$ and $e = (X - \hat{X})\beta^* + \boldsymbol{\eta}\beta^* + \epsilon$. Suppose the random matrix $\hat{\Gamma}$ satisfies the RE1 condition (3) with $\gamma = 3$ and the vector β^* is supported on a subset $J(\beta^*) \subseteq \{1, 2, ...p\}$ with its cardinality $|J(\beta^*)| \leq k$. If we solve the Lasso (5) with the choice of

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

then for any optimal solution $\hat{\beta}_{H2SLS}$, there is a constant $c > 0$ such that

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

Remarks

Notice that the choice of λ_n in Lemma 3.1 depends on unknown quantities and therefore, Lemma 3.1 does not provide guidance to practical implementation. Rather, it should be viewed as an intermediate lemma for proving consistency of the two-stage estimator later on. The idea of the choice of the tuning parameter λ_n is that it should "overrule" the empirical process part $\frac{1}{n}\hat{X}^T e$ so that we can work with deterministic argument (this type of approach is standard in the high-dimensional statistics literature). As will become clear in the subsequent sections, we can bound the term $|\frac{1}{n}\hat{X}^T e|_\infty$ from above and the order of the resulting upper bound can be used to set the tuning parameter λ_n. In order to apply Lemma 3.1 to prove consistency, we need to show (i) $\hat{\Gamma} = \hat{X}^T\hat{X}$ satisfies the RE1 condition (3) with $\gamma = 3$ and (ii) the term $|\frac{1}{n}\hat{X}^T e|_\infty \precsim \sqrt{\frac{\log p}{n}}$ with high probability, then we can show

$$|\hat{\beta}_{H2SLS} - \beta^*|_2 \precsim \sqrt{\frac{k \log p}{n}}$$

by choosing $\lambda_n \asymp \sqrt{\frac{\log p}{n}}$. The assumption $\frac{k \log p}{n} \to 0$ will therefore imply the l_2-consistency of $\hat{\beta}_{H2SLS}$. Applying Lemma 3.1 to the triangular simultaneous equations model (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 RE1 condition as an assumption on the design matrix (e.g., Belloni, et. al. 2010; Belloni and Chernozhukov; 2011b) in analyzing the statistical properties of the Lasso, a substantial amount of analysis is needed in this paper to verify that $\hat{X}^T\hat{X}$ satisfies the RE1 condition because $\hat{X}$ is obtained from a first-stage estimation and there is no guarantee that the random matrix $\hat{X}^T\hat{X}$ would automatically satisfy the RE1 condition. To the best of my knowledge, previous literature has not dealt with this issue directly. Consequently, the RE analysis introduced in this paper 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 paper 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 sparsity case.

3.1 Estimation consistency for the sparsity case

Assumption 3.1: The numbers of regressors $p(=p_n)$ and $d_j(=d_{jn})$ for every $j=1,...,p$ in model (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 d_j, p, and n, for $j=1,...,p$.

Assumption 3.2: ϵ (and η_j for $j=1,...,p$) is an *i.i.d.* zero-mean sub-Gaussian vector with the parameter σ_ϵ^2 (and respectively σ_η^2). The random matrix $\mathbf{z}_j \in \mathbb{R}^{n\times d_j}$ is sub-Gaussian with parameters at most $(\Sigma_{Z_j}, \sigma_Z^2)$ for all $j=1,...,p$.

Assumption 3.3: For every $j=1,...,p$, $\mathbf{x}_j^* := \mathbf{z}_j\pi_j^*$. $X^* \in \mathbb{R}^{n\times p}$ is a sub-Gaussian matrix with parameters at most $(\Sigma_{X^*}, \sigma_{X^*}^2)$ where the jth column of X^* is $\mathbf{x}_j^*$.

Assumption 3.4: For every $j=1,...,p$, $\mathbf{w}_j := \mathbf{z}_j v_j$ where $v_j \in \mathbb{K}(k_1, d_j) := \mathbb{B}_0^{d_j}(k_1) \cap \mathbb{B}_2^{d_j}(1)$. $W \in \mathbb{R}^{n\times p}$ is a sub-Gaussian matrix with parameters at most (Σ_W, σ_W^2) where the jth column of W is $\mathbf{w}_j$.

Assumption 3.5a: For every $j=1,...,p$, the first stage estimator $\hat{\pi}_j$ satisfies the bound $|\hat{\pi}_j - \pi_j|_1 \precsim \frac{c\sigma_\eta}{\lambda_{\min}(\Sigma_Z)} k_1 \sqrt{\frac{\log d_j}{n}}$ with probability close to 1 (e.g., Lasso, Dantzig selector), where $\lambda_{\min}(\Sigma_Z) = \min_{j=1,...,p} \lambda_{\min}(\Sigma_{Z_j})$.

Assumption 3.5b: For every $j=1,...,p$, the first stage estimator $\hat{\pi}_j$ satisfies the bound $|\hat{\pi}_j - \pi_j|_2 \precsim \frac{c\sigma_\eta}{\lambda_{\min}(\Sigma_Z)} \sqrt{\frac{k_1 \log d_j}{n}}$ with probability close to 1 (e.g., Lasso, Dantzig selector), where $\lambda_{\min}(\Sigma_Z) = \min_{j=1,...,p} \lambda_{\min}(\Sigma_{Z_j})$.

Assumption 3.6: For every $j=1,...,p$, with probability close to 1, 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 π_j^* where $k_j^* \ll n$. 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 3.1 is standard in the literature on sparsity in high-dimensional linear models. Assumption 3.2 is common in the literature (see, Rosenbaum and Tsybakov, 2010; Negahban, et.

al 2010; Loh and Wainwright, 2012). This type of assumptions allow us to evoke large-deviation bounds of the Bernstein type (see, Vershynin) based on sub-exponential random variables. In particular, if U is a zero-mean sub-Gaussian random variable with parameter σ, then the random variable $Y = U^2 - \mathbb{E}(U^2)$ is sub-exponential[1] with parameter at most $2\sigma^2$ (see Vershynin). Loh and Wainwright, 2012 extends this type of analysis from the sub-Gaussian variable to the sub-Gaussian matrix.

Based on the second part of Assumption 3.2 that $\mathbf{z}_j \in \mathbb{R}^{n\times d_j}$ is sub-Gaussian with parameters at most $(\Sigma_{Z_j}, \sigma_Z^2)$ for all $j = 1, ..., p$, we have that $\mathbf{z}_j\pi_j^* := \mathbf{x}_j^*$ and $\mathbf{z}_j v_j := \mathbf{w}_j$ (where $v_j \in \mathbb{K}(k_1, d_j) := \mathbb{B}_0^{d_j}(k_1) \cap \mathbb{B}_2^{d_j}(1)$) are sub-Gaussian vectors. Therefore, the conditions that $X^* \in \mathbb{R}^{n\times p}$ is a sub-Gaussian matrix with parameters at most $(\Sigma_{X^*}, \sigma_{X^*}^2)$ where the jth column of X^* is $\mathbf{x}_j^*$(Assumption 3.3) and $W \in \mathbb{R}^{n\times p}$ is a sub-Gaussian matrix with parameters at most (Σ_W, σ_W^2) where the jth column of W is $\mathbf{w}_j$ (Assumption 3.4) are mild extensions.

For Assumptions 3.5a(b), 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. 2010), STIV estimator (Gautier and Tsybakov, 2011), etc., simultaneously satisfy the error bounds $|\hat{\pi}_j - \pi_j|_1 \precsim \frac{c\sigma_\eta}{\lambda_{\min}(\Sigma_Z)} k_1 \sqrt{\frac{\log d_j}{n}}$ (Assumption 3.5a) and $|\hat{\pi}_j - \pi_j|_2 \precsim \frac{c\sigma_\eta}{\lambda_{\min}(\Sigma_Z)} \sqrt{\frac{k_1 \log d_j}{n}}$ (Assumption 3.5b) with high probability. The reason I introduce Assumptions 3.5a and 3.5b separately is explained in the following paragraph.

Assumption 3.6 says that the first-stage estimators correctly select the non-zero coefficients with probability close to 1. It is known that under some stringent conditions such as the *irrepresentable condition* (Bühlmann and van de Geer, 2011) and the *mutual incoherence condition* (Wainwright, 2009), 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 3.6 is the key condition that differentiates the upper bounds in the two theorems to be presented immediately.

First, I present two results for the case where $p \gg n$ and $d_j \gg n$ for at least one j under the sparsity condition. As discussed earlier, 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 3.6. With this assumption, the estimation error of the parameters of interests in the main equation can be bounded by the first-stage estimation error in l_2-norm. However, in the absence of the selection-consistency in the first-stage estimation, the first-stage statistical error in l_2-norm is not enough for bounding the estimation error in the second-stage estimation. Rather, the estimation error of the parameters of interests in the second-stage estimation needs to be bounded by the first-stage estimation error

[1] A random variable U with mean $\mu = \mathbb{E}(U)$ is sub-exponential if there are non-negative parameters (φ, b) such that $\mathbb{E}(\exp(t(U-\mu))] \leq \exp(\varphi^2 t^2/2)$ for all $|t| < \frac{1}{b}$.

in l_1-norm. As discussed in the previous remarks, for many known high-dimensional estimation procedures such as Lasso, Dantzig selector, and STIV estimator, the upper bounds on the l_1-error usually equal the upper bounds on the l_2-error multiplied by a factor of $\sqrt{k_1}$.

Theorem 3.2 (Upper bound on the l_2-error and estimation consistency): Suppose Assumptions 1.1, 3.1-3.3, and 3.5a hold. Let $d = \max_{j=1,\ldots,p} d_j$. Then, under the scaling

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

and the choice of the tuning parameter

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

we have, 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,

$$|\hat{\beta}_{H2SLS} - \beta^*|_2 \precsim \psi_1 |\beta^*|_1 \max\left\{ \sqrt{k_1 k_2} \sqrt{\frac{k_1 \log d}{n}},\, \sqrt{\frac{k_2 \log p}{n}} \right\},$$

where

$$\psi_1 = \max\left\{ \frac{\sigma_\eta \max_{j,j'} |\mathrm{cov}(x^*_{1j'},\, \mathbf{z}_{1j})|_\infty}{\lambda_{\min}(\Sigma_Z)\lambda_{\min}(\Sigma_{X^*})},\, \frac{\sigma_{X^*} \max(\sigma_\epsilon,\, \sigma_\eta)}{\lambda_{\min}(\Sigma_{X^*})} \right\}.$$

If we also have $k_2 k_1 \sqrt{\frac{k_2 \log d}{n}} = o(1)$ and $k_2 \sqrt{\frac{k_2 \log p}{n}} = o(1)$, then[2] the two-stage estimator $\hat{\beta}_{H2SLS}$ is l_2-consistent for β^*.

Theorem 3.3 (An improved upper bound on the l_2-error and estimation consistency): Suppose Assumptions 1.1, 3.1-3.4, 3.5b, and 3.6 hold. Let $d = \max_{j=1,\ldots,p} d_j$. Then, under the scaling

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

and the choice of the tuning parameter

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

we have, with probability at least $1 - c_1 \exp(-c_2 \log \max(\min(p,\, d),\, n))$ for some universal positive

[2]The extra factor of k_2 in front of these scaling conditions in Theorem 3.2 (as well as in the subsequent theorems 3.2-3.6) comes from the simple inequality $|\beta^*|_1 \leq k_2 \max_j \beta^*_j$.

constants c_1 and c_2,

$$|\hat{\beta}_{H2SLS} - \beta^*|_2 \precsim \psi_2 |\beta^*|_1 \max\left\{\sqrt{k_2}\sqrt{\frac{k_1 \log d}{n}},\ \sqrt{\frac{k_2 \log p}{n}}\right\},$$

where

$$\psi_2 = \max\left\{\frac{\sigma_\eta \max_{j',j} \sup_{v^j \in \mathbb{K}(k_1, d_j)} \left|\mathrm{cov}(x^*_{1j'},\ \mathbf{z}_{1j} v^j)\right|}{\lambda_{\min}(\Sigma_Z)\lambda_{\min}(\Sigma_{X^*})},\ \frac{\sigma_{X^*} \max(\sigma_\epsilon,\ \sigma_\eta)}{\lambda_{\min}(\Sigma_{X^*})}\right\}.$$

If we also have $k_2\sqrt{\frac{k_1 k_2 \log d}{n}} = o(1)$ and $k_2\sqrt{\frac{k_2 \log p}{n}} = o(1)$, then the two-stage estimator $\hat{\beta}_{H2SLS}$ is l_2-consistent for β^*.

Remarks

The proofs for Theorems 3.2 and 3.3 consist of two parts. The first part is to show $\hat{X}^T\hat{X}$ satisfies the RE1 condition (3) and the second part is to bound the term $|\frac{1}{n}\hat{X}^T e|_\infty$ from above. Based on Lemma 3.1, the upper bound on $|\frac{1}{n}\hat{X}^T e|_\infty$ pins down the scaling requirement of λ_n, as mentioned previously. The scaling requirement of n and λ_n depends on the sparsity parameters k_1 and k_2, which are typically unknown. Nevertheless, I will assume that upper bounds on k_1 and k_2 are available, i.e., we know that $k_1 \leq \bar{k}_1$ and $k_2 \leq \bar{k}_2$ for some integers $\bar{k}_1$ and $\bar{k}_2$ that grow with n just like k_1 and k_2. Meaningful values of $\bar{k}_1$ and $\bar{k}_2$ are small relative to n presuming that only a few regressors are relevant. This type of upper bound assumption on the sparsity is called *sparsity certificate* in the literature (see, e.g., Gautier and Tsybakov, 2011).

Under the assumption that the first-stage estimators correctly select the non-zero coefficients with high probability (Assumption 3.6), the scaling of the smallest sample size n in Theorem 3.3 is guaranteed to be no greater than that in Theorem 3.2. The optimal choice of r in the scaling requirement depends on the combinations of d, p, k_1, and k_2. In practice, it is not always necessary for the researcher to determine the optimal r to evaluate whether his/her sample size is large enough. For instance, if $k_1^3 \leq k_2$ is known *a priori* to the researcher and also $p \leq d$, then by letting $r = 0$,

$$\max\left\{k_1^3 \log d,\ k_2 \log d,\ k_2 \log p\right\} = k_2 \log d \leq \max\left\{k_1^2 k_2^2 \log d,\ k_2 \log p\right\} = k_1^2 k_2^2 \log d.$$

In this example, Theorem 3.2 suggests that the choice of sample size needs to satisfy $\frac{k_1^2 k_2^2 \log d}{n} = o(1)$ while Theorem 3.3 suggests that the choice of sample size only needs to satisfy $\frac{k_2 \log d}{n} = o(1)$.

From Theorem 3.2 (respectively, Theorem 3.3), we see that the estimation error of the parameters of interests in the main equation is of the order of the maximum of the first-stage estimation error in l_2-norm multiplied by a factor of $\sqrt{k_1 k_2}$ (respectively, $\sqrt{k_2}$) and the second-stage estimation error. These results partially agree[3] with the speculation in Gautier and Tsybakov (2011) (Section 7.2)

[3] To verify whether the rate $\sqrt{\frac{\log p}{n}}$ is achievable (or not) by any procedure for the triangular simultaneous linear equations models, a minimax lower bound result needs to be established in future work.

that the two-stage estimation procedures can reduce the estimation error from an order of $\sqrt{\frac{\log L}{n}}$ to $\sqrt{\frac{\log p}{n}}$, where in their notations, L denotes the number of instruments and $L \geq p \gg n$. My results show that the reduction from $\sqrt{\frac{\log L}{n}}$ to $\sqrt{\frac{\log p}{n}}$ occurs when when the second-stage estimation error dominates the first-stage estimation error. On the other hand, if $d = O(L)$ and the first-stage estimation error dominates the second-stage estimation error, then the estimation error cannot be reduced from $\sqrt{\frac{\log L}{n}}$ to $\sqrt{\frac{\log p}{n}}$.

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 in the context of the high-dimensional linear regression model without endogeneity, i.e., $\sqrt{\frac{k_2 \log p}{n}}$. A special case occurs when the number of regressors in the main equation is large relative to the number of regressors in each of the first-stage equations and the result is formally stated in Corollary 3.4 below.

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 and 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 on the estimated supports 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, 2010).

Under this three-stage procedure, Corollary 3.4 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 achieve the 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 becomes invalid. 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 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) exhibit a high amount of dependency among the covariates.

For Theorem 3.2 (respectively, Theorem 3.3), we give an explicit form of the first-stage estimation

error in Assumptions 3.5a (respectively, 3.5b) and as discussed earlier, Lasso types of techniques yield these estimation errors. However, the claim that, the estimation error of the parameters of interests in the main equation can be bounded by the maximum of the first-stage estimation error in l_2-norm multiplied by a factor of $\sqrt{k_1 k_2}$ (or $\sqrt{k_2}$ if the first-stage estimators correctly select the non-zero coefficients with probability close to 1) and the second-stage estimation error, can be made for general first-stage estimation errors. This claim is formally stated in Theorems 3.5 and 3.6 below.

Upon an additional condition that the first-stage estimators correctly select the non-zero coefficients with probability close to 1, note that the statistical error of the high-dimensional two-stage estimator $\hat{\beta}_{H2SLS}$ in Theorem 3.3 (Theorem 3.6) is improved upon that in Theorem 3.2 (Theorem 3.5) by a factor of $\sqrt{k_1}$ if the first term in the braces dominates the second one. The improvement of the scaling of n for general first-stage estimation errors is also observed in Theorem 3.6 when it is compared to Theorem 3.5.

The factor of $\sqrt{k_1}$ improvement in the estimation error comes from the fact that, when the first-stage estimators correctly select the non-zero coefficients with high probability, the estimation error of the parameters of interests in the main equation is bounded by the first-stage estimation error in l_2-norm. As discussed earlier, in the absence of the selection-consistency in the first-stage estimation, the estimation error of the parameters of interests in the second-stage estimation needs to be bounded by the first-stage estimation error in l_1-norm.

In Theorems 3.2 and 3.3, we see that the l_2-error also depends on a quantity involving σ_η, σ_ϵ, σ_{X^*}, $\lambda_{\min}(\Sigma_Z)$, $\lambda_{\min}(\Sigma_{X^*})$, $\max_{j,j'} |\text{cov}(x^*_{1j'}, \mathbf{z}_{1j})|_\infty$, or $\max_{j',j} \sup_{v^j \in \mathbb{K}(k_1, d_j)} \left|\text{cov}(x^*_{1j'}, \mathbf{z}_{1j} v^j)\right|$. In the simple case of $\boldsymbol{\eta} = \mathbf{0}$ with probability 1 (i.e., high-dimensional linear regression models without endogeneity), one has $\sigma_\eta = 0$ and therefore the multipliers ψ_1 in Theorem 3.2 and ψ_2 in Theorem 3.3 reduce to $\frac{\sigma_{X^*}\sigma_\epsilon}{\lambda_{\min}(\Sigma_{X^*})}$, a factor that has a natural interpretation of an inverse signal-to-noise ratio. For instance, when X^* is a zero-mean Gaussian matrix with covariance $\Sigma_{X^*} = \sigma_{X^*} I$, 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 signal-to-noise of the regressors in a high-dimensional linear regression model without endogeneity.

The terms $\max_{j,j'} |\text{cov}(x^*_{1j'}, \mathbf{z}_{1j})|_\infty$ (Theorem 3.2) and $\max_{j',j} \sup_{v^j \in \mathbb{K}(k_1, d_j)} \left|\text{cov}(x^*_{1j'}, \mathbf{z}_{1j} v^j)\right|$ (Theorem 3.3) are related to the degree of multi-collinearity between the columns of the design matrices in the first-stage regressions. For instance, for any $l = 1, ..., d_j$ and $j = 1, ..., p$, notice that

$$\text{cov}(x^*_{1j}, z_{1jl}) = \text{cov}(\mathbf{z}_{1j}\pi^*_j, z_{1jl}).$$

The greater $\max_l \text{cov}(x^*_{1j}, z_{1jl})$ is, the more multi-collinearity between the columns of the design matrix $\mathbf{z}_j$ we would expect, and the harder the estimation problem becomes. In the special case of $\max_{j,j'} |\text{cov}(x^*_{1j'}, \mathbf{z}_{1j})|_\infty = \sigma^2_Z$, $\lambda_{\min}(\Sigma_Z) = \sigma^2_Z$, and $\lambda_{\min}(\Sigma_{X^*}) = \sigma^2_{X^*}$, the first term in the

maximum expression of ψ_1 in Theorem 3.2 reduces to $\frac{\sigma_\eta}{\sigma^2_{X^*}} = \frac{1}{\sigma_{X^*}}\left(\frac{\sigma_\eta}{\sigma_{X^*}}\right)$, which can be related to the ratio of the signal of the true fitted regressors to the noise of the first-stage error terms. The term $\max_{j',j} \sup_{v^j \in \mathbb{K}(k_1,d_j)} \left|\text{cov}(x^*_{1j'}, \mathbf{z}_{1j} v^j)\right|$ can be interpreted in a similar way.

In analogy to the various sparsity assumptions on the true parameters in the high-dimensional statistics literature (including the case of 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 3.6 can be interpreted as a sparsity constraint on the first-stage estimate $\hat{\pi}_j$ for $j = 1, ..., p$, in terms of the l_0-ball, given by

$$\mathbb{B}^d_0(k_1) := \left\{ \hat{\pi}_j \in \mathbb{R}^d \,|\, \sum_{l=1}^{d} 1\{\hat{\pi}_{jl} \neq 0\} \leq k_1 \right\} \text{ for } j = 1, ..., p.$$

As discussed earlier, the sparsity constraint (namely, the selection consistency) of these first-stage estimators is guaranteed under some conditions that may be violated in many problems. It is plausible to extend Assumption 3.6 to the following approximate sparsity constraint on the first-stage estimates in terms of l_1- balls, given by

$$\mathbb{B}^d_1(R_j) := \left\{ \hat{\pi}_j \in \mathbb{R}^d \,|\, |\hat{\pi}_j|_1 = \sum_{l=1}^{d} |\hat{\pi}_{jl}| \leq R_j \right\} \text{ for } j = 1, ..., p.$$

If the first-stage estimation employs a Lasso or Dantzig procedure, then we are guaranteed to have $\hat{\pi}_j \in \mathbb{B}^d_1(R_j)$ for every $j = 1, ..., p$. Depending on the type of sparsity assumptions imposed on the first-stage estimates, the statistical error of the high-dimensional two-stage estimator $\hat{\beta}_{H2SLS}$ in l_2-norm and the requirement of the smallest sample size differ. An inspection of the proof for Theorem 3.2 reveals that the error bound and requirement of the smallest sample size in Theorem 3.2 will hold regardless of the sparsity assumption on the first-stage estimates. However, under these special structures that impose a certain decay rate on the ordered entries of the first-stage estimates, the bound and scaling of the smallest sample size in Theorem 3.2 is likely to be suboptimal. In order to obtain sharper results, the proof technique adopted for showing Theorem 3.3 seems more appropriate. I provide a heuristic truncation argument to illustrate how the proof for Theorem 3.3 can be extended to allow the weaker sparsity constraint (in terms of l_1-balls) on the first-stage Lasso estimates. Suppose for every $j = 1, ..., p$, we choose the top k^j coefficients of $\hat{\pi}_j$ in absolute value, then the fast decay imposed by the l_1- ball condition on $\hat{\pi}_j$ arising from the Lasso procedure would mean that the remaining $d_j - k^j$ coefficients would have relatively little impact. With this intuition, the proof follows as if Assumption 3.6 were imposed with the only exception that we also need to take into account the approximation error arising from the the remaining $d_j - k^j$ coefficients of $\hat{\pi}_j$.

Corollary 3.4 (First-stage estimation in the low-dimensional setting): Assume the number of non-zero components in β^* is at most k_2 and let $d = \max_{j=1,...,p} |\pi_j| \ll n$ where $|\cdot|$ denotes the number

of components in π_j. Assume that, for every $j = 1, ..., p$, the first-stage estimator $\hat{\pi}_j$ satisfies the bound $|\hat{\pi}_j - \pi_j|_2 \precsim \sqrt{\frac{1}{n}}$ with probability close to 1. Suppose Assumptions 1.1, 3.2, and 3.3 hold. Then, under the scaling

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

and the choice of the tuning parameter

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

we have

$$|\hat{\beta}_{H2SLS} - \beta^*|_2 \precsim \frac{|\beta^*|_1 \sigma_{X^*} \max(\sigma_\epsilon, \sigma_\eta)}{\lambda_{\min}(\Sigma_{X^*})} \sqrt{\frac{k_2 \log p}{n}},$$

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

Theorem 3.5: Suppose Assumptions 1.1 and 3.1-3.3 hold. Also, for every $j = 1, ..., p$, let the first-stage estimator $\hat{\pi}_j$ satisfies the bound $|\hat{\pi}_j - \pi_j|_1 \leq \sqrt{k_1} M(d, k_1, n)$ with probability close to 1. Then, under the scaling

$$\max \left\{ k_2 \sqrt{k_1} M(d, k_1, n), \frac{k_2 \log d}{n}, \frac{k_2 \log p}{n} \right\} = o(1),$$

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

we have

$$|\hat{\beta}_{H2SLS} - \beta^*|_2 \precsim \frac{|\beta^*|_1}{\lambda_{\min}(\Sigma_{X^*})} \max \left\{ \sqrt{k_2} \sqrt{k_1} M(d, k_1, n), \sigma_{X^*} \max(\sigma_\epsilon, \sigma_\eta) \sqrt{\frac{k_2 \log p}{n}} \right\},$$

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 $k_2 \max \left\{ \sqrt{k_2} \sqrt{k_1} M(d, k_1, n), \sqrt{\frac{k_2 \log p}{n}} \right\} = o(1)$, then the two-stage estimator $\hat{\beta}_{H2SLS}$ is l_2-consistent for β^*.

Theorem 3.6: Suppose Assumptions 1.1, 3.1-3.4, and 3.6 hold. Also, for every $j = 1, ..., p$, let the first stage estimator $\hat{\pi}_j$ satisfies the bound $|\hat{\pi}_j - \pi_j|_2 \leq M(d, k_1, n)$ with probability close to 1. Then, under the scaling

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

$$= o(1),$$

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

we have

$$|\hat{\beta}_{H2SLS} - \beta^*|_2 \precsim \frac{|\beta^*|_1}{\lambda_{\min}(\Sigma_{X^*})} \max\left\{ \sqrt{k_2} M(d,\, k_1,\, n),\, \sigma_{X^*} \max(\sigma_\epsilon,\, \sigma_\eta) \sqrt{\frac{k_2 \log p}{n}} \right\},$$

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 $k_2 \max\left\{ \sqrt{k_2} M(d,\, k_1,\, n),\, \sqrt{\frac{k_2 \log p}{n}} \right\} = o(1)$, then the two-stage estimator $\hat{\beta}_{H2SLS}$ is l_2-consistent for β^*.

3.2 Variable-selection consistency

In this subsection, I address the following question: given an optimal two-stage Lasso solution $\hat{\beta}_{H2SLS}$, when is its support set, $J(\hat{\beta}_{H2SLS})$, exactly equal to the true support $J(\beta^*)$? That is, when can we conclude $\hat{\beta}_{H2SLS}$ correctly selects the non-zero coefficients in the main equation with high probability? I refer to this property 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" (Zhao and Yu, 2006), is sufficient and necessary. A further refined analysis is given in Wainwright (2009), which presents under certain incoherence conditions the smallest sample size needed to recover a sparse signal. In this paper, I adopt the analysis by Wainwright (2009), and Ravikumar, Wainwright, and Lafferty (2009) to analyze the selection consistency of $\hat{\beta}_{H2SLS}$. In particular, I need the following assumptions.

Assumption 3.7: $\left\| \mathbb{E}\left[X^{*T}_{1,J(\beta^*)^c} X^*_{1,J(\beta^*)} \right] \left[\mathbb{E}(X^{*T}_{1,J(\beta^*)} X^*_{1,J(\beta^*)}) \right]^{-1} \right\|_1 \leq 1 - \phi$ for some $\phi \in (0,\, 1]$.

Assumption 3.8: The smallest eigenvalue of the submatrix $X^*_{J(\beta^*)}$ satisfies the bound

$$\lambda_{\min}\left(\mathbb{E}\left[X^{*T}_{1,J(\beta^*)} X^*_{1,J(\beta^*)} \right] \right) \geq C_{\min} > 0.$$

Remarks

Assumption 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.

Notice that in order for Assumption 3.7 to be scale invariant so that the quantity on the left-hand-side always falls in [0, 1), one needs to have some type of normalization on the matrix $X_j^* = (X_{1j}^*, ..., X_{nj}^*)^T$ for all $j = 1, ..., p$. One possibility is to impose a column normalization as follows

$$\max_{j=1,...,p} \frac{|X_j^*|_2}{\sqrt{n}} \leq \kappa_c, \quad 0 < \kappa_c < \infty.$$

Under Assumptions 1.1 and 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, the normalization above follows from a standard bound for the norms of zero-mean sub-Gaussian vectors and a union bound

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

where the last inequality follows from $n \gg \log p$. For example, if X^* has a Gaussian design (Raskutti, et. al, 2011), then we have

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

where $\max_{j=1,..,p} \Sigma_{jj}$ corresponds to the maximal variance of any element of X^*.

Assumption 3.8 is required to ensure that the model is identifiable if the support set $J(\beta^*)$ were known *a priori*. Assumption 3.8 is relatively mild compared to Assumption 3.7.

Theorem 3.7 (Selection consistency): Suppose Assumptions 1.1, 3.2, 3.3, 3.5a, 3.7, and 3.8 hold. If we solve the Lasso program (5) with the scaling of the tuning parameter

$$\lambda_n \asymp |\beta^*|_1 \max\left\{k_1\sqrt{\frac{\log d}{n}}, \sqrt{\frac{\log p}{n}}\right\},$$

and $k_2 \max\left\{k_1\sqrt{\frac{\log d}{n}}, \sqrt{\frac{\log p}{n}}\right\} = o(1)$, then, with probability at least $1 - c_1\exp(-c_2 \log \max(\min(p, d), n))$, we have: (a) The Lasso has a unique optimal solution $\hat{\beta}_{H2SLS}$. (b) The support $J(\hat{\beta}_{H2SLS}) \subseteq J(\beta^*)$.

(c) $$|\hat{\beta}_{H2SLS,\, J(\beta^*)} - \beta^*_{H2SLS,\, J(\beta^*)}|_\infty \leq \left[b|\beta^*|_1 \max\left\{k_1\sqrt{\frac{\log d}{n}}, \sqrt{\frac{\log p}{n}}\right\} + \lambda_n\right] \frac{4\sqrt{k_2}}{C_{\min}} := B_1$$

for some constant b. (d) If $\min_{j\in J(\beta^*)}|\beta_j^*| > B_1$, then $J(\hat{\beta}_{H2SLS}) \supseteq J(\beta^*)$ and hence $\hat{\beta}_{H2SLS}$ is variable-selection consistent, i.e., $J(\hat{\beta}_{H2SLS}) = J(\beta^*)$.

Theorem 3.8 (Selection consistency): Suppose Assumptions 1.1, 3.2-3.4, 3.5b, 3.7, and 3.8 hold. If we solve the Lasso program (5) with the scaling of the tuning parameter

$$\lambda_n \asymp |\beta^*|_1 \max\left\{\sqrt{\frac{k_1 \log d}{n}}, \sqrt{\frac{\log p}{n}}\right\},$$

and $k_2 \max\left\{\sqrt{\frac{k_1 \log d}{n}}, \sqrt{\frac{\log p}{n}}\right\} = o(1)$, then, with probability at least $1-c_1\exp(-c_2 \log\max(\min(p,\, d),\, n))$, we have: (a) The Lasso has a unique optimal solution $\hat{\beta}_{H2SLS}$. (b) The support $J(\hat{\beta}_{H2SLS}) \subseteq J(\beta^*)$.

$$\text{(c)}\quad |\hat{\beta}_{H2SLS,\, J(\beta^*)} - \beta^*_{H2SLS,\, J(\beta^*)}|_\infty \leq \left[b|\beta^*|_1 \max\left\{\sqrt{\frac{k_1 \log d}{n}}, \sqrt{\frac{\log p}{n}}\right\} + \lambda_n\right]\frac{4\sqrt{k_2}}{C_{\min}} := B_2$$

for some constant b. (d) If $\min_{j\in J(\beta^*)}|\beta_j^*| > B_2$, then $J(\hat{\beta}_{H2SLS}) \supseteq J(\beta^*)$ and hence $\hat{\beta}_{H2SLS}$ is variable-selection consistent, i.e., $J(\hat{\beta}_{H2SLS}) = J(\beta^*)$.

Remarks

The proofs for Theorems 3.7 and 3.8 are based on a constructive procedure called Primal-Dual Witness (PDW) method developed by Wainwright (2009). This procedure 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.

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

2. 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 $|\cdot|_1 : \mathbb{R}^p \to \mathbb{R}$.

3. 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.

Theorems 3.7 and 3.8 include four parts. Part (a) guarantees the uniqueness of the optimal solution of the two-stage Lasso, $\hat{\beta}_{H2SLS}$ (in fact, from the proofs for Theorems 3.7 and 3.8, we have that $\hat{\beta}_{H2SLS} = (\hat{\beta}_{J(\beta^*)}, \mathbf{0})$ where $\hat{\beta}_{J(\beta^*)}$ is the solution obtained in step 2 of the PDW construction above). Based on this uniqueness claim, one can then talk unambiguously about the support of the 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. Notice that the $l_\infty-$bound in Part (c) of Theorem 3.8 is improved by a factor of $\sqrt{k_1}$ upon that in Part (c) of Theorem 3.7 if the first term in the braces dominates the second one. Also, the scaling conditions in Theorem 3.8 are improved by a factor of $\sqrt{k_1}$ upon those in Theorem 3.7 if the first term in the braces dominates the second one. Recall that similar observations were made earlier when we compared the bound in Theorem 3.2 with the bound in Theorem 3.3 (or, the bound in Theorem 3.5 with the bound in Theorem 3.6). Again, these observations are attributed to that the additional assumption of the first-stage estimators correctly selecting the non-zero coefficients (Assumption 3.6) is imposed in Theorem 3.8 but not in Theorem 3.7. 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 Lasso correctly selects the non-zero coefficients 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).

The proof for Theorems 3.7 and 3.8 hinges on an intermediate result that shows the mutual incoherence assumption on $\mathbb{E}[X^{*T}X^*]$ (the population version of the matrix $X^{*T}X^*$) guarantees that, with high probability, analogous conditions hold for the estimated quantities, $\hat{X}^T\hat{X}$, which is based on the first-stage regression. This result is established in Lemma 6.5 in Section 6.5.

4 Simulations [Incomplete, add signal-to-noise ratio experiments]

In this section, simulations are conducted to gain insight on the finite sample performance of the regularized two-stage estimators. I consider the following model:

$$\begin{aligned} y_i &= \sum_{j=1}^{p} x_{ij}\beta^*_j + \epsilon_i, \\ x_{ij} &= \sum_{l=1}^{d} z_{ijl}\pi^*_{jl} + \eta_{ij}, \quad j = 1, ..., p, \end{aligned}$$

where $(y_i,\ \mathbf{x}_i^T,\ \mathbf{z}_i^T,\ \epsilon_i,\ \boldsymbol{\eta}_i)$ are *i.i.d.*, and $(\epsilon_i,\ \boldsymbol{\eta}_i)$ have the joint normal distribution

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

$\mathbf{z}_i^T$ is a $p \times d$ matrix of independent standard normal random variables, and $\mathbf{z}_{ij}^T$ is independent of $(\epsilon_i,\ \eta_{i1}, ..., \eta_{ip})$ for all $j = 1, ..., p$. In particular, I choose $\sigma_\epsilon = 0.15$, $\sigma_\eta = 0.3$, and $\rho = 0.1$. With this setup, I simulate 1000 data 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 data set. I perform 18 Monte Carlo simulation experiments constructed from various combinations of sample sizes n, model parameters (including d, k_1, p, k_2, and β^*), as well as the types of first-stage and second-stage estimators employed (Lasso vs. OLS). In each experiment, I compute 1000 estimates of the main-equation parameters β^*, l_2-errors of these estimates, $|\hat{\beta} - \beta^*|_2$, and selection percentages of the supports of $\hat{\beta}$ (computed by the number of the entries in $\hat{\beta}$ sharing the same sign as the entries in β^*, divided by the total number of entries in β^*). Table 4.1 displays the designs of the 18 experiments.

It is easy to see from Table 4.1 that, for each experiment, only the *first 4* parameters in each of the first-stage equations are non-zero and only the *first 5* parameters in the main equation are non-zero. Experiments 0-0, 0-1, and 0-2 concern 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*. These experiments serve as a benchmark for the two-stage Lasso procedure in Experiments 1-0, 1-1, and 1-2. Notice that there are 3 different sample sizes: 47, 470, and 4700. The smallest sample size 47 is chosen according to the scaling requirement in Theorem 3.3. Given $d_j = d = 100$, $k_{1j} = k_1 = 4$ for all $j = 1, ..., p$, $p = 50$, and $k_2 = 5$ in experiments 1-0 and 6-0, these experiments are in the high-dimensional setting under the sparsity scenario in terms of both the first-stage equations and the main equation. Given $d_j = d = k_{1j} = k_1 = 4$ for all $j = 1, ..., p$, $p = 50$, and $k_2 = 5$ in experiment 5-0, this experiment is in the high-dimensional setting under the sparsity scenario in terms of the main equation (with the first-stage equations being in the low-dimensional setting). All experiments are performed over 1000 data sets. In particular, experiments 1-0, 5-0, and 6-0 use the same 1000 data sets with each data set consisted of 47 data points; experiments 1-1, 2-1, 3-1, 4-1, 5-1, and 6-1 use the same 1000 data sets with each data set consisted of 470 data points; experiments 1-2, 2-2, 3-2, 4-2, 5-2, and 6-2 use the same 1000 data sets with each data set consisted of 4700 data points.

The tuning parameters λ_{1n} in the first-stage Lasso procedure (in experiments 1-0, 1-1, 1-2, 3-1, 3-2, 5-0, 5-1, 5-2, 6-0, 6-1, and 6-2) are chosen according to the standard Lasso theory of high-dimensional estimation techniques (e.g., Bickel, 2009); in particular, $\lambda_{1n} = \sigma_\eta\sqrt{\frac{\log d}{n}}$. The tuning

parameters λ_{2n} in the second-stage Lasso procedure (in experiments 1-0, 1-1, 1-2, 2-1, 2-2, 6-0, 6-1, and 6-2) are chosen according to the scaling requirement in Theorem 3.3; The tuning parameters λ_{2n} in the second-stage Lasso procedure (in experiments 5-0, 5-1, and 5-2) are chosen according to the scaling requirement in Corollary 3.4. 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 1-0, 1-1, 1-2, 2-1, and 2-2; $\lambda_{2n} = 0.1 \cdot k_2 \sqrt{\frac{\log p}{n}}$ in experiments 5-0, 5-1, and 5-2; and $\lambda_{2n} = 0.001 \cdot k_2 \max\left\{\sqrt{\frac{k_1 \log d}{n}}, \sqrt{\frac{\log p}{n}}\right\}$ in experiments 6-0, 6-1, and 6-2. The value of λ_{2n} in experiments 1-0, 1-1, 1-2, 2-1, 2-2, 5-0, 5-1, and 5-2 exceeds the value of λ_{2n} in experiments 6-0, 6-1, and 6-2 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 1-0, 1-1, 1-2, 2-1, 2-2, 5-0, 5-1, and 5-2 exceed the non-zero parameters $(\beta_1, ..., \beta_5) = (0.01, \ldots, 0.01)$ in experiments 6-0, 6-1, and 6-2 by a factor of 0.01.

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

Experiment #	0-0	0-1	0-2	1-0	1-1	1-2	2-1	2-2	3-1	3-2	4-1	4-2	5-0	5-1	5-2	6-0	6-1	6-2
n	47	470	4700	47	470	4700	470	4700	470	4700	470	4700	47	470	4700	47	470	4700
d	4	4	4	100	100	100	100	100	100	100	100	100	4	4	4	100	100	100
k_1	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
p	5	5	5	50	50	50	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	5	5	5	5
$(\beta_1, ..., \beta_5)$	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0.01**	**0.01**	**0.01**
$(\beta_6, ..., \beta_{50})$	NA	NA	NA	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
$(\pi_{j1}, ..., \pi_{j4})$ for all j	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
$(\pi_{j5}, ..., \pi_{j100})$ for all j	NA	NA	NA	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	NA	NA	NA	**0**	**0**	**0**
1st-stage estimation	OLS	OLS	OLS	Lasso	Lasso	Lasso	OLS	OLS	Lasso	Lasso	OLS	OLS	Lasso	Lasso	Lasso	Lasso	Lasso	Lasso
2nd-stage estimation	OLS	OLS	OLS	Lasso	Lasso	Lasso	Lasso	Lasso	OLS	OLS	OLS	OLS	Lasso	Lasso	Lasso	Lasso	Lasso	Lasso

With the 1000 estimates $\hat{\beta}_{H2SLS}$ of the main-equation parameters from experiments 0-0, 0-1, and 0-2 (experiments 1-0, 1-1, and 1-2), Table 4.2 (respectively, Table 4.3) displays the 5^{th} percentile, the median, the 95^{th} percentile, and the mean of these estimates. Table 4.4 (Table 4.5) shows the 5^{th} percentile, the median, the 95^{th} percentile, and the mean of the l_2-errors of these estimates, $|\hat{\beta}_{H2SLS} - \beta^*|_2$ and $|\hat{\beta}_{2SLS} - \beta^*|_2$ (respectively, the selection percentages of the supports of $\hat{\beta}_{H2SLS}$). The selection percentages of the supports of $\hat{\beta}_{2SLS}$ concerning the classical (low-dimensional) simultaneous linear equations model (experiments 0-0, 0-1, 0-2) are exactly 100%, as expected (hence, these statistics are not tabulated). In addition, for each of the first-stage equations $j = 1, ..., p$, I compute the 1000 estimates $\hat{\pi}_{j,\,Lasso} = (\hat{\pi}_{j1,\,Lasso}, \, ... \, \hat{\pi}_{j100,\,Lasso})$ of the corresponding nuisance parameters $\pi_j^* = (\pi_{j1}^*, \, ... \, \pi_{j100}^*)$ as well as the l_2-errors and the selection percentages of the supports of these estimates, for $j = 1, ..., p$. The 5^{th} percentile, the median, the 95^{th} percentile, and the mean of the l_2-errors and the selection percentages of the supports of these estimates are also computed for every $j = 1, ..., p$ (notice that this yields a 50×4 matrix where each row contains information with respect to the l_2-errors or the selection percentages of a single first-stage equation parameters' estimate). To provide a sense of how good these first-stage estimates are, Table 4.6 averages these quantile and mean statistics over the 50 rows, i.e., $j = 1, ..., p$.

Comparing Table 4.3 with Table 4.2, notice that in the high-dimensional setting (i.e., when $n = 47$), the two-stage Lasso estimator performs well in estimating β^*. From Table 4.4, we see that when $n = 47$, the l_2-errors of the main-equation estimate $\hat{\beta}_{H2SLS}$ are similar to those of $\hat{\beta}_{2SLS}$. When $n = 470$ and $n = 4700$, the l_2-errors of $\hat{\beta}_{H2SLS}$ and $\hat{\beta}_{2SLS}$ are very close to each other; in addition, these errors shrink as the sample size increases. $|\hat{\beta}_{2SLS} - \beta^*|_2$ being proportional to $\frac{1}{\sqrt{n}}$ is a known fact in low-dimensional settings. From Section 3.1, we also have that the upper bounds for $|\hat{\beta}_{H2SLS} - \beta^*|_2$ are proportional to $\frac{1}{\sqrt{n}}$ up to some factors involving $\log d$, $\log p$, k_1, and k_2. That $|\hat{\beta}_{2SLS} - \beta^*|_2$ and $|\hat{\beta}_{H2SLS} - \beta^*|_2$ decrease (as n increases) in the highly similar fashion suggests that $\hat{\beta}_{H2SLS}$ behaves more and more like $\hat{\beta}_{2SLS}$ as the sample size increases. The selection percentages of the main-equation estimate$\hat{\beta}_{H2SLS}$ and the first-stage estimate $\hat{\pi}_{Lasso}$ are also high (Table 4.5 and Table 4.6). As the sample size increases, notice that the estimate $\hat{\beta}_{H2SLS}$ (the selection percentages of $\hat{\beta}_{H2SLS}$ and $\hat{\pi}_{Lasso}$) are getting closer and closer to the truth β^*(respectively, the perfect selection). Additionally, from Table 4.6, we see that the l_2-errors of the first-stage estimate$\hat{\pi}_{Lasso}$ also shrink as n increases, an outcome to be expected from the standard Lasso theory (e.g., Bickel, 2009).

Table 4.2: Estimates of the main-equation parameters by the two-stage OLS procedure, Experiments 0-0, 0-1, 0-2

(β^*)	Expr. 0-0, $\hat{\beta}_{2SLS}$, $n = 47$				Expr. 0-1, $\hat{\beta}_{2SLS}$, $n = 470$				Expr. 0-2, $\hat{\beta}_{2SLS}$, $n = 4700$			
	5^{th}	Median	95^{th}	Mean	5^{th}	Median	95^{th}	Mean	5^{th}	Median	95^{th}	Mean
$\beta_1^* = 1$	0.827	0.999	1.181	1.000	0.954	1.000	1.045	1.000	0.986	1.000	1.014	1.000
$\beta_2^* = 1$	0.825	1.004	1.166	1.002	0.955	1.000	1.047	1.000	0.985	1.000	1.014	1.000
$\beta_3^* = 1$	0.827	1.005	1.174	1.002	0.955	1.000	1.045	0.999	0.986	1.000	1.015	1.000
$\beta_4^* = 1$	0.841	0.993	1.165	0.998	0.955	1.000	1.046	1.001	0.987	1.001	1.015	1.000
$\beta_5^* = 1$	0.816	0.995	1.166	0.991	0.952	1.001	1.047	1.000	0.985	1.001	1.014	1.000

Table 4.3: Estimates of the main-equation parameters by the two-stage Lasso procedure, Experiments 1-0, 1-1, 1-2

(β^*)	Expr. 1-0, $\hat{\beta}_{H2SLS}$, $n = 47$				Expr. 1-1, $\hat{\beta}_{H2SLS}$, $n = 470$				Expr. 1-2, $\hat{\beta}_{H2SLS}$, $n = 4700$			
	5^{th}	Median	95^{th}	Mean	5^{th}	Median	95^{th}	Mean	5^{th}	Median	95^{th}	Mean
$\beta_1^* = 1$	0.750	0.909	1.037	0.905	0.953	0.979	1.007	0.980	0.986	0.994	1.002	0.994
$\beta_2^* = 1$	0.747	0.910	1.029	0.903	0.952	0.979	1.005	0.979	0.986	0.994	1.002	0.994
$\beta_3^* = 1$	0.749	0.909	1.040	0.904	0.954	0.980	1.005	0.980	0.985	0.994	1.002	0.994
$\beta_4^* = 1$	0.756	0.909	1.032	0.901	0.953	0.979	1.003	0.979	0.986	0.994	1.001	0.994
$\beta_5^* = 1$	0.764	0.910	1.041	0.906	0.952	0.977	1.006	0.978	0.986	0.994	1.002	0.994
$\beta_6^* = 0$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$\beta_7^* = 0$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$\beta_8^* = 0$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\beta_{48}^* = 0$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$\beta_{49}^* = 0$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$\beta_{50}^* = 0$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

At the 5^{th} percentile, the median, and the 95^{th}percentile, the estimates of $(\beta_6^*, ..., \beta_{50}^*)$ are exactly 0.

The mean statistics of the estimates of $(\beta_6^*, ..., \beta_{50}^*)$ range from -5.066×10^{-4} to 5.550×10^{-4} when $n = 47$, -3.833×10^{-5} to 2.300×10^{-5} when $n = 470$, and -1.328×10^{-5} to 1.439×10^{-5} when $n = 4700$.

Table 4.4: l_2-errors of the estimates of the main-equation parameters by the two-stage Lasso and OLS procedures

Experiments 0-0, 1-0, 0-1, 1-1, 0-2, 1-2

l_2-error	5^{th}		Median		95^{th}		Mean	
	$\hat{\beta}_{H2SLS}$	$\hat{\beta}_{2SLS}$	$\hat{\beta}_{H2SLS}$	$\hat{\beta}_{2SLS}$	$\hat{\beta}_{H2SLS}$	$\hat{\beta}_{2SLS}$	$\hat{\beta}_{H2SLS}$	$\hat{\beta}_{2SLS}$
$n = 47$	0.123	0.093	0.251	0.205	0.493	0.391	0.270	0.217
$n = 470$	0.033	0.031	0.057	0.058	0.082	0.093	0.057	0.059
$n = 4700$	0.010	0.009	0.017	0.018	0.025	0.030	0.017	0.019

$\hat{\beta}_{H2SLS}$ denote estimates from the two-stage Lasso procedure (Expr. 1-0, 1-1, and 1-2).

$\hat{\beta}_{2SLS}$ denote estimates from the two-stage OLS procedure (Expr. 0-0, 0-1, and 0-2).

Table 4.5: Selection percentages of the estimates of the main-equation parameters by the two-stage Lasso procedure

Experiments 1-0, 1-1, and 1-2

selection % ($\hat{\beta}_{H2SLS}$)	5^{th}	Median	95^{th}	Mean
Expr. 1-0, $n = 47$	96.0	100	100	98.9
Expr. 1-1, $n = 470$	98.0	100	100	99.8
Expr. 1-2, $n = 4700$	98.0	100	100	99.9

Table 4.6: l_2-errors and selection percentages of the 1st-stage estimates by the Lasso procedure

Experiments 1-0, 1-1, 1-2

Average*	Expr. 1-0, $\hat{\pi}_{Lasso}$, $n = 47$				Expr. 1-1, $\hat{\pi}_{Lasso}$, $n = 470$				Expr. 1-2, $\hat{\pi}_{Lasso}$, $n = 4700$			
Statistics	5^{th}	Median	95^{th}	Mean	5^{th}	Median	95^{th}	Mean	5^{th}	Median	95^{th}	Mean
l_2-error	0.157	0.253	0.396	0.262	0.044	0.066	0.090	0.066	0.014	0.021	0.028	0.021
selection %	95.0	98.0	100	97.7	96.0	99.0	100	98.4	96.0	99.0	100	98.5

*: Averages are taken over $j = 1, ..., 50$.

In the following, I compare, in the relatively large sample size settings (i.e., when $n = 470$ and $n = 4700$) under the sparsity scenario, the performance of the two-stage Lasso estimator with the performance of the "partially" regularized or non-regularized estimators: first-stage-OLS-second-stage-Lasso (experiments 2-1 and 2-2), first-stage-Lasso-second-stage-OLS (experiments 3-1 and 3-2), and first-stage-OLS-second-stage-OLS (experiments 4-1 and 4-2). Clearly, it makes little sense to consider these "partially" regularized or non-regularized estimators in the high-dimensional setting (i.e., when $n = 47$). The l_2-errors and selection percentages of the main-equation estimates from these "partially" regularized or non-regularized estimators are displayed in Tables 4.7 and 4.8. In a similar fashion as Table 4.6, Table 4.9 reports the average (over the 50 first-stage equations') quantile and mean statistics of the l_2-errors and selection percentages of the first-stage OLS estimates appearing in experiments 2-1, 2-2, 4-1, and 4-2.

From Table 4.7, we again see that the l_2-errors of the main-equation estimates shrink as n increases. Comparing Table 4.7 with the last two rows of Table 4.4, we see that the two-stage Lasso estimator achieves the smallest l_2-errors of the main-equation estimates among all the estimators considered here. The fact that the l_2-errors (of the main-equation estimates) of the two-stage Lasso estimator are 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. First, comparing Table 4.9 with Table 4.6, we see that $\hat{\pi}_{Lasso}$ outperforms $\hat{\pi}_{OLS}$ in both estimation errors and variable selections even in the relatively large sample size settings with sparsity (an expected outcome attributed to the so-called "oracle property" of the Lasso technique; e.g., Bühlmann and van de Geer, 2011). Second, recall in Section 3, we have seen that, (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 (with the choices of p, d, k_1, and k_2 in experiment 1-0 through experiment 4-2, according to the theorems in Section 3, we should expect the estimation error of the parameters of interests in the main equation to be bounded by the term involving the first-stage estimation error); (2) upon the first-stage estimator correctly selecting the non-zero coefficients with high probability, the statistical error of the two-stage estimator $\hat{\beta}_{H2SLS}$ in Theorem 3.6 is smaller relative to that in Theorem 3.5 (where the first-stage selection-consistency condition is absent) when the error arising from the first-stage estimation dominates the second-stage error (which is indeed the case here). Given the choices of p, d, k_1, and k_2 in experiment 1-0 through experiment 4-2, these experiments seem to agree with the theorems in Section 3.1. The fact that the l_2-errors (of the main-equation estimates) of the two-stage Lasso estimator are smaller than the l_2-errors of the first-stage-Lasso-second-stage-OLS estimator and the first-stage-OLS-second-stage-OLS estimator can be explained via the standard Lasso theory: the Lasso reduces the l_2-error of the OLS from $\sqrt{\frac{p}{n}}$ to $\sqrt{\frac{\log p}{n}}$ (e.g., Bickel, 2009).

Comparing Table 4.8 with the last two rows of Table 4.5, we see that the two-stage Lasso estimator achieves higher selection percentages of the main-equation estimates relative to the first-stage-Lasso-second-stage-OLS estimator and the first-stage-OLS-second-stage-OLS estimator (again, this outcome is expected due to the so-called "oracle property" of the Lasso technique.). The selection percentages of the two-stage Lasso estimator and the first-stage-OLS-second-stage-Lasso, on the other hand, are comparable.

Table 4.7: l_2-errors of the estimates of the main-equation parameters by the "partially" regularized or non-regularized procedures, Experiments 2-1, 2-2, 3-1, 3-2, 4-1, 4-2

l_2 error	$\hat{\beta}$, $n = 470$				l_2 error	$\hat{\beta}$, $n = 4700$			
(Expr. #)	5^{th}	Median	95^{th}	Mean	(Expr. #)	5^{th}	Median	95^{th}	Mean
2-1	0.090	0.115	0.141	0.115	2-2	0.029	0.036	0.044	0.036
	(175.7%)	(102.7%)	(71.2%)	(101.1%)		(177.7%)	(112.8%)	(77.3%)	(110.7%)
3-1	0.120	0.142	0.165	0.142	3-2	0.036	0.043	0.050	0.043
	(268.9%)	(149.7%)	(101.4%)	(148.7%)		(245.3%)	(153.3%)	(100.8%)	(148.4%)
4-1	0.090	0.108	0.129	0.109	4-2	0.030	0.036	0.043	0.036
	(178.3%)	(91.2%)	(57.1%)	(90.1%)		(193.4%)	(114.2%)	(72.5%)	(111.3%)

The upper numbers are the actual quantile or mean statistics; the lower numbers in paratenses corresponding to expr. 2-1, 3-1, and 4-1 (expr. 2-2, 3-2, and 4-2) are percent changes relative to expr. 1-1 (respectively, expr. 1-2).

Table 4.8: Selection percentages of the estimates of the main-equation parameters by the "partially" regularized or non-regularized procedures, Experiments 2-1, 2-2, 3-1, 3-2, 4-1, 4-2

selection %	$\hat{\beta}$, $n = 470$				selection %	$\hat{\beta}$, $n = 4700$			
(Expr. #)	5^{th}	Median	95^{th}	Mean	(Expr. #)	5^{th}	Median	95^{th}	Mean
2-1	100 (2.0%)	100 (0%)	100 (0%)	99.9 (0.1%)	2-2	98.0 (0%)	100 (0%)	100 (0%)	99.9 (0%)
3-1	44.0 (-55.1%)	56.0 (-44.0%)	66.0 (-34.0%)	55.0 (-44.8%)	3-2	44.0 (-55.1%)	54.0 (-46.0%)	66.0 (-34.0%)	54.8 (-45.1%)
4-1	44.0 (-55.1%)	54.0 (-46.0%)	64.0 (-36.0%)	54.3 (-45.6%)	4-2	44.0 (-55.1%)	54.0 (-46.0%)	66.0 (-34.0%)	54.6 (-45.3%)

The upper numbers are the actual quantile or mean statistics; the lower numbers in paratenses corresponding to expr. 2-1, 3-1, and 4-1 (expr. 2-2, 3-2, and 4-2) are percent changes relative to expr. 1-1 (respectively, expr. 1-2).

Table 4.9: l_2-errors and selection percentages of the 1st-stage estimates by the OLS procedure Experiments 2-1, 2-2, 4-1, 4-2

Average*	$\hat{\pi}_{OLS}$, $n = 470$				$\hat{\pi}_{OLS}$, $n = 4700$			
Statistics	5^{th}	Median	95^{th}	Mean	5^{th}	Median	95^{th}	Mean
l_2-error	0.136 (207.2%)	0.155 (134.9%)	0.177 (97.3%)	0.156 (134.3%)	0.039 (179.5%)	0.044 (113.7%)	0.049 (79.0%)	0.044 (113.3%)
selection %	44.0 (-54.2%)	52.0 (-47.5%)	60.0 (-40.0%)	52.0 (-47.2%)	43.9 (-54.2%)	52.0 (-47.4%)	60.1 (-39.9%)	52.0 (-47.2%)

*: Averages are taken over $j = 1, ..., 50$.

The upper numbers are the actual quantile or mean statistics; the lower numbers in paratenses are percent changes relative to the 1st-stage estimates by the Lasso in expr. 1-1 and 1-2.

In the following 3 experiments (5-0, 5-1, and 5-2) where the first-stage equations are in the low-dimensional setting and the main equation is in the high-dimensional setting, I compute the l_2-errors of the main-equation estimates (see Table 4.10) following the two-stage Lasso procedure as in experiments 1-0, 1-1, and 1-2. In a similar fashion as Table 4.6, Table 4.11 reports the average (over the 50 first-stage equations') quantile and mean statistics of the l_2-errors and selection percentages of the first-stage Lasso estimate appearing in experiments 5-0, 5-1, and 5-2. Given the choices of p, d, k_1, and k_2 in experiments 5-0, 5-1, and 5-2, according to Corollary 3.4, I expect these experiments to yield smaller l_2-errors of the main-equation estimates relative to experiments 1-0, 1-1, and 1-2, respectively. Comparing Table 4.4 with Table 4.10, we notice that this is indeed the case.

Table 4.10: l_2-errors of the estimates of the main-equation parameters by the two-stage Lasso procedure (low-dimensional 1st-stage), Experiments 5-0, 5-1, 5-2

l_2-error ($\hat{\beta}_{H2SLS}$)	5^{th}	Median	95^{th}	Mean
Expr. 5-0, $n = 47$	0.105	0.188	0.340	0.201
	(-14.6%)	(-25.0%)	(-31.1%)	(-25.6%)
Expr. 5-1, $n = 470$	0.028	0.045	0.065	0.046
	(-13.1%)	(-20.1%)	(-21.2%)	(-19.5%)
Expr. 5-2, $n = 4700$	0.009	0.014	0.020	0.014
	(-15.6%)	(-16.8%)	(-17.8%)	(-17.2%)

The upper numbers are the actual quantile or mean statistics; the lower numbers in paratenses corresponding to expr. 5-0, 5-1, and 5-2 are percent changes relative to expr. 1-0, 1-1, and 1-2, respectively.

Table 4.11: l_2-errors and selection percentages of the 1st-stage estimates by the Lasso procedure (low-dimensional 1st-stage), Experiments 5-0, 5-1, 5-2

Average*	Expr. 5-0, $\hat{\pi}_{Lasso}$, $n = 47$				Expr. 5-1, $\hat{\pi}_{Lasso}$, $n = 470$				Expr. 5-2, $\hat{\pi}_{Lasso}$, $n = 4700$			
Statistics	5^{th}	Median	95^{th}	Mean	5^{th}	Median	95^{th}	Mean	5^{th}	Median	95^{th}	Mean
l_2-error	0.071	0.140	0.229	0.144	0.021	0.041	0.063	0.041	0.007	0.013	0.020	0.013
	(-54.5%)	(-44.7%)	(-42.2%)	(-45.1%)	(-52.0%)	(-38.2%)	(-29.9%)	(-37.8%)	(-51.2%)	(-37.7%)	(-29.1%)	(-37.3%)

*: Averages are taken over $j = 1, ..., 50$.

The upper numbers are the actual quantile or mean statistics; the lower numbers in paratenses corresponding to expr. 5-0, 5-1, and 5-2 are percent changes relative to the 1st-stage estimates in expr. 1-0, 1-1, and 1-2, respectively.

In the following experiments (6-0, 6-1, and 6-2) where $(\beta_1^*, ..., \beta_5^*) = (0.01, ..., 0.01)$ (as opposed to $(\beta_1^*, ..., \beta_5^*) = (1, ..., 1)$ in the previous experiments), following the two-stage Lasso procedure as in experiments 1-0, 1-1, and 1-2, I compute the number of occurrences that each estimate $\hat{\beta}_{H2SLS,1}, ..., \hat{\beta}_{H2SLS,5}$ takes on the 0 value over the 1000 replications (Table 4.12), as well as the overall selection percentages of the main-equation estimates (Table 4.13). Because the non-zero parameters are reduced by a factor of 100, I expect it more difficult for the two-stage Lasso procedure to distinguish the non-zero coefficients from the zero coefficients. From Tables 4.12 and 4.13, we notice that this is indeed the case.

Table 4.12: Number of occurrences that each estimate $\hat{\beta}_{H2SLS,1}$, ..., $\hat{\beta}_{H2SLS,5}$ takes on the 0 value over the 1000 replications

$(\beta_1^*, .., \beta_5^*) = (0.01, ..., 0.01)$, Experiments 6-0, 6-1, 6-2

$\hat{\beta}_{H2SLS}$	Expr. 6-0 $n = 47$	Expr. 6-1 $n = 470$	Expr. 6-2 $n = 4700$	Other* expr.
$\hat{\beta}_{H2SLS,1}$	285	4	0	0
$\hat{\beta}_{H2SLS,2}$	265	5	0	0
$\hat{\beta}_{H2SLS,3}$	251	7	0	0
$\hat{\beta}_{H2SLS,4}$	246	3	0	0
$\hat{\beta}_{H2SLS,5}$	266	5	0	0

*: "Other" includes all the previous experiments 0-0 to 5-2.

Table 4.13: Selection percentages of the estimates of the main-equation parameters by the two-stage Lasso procedure

$(\beta_1^*, .., \beta_5^*) = (0.01, ..., 0.01)$, Experiments 6-0, 6-1, 6-2

selection % ($\hat{\beta}_{H2SLS}$)	5^{th}	Median	95^{th}	Mean
Expr. 6-0	54.0	64.0	76.0	64.2
$n = 47$	(-43.7%)	(-36.0%)	(-24.0%)	(-35.1%)
Expr. 6-1	50.0	60.0	70.0	60.4
$n = 470$	(-49.0%)	(-40.0%)	(-30.0%)	(-39.5%)
Expr. 6-2	50.0	60.0	72.0	60.5
$n = 4700$	(-49.0%)	(-40.0%)	(-28.0%)	(-39.4%)

The upper numbers are the actual quantile or mean statistics; the lower numbers in paratenses corresponding to expr. 6-0, 6-1, and 6-2 are percent changes relative to expr. 1-0, 1-1, and 1-2, respectively.

5 Conclusion and future work

This paper explores the validity of the two-stage estimation procedures for triangular simultaneous linear equations models when the number(s) of the first and/or second-stage regressors grow with and exceed the sample size n. In particular, the number of endogenous regressors in the main equation can also grow with and exceed n. Sufficient conditions for estimation consistency in l_2-norm and variable-selection consistency of the two-stage high-dimensional estimators are established. Depending on the underlying sufficient conditions that are imposed, the rates of convergence in terms of the l_2-error and the smallest sample size required to obtain these consistency results differ by factors involving the sparsity parameters k_1 and/or k_2. Simulations are conducted to gain insight on the finite sample performance of these two-stage high-dimensional estimators as well as confirm the theoretical results. Several extensions are briefly discussed in the following.

The approximate sparsity case. First, 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. We 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. When the first-stage equations parameters are approximately sparse, the argument in the proof for Theorem 3.2 can still be carried through while Theorem 3.3 is no longer meaningful.

Control function approach in high-dimensional settings. Also, it is 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 $j = 1, ..., p$), the 2SLS estimation and control function approach yield algebraically equivalent estimators. Such equivalence no longer holds in high-dimensional settings because the regularization employed destroys the projection algebra. The extension for the 2SLS estimator from low-dimensional settings to high-dimensional settings seems somewhat more natural than the extension for the two-stage estimator based on the control function approach. One question to ask is: under what conditions can we translate the sparsity or approximate sparsity assumption on the coefficients β^* in the following triangular simultaneous equations model

$$\begin{aligned} y_i &= \mathbf{x}_i^T \beta^* + \epsilon_i, \qquad i = 1, ..., n, \\ x_{ij} &= \mathbf{z}_{ij}^T \pi_j^* + \eta_{ij}, \qquad i = 1, ..., n, \ j = 1, ..., p, \end{aligned}$$

to the sparsity or approximate sparsity assumption on the coefficients β^* and α^* in the model $y_i = \mathbf{x}_i^T \beta^* + \boldsymbol{\eta}_i^T \alpha^* + \xi_i$? A simple sufficient condition for such a translation is to impose the joint normality assumption of the error terms ϵ_i and $\boldsymbol{\eta}_i = (\eta_{i1}, ..., \eta_{ip})^T$. Then, by the property of multivariate normal distributions, we have

$$\mathbb{E}(\epsilon_i | \boldsymbol{\eta}_i) = \Sigma_{\epsilon\eta} \Sigma_{\eta\eta}^{-1} \boldsymbol{\eta}_i.$$

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 $y_i = \mathbf{x}_i^T \beta^* + \boldsymbol{\eta}_i^T \alpha^* + \xi_i$. 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 (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 $\hat{\beta}_{H2SLS}$.

Minimax lower bounds for the triangular simultaneous linear equation models. Furthermore, it is worthwhile to establish the minimax lower bounds on the parameters in the main equation for the triangular simultaneous linear equations models. In particular, my 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. To the best of my knowledge, in econometric literature, there has been only limited attention given to the minimax rates of linear models with endogeneity in high-dimensional settings.

6 Appendix: Proofs

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 means). Also, as a general rule for my proofs, b constants denote constants that are independent of n, p, d, k_1 and k_2 but possibly depend on the sub-Gaussian parameters; c constants denote universal 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. In addition, for notational simplicity, I assume the regime of interest is $p \geq n$ or $p \gg n$, as in most high-dimensional statistics literature. The modification to allow $p < n$ is trivial.

6.1 Lemma 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

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

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

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

Since we assume in Lemma 3.1 that the random matrix $\hat{\Gamma} = \hat{X}^T\hat{X}$ satisfies the RE1 condition (3) with $\gamma = 3$, we have

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

6.2 Theorem 3.2

As discussed in Section 3, the l_2-consistency of $\hat{\beta}_{H2SLS}$ requires verifications of two conditions: (i) $\hat{\Gamma} = \hat{X}^T\hat{X}$ satisfies the RE1 condition (3) with $\gamma = 3$, and (ii) the term $|\frac{1}{n}\hat{X}^T e|_\infty \precsim \sqrt{\frac{\log p}{n}}$ with high probability. This is done via Lemmas 6.1 and 6.2.

Lemma 6.1 (RE condition): Under Assumptions 1.1, 3.1-3.3, and 3.5a, with the scaling $n \succsim \max(k_1^2 \log d, \log p)$, we have

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

with probability at least $1 - c_1\exp(-c_2 \log\max(p, d))$ for some universal constants c_1 and c_2, where κ_1 and κ_2 are constants depending on $\lambda_{\min}(\Sigma_{X^*})$, $\lambda_{\min}(\Sigma_Z)$, σ_η, σ_{X^*}, and $\max_{j',j}|\mathbb{E}(x^*_{ij'}, \mathbf{z}_{ij})|_\infty$.

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

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

To bound the term $\left|\frac{X^{*T}(\hat{X}-X^*)}{n}\right|_\infty$, let us first fix $(j^{'},\, j)$ and bound the $(j^{'},\, j)$ element of the matrix $\frac{X^{*T}(\hat{X}-X^*)}{n}$. Notice that

$$\begin{aligned}\left|\frac{1}{n}\mathbf{x}_{j^{'}}^{*T}(\hat{\mathbf{x}}_j-\mathbf{x}_j^*)\right| &= \left|\left(\frac{1}{n}\sum_{i=1}^{n}x_{ij^{'}}^*\mathbf{z}_{ij}\right)(\hat{\pi}_j-\pi_j^*)\right| \\ &\leq \left|\hat{\pi}_j-\pi_j^*\right|_1\left|\frac{1}{n}\sum_{i=1}^{n}x_{ij^{'}}^*\mathbf{z}_{ij}\right|_\infty\end{aligned}$$

Under Assumptions 3.2 and 3.3, we have that the random matrix $\mathbf{z}_j\in\mathbb{R}^{n\times d_j}$ is a sub-Gaussian with parameters at most $(\Sigma_{Z_j},\,\sigma_Z^2)$ for all $j=1,...,p$, and $x_{j^{'}}^*$ is a sub-Gaussian vector with a parameter at most σ_{X^*} for every $j^{'}=1,...,p$. Denote $Z_j^T=\left(\mathbf{z}_{1j}^T,\,...\,\mathbf{z}_{nj}^T\right)\in\mathbb{R}^{d_j\times n}$ for every $j=1,...,p$. Therefore, by Lemma 6.8 and an application of union bound, we have

$$\mathbb{P}\left[\max_{j^{'},j}|\frac{1}{n}\mathbf{x}_{j^{'}}^{*T}Z_j-\mathbb{E}(x_{ij^{'}}^*,\,\mathbf{z}_{ij})|_\infty\geq t\right]\leq 6p^2d\exp(-cn\min\{\frac{t^2}{\sigma_{X^*}^2\sigma_Z^2},\,\frac{t}{\sigma_{X^*}\sigma_Z}\}),$$

and consequently as long as $n\gtrsim\log\max(p,\,d)$,

$$\mathbb{P}\left[\max_{j^{'},j}|\frac{1}{n}\mathbf{x}_{j^{'}}^{*T}Z_j-\mathbb{E}(x_{ij^{'}}^*,\,\mathbf{z}_{ij})|_\infty\geq c_0\sigma_{X^*}\sigma_Z\sqrt{\frac{\log\max(p,\,d)}{n}}\right]\leq c_1\exp(-c_2\log\max(p,\,d)),$$

where c_0, c_1, and c_2 are some universal constants. Hence, under Assumption 3.5a, we have, with probability at least $1-c_1\exp(-c_2\log\max(p,\,d))$,

$$\begin{aligned}\left|\frac{X^{*T}(\hat{X}-X^*)}{n}\right|_\infty &\leq \frac{c\sigma_\eta}{\lambda_{\min}(\Sigma_Z)}k_1\sqrt{\frac{\log d}{n}}\left(\max_{j^{'},j}|\mathbb{E}(x_{ij^{'}}^*,\,\mathbf{z}_{ij})|_\infty+c_0\sigma_{X^*}\sigma_Z\sqrt{\frac{\log\max(p,\,d)}{n}}\right) \\ &\leq c_3\frac{\sigma_\eta\max_{j^{'},j}|\mathbb{E}(x_{ij^{'}}^*,\,\mathbf{z}_{ij})|_\infty}{\lambda_{\min}(\Sigma_Z)}k_1\sqrt{\frac{\log d}{n}},\end{aligned}$$

where c_3 is some positive constant chosen to be sufficiently large.

To bound the term $\left|\frac{(\hat{X}-X^*)^T(\hat{X}-X^*)}{n}\right|_\infty$, again let us first fix $(j^{'},\, j)$ and bound the $(j^{'},\, j)$ element of the matrix $\frac{(\hat{X}-X^*)^T(\hat{X}-X^*)}{n}$. Using the similar argument as above, we have, with probability at

least $1 - c_1 \exp(-c_2 \log \max(p,\, d))$ for some universal constants c_1 and c_2,

$$\begin{aligned}
\left| \frac{(\hat{X} - X^*)^T (\hat{X} - X^*)}{n} \right|_\infty &= \max_{j',j} \left| (\hat{\pi}_{j'} - \pi^*_{j'})^T \left(\frac{1}{n} \sum_{i=1}^n \mathbf{z}_{ij'}^T \mathbf{z}_{ij} \right) (\hat{\pi}_j - \pi^*_j) \right| \\
&\leq \max_{j',j} \left(\left| \hat{\pi}_{j'} - \pi^*_{j'} \right|_1 \left| \hat{\pi}_j - \pi^*_j \right|_1 \left| \frac{1}{n} \sum_{i=1}^n \mathbf{z}_{ij'}^T \mathbf{z}_{ij} \right|_\infty \right) \\
&\leq \left(\frac{c\sigma_\eta}{\lambda_{\min}(\Sigma_Z)} k_1 \sqrt{\frac{\log d}{n}} \right)^2 \left(\max_{j',j} |\mathbb{E}(\mathbf{z}_{ij'},\, \mathbf{z}_{ij})|_\infty + c_0 \sigma_Z^2 \sqrt{\frac{\log \max(p,\, d)}{n}} \right) \\
&\leq c_3 \frac{\sigma_\eta^2 \max_{j',j} |\mathbb{E}(\mathbf{z}_{ij'},\, \mathbf{z}_{ij})|_\infty}{\lambda^2_{\min}(\Sigma_Z)} k_1^2 \frac{\log d}{n},
\end{aligned}$$

where c_3 is some positive constant chosen to be sufficiently large.

Putting everything together, under the scaling $n \gtrsim \max(k_1^2 \log d,\, \log p)$ and by Lemma 6.10, we have

$$\begin{aligned}
\left| v^{0T} \frac{\hat{X}^T \hat{X}}{n} v^0 \right| &\geq \left| v^{0T} \frac{X^{*T} X^*}{n} v^0 \right| \\
&- \left(2c_3 \frac{\sigma_\eta \max_{j',j} |\mathbb{E}(x^*_{ij'},\, \mathbf{z}_{ij})|_\infty}{\lambda_{\min}(\Sigma_Z)} k_1 \sqrt{\frac{\log d}{n}} + c_4 \frac{\sigma_\eta^2 \max_{j',j} |\mathbb{E}(\mathbf{z}_{ij'},\, \mathbf{z}_{ij})|_\infty}{\lambda^2_{\min}(\Sigma_Z)} k_1^2 \frac{\log d}{n} \right) |v^0|_1^2 \\
&\geq \left| v^{0T} \frac{X^{*T} X^*}{n} v^0 \right| - \left(c_5 \frac{\sigma_\eta \max_{j',j} |\mathbb{E}(x^*_{ij'},\, \mathbf{z}_{ij})|_\infty}{\lambda_{\min}(\Sigma_Z)} k_1 \sqrt{\frac{\log d}{n}} \right) |v^0|_1^2 \\
&\geq \frac{\lambda_{\min}(\Sigma_{X^*})}{2} |v^0|_2^2 - c_0 \lambda_{\min}(\Sigma_{X^*}) \max \left\{ \frac{\sigma^4_{X^*}}{\lambda^2_{\min}(\Sigma_{X^*})},\, 1 \right\} \frac{\log \max(p,\, d)}{n} |v^0|_1^2 \\
&- \left(c_5 \frac{\sigma_\eta \max_{j',j} |\mathbb{E}(x^*_{ij'},\, \mathbf{z}_{ij})|_\infty}{\lambda_{\min}(\Sigma_Z)} k_1 \sqrt{\frac{\log d}{n}} \right) |v^0|_1^2
\end{aligned}$$

where c_5 is some positive constant chosen to be sufficiently large. Notice the last inequality can be written in the form

$$\left| v^{0T} \frac{\hat{X}^T \hat{X}}{n} v^0 \right| \geq \kappa_1 |v^0|_2^2 - \kappa_2 \max \left\{ k_1 \sqrt{\frac{\log d}{n}},\, \frac{\log d}{n},\, \frac{\log p}{n} \right\} |v^0|_1^2.$$

□

In proving Lemma 3.1, upon our choice λ_n, we have shown

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

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

$$\frac{1}{n}\max\left\{\frac{c_5\left[\sigma_\eta \max_{j',j}|\mathbb{E}(x^*_{ij'}, \mathbf{z}_{ij})|_\infty\right]^2}{\lambda^2_{\min}(\Sigma_{X^*})\lambda^2_{\min}(\Sigma_Z)}k_1^2k_2^2\log d, \quad c_0\lambda_{\min}(\Sigma_{X^*})\max\left\{\frac{\sigma^4_{X^*}}{\lambda^2_{\min}(\Sigma_{X^*})}, 1\right\}k_2\log\max(p, d)\right\} = o(1),$$

i.e., the scaling $\frac{\max(k_1^2k_2^2\log d, k_2\log p)}{n} = o(1)$, then,

$$\left|\hat{v}^{0T}\frac{\hat{X}^T\hat{X}}{n}\hat{v}^0\right| \geq \frac{\lambda_{\min}(\Sigma_{X^*})}{4}\left|\hat{v}^0\right|_2^2,$$

which implies RE1 (3).

Lemma 6.2 (Upper bound on $|\frac{1}{n}\hat{X}^Te|_\infty$): Under Assumptions 1.1, 3.2, 3.3, and 3.5a, with the scaling $n \gtrsim \max(k_1^2\log d, \log p)$, we have

$$|\frac{1}{n}\hat{X}^Te|_\infty \precsim |\beta^*|_1\max\left\{k_1\sqrt{\frac{\log d}{n}}, \sqrt{\frac{\log p}{n}}\right\}$$

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

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

Hence,

$$\begin{aligned}|\frac{1}{n}\hat{X}^Te|_\infty &\leq |\frac{1}{n}X^{*T}(\hat{X} - X^*)\beta^*|_\infty + |\frac{1}{n}X^{*T}\boldsymbol{\eta}\beta^*|_\infty + |\frac{1}{n}X^{*T}\epsilon|_\infty \\ &+ |\frac{1}{n}(\hat{X} - X^*)^T(\hat{X} - X^*)\beta^*|_\infty + |\frac{1}{n}(X^* - \hat{X})^T\boldsymbol{\eta}\beta^*|_\infty + |\frac{1}{n}(X^* - \hat{X})^T\epsilon|_\infty.\end{aligned} \tag{6}$$

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 6.1, we have shown, with probability at least $1 - c_1 \exp(-c_2 \log \max(p,\, d))$,

$$\left| \frac{X^{*T}(\hat{X} - X^*)}{n} \right|_\infty \leq c \frac{\sigma_\eta \max_{j', j} |\mathbb{E}(x_{ij'}^*,\, \mathbf{z}_{ij})|_\infty}{\lambda_{\min}(\Sigma_Z)} k_1 \sqrt{\frac{\log d}{n}},$$

therefore,

$$|\frac{1}{n} X^{*T}(\hat{X} - X^*)\beta^*|_\infty \leq c \frac{\sigma_\eta \max_{j', j} |\mathbb{E}(x_{ij'}^*,\, \mathbf{z}_{ij})|_\infty}{\lambda_{\min}(\Sigma_Z)} |\beta^*|_1 k_1 \sqrt{\frac{\log d}{n}}.$$

The term $|\frac{1}{n}(\hat{X} - X^*)^T(\hat{X} - X^*)\beta^*|_\infty$ can be bounded using a similar argument and we have, with probability at least $1 - c_1 \exp(-c_2 \log \max(p,\, d))$,

$$|\frac{1}{n}(\hat{X} - X^*)^T(\hat{X} - X^*)\beta^*|_\infty \leq c \frac{\sigma_\eta^2 \max_{j', j} |\mathbb{E}(\mathbf{z}_{ij'},\, \mathbf{z}_{ij})|_\infty}{\lambda_{\min}^2(\Sigma_Z)} |\beta^*|_1 k_1^2 \frac{\log d}{n}.$$

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 6.8 and the assumption that $\mathbb{E}(\mathbf{z}_{ij'} \eta_{ij}) = \mathbf{0}$ for all $j^{'}$, j as well as Assumption 3.2 that η_j is an *i.i.d.* zero-mean sub-Gaussian vector with the parameter σ_η^2 for $j = 1, ..., p$, and the random matrix $\mathbf{z}_j \in \mathbb{R}^{n \times d_j}$ is a sub-Gaussian with parameters at most $(\Sigma_{Z_j},\, \sigma_Z^2)$ for all $j = 1, ..., p$. For the term $|\frac{1}{n}(X^* - \hat{X})^T \boldsymbol{\eta} \beta^*|_\infty$, we have, with probability at least $1 - c_1 \exp(-c_2 \log \max(p,\, d)))$,

$$\begin{aligned} |\frac{1}{n}(X^* - \hat{X})^T \boldsymbol{\eta} \beta^*|_\infty &\leq \max_{j'} |\hat{\pi}_{j'} - \pi_{j'}^*|_1 \max_{j', j} |\frac{1}{n} \sum_{i=1}^{n} \mathbf{z}_{ij'}^T \eta_{ij}|_\infty |\beta^*|_1 \\ &\leq c \frac{\sigma_Z \sigma_\eta^2 \max_{j', j} |\mathbb{E}(x_{ij'}^*,\, \mathbf{z}_{ij})|_\infty}{\lambda_{\min}(\Sigma_Z)} |\beta^*|_1 k_1 \sqrt{\frac{\log d}{n}} \sqrt{\frac{\log \max(p,\, d)}{n}}, \end{aligned}$$

with c chosen to be sufficiently large. Again, the last inequality follows from Lemma 6.8 and the

assumption that $\mathbb{E}(\mathbf{z}_{ij'}\eta_{ij}) = \mathbf{0}$ for all $j^{'}$, j as well as Assumption 3.2.

To bound the term $|\frac{1}{n}X^{*T}\epsilon|_\infty$, note under Assumptions 3.2 and 3.3 as well as the assumption $\mathbb{E}(\mathbf{z}_{ij}\epsilon_i) = \mathbf{0}$ for all $j = 1, ..., p$, again by Lemma 6.8,

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

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

$$\begin{aligned}|\frac{1}{n}(X^* - \hat{X})^T\epsilon|_\infty &\leq \max_j|\hat{\pi}_j - \pi_j^*|_1 \max_j|\frac{1}{n}\sum_{i=1}^n \mathbf{z}_{ij}^T\epsilon_i|_\infty \\ &\leq c\frac{\sigma_Z\sigma_\epsilon\sigma_\eta \max_{j',j}|\mathbb{E}(x_{ij'}^*, \mathbf{z}_{ij})|_\infty}{\lambda_{\min}(\Sigma_Z)}k_1\sqrt{\frac{\log d}{n}}\sqrt{\frac{\log p}{n}},\end{aligned}$$

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

Putting everything together, the claim in Lemma 6.2 follows. □

Under the scaling $\frac{\max(k_1^2k_2^2\log d, k_2\log p)}{n} = o(1)$ and $\lambda_n \asymp k_2\max\left\{k_1\sqrt{\frac{\log d}{n}}, \sqrt{\frac{\log p}{n}}\right\}$ (the k_2 factor in the choice of λ_n comes from the simple inequality $|\beta^*|_1 \leq k_2\max_{j=1,...,p}\beta^*$ by exploring the sparsity of β^*), combining Lemmas 3.1, 6.1, and 6.2, we have

$$|\hat{\beta}_{H2SLS} - \beta^*|_2 \precsim \psi_1|\beta^*|_1\max\left\{\sqrt{k_1k_2}\sqrt{\frac{k_1\log d}{n}}, \sqrt{\frac{k_2\log p}{n}}\right\},$$

where $\psi_1 = \max\left\{\frac{\sigma_\eta\max_{j,j'}|\mathrm{cov}(x_{1j'}^*, \mathbf{z}_{1j})|_\infty}{\lambda_{\min}(\Sigma_Z)\lambda_{\min}(\Sigma_{X^*})}, \frac{\sigma_{X^*}\max(\sigma_\epsilon, \sigma_\eta)}{\lambda_{\min}(\Sigma_{X^*})}\right\}$, which proves Theorem 3.2. □

6.3 Theorem 3.3

Again, we verify the conditions: i) $\hat{\Gamma} = \hat{X}^T\hat{X}$ satisfies the RE1 condition (3) with $\gamma = 3$, and (ii) the term $|\frac{1}{n}\hat{X}^T\hat{\epsilon}|_\infty \precsim \sqrt{\frac{\log p}{n}}$ with high probability. This is done via Lemmas 6.3 and 6.4.

Lemma 6.3 (RE condition): Let $r \in [0, 1]$. Under Assumptions 1.1, 3.1, 3.3, 3.4, 3.5b, and 3.6, with the scaling $n \succsim k_1^{3-2r}\log d$ and some universal constant c, we have

$$\frac{|\hat{X}v^0|_2^2}{n} \geq \left(\kappa_1 - c\max(b_2b_1^{-1}, b_3b_1^{-2})k_1^{1-r}\sqrt{\frac{k_1\log d}{n}}\right)|v^0|_2^2 - \kappa_2\frac{k_1^r\log\max(p, d)}{n}|v^0|_1^2, \qquad \text{for all } v^0 \in \mathbb{R}^p,$$

with probability at least $1 - c_1 \exp(-c_2 n)$ for some universal constants c_1 and c_2, where

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

Proof. First notice that

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

To bound the above terms, I now apply a discretization argument similar to 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 spaces $\mathbb{K}(2s,\, p) \times \mathbb{K}(k_1,\, d_1) \times ... \times \mathbb{K}(k_1,\, d_p)$ 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 $s := \frac{1}{c} \frac{n}{\log \max(p,\, d)} \min\left\{ \frac{\lambda_{\min}^2(\Sigma_{X^*})}{\sigma_{X^*}^4}, 1 \right\}$. The choice of s is explained in the proof for Lemma 6.10. 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 $\frac{1}{9}$-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 $t_{i'}^0 \in S_{V^0}$ such that

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

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 3.5b and 3.6, $|\hat{\pi}_j - \pi_j^*|_2 b_1\sqrt{\frac{n}{k_1 \log d}} \leq 1$ and $|\text{supp}(\hat{\pi}_j - \pi_j^*)| \leq k_1$ for every $j = 1, ..., p$. Define $\bar{\pi}_j = (\hat{\pi}_j - \pi_j^*)b_1\sqrt{\frac{n}{k_1 \log d}}$ 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} and $t_i^j \in S_{V^j}$ such that

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

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 $A_{j^{'}j}$. Write $v = (v^0, v^1, ..., v^p) \in S_V := S_{V^0} \times S_{V^1} \times ... \times S_{V^p}$. 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 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 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| + \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| \\
&+ \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| + \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| \\
&+ \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| + \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 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| + \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| \\
&+ \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| + \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| \\
&+ \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| + \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,

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

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

Under Assumptions 3.3 and 3.4, we have that $x_{j'}^*$ is a sub-Gaussian vector with a parameter at most σ_{X^*} for every $j' = 1,...,p$, and $\mathbf{z}_j t_i^j := \mathbf{w}_j$ is a sub-Gaussian vector with a parameter at most σ_{W^*}. An application of Lemma 6.8 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. 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

$$\mathbb{P}\left(\sup_{v\in\mathbb{K}(2s,p)\times\mathbb{K}(k_1,d_1)\times...\times\mathbb{K}(k_1,d_p)}\left|v^{0T}\left[\frac{1}{n}x_{j'}^{*T}\mathbf{z}_j v^j\right]v^0 - v^{0T}\left[\mathbb{E}(x_{1j'}^*\mathbf{z}_{1j}v^j)\right]v^0\right|\geq t\right)$$

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

With the choice of $s = s(r) := \frac{1}{c}\frac{n}{k_1^r\log\max(p,d)}\min\left\{\frac{\lambda_{\min}^2(\Sigma_{X^*})}{\sigma_{X^*}^4},1\right\}$, $r\in[0,1]$ from the proof for Lemma 6.10 and $t = c'k_1\sigma_{X^*}\sigma_W$ for some universal constant $c'\geq 1$, we have

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

$$
\begin{aligned}
&\leq \left(\sup_{v \in \mathbb{K}(2s,p) \times \mathbb{K}(k_1,d_1) \times \ldots \times \mathbb{K}(k_1,d_p)} \left| v^{0T} \left[\frac{1}{n} x_{j'}^{*T} \mathbf{z}_j v^j - \mathbb{E}(x_{1j'}^* \mathbf{z}_{1j} v^j) \right] v^0 \right| \right) b_1^{-1} \sqrt{\frac{k_1 \log d}{n}} \\
&\leq c' b_1^{-1} k_1^{1-r} \sqrt{\frac{k_1 \log d}{n}} \sigma_{X^*} \sigma_W
\end{aligned}
$$

with probability at least $1 - c_1 \exp(-c_2 n k_1)$ for some universal constants c_1 and c_2 chosen to be sufficiently large. Therefore, we have

$$
\begin{aligned}
\left| v^{0T} \frac{X^{*T}(\hat{X} - X^*)}{n} v^0 \right| &\leq \left(\sup_{v \in \mathbb{K}(2s,p) \times \mathbb{K}(k_1,d_1) \times \ldots \times \mathbb{K}(k_1,d_p)} \left| v^{0T} \left[\mathbb{E}(x_{1j'}^* \mathbf{z}_{1j} v^j) \right] v^0 \right| \right) b_1^{-1} \sqrt{\frac{k_1 \log d}{n}} \\
&+ c' b_1^{-1} k_1^{3/2-r} \sqrt{\frac{\log d}{n}} \sigma_{X^*} \sigma_W \\
&\leq c' b_2 b_1^{-1} k_1^{3/2-r} \sqrt{\frac{\log d}{n}},
\end{aligned}
$$

where $b_2 = \max \left\{ \sigma_{X^*} \sigma_W,\ \sup_{v \in \mathbb{K}(2s,p) \times \mathbb{K}(k_1,d_1) \times \ldots \times \mathbb{K}(k_1,d_p)} \left| v^{0T} \left[\mathbb{E}(x_{1j'}^* \mathbf{z}_{1j} v^j) \right] v^0 \right| \right\}$. Notice that the term

$$
\sup_{v \in \mathbb{K}(2s,p) \times \mathbb{K}(k_1,d_1) \times \ldots \times \mathbb{K}(k_1,d_p)} \left| v^{0T} \left[\mathbb{E}(x_{1j'}^* \mathbf{z}_{1j} v^j) \right] v^0 \right|
$$

is bounded above by the spectral norm of the matrix $\left[\mathbb{E}(x_{1j'}^* \mathbf{z}_{1j} v^j) \right]$ for all $v^j \in \mathbb{K}(k_1, d_j)$ and j', $j = 1, ..., p$.

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

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

Combining with

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

after some tedious algebra, we obtain

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

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

Hence,

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

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

An application of Lemma 6.8 and a sequence of union bounds yields

$$\mathbb{P}\left(\sup_{v \times v' \in \mathbb{K}(2s,p) \times \mathbb{K}^2(k_1,d_1) \times \ldots \times \mathbb{K}^2(k_1,d_p)} \left| v^{0T} \left[\frac{1}{n} v^{j'} \mathbf{z}_{j'}^T \mathbf{z}_j v^j \right] v^0 - v^{0T} \left[\mathbb{E}(v^{j'} \mathbf{z}_{1j'}^T \mathbf{z}_{1j} v^j) \right] v^0 \right| \geq t \right)$$

$$\leq 2 \exp(-cn \min(\frac{t^2}{\sigma_W^4}, \frac{t}{\sigma_W^2}) + 4sk_1 \log d + 2s \log p).$$

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 6.10 and $t = c'' k_1 \sigma_W^2$ for some universal constant $c'' \geq 1$, we have,

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

$$\begin{aligned} &\leq \left(\sup_{v \in \mathbb{K}(2s,p) \times \mathbb{K}^2(k_1,d_1) \times \ldots \times \mathbb{K}^2(k_1,d_p)} \left| v^{0T} \left[\frac{1}{n} v^{j'} \mathbf{z}_{j'}^T \mathbf{z}_j v^j \right] v^0 - v^{0T} \left[\mathbb{E}(v^{j'} \mathbf{z}_{1j'}^T \mathbf{z}_{1j} v^j) \right] v^0 \right| \right) b_1^{-2} \frac{k_1 \log d}{n} \\ &\leq c'' b_1^{-2} \frac{k_1^{2-r} \log d}{n} \sigma_W^2 \end{aligned}$$

with probability at least $1 - c_1 \exp(-c_2 n k_1)$ for some universal constants c_1 and c_2 chosen to be sufficiently large. Therefore, we have

$$\begin{aligned}\left| v^{0T} \frac{(\hat{X} - X^*)^T (\hat{X} - X^*)}{n} v^0 \right| &\leq \left(\sup_{v \in \mathbb{K}(2s,p) \times \mathbb{K}^2(k_1,d_1) \times \ldots \times \mathbb{K}^2(k_1,d_p)} \left| v^{0T} \left[\mathbb{E}(v^{j'} \mathbf{z}_{1j'}^T \mathbf{z}_{1j} v^j) \right] v^0 \right| \right) b_1^{-2} \frac{k_1 \log d}{n} \\ &+ c'' b_1^{-2} \frac{k_1^{2-r} \log d}{n} \sigma_W^2 \\ &\leq c'' b_3 b_1^{-2} \frac{k_1^{2-r} \log d}{n},\end{aligned}$$

where $b_3 = \max\left\{ \sigma_W^2, \sup_{v \in \mathbb{K}(2s,p) \times \mathbb{K}^2(k_1,d_1) \times \ldots \times \mathbb{K}^2(k_1,d_p)} \left| v^{0T} \left[\mathbb{E}(v^{j'} \mathbf{z}_{1j'}^T \mathbf{z}_{1j} v^j) \right] v^0 \right| \right\}$. Notice that the term

$$\sup_{v \in \mathbb{K}(2s,p) \times \mathbb{K}^2(k_1,d_1) \times \ldots \times \mathbb{K}^2(k_1,d_p)} \left| v^{0T} \left[\mathbb{E}(v^{j'} \mathbf{z}_{1j'}^T \mathbf{z}_{1j} v^j) \right] v^0 \right|$$

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 all $v^j \in \mathbb{K}(k_1, d_j)$ and $j = 1, ..., p$.

By Lemma 6.9, the bound

$$\left| v^{0T} \frac{X^{*T} (\hat{X} - X^*)}{n} v^0 \right| \leq c' b_2 b_1^{-1} k_1^{3/2-r} \sqrt{\frac{\log d}{n}} \qquad \forall v^0 \in \mathbb{K}(2s, p)$$

implies

$$\left| v^{0T} \frac{X^{*T} (\hat{X} - X^*)}{n} v^0 \right| \leq 27 c' b_2 b_1^{-1} k_1^{3/2-r} \sqrt{\frac{\log d}{n}} (|v^0|_2^2 + \frac{1}{s} |v^0|_1^2) \qquad \forall v^0 \in \mathbb{R}^p. \tag{7}$$

Similarly, the bound

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

implies

$$\left| v^{0T} \frac{(\hat{X} - X^*)^T (\hat{X} - X^*)}{n} v^0 \right| \leq 27 c'' b_3 b_1^{-2} \frac{k_1^{2-r} \log d}{n} (|v^0|_2^2 + \frac{1}{s} |v^0|_1^2) \qquad \forall v^0 \in \mathbb{R}^p. \tag{8}$$

Therefore, 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 scaling

$n \gtrsim k_1^{3-2r} \log d$, we have

$$\begin{aligned}
\left| v^{0T} \frac{\hat{X}^T \hat{X}}{n} v^0 \right| &\geq \left| v^{0T} \frac{X^{*T} X^*}{n} v^0 \right| - c \max(b_2 b_1^{-1},\, b_3 b_1^{-2}) k_1^{3/2-r} \sqrt{\frac{\log d}{n}} (|v^0|_2^2 + \frac{1}{s} |v^0|_1^2) \\
&\geq \left(\frac{\lambda_{\min}(\Sigma_{X^*})}{2} - c \max(b_2 b_1^{-1},\, b_3 b_1^{-2}) k_1^{3/2-r} \sqrt{\frac{\log d}{n}} \right) |v^0|_2^2 \\
&\quad - c^{'} \lambda_{\min}(\Sigma_{X^*}) \max \left\{ \frac{\sigma_{X^*}^4}{\lambda_{\min}^2(\Sigma_{X^*})},\, 1 \right\} \frac{k_1^r \log \max(p, d)}{n} |v^0|_1^2,
\end{aligned}$$

which can be written in the form

$$\frac{|\hat{X} v^0|_2^2}{n} \geq \left(\kappa_1 - c \max(b_2 b_1^{-1},\, b_3 b_1^{-2}) k_1^{3/2-r} \sqrt{\frac{\log d}{n}} \right) |v^0|_2^2 - \kappa_2 \frac{k_1^r \log \max(p,\, d)}{n} |v^0|_1^2, \qquad \text{for all } v^0 \in \mathbb{R}^p.$$

□

Again, recall in proving Lemma 3.1, upon our choice λ_n, we have shown

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

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

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

then,

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

which implies RE1 (3). Because the argument for showing Lemma 6.1 and that it implies RE1 (3) also works under the assumptions of Lemma 6.3, we can combine the scaling $\frac{\max(k_1^2 k_2^2 \log d,\, k_2 \log p)}{n} = o(1)$ from the proof for Lemma 6.1 with the scaling $\frac{\min_{r \in [0,\, 1]} \max \left\{ k_1^{3-2r} \log d,\, k_1^r k_2 \log d,\, k_1^r k_2 \log p \right\}}{n} = o(1)$ from above to obtain a more optimal scaling requirement of the smallest sample size

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

which implies RE1 (3).

Lemma 6.4 (Upper bound on $|\frac{1}{n} \hat{X}^T e|_\infty$): Under Assumptions 1.1, 3.2-3.4, 3.5b, and 3.6, with

the scaling $n \gtrsim \max(k_1 \log d,\, \log p)$, we have

$$|\frac{1}{n}\hat{X}^T e|_\infty \precsim |\beta^*|_1 \max\left\{\sqrt{\frac{k_1 \log d}{n}},\, \sqrt{\frac{\log p}{n}}\right\}$$

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

Proof. Recall (6) from the proof for Lemma 6.2. Let us first bound $|\frac{1}{n}X^{*T}(\hat{X} - X^*)\beta^*|_\infty$. 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 6.3, we have shown that with a covering subset argument, the $(j^{'}, j)$ element of the matrix $\frac{X^{*T}(\hat{X} - X^*)}{n}$ can be rewritten as follows.

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

where we define $\bar{\pi}_j = (\hat{\pi}_j - \pi_j^*) b_1 \sqrt{\frac{n}{k_1 \log d}}$. Hence,

$$\left|\frac{1}{n}\mathbf{x}_{j^{'}}^{*T}(\hat{\mathbf{x}}_j - \mathbf{x}_j^*) - \mathbb{E}(\frac{1}{n}\mathbf{x}_{j^{'}}^{*T}(\hat{\mathbf{x}}_j - \mathbf{x}_j^*))\right|$$

$$\begin{aligned} &\leq \sup_{v^j \in S_{Vj}} b_1^{-1}\sqrt{\frac{k_1 \log d}{n}}\left|\frac{1}{n} x_{j^{'}}^{*T}\mathbf{z}_j v^j - \mathbb{E}(x_{1j^{'}}^* \mathbf{z}_{1j} v^j)\right| \\ &\leq b_1^{-1}\sqrt{\frac{k_1 \log d}{n}}\{\max_i \left|\frac{1}{n} x_{j^{'}}^{*T}\mathbf{z}_j t_i^j - \mathbb{E}(x_{1j^{'}}^* \mathbf{z}_{1j} t_i^j)\right| + \sup_{v^j \in S_{Vj}} \left|\frac{1}{n} x_{j^{'}}^{*T}\mathbf{z}_j \triangle v^j - \mathbb{E}(x_{1j^{'}}^* \mathbf{z}_{1j} \triangle v^j)\right|\} \\ &\leq b_1^{-1}\sqrt{\frac{k_1 \log d}{n}}\{\max_i \left|\frac{1}{n} x_{j^{'}}^{*T}\mathbf{z}_j t_i^j - \mathbb{E}(x_{1j^{'}}^* \mathbf{z}_{1j} t_i^j)\right| + \sup_{v^j \in S_{Vj}} \frac{1}{9}\left|\frac{1}{n} x_{j^{'}}^{*T}\mathbf{z}_j v^j - \mathbb{E}(x_{1j^{'}}^* \mathbf{z}_{1j} v^j)\right|\}. \end{aligned}$$

With a similar argument as in the proof for Lemma 6.3, we obtain

$$\mathbb{P}\left(\max_{j^{'}, j} \sup_{v^j \in \mathbb{K}(k_1, d_j)} \left|\frac{1}{n} x_{j^{'}}^{*T}\mathbf{z}_j v^j - \mathbb{E}(x_{1j^{'}}^* \mathbf{z}_{1j} v^j)\right| \geq t\right)$$

$$
\begin{aligned}
&\leq p^2 d^{k_1} \cdot 2\exp(-cn\min(\frac{t^2}{\sigma_{X^*}^2\sigma_W^2},\ \frac{t}{\sigma_{X^*}\sigma_W})) \\
&= 2\exp(-cn\min(\frac{t^2}{\sigma_{X^*}^2\sigma_W^2},\ \frac{t}{\sigma_{X^*}\sigma_W}) + k_1\log d + 2\log p).
\end{aligned}
$$

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

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

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

which implies, under the scaling

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

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

$$
\begin{aligned}
\left|\frac{X^{*T}(\hat{X}-X^*)}{n}\right|_\infty &\leq \left(\max_{j',j}\ \sup_{v^j\in\mathbb{K}(k_1,\,d_j)}\left|\mathbb{E}(x_{1j'}^*\mathbf{z}_{1j}v^j)\right|\right) b_1^{-1}\sqrt{\frac{k_1\log d}{n}} \\
&\quad + c^{'} b_1^{-1}\sqrt{\frac{k_1\log d}{n}}\sigma_{X^*}\sigma_W \max\left\{\sqrt{\frac{k_1\log d}{n}},\ \sqrt{\frac{\log p}{n}}\right\} \qquad (9) \\
&\leq c^{''} b_4 b_1^{-1}\sqrt{\frac{k_1\log d}{n}},
\end{aligned}
$$

where $b_4 = \max\left\{\sigma_{X^*}\sigma_W,\ \max_{j',\,j}\sup_{v^j\in\mathbb{K}(k_1,\,d_j)}\left|\mathbb{E}(x_{1j'}^*\mathbf{z}_{1j}v^j)\right|\right\}$.

To bound the term $\left|\frac{(\hat{X}-X^*)^T(\hat{X}-X^*)}{n}\right|_\infty$, again let us first fix $(j^{'},\, j)$ and bound the $(j^{'},\, j)$ element of the matrix $\frac{(\hat{X}-X^*)^T(\hat{X}-X^*)}{n}$. Recall from the proof for Lemma 6.3, with a covering subset argument, the $(j^{'},\, j)$ element of the matrix $\frac{(\hat{X}-X^*)^T(\hat{X}-X^*)}{n}$ can be rewritten as follows.

$$
\frac{1}{n}(\hat{\mathbf{x}}_{j'}-\mathbf{x}_{j'}^*)^T(\hat{\mathbf{x}}_j-\mathbf{x}_j^*) = b_1^{-2}\frac{k_1\log d}{n}\{\frac{1}{n}t_{i'}^{j'T}\mathbf{z}_{j'}^T\mathbf{z}_j t_i^j + \frac{1}{n}\triangle v^{j'T}\mathbf{z}_{j'}^T\mathbf{z}_j t_i^j + \frac{1}{n}t_{i'}^{j'T}\mathbf{z}_{j'}^T\mathbf{z}_j\triangle v^j + \frac{1}{n}\triangle v^{j'T}\mathbf{z}_{j'}^T\mathbf{z}_j\triangle v^j\}.
$$

With a similar argument as in the proof for Lemma 6.3, we obtain

$$\mathbb{P}\left(\max_{j',j}\sup_{v^{j'}\in\mathbb{K}(k_1,d_{j'}),v^j\in\mathbb{K}(k_1,d_j)}\left|\frac{1}{n}v^{j'}\mathbf{z}_{j'}^T\mathbf{z}_j v^j-\mathbb{E}(v^{j'}\mathbf{z}_{1j'}^T\mathbf{z}_{1j}v^j)\right|\geq t\right)$$

$$\begin{aligned}&\leq\ p^2d^{2k_1}\cdot 2\exp(-cn\min(\frac{t^2}{\sigma_W^4},\ \frac{t}{\sigma_W^2}))\\&=\ 2\exp(-cn\min(\frac{t^2}{\sigma_W^4},\ \frac{t}{\sigma_W^2})+2k_1\log d+2\log p).\end{aligned}$$

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

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

$$\begin{aligned}&\leq\ \left(\max_{j',j}\sup_{v^{j'}\in\mathbb{K}(k_1,d_{j'}),v^j\in\mathbb{K}(k_1,d_j)}\left|\frac{1}{n}v^{j'}\mathbf{z}_{j'}^T\mathbf{z}_j v^j-\mathbb{E}(v^{j'}\mathbf{z}_{1j'}^T\mathbf{z}_{1j}v^j)\right|\right)b_1^{-2}\frac{k_1\log d}{n}\\&\leq\ c^{'}b_1^{-2}\frac{k_1\log d}{n}\sigma_W^2\max\left\{\sqrt{\frac{k_1\log d}{n}},\ \sqrt{\frac{\log p}{n}}\right\},\end{aligned}$$

which implies, under the scaling

$$n\gtrsim\min\left\{\max\left\{k_1^2k_2^2\log d,\ k_2\log p\right\},\ \min_{r\in[0,1]}\max\left\{k_1^{3-2r}\log d,\ k_1^rk_2\log d,\ k_1^rk_2\log p\right\}\right\},$$

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

$$\begin{aligned}\left|\frac{(\hat{X}-X^*)^T(\hat{X}-X^*)}{n}\right|_\infty\ &\leq\ \left(\max_{j',j}\sup_{v^{j'}\in\mathbb{K}(k_1,d_{j'}),v^j\in\mathbb{K}(k_1,d_j)}\left|\mathbb{E}(v^{j'}\mathbf{z}_{1j'}^T\mathbf{z}_{1j}v^j)\right|\right)b_1^{-2}\frac{k_1\log d}{n}\\&+\ c^{'}b_1^{-2}\frac{k_1\log d}{n}\sigma_W^2\max\left\{\sqrt{\frac{k_1\log d}{n}},\ \sqrt{\frac{\log p}{n}}\right\}\qquad(10)\\&\leq\ c^{''}b_5b_1^{-2}\frac{k_1\log d}{n},\end{aligned}$$

where $b_5=\max\left\{\sigma_W^2,\ \max_{j',j}\sup_{v^{j'}\in\mathbb{K}(k_1,d_{j'}),v^j\in\mathbb{K}(k_1,d_j)}\left|\mathbb{E}(v^{j'}\mathbf{z}_{1j'}^T\mathbf{z}_{1j}v^j)\right|\right\}$.

Therefore,

$$|\frac{1}{n}X^{*T}(\hat{X}-X^*)\beta^*|_\infty \leq cb_4b_1^{-1}\sqrt{\frac{k_1\log d}{n}}|\beta^*|_1,$$

and

$$|\frac{1}{n}(\hat{X}-X^*)^T(\hat{X}-X^*)\beta^*|_\infty \leq cb_5b_1^{-2}\frac{k_1\log d}{n}|\beta^*|_1.$$

With exactly the same discretization argument as above, we can show that, with probability at least $1-c_1\exp(-c_2\log\min(p,\,d))$,

$$\begin{aligned}
|\frac{1}{n}(X^*-\hat{X})^T\boldsymbol{\eta}\beta^*|_\infty &\leq c^{'}b_1^{-1}\sigma_\eta\sigma_W|\beta^*|_1\sqrt{\frac{k_1\log d}{n}}\max\left\{\sqrt{\frac{k_1\log d}{n}},\,\sqrt{\frac{\log p}{n}}\right\},\\
|\frac{1}{n}(X^*-\hat{X})^T\epsilon|_\infty &\leq c^{''}b_1^{-1}\sigma_\epsilon\sigma_W\sqrt{\frac{k_1\log d}{n}}\max\left\{\sqrt{\frac{k_1\log d}{n}},\,\sqrt{\frac{\log p}{n}}\right\}.
\end{aligned}$$

For the rest of terms in (6), we can use the bounds provided in the proof for Lemma 6.2. In particular, recall we have, with probability at least $1-c_1\exp(-c_2\log p)$,

$$\begin{aligned}
|\frac{1}{n}X^{*T}\boldsymbol{\eta}\beta^*|_\infty &\leq c^{'}\sigma_{X^*}\sigma_\eta|\beta^*|_1\sqrt{\frac{\log p}{n}},\\
|\frac{1}{n}X^{*T}\epsilon|_\infty &\leq c^{''}\sigma_{X^*}\sigma_\epsilon\sqrt{\frac{\log p}{n}},
\end{aligned}$$

Putting everything together, the claim in Lemma 6.4 follows. □

Under the scaling

$$\frac{1}{n}\min\left\{\max\left\{k_1^2k_2^2\log d,\,k_2\log p\right\},\,\min_{r\in[0,\,1]}\max\left\{k_1^{3-2r}\log d,\,k_1^rk_2\log d,\,k_1^rk_2\log p\right\}\right\}=o(1),$$

and

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

(9) yields

$$\left|\frac{X^{*T}(\hat{X}-X^*)}{n}\right|_\infty \leq c\left(\max_{j^{'},\,j}\sup_{v^j\in\mathbb{K}(k_1,\,d_j)}\left|\mathbb{E}(x^*_{1j^{'}}\mathbf{z}_{1j}v^j)\right|\right)b_1^{-1}\sqrt{\frac{k_1\log d}{n}},$$

(10) yields

$$\left|\frac{(\hat{X}-X^*)^T(\hat{X}-X^*)}{n}\right|_\infty \leq c^{'}\left(\max_{j^{'},\,j}\sup_{v^{j^{'}}\in\mathbb{K}(k_1,\,d_{j^{'}}),\,v^j\in\mathbb{K}(k_1,\,d_j)}\left|\mathbb{E}(v^{j^{'}}\mathbf{z}^T_{1j^{'}}\mathbf{z}_{1j}v^j)\right|\right)b_1^{-2}\frac{k_1\log d}{n},$$

and furthermore,

$$\begin{aligned} |\frac{1}{n}X^{*T}\boldsymbol{\eta}\beta^*|_\infty + |\frac{1}{n}(X^* - \hat{X})^T\boldsymbol{\eta}\beta^*|_\infty &\leq c''\sigma_{X^*}\sigma_\eta|\beta^*|_1\sqrt{\frac{\log p}{n}}, \\ |\frac{1}{n}X^{*T}\epsilon|_\infty + |\frac{1}{n}(X^* - \hat{X})^T\epsilon|_\infty &\leq c'''\sigma_{X^*}\sigma_\epsilon\sqrt{\frac{\log p}{n}}. \end{aligned}$$

Combining the bounds above with Lemmas 3.1 and 6.3, we have

$$|\hat{\beta}_{H2SLS} - \beta^*|_2 \precsim \psi_2|\beta^*|_1 \max\left\{\sqrt{k_2}\sqrt{\frac{k_1\log d}{n}}, \sqrt{\frac{k_2\log p}{n}}\right\},$$

where $\psi_2 = \max\left\{\frac{\sigma_\eta \max_{j',j}\sup_{v^j\in\mathbb{K}(k_1,\,d_j)}\left|\mathrm{cov}(x^*_{1j'}, \mathbf{z}_{1j}v^j)\right|}{\lambda_{\min}(\Sigma_Z)\lambda_{\min}(\Sigma_{X^*})}, \frac{\sigma_{X^*}\max(\sigma_\epsilon,\sigma_\eta)}{\lambda_{\min}(\Sigma_{X^*})}\right\}$, which proves Theorem 3.3.□

6.4 Corollary 3.4, Theorems 3.5, and 3.6

Corollary 3.4 is obvious from inspecting the form of the bounds in Theorems 3.2 and 3.3. The proof for Theorem 3.5 is completely identical to that for Theorem 3.2 except we will replace the inequality $|\hat{\pi}_j - \pi^*_j|_1 \leq \frac{c\sigma_\eta}{\lambda_{\min}(\Sigma_Z)}k_1\sqrt{\frac{\log d_j}{n}}$ by $|\hat{\pi}_j - \pi_j|_1 \leq \sqrt{k_1}M(d,\,k_1,\,n)$. Also, the proof for Theorem 3.6 is completely identical to that for Theorem 3.3 except we will replace the inequality $|\hat{\pi}_j - \pi^*_j|_2 \leq \frac{c\sigma_\eta}{\lambda_{\min}(\Sigma_Z)}\sqrt{\frac{k_1\log d}{n}}$ by $|\hat{\pi}_j - \pi_j|_2 \leq M(d,\,k_1,\,n)$.

6.5 Lemma 6.5

Lemma 6.5: Suppose the assumptions in Lemmas 6.1 and 6.2 (or, Lemmas 6.3 and 6.4) and Assumptions 3.7 and 3.8 hold. 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 $\tilde{\Sigma} := \frac{1}{n}\hat{X}^T_{K^c}\hat{X}_K$. 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 $\tilde{\Sigma}_{KK} := \frac{1}{n}\hat{X}^T_K\hat{X}_K$. Then, the sample matrix $\frac{1}{n}\hat{X}^T\hat{X}$ satisfies an analogous version of the mutual incoherence assumption, with high probability in the sense that

$$\mathbb{P}\left[\left\|\frac{1}{n}\hat{X}^T_{K^c}\hat{X}_K\left(\frac{1}{n}\hat{X}^T_K\hat{X}_K\right)^{-1}\right\|_1 \geq 1 - \frac{\phi}{4}\right] \leq O\left(\exp(-b\frac{n}{k_2^3} + \log p)\right).$$

for some constant b.

Proof. I adopt the method used in Ravikumar, et. al. (2009), Lemma 6. Note that we can perform the following decomposition

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

where

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

By the incoherence assumption, we have

$$\left\|\Sigma_{K^cK}\Sigma_{KK}^{-1}\right\|_1 \leq 1 - \phi.$$

Hence, it suffices to show that $||R_i||_1 \leq \frac{\phi}{6}$ for $i = 1, ..., 3$ and $||R_i||_1 \leq \frac{\phi}{12}$ 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 ($||AB||_1 \leq ||A||_1||B||_1$) and the elementary inequality $||A||_1 \leq \sqrt{a}||A||_2$ for any symmetric matrix $A \in \mathbb{R}^{a\times a}$, we can bound R_1 as follows:

$$\begin{aligned}
||R_1||_1 &\leq \left\|\Sigma_{K^cK}\Sigma_{KK}^{-1}\right\|_1 \left\|\hat{\Sigma}_{KK} - \Sigma_{KK}\right\|_1 \left\|\hat{\Sigma}_{KK}^{-1}\right\|_1 \\
&\leq (1-\phi)\left\|\hat{\Sigma}_{KK} - \Sigma_{KK}\right\|_1 \sqrt{k_2}\left\|\hat{\Sigma}_{KK}^{-1}\right\|_2,
\end{aligned}$$

where the last inequality follows from the incoherence assumption. Using bound (16) from Lemma 6.11, we have

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

with probability at least $1 - c_1\exp(-c_2\frac{n}{k_2})$. Next, applying bound (12) from Lemma 6.11 with $\varepsilon = \frac{c}{\sqrt{k_2}}$, with probability at least $1 - 2\exp(-b\frac{n}{k_2^3} + 2\log k_2)$, we have

$$\left\|\hat{\Sigma}_{KK} - \Sigma_{KK}\right\|_1 \leq \frac{c}{\sqrt{k_2}}.$$

By choosing the constant $c > 0$ sufficiently small, we are guaranteed that

$$\mathbb{P}[||R_1||_1 \geq \frac{\phi}{6}] \leq 2\exp(-b\frac{n}{k_2^3} + \log k_2).$$

For the second term R_2, we first write

$$\begin{aligned} ||R_2||_1 &\leq \sqrt{k_2} \left\|\Sigma_{KK}^{-1}\right\|_2 \left\|\hat{\Sigma}_{K^cK} - \Sigma_{K^cK}\right\|_1 \\ &\leq \frac{\sqrt{k_2}}{\lambda_{\min}(\Sigma_{KK})} \left\|\hat{\Sigma}_{K^cK} - \Sigma_{K^cK}\right\|_1 . \end{aligned}$$

An application of bound (11) from Lemma 6.11 with $\varepsilon = \frac{\phi}{6}\frac{\lambda_{\min}(\Sigma_{KK})}{\sqrt{k_2}}$ yields

$$\mathbb{P}[||R_2||_1 \geq \frac{\phi}{6}] \leq 2\exp(-b\frac{n}{k_2^3} + \log(p-k_2) + \log k_2).$$

For the third term R_3, by applying bounds (11) and (13), with $\varepsilon = \frac{\phi}{6}\frac{\lambda_{\min}(\Sigma_{KK})}{b'}$ for (11) with b' chosen to be sufficiently large, and the fact that $\log k_2 \leq \log(p-k_2)$ yields

$$||R_3||_1 \leq \frac{\phi}{6}$$

with probability at least $1 - c\exp(-b\frac{n}{k_2^2} + \log(p-k_2))$.

Putting everything together, we conclude that

$$\mathbb{P}[||\hat{\Sigma}_{K^cK}\hat{\Sigma}_{KK}^{-1}||_1 \geq 1 - \frac{\phi}{2}] \leq O\left(\exp(-b\frac{n}{k_2^3} + \log p)\right).$$

For the fourth term R_4, we have, with probability at least $1 - b'\exp(-b\frac{n}{k_2^3} + \log p)$,

$$\begin{aligned} ||R_4||_1 &\leq \left\|\hat{\Sigma}_{K^cK}\hat{\Sigma}_{KK}^{-1}\right\|_1 \left\|\tilde{\Sigma}_{KK} - \hat{\Sigma}_{KK}\right\|_1 \left\|\tilde{\Sigma}_{KK}^{-1}\right\|_1 \\ &\leq (1 - \frac{\phi}{2}) \left\|\tilde{\Sigma}_{KK} - \hat{\Sigma}_{KK}\right\|_1 \sqrt{k_2} \left\|\tilde{\Sigma}_{KK}^{-1}\right\|_2 , \end{aligned}$$

where the last inequality follows from the bound on $||\hat{\Sigma}_{K^cK}\hat{\Sigma}_{KK}^{-1}||_1$ established previously. Using bounds (16) and (23) (or (25)) from Lemma 6.12, we have

$$\left\|\tilde{\Sigma}_{KK}^{-1}\right\|_2 \leq \frac{2}{\lambda_{\min}(\hat{\Sigma}_{KK})} \leq \frac{4}{\lambda_{\min}(\Sigma_{KK})}$$

with probability at least $1 - c_1\exp(-c_2\log\max(p,\, d))$ (or, $1 - c_1\exp(-c_2 n)$). Next, applying bound (18) (or (21)) from Lemma 6.12 with $\varepsilon' = \frac{c}{\sqrt{k_2}}$, with probability at least $1 - 2\exp(-b\frac{n}{k_2^3} + 2\log k_2)$, we have

$$\left\|\tilde{\Sigma}_{KK} - \hat{\Sigma}_{KK}\right\|_2 \leq \frac{c}{\sqrt{k_2}},$$

By choosing the constant $c > 0$ sufficiently small, we are guaranteed that

$$\mathbb{P}[||R_4||_1 \geq \frac{\phi}{12}] \leq 2\exp(-b\frac{n}{k_2^3} + 2\log k_2).$$

For the fifth term R_5, we first write

$$\begin{aligned} ||R_5||_1 &\leq \sqrt{k_2} \left\| \hat{\Sigma}_{KK}^{-1} \right\|_2 \left\| \tilde{\Sigma}_{K^cK} - \hat{\Sigma}_{K^cK} \right\|_1 \\ &\leq \frac{4\sqrt{k_2}}{\lambda_{\min}(\Sigma_{KK})} \left\| \tilde{\Sigma}_{K^cK} - \hat{\Sigma}_{K^cK} \right\|_1 . \end{aligned}$$

An application of bound (17) (or (20)) from Lemma 6.12 with $\varepsilon = \frac{\phi}{12} \frac{\lambda_{\min}(\Sigma_{KK})}{\sqrt{k_2}}$ yields

$$\mathbb{P}[||R_5||_1 \geq \frac{\phi}{12}] \leq 2\exp(-b\frac{n}{k_2^3} + \log(p-k_2) + \log k_2).$$

For the sixth term R_6, by applying bounds (17) and (19) (or, (20) and (22)), with $\varepsilon = \frac{\phi}{12} \frac{\lambda_{\min}(\Sigma_{KK})}{b'}$ for (17) (or (20)) with b' chosen to be sufficiently large, and the fact that $\log k_2 \leq \log(p-k_2)$, we are guaranteed that

$$||R_6||_1 \leq \frac{\phi}{12}$$

with probability at least $1 - c\exp(-b\frac{n}{k_2^2} + \log(p-k_2))$.

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

$$\mathbb{P}[||\tilde{\Sigma}_{K^cK}\tilde{\Sigma}_{KK}^{-1}||_1 \geq 1 - \frac{\phi}{4}] \leq O\left(\exp(-b\frac{n}{k_2^3} + \log p)\right).$$

□

6.6 Theorems 3.7-3.8

The proof for the first claim in Theorems 3.7 and 3.8 is established in Lemma 6.6, which shows that $\hat{\beta}_{H2SLS} = (\hat{\beta}_{J(\beta^*)}, \mathbf{0})$ where $\hat{\beta}_{J(\beta^*)}$ is the solution obtained in step 2 of the PDW construction. The second and third claims are proved using Lemma 6.7. The last claim is a consequence of the third claim.

Lemma 6.6: If the PDW construction succeeds, then under Assumption 3.8, the vector $(\hat{\beta}_{J(\beta^*)}, \mathbf{0}) \in \mathbb{R}^p$ is the unique optimal solution of the Lasso.

Remark: The proof for Lemma 6.6 is given in Lemma 1 in Wainwright [add reference].

Lemma 6.7: Suppose Assumptions 1.1, 3.2, 3.3, 3.5a, 3.7, and 3.8 hold. With the choice of

the tuning parameter

$$\begin{aligned}\lambda_n &\geq \frac{48(2-\frac{\phi}{4})}{\phi} b\sigma_{X^*}\sigma_Z \max_{j',j} |\mathrm{cov}(x^*_{ij'}, \mathbf{z}_{ij})|_\infty |\beta^*|_1 \max\left\{k_1\sqrt{\frac{\log d}{n}}, \sqrt{\frac{\log p}{n}}\right\} \\ &\asymp k_2 \max\left\{k_1\sqrt{\frac{\log d}{n}}, \sqrt{\frac{\log p}{n}}\right\},\end{aligned}$$

and under the condition $k_2 \max\left\{k_1\sqrt{\frac{\log d}{n}}, \sqrt{\frac{\log p}{n}}\right\} \to 0$, 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,

$$|\hat{\beta}_K - \beta^*_K|_\infty \leq \left[b|\beta^*|_1 \max\left\{k_1\sqrt{\frac{\log d}{n}}, \sqrt{\frac{\log p}{n}}\right\} + \lambda_n\right] \frac{4\sqrt{k_2}}{\lambda_{\min}(\Sigma_{KK})},$$

with probability at least $1 - c_1\exp(-c_2 \log\min(p, d))$. If Assumptions 1.1, 3.2-3.4, 3.5b, 3.6, and 3.8 hold, then with the choice of tuning parameter

$$\begin{aligned}\lambda_n &\geq \frac{48(2-\frac{\phi}{4})}{\phi} b\sigma_{X^*}\sigma_Z \max_{j',j} \sup_{v^j \in \mathbb{K}(k_1, d_j)} \left|\mathbb{E}(x^*_{1j'}\mathbf{z}_{1j}v^j)\right| |\beta^*|_1 \max\left\{\sqrt{\frac{k_1\log d}{n}}, \sqrt{\frac{\log p}{n}}\right\} \\ &\asymp k_2 \max\left\{\sqrt{\frac{k_1\log d}{n}}, \sqrt{\frac{\log p}{n}}\right\}\end{aligned}$$

and under the condition $k_2 \max\left\{\sqrt{\frac{k_1\log d}{n}}, \sqrt{\frac{\log p}{n}}\right\} \to 0$, we have $|\hat{\mu}_{K^c}|_\infty \leq 1 - \frac{\phi}{8}$ with probability at least $1 - c_1\exp(-c_2 \log\min(p, d))$, and

$$|\hat{\beta}_K - \beta^*_K|_\infty \leq \left[b|\beta^*|_1 \max\left\{\sqrt{\frac{k_1\log d}{n}}, \sqrt{\frac{\log p}{n}}\right\} + \lambda_n\right] \frac{4\sqrt{k_2}}{\lambda_{\min}(\Sigma_{KK})},$$

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-gradient condition in the PDW construction. Recall $e := (X - \hat{X})\beta^* + \boldsymbol{\eta}\beta^* + \epsilon$ from Lemma 3.1. With the fact that $\hat{\beta}_{K^c} = \beta^*_{K^c} = 0$, we have

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

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

$$|\hat{\mu}_{K^c}|_\infty \leq \left\|\tilde{\Sigma}_{K^cK}\tilde{\Sigma}_{KK}^{-1}\right\|_1 + \left|\hat{X}_{K^c}^T\frac{e}{n\lambda_n}\right|_\infty + \left\|\tilde{\Sigma}_{K^cK}\tilde{\Sigma}_{KK}^{-1}\right\|_1\left|\hat{X}_K^T\frac{e}{n\lambda_n}\right|_\infty,$$

where I used the fact that $|\hat{\mu}_K|_\infty \leq 1$. By Lemma 6.5, we have $\left\|\tilde{\Sigma}_{K^cK}\tilde{\Sigma}_{KK}^{-1}\right\|_1 \leq 1 - \frac{\phi}{4}$ with probability at least $1 - c\exp(-b\frac{n}{k_2^3} + \log p)$. Hence,

$$\begin{aligned}
|\hat{\mu}_{K^c}|_\infty &\leq 1 - \frac{\phi}{4} + \left|\hat{X}_{K^c}^T\frac{e}{n\lambda_n}\right|_\infty + \left\|\tilde{\Sigma}_{K^cK}\tilde{\Sigma}_{KK}^{-1}\right\|_1\left|\hat{X}_K^T\frac{e}{n\lambda_n}\right|_\infty \\
&\leq 1 - \frac{\phi}{4} + \left(2 - \frac{\phi}{4}\right)\left|\hat{X}^T\frac{e}{n\lambda_n}\right|_\infty.
\end{aligned}$$

Therefore, in order to show it suffices to show that $\left(2 - \frac{\phi}{4}\right)\left|\hat{X}^T\frac{e}{n\lambda_n}\right|_\infty \leq \frac{\phi}{8}$ with high probability. This result is established in Lemma 6.12. Thus, we have $|\hat{\mu}_{K^c}|_\infty \leq 1 - \frac{\phi}{8}$ 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

$$\begin{aligned}
|\hat{\beta}_K - \beta_K^*|_\infty &\leq \left|\left(\frac{\hat{X}_K^T\hat{X}_K}{n}\right)^{-1}\frac{\hat{X}_K^T e}{n}\right|_\infty + \lambda_n\left\|\left(\frac{\hat{X}_K^T\hat{X}_K}{n}\right)^{-1}\right\|_1 \\
&\leq \left\|\left(\frac{\hat{X}_K^T\hat{X}_K}{n}\right)^{-1}\right\|_1\left|\frac{\hat{X}_K^T e}{n}\right|_\infty + \lambda_n\left\|\left(\frac{\hat{X}_K^T\hat{X}_K}{n}\right)^{-1}\right\|_1,
\end{aligned}$$

Using bounds (16) and (23) (or (25)) from Lemma 6.11, we have

$$\left\|\left(\frac{\hat{X}_K^T\hat{X}_K}{n}\right)^{-1}\right\|_1 \leq \frac{2\sqrt{k_2}}{\lambda_{\min}(\hat{\Sigma}_{KK})} \leq \frac{4\sqrt{k_2}}{\lambda_{\min}(\Sigma_{KK})}.$$

By Lemma 6.2, we have, with probability at least $1 - c_1 \exp(-c_2 \log \min(p,\, d))$,

$$|\frac{1}{n}\hat{X}^T e|_\infty \leq b|\beta^*|_1 \max\left\{k_1\sqrt{\frac{\log d}{n}},\, \sqrt{\frac{\log p}{n}}\right\}.$$

By Lemma 6.4, we have, with probability at least $1 - c_1 \exp(-c_2 \log \min(p,\, d)))$,

$$|\frac{1}{n}\hat{X}^T e|_\infty \leq b|\beta^*|_1 \max\left\{\sqrt{\frac{k_1 \log d}{n}},\, \sqrt{\frac{\log p}{n}}\right\}.$$

Putting everything together, we obtain

$$|\hat{\beta}_K - \beta^*_K|_\infty \leq \left[b|\beta^*|_1 \max\left\{k_1\sqrt{\frac{\log d}{n}},\, \sqrt{\frac{\log p}{n}}\right\} + \lambda_n\right] \frac{4\sqrt{k_2}}{\lambda_{\min}(\Sigma_{KK})},$$

or,

$$|\hat{\beta}_K - \beta^*_K|_\infty \leq \left[b|\beta^*|_1 \max\left\{\sqrt{\frac{k_1 \log d}{n}},\, \sqrt{\frac{\log p}{n}}\right\} + \lambda_n\right] \frac{4\sqrt{k_2}}{\lambda_{\min}(\Sigma_{KK})},$$

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

6.7 Lemmas 6.8-6.13

Lemma 6.8: 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

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

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

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

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

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

Remark. Lemma 6.8 is Lemma 14 in Loh and Wainwright (2012).

Lemma 6.9: 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

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

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 6.9 is Lemma 12 in Loh and Wainwright (2012).

Lemma 6.10: Under Assumption 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. Under Assumption 3.3, we have that X^* is sub-Gaussian with parameters $(\Sigma_{X^*},\, \sigma_{X^*})$ where $\Sigma_{X^*} = \mathbb{E}(X^{*T} X^*)$. Therefore, by Lemma 6.8 and a discretization argument, we have

$$\mathbb{P}\left(\sup_{v^0 \in \mathbb{K}(2s,\, p)} \left| v^{0T} \left(\frac{X^{*T} X^*}{n} - \Sigma_{X^*} \right) v^0 \right| \geq t \right) \leq 2 \exp(-c^{'} n \min(\frac{t^2}{\sigma_{X^*}^4},\, \frac{t}{\sigma_{X^*}^2}) + 2s \log p),$$

for some universal constants $c^{'} > 0$. By choosing $t = \frac{\lambda_{\min}(\Sigma_{X^*})}{54}$ and let

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

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 6.9 and the following substitutions

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

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

$$v^{0T}\frac{X^{*T}X^*}{n}v^0 \geq v^{0T}\Sigma_{X^*}v^0 - \frac{\lambda_{\min}(\Sigma_{X^*})}{2}\left(||v^0||_2^2 + \frac{1}{s}||v^0||_1^2\right).$$

Again, with the choice of

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

where c' is chosen sufficiently small so $s \geq 1$, the claim follows. $\square$

Lemma 6.11: Suppose Assumptions 1.1 and 3.8 hold. For any $\varepsilon > 0$ and constant c, we have

$$\mathbb{P}\left\{\left\|\hat{\Sigma}_{K^cK} - \Sigma_{K^cK}\right\|_1 \geq \varepsilon\right\} \leq (p-k_2)k_2 \cdot 2\exp(-c\min\{\frac{n\varepsilon^2}{4k_2^2\sigma_{X^*}^4},\, \frac{n\varepsilon}{2k_2\sigma_{X^*}^2}\}), \tag{11}$$

$$\mathbb{P}\left\{\left\|\hat{\Sigma}_{KK} - \Sigma_{KK}\right\|_1 \geq \varepsilon\right\} \leq k_2^2 \cdot 2\exp(-c\min\{\frac{n\varepsilon^2}{4k_2^2\sigma_{X^*}^4},\, \frac{n\varepsilon}{2k_2\sigma_{X^*}^2}\}). \tag{12}$$

Furthermore, under the scaling $n \gtrsim k_2 \log p$, for constants b_1 and b_2, we have

$$\left\|\hat{\Sigma}_{KK}^{-1} - \Sigma_{KK}^{-1}\right\|_1 \leq \frac{1}{\lambda_{\min}(\Sigma_{KK})} \quad \text{with probability at least } 1 - b_1\exp(-b_2\frac{n}{k_2}). \tag{13}$$

Proof. Denote the element $(j',\, j)$ of the matrix difference $\hat{\Sigma}_{K^cK} - \Sigma_{K^cK}$ by $u_{j'j}$. By the definition of the l_1- operator norm, we have

$$\begin{aligned}
\mathbb{P}\left\{\left\|\hat{\Sigma}_{K^cK} - \Sigma_{K^cK}\right\|_1 \geq \varepsilon\right\} &= \mathbb{P}\left\{\max_{j'\in K^c}\sum_{j\in K}|u_{j'j}| \geq \varepsilon\right\} \\
&\leq (p-k_2)\mathbb{P}\left\{\sum_{j\in K}|u_{j'j}| \geq \varepsilon\right\} \\
&\leq (p-k_2)\mathbb{P}\left\{\exists j \in K \,|\, |u_{j'j}| \geq \frac{\varepsilon}{k_2}\right\} \\
&\leq (p-k_2)k_2\mathbb{P}\left\{|u_{j'j}| \geq \frac{\varepsilon}{k_2}\right\} \\
&\leq (p-k_2)k_2 \cdot 2\exp(-c\min\{\frac{n\varepsilon^2}{4k_2^2\sigma_{X^*}^4},\, \frac{n\varepsilon}{2k_2\sigma_{X^*}^2}\}),
\end{aligned}$$

where the last inequality the deviation bound for sub-exponential random variables. Bound (12) 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 (13), write

$$\begin{aligned}
\left\|\hat{\Sigma}_{KK}^{-1} - \Sigma_{KK}^{-1}\right\|_1 &= \left\|\Sigma_{KK}^{-1}\left[\Sigma_{KK} - \hat{\Sigma}_{KK}\right]\hat{\Sigma}_{KK}^{-1}\right\|_1 \\
&= \sqrt{k_2}\left\|\Sigma_{KK}^{-1}\left[\Sigma_{KK} - \hat{\Sigma}_{KK}\right]\hat{\Sigma}_{KK}^{-1}\right\|_2 \\
&= \sqrt{k_2}\left\|\Sigma_{KK}^{-1}\right\|_2\left\|\Sigma_{KK} - \hat{\Sigma}_{KK}\right\|_2\left\|\hat{\Sigma}_{KK}^{-1}\right\|_2 \\
&\leq \frac{\sqrt{k_2}}{\lambda_{\min}(\Sigma_{KK})}\left\|\Sigma_{KK} - \hat{\Sigma}_{KK}\right\|_2\left\|\hat{\Sigma}_{KK}^{-1}\right\|_2. \qquad (14)
\end{aligned}$$

To bound the terms $\left\|\Sigma_{KK} - \hat{\Sigma}_{KK}\right\|_2$ and $\left\|\hat{\Sigma}_{KK}^{-1}\right\|_2$ in (14), note that we can write

$$\begin{aligned}
\lambda_{\min}(\Sigma_{KK}) &= \min_{||h||_2=1} h^T\Sigma_{KK}h \\
&= \min_{||h||_2=1} h^T\hat{\Sigma}_{KK}h + h^T(\Sigma_{KK} - \hat{\Sigma}_{KK})h \\
&\leq h^T\hat{\Sigma}_{KK}h + h^T(\Sigma_{KK} - \hat{\Sigma}_{KK})h \qquad (15)
\end{aligned}$$

where $h \in \mathbb{R}^{k_2}$ is a unit-norm minimal eigenvector of $\hat{\Sigma}_{KK}$. With a slight modification of Lemma 6.9 by replacing $\frac{X^{*T}X^*}{n} - \Sigma_{X^*}$ with $\Sigma_{KK} - \hat{\Sigma}_{KK}$, and setting

$$s := \frac{1}{b}\frac{n}{k_2\log p}\min\left\{\frac{\lambda_{\min}^2(\Sigma_{KK})}{\sigma_{X^*}^4}, 1\right\},$$

with b chosen sufficiently small so $s \geq 1$, we have

$$h^T\left(\Sigma_{KK} - \hat{\Sigma}_{KK}\right)h \leq \frac{\lambda_{\min}(\Sigma_{KK})}{4\sqrt{k_2}}\left(1 + \frac{k_2\log p}{n}\right)$$

with probability at least $1 - b_1\exp(-b_2\frac{n}{k_2})$. Therefore, if $n \gtrsim k_2\log p$,

$$\begin{aligned}
h^T\left(\Sigma_{KK} - \hat{\Sigma}_{KK}\right)h &\leq \frac{\lambda_{\min}(\Sigma_{KK})}{4\sqrt{k_2}}\left(1 + \frac{k_2\log p}{n}\right) \leq \frac{\lambda_{\min}(\Sigma_{KK})}{2\sqrt{k_2}}, \\
\lambda_{\min}(\hat{\Sigma}_{KK}) &\geq \lambda_{\min}(\Sigma_{KK}) - \frac{\lambda_{\min}(\Sigma_{KK})}{4\sqrt{k_2}}\left(1 + \frac{k_2\log p}{n}\right) \geq \frac{\lambda_{\min}(\Sigma_{KK})}{2} \qquad (16)
\end{aligned}$$

and consequently,

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

Putting everything together, we have

$$\left\|\hat{\Sigma}_{KK}^{-1} - \Sigma_{KK}^{-1}\right\|_1 \leq \frac{\sqrt{k_2}}{\lambda_{\min}(\Sigma_{KK})} \frac{\lambda_{\min}(\Sigma_{KK})}{2\sqrt{k_2}} \frac{2}{\lambda_{\min}(\Sigma_{KK})} = \frac{1}{\lambda_{\min}(\Sigma_{KK})}.$$

with probability at least $1 - b_1 \exp(-b_2 \frac{n}{k_2})$. $\square$

Lemma 6.12: (i) Suppose the assumptions in Lemmas 6.1 and 6.2 hold. For any $\varepsilon^{'} > 0$ and constants c and $c^{'}$, under the condition $k_1\sqrt{\frac{\log d}{n}} \to 0$, we have

$$\mathbb{P}\left\{\left\|\tilde{\Sigma}_{K^cK} - \hat{\Sigma}_{K^cK}\right\|_1 \geq \varepsilon^{'}\right\} \leq (p-k_2)k_2 \cdot \exp(-cn\min\{\frac{\varepsilon^{'2}}{k_2^2\sigma_{X^*}^2\sigma_Z^2}, \frac{\varepsilon^{'}}{k_2\sigma_{X^*}\sigma_Z}\} + c^{'}\log p + \log d), \quad (17)$$

$$\mathbb{P}\left\{\left\|\tilde{\Sigma}_{KK} - \hat{\Sigma}_{KK}\right\|_1 \geq \varepsilon^{'}\right\} \leq k_2^2 \cdot \exp(-cn\min\{\frac{\varepsilon^{'2}}{k_2^2\sigma_{X^*}^2\sigma_Z^2}, \frac{\varepsilon^{'}}{k_2\sigma_{X^*}\sigma_Z}\} + c^{'}\log p + \log d). \quad (18)$$

Furthermore, if $n \gtrsim k_1^2 k_2^3 \log d$ and $k_1 k_2 \sqrt{\frac{\log d}{n}} \to 0$, for constants b, c_1, and c_2, we have

$$\left\|\tilde{\Sigma}_{KK}^{-1} - \hat{\Sigma}_{KK}^{-1}\right\|_1 \leq \frac{b}{\lambda_{\min}(\Sigma_{KK})} \quad \text{with probability at least } 1 - c_1\exp(-c_2\log\max(p,\, d)). \quad (19)$$

(ii) Suppose the assumptions in Lemmas 6.3 and 6.4 hold. For any $\varepsilon^{'} > 0$ and constant c, under the condition $\sqrt{\frac{k_1 \log d}{n}} \to 0$, we have

$$\mathbb{P}\left\{\left\|\tilde{\Sigma}_{K^cK} - \hat{\Sigma}_{K^cK}\right\|_1 \geq \varepsilon^{'}\right\} \leq (p-k_2)k_2 \cdot 2\exp(-cn\min(\frac{\varepsilon^{'2}}{k_2^2\sigma_{X^*}^2\sigma_W^2}, \frac{\varepsilon^{'}}{k_2\sigma_{X^*}\sigma_W}) + k_1\log d + 2\log p), \quad (20)$$

$$\mathbb{P}\left\{\left\|\tilde{\Sigma}_{KK} - \hat{\Sigma}_{KK}\right\|_1 \geq \varepsilon^{'}\right\} \leq k_2^2 \cdot 2\exp(-cn\min(\frac{\varepsilon^{'2}}{k_2^2\sigma_{X^*}^2\sigma_W^2}, \frac{\varepsilon^{'}}{k_2\sigma_{X^*}\sigma_W}) + k_1\log d + 2\log p). \quad (21)$$

Furthermore, if $n \gtrsim \max\{k_1^2 k_2 \log d,\, k_2^{3/2}\log d,\, k_2^{3/2}\log p\}$ and $\max\{k_1\sqrt{\frac{\log d}{n}},\, k_2\frac{\log\max(p,\,d)}{n}\} \to 0$, for constants b, c_1, and c_2, we have

$$\left\|\tilde{\Sigma}_{KK}^{-1} - \hat{\Sigma}_{KK}^{-1}\right\|_1 \leq \frac{b}{\lambda_{\min}(\Sigma_{KK})} \quad \text{with probability at least } 1 - c_1\exp(-c_2 n). \quad (22)$$

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 6.11, under the condition $k_1\sqrt{\frac{\log d}{n}} \to 0$, we have

$$\begin{aligned}
\mathbb{P}\left\{\left\|\tilde{\Sigma}_{K^cK} - \hat{\Sigma}_{K^cK}\right\|_1 \geq \varepsilon^{'}\right\} &\leq (p-k_2)k_2\mathbb{P}\left\{|w_{j'j}| \geq \frac{\varepsilon^{'}}{k_2}\right\} \\
&\leq (p-k_2)k_2 \cdot \exp(-cn\min\{\frac{\varepsilon^{'2}}{\sigma_{X^*}^2\sigma_Z^2}, \frac{\varepsilon^{'}}{\sigma_{X^*}\sigma_Z}\} + c^{'}\log p + \log d),
\end{aligned}$$

where the last inequality follows from the bounds on $\left|\frac{(\hat{X}-X^*)^TX^*}{n}\right|_1$ and $\left|\frac{(\hat{X}-X^*)^T(\hat{X}-X^*)}{n}\right|_1$ in the proof for Lemma 6.1 and the identity

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

Bound (18) can be obtained in a similar way except that the pre-factor $(p-k_2)$ is replaced by k_2.

To prove bound (19), by applying the same argument as in Lemma 6.11, we have

$$\begin{aligned}
\left\|\hat{\Sigma}_{KK}^{-1} - \Sigma_{KK}^{-1}\right\|_1 &\leq \frac{\sqrt{k_2}}{\lambda_{\min}(\hat{\Sigma}_{KK})}\left\|\hat{\Sigma}_{KK} - \tilde{\Sigma}_{KK}\right\|_2\left\|\tilde{\Sigma}_{KK}^{-1}\right\|_2 \\
&\leq \frac{2\sqrt{k_2}}{\lambda_{\min}(\Sigma_{KK})}\left\|\hat{\Sigma}_{KK} - \tilde{\Sigma}_{KK}\right\|_2\left\|\tilde{\Sigma}_{KK}^{-1}\right\|_2,
\end{aligned}$$

where the last inequality comes from bound (16). To bound the terms $\left\|\hat{\Sigma}_{KK} - \tilde{\Sigma}_{KK}\right\|_2$ and $\left\|\tilde{\Sigma}_{KK}^{-1}\right\|_2$, we have, again

$$\begin{aligned}
\lambda_{\min}(\hat{\Sigma}_{KK}) &\leq h^T\tilde{\Sigma}_{KK}h + h^T(\hat{\Sigma}_{KK} - \tilde{\Sigma}_{KK})h \\
&\leq h^T\tilde{\Sigma}_{KK}h + k_2\left|\hat{\Sigma}_{KK} - \tilde{\Sigma}_{KK}\right|_\infty \\
&\leq h^T\tilde{\Sigma}_{KK}h + bk_1k_2\sqrt{\frac{\log d}{n}},
\end{aligned}$$

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 $\left|\frac{(\hat{X}-X^*)^TX^*}{n}\right|_\infty$ and $\left|\frac{(\hat{X}-X^*)^T(\hat{X}-X^*)}{n}\right|_\infty$ from the proof for Lemma 1 with probability at least $1 - c_1\exp(-c_2\log\max(p, d))$. Therefore, if $n \gtrsim k_1^2k_2^3\log d$ and $k_1k_2\sqrt{\frac{\log d}{n}} \to 0$, then we have

$$\lambda_{\min}(\tilde{\Sigma}_{KK}) \geq \lambda_{\min}(\hat{\Sigma}_{KK}) - k_1k_2\sqrt{\frac{\log d}{n}} \geq \frac{\lambda_{\min}(\hat{\Sigma}_{KK})}{2}$$

$$\Longrightarrow \left\|\tilde{\Sigma}_{KK}^{-1}\right\|_2 \leq \frac{2}{\lambda_{\min}(\hat{\Sigma}_{KK})}, \tag{23}$$

$$\text{and} \quad \left\|\hat{\Sigma}_{KK} - \tilde{\Sigma}_{KK}\right\|_2 \leq \frac{b\lambda_{\min}(\hat{\Sigma}_{KK})}{\sqrt{k_2}}. \tag{24}$$

Putting everything together, we have

$$\left\|\hat{\Sigma}_{KK}^{-1} - \tilde{\Sigma}_{KK}^{-1}\right\|_1 \leq \frac{2\sqrt{k_2}}{\lambda_{\min}(\Sigma_{KK})} \frac{b\lambda_{\min}(\hat{\Sigma}_{KK})}{\sqrt{k_2}} \frac{2}{\lambda_{\min}(\hat{\Sigma}_{KK})} = \frac{b'}{\lambda_{\min}(\Sigma_{KK})}.$$

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

For Part (ii) of Lemma 6.12, we can bound the terms using results from Lemma 6.3 instead of Lemma 6.1. 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 6.11, under the condition $\sqrt{\frac{k_1 \log d}{n}} \to 0$, we have

$$\begin{aligned}
\mathbb{P}\left\{\left\|\tilde{\Sigma}_{K^cK} - \hat{\Sigma}_{K^cK}\right\|_1 \geq \varepsilon^{'}\right\} &\leq (p-k_2)k_2 \mathbb{P}\left\{|w_{j'j}| \geq \frac{\varepsilon^{'}}{k_2}\right\} \\
&\leq (p-k_2)k_2 \cdot 2\exp(-cn\min(\frac{\varepsilon^{'2}}{k_2^2\sigma_{X^*}^2\sigma_W^2},\, \frac{\varepsilon^{'}}{k_2\sigma_{X^*}\sigma_W}) + k_1 \log d + 2\log p),
\end{aligned}$$

where the last inequality follows from the bounds (9)-(10) and the identity

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

Bound (21) 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 (22), by applying the same argument as in Lemma 6.11, we have

$$\begin{aligned}
\left\|\hat{\Sigma}_{KK}^{-1} - \Sigma_{KK}^{-1}\right\|_1 &\leq \frac{\sqrt{k_2}}{\lambda_{\min}(\hat{\Sigma}_{KK})} \left\|\hat{\Sigma}_{KK} - \tilde{\Sigma}_{KK}\right\|_2 \left\|\tilde{\Sigma}_{KK}^{-1}\right\|_2 \\
&\leq \frac{2\sqrt{k_2}}{\lambda_{\min}(\Sigma_{KK})} \left\|\hat{\Sigma}_{KK} - \tilde{\Sigma}_{KK}\right\|_2 \left\|\tilde{\Sigma}_{KK}^{-1}\right\|_2,
\end{aligned}$$

where the last inequality comes from bound (16). Again, we have

$$\lambda_{\min}(\hat{\Sigma}_{KK}) \leq h^T\tilde{\Sigma}_{KK}h + h^T(\hat{\Sigma}_{KK} - \tilde{\Sigma}_{KK})h$$

where $h \in \mathbb{R}^{k_2}$ is a unit-norm minimal eigenvector of $\tilde{\Sigma}_{KK}$. By bounds (7) and (8) and choosing

$s = \frac{1}{c}\frac{n}{\log\max(p,d)}\min\left\{\frac{\lambda^2_{\min}(\Sigma_{X^*})}{\sigma^4_{X^*}}, 1\right\}$, we have

$$\begin{aligned} h^T\left(\hat{\Sigma}_{KK} - \tilde{\Sigma}_{KK}\right)h &\leq 27bk_1\sqrt{\frac{\log d}{n}}(|h|_2^2 + \frac{1}{s}|h|_1^2) \\ &\leq b_1\max\{k_1\sqrt{\frac{\log d}{n}}, k_2\frac{\log\max(p,\, d)}{n}\}|h|_2^2 \end{aligned}$$

with probability at least $1 - c_1\exp(-c_2 n)$. Therefore, if $n \gtrsim \max\{k_1^2 k_2 \log d,\, k_2^{3/2}\log d,\, k_2^{3/2}\log p\}$ and $\max\{k_1\sqrt{\frac{\log d}{n}}, k_2\frac{\log\max(p,d)}{n}\} \to 0$, then we have

$$\lambda_{\min}(\tilde{\Sigma}_{KK}) \geq \lambda_{\min}(\hat{\Sigma}_{KK}) - b_1\max\{k_1\sqrt{\frac{\log d}{n}}, k_2\frac{\log\max(p,\, d)}{n}\} \geq \frac{\lambda_{\min}(\hat{\Sigma}_{KK})}{2}$$

$$\begin{aligned} \Longrightarrow \left\|\tilde{\Sigma}_{KK}^{-1}\right\|_2 &\leq \frac{2}{\lambda_{\min}(\hat{\Sigma}_{KK})}, &(25)\\ \text{and} \quad \left\|\hat{\Sigma}_{KK} - \tilde{\Sigma}_{KK}\right\|_2 &\leq \frac{b_2\lambda_{\min}(\hat{\Sigma}_{KK})}{\sqrt{k_2}}. &(26) \end{aligned}$$

Putting everything together, we have

$$\left\|\hat{\Sigma}_{KK}^{-1} - \tilde{\Sigma}_{KK}^{-1}\right\|_1 \leq \frac{2\sqrt{k_2}}{\lambda_{\min}(\Sigma_{KK})}\frac{b_2\lambda_{\min}(\hat{\Sigma}_{KK})}{\sqrt{k_2}}\frac{2}{\lambda_{\min}(\hat{\Sigma}_{KK})} = \frac{b_3}{\lambda_{\min}(\Sigma_{KK})}.$$

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

Lemma 6.13: (i) Suppose Assumptions 1.1, 3.2, 3.3, and 3.5a hold. With the choice of the tuning parameter

$$\begin{aligned} \lambda_n &\geq \frac{48(2-\frac{\phi}{4})}{\phi}b\sigma_{X^*}\sigma_Z\max_{j',j}|\mathrm{cov}(x^*_{ij'},\, \mathbf{z}_{ij})|_\infty|\beta^*|_1\max\left\{k_1\sqrt{\frac{\log d}{n}}, \sqrt{\frac{\log p}{n}}\right\} \\ &\asymp k_2\max\left\{k_1\sqrt{\frac{\log d}{n}}, \sqrt{\frac{\log p}{n}}\right\} \end{aligned}$$

and under the condition $k_2\max\left\{k_1\sqrt{\frac{\log d}{n}}, \sqrt{\frac{\log p}{n}}\right\} \to 0$, 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 $1 - c_1\exp(-c_2\log\max(p,\, d))$. (ii) Suppose Assumptions 1.1, 3.2-3.4, 3.5a,

and 3.6 hold. Then the same result can be obtained with the choice of tuning parameter

$$\begin{aligned}\lambda_n &\geq \frac{48(2-\frac{\phi}{4})}{\phi} b\sigma_{X^*}\sigma_Z \max_{j',j} \sup_{v^j \in \mathbb{K}(k_1,d_j)} \left|\mathbb{E}(x^*_{1j'}\mathbf{z}_{1j}v^j)\right| |\beta^*|_1 \max\left\{\sqrt{\frac{k_1 \log d}{n}}, \sqrt{\frac{\log p}{n}}\right\} \\ &\asymp k_2 \max\left\{\sqrt{\frac{k_1 \log d}{n}}, \sqrt{\frac{\log p}{n}}\right\}\end{aligned}$$

and the condition $k_2 \max\left\{\sqrt{\frac{k_1 \log d}{n}}, \sqrt{\frac{\log p}{n}}\right\} \to 0$.

Proof. Recall from the proof for Lemma 6.2,

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

Hence,

$$\begin{aligned}|\frac{1}{n\lambda_n}\hat{X}^T e|_\infty &\leq |\frac{1}{n\lambda_n}X^{*T}(\hat{X}-X^*)\beta^*|_\infty + |\frac{1}{n\lambda_n}X^{*T}\boldsymbol{\eta}\beta^*|_\infty + |\frac{1}{n\lambda_n}X^{*T}\epsilon|_\infty \\ &+ |\frac{1}{n\lambda_n}(\hat{X}-X^*)^T(\hat{X}-X^*)\beta^*|_\infty + |\frac{1}{n\lambda_n}(X^*-\hat{X})^T\boldsymbol{\eta}\beta^*|_\infty + |\frac{1}{n\lambda_n}(X^*-\hat{X})^T\epsilon|_\infty.\end{aligned} \tag{27}$$

Again, for any $j' = 1, ..., p$, we have

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

Hence, by applying the same argument as in the proof for Lemma 6.1, we have

$$\mathbb{P}\left[\max_{j',j}|\frac{1}{n}\mathbf{x}^{*T}_{j'}Z_j - \mathrm{cov}(x^*_{ij'}, \mathbf{z}_{ij})|_\infty \geq \lambda_n\right] \leq 6p^2 d\exp(-cn\min\{\frac{\lambda_n^2}{\sigma^2_{X^*}\sigma^2_Z}, \frac{\lambda_n}{\sigma_{X^*}\sigma_Z}\}).$$

Therefore as long as

$$\begin{aligned}\lambda_n &\geq \frac{48(2-\frac{\phi}{4})}{\phi} b\sigma_{X^*}\sigma_Z \max_{j',j} |\mathrm{cov}(x^*_{ij'}, \mathbf{z}_{ij})|_\infty |\beta^*|_1 \max\left\{k_1\sqrt{\frac{\log d}{n}}, \sqrt{\frac{\log p}{n}}\right\} \\ &\asymp k_2 \max\left\{k_1\sqrt{\frac{\log d}{n}}, \sqrt{\frac{\log p}{n}}\right\},\end{aligned}$$

with b chosen to be sufficiently large, under the scaling $k_2 \max\left\{k_1\sqrt{\frac{\log d}{n}}, \sqrt{\frac{\log p}{n}}\right\} \to 0$, we have

$$|\frac{1}{n\lambda_n} X^{*T}(\hat{X} - X^*)\beta^*|_\infty \leq \frac{\phi}{48(2-\frac{\phi}{4})},$$

with probability at least $1 - c_1\exp(-c_2 \log\max(p, d))$. With the same choice of λ_n and under the scaling $k_2 \max\left\{k_1\sqrt{\frac{\log d}{n}}, \sqrt{\frac{\log p}{n}}\right\} \to 0$, we also have

$$|\frac{1}{n\lambda_n}(\hat{X} - X^*)^T(\hat{X} - X^*)\beta^*|_\infty \precsim \sqrt{\frac{\log d}{n}} = o(1) \leq \frac{\phi}{48(2-\frac{\phi}{4})}.$$

For the term $|\frac{1}{n\lambda_n} X^{*T}\boldsymbol{\eta}\beta^*|_\infty$, we have

$$|\frac{1}{n\lambda_n} X^{*T}\boldsymbol{\eta}\beta^*|_\infty \leq \max_{j',j} |\frac{1}{n\lambda_n}\sum_{i=1}^n x^*_{ij'}\eta_{ij}||\beta^*|_1.$$

Notice that

$$\mathbb{P}\left[\max_{j',j} \frac{1}{n}\mathbf{x}^{*T}_{j'}\boldsymbol{\eta}_j \geq \lambda_n\right] \leq 6p^2\exp(-cn\min\{\frac{\lambda_n^2}{\sigma^2_{X^*}\sigma^2_\eta}, \frac{\lambda_n}{\sigma_{X^*}\sigma_\eta}\}).$$

where I have used the assumption that $\mathbb{E}(\mathbf{z}_{ij'}\eta_{ij}) = \mathbf{0}$ for all j', j. With the same choice of λ_n and under the scaling $k_2 \max\left\{k_1\sqrt{\frac{\log d}{n}}, \sqrt{\frac{\log p}{n}}\right\} \to 0$, we have

$$|\frac{1}{n\lambda_n} X^{*T}\boldsymbol{\eta}\beta^*|_\infty \leq \frac{\phi}{48(2-\frac{\phi}{4})},$$

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

$$|\frac{1}{n\lambda_n}(X^* - \hat{X})^T\boldsymbol{\eta}\beta^*|_\infty \precsim \sqrt{\frac{\log d}{n}} = o(1) \leq \frac{\phi}{48(2-\frac{\phi}{4})}.$$

Similarly, for the term $|\frac{1}{n\lambda_n} X^{*T}\epsilon|_\infty$ and $|\frac{1}{n\lambda_n}(X^* - \hat{X})^T\epsilon|_\infty$, we can show

$$\begin{aligned} |\frac{1}{n} X^{*T}\epsilon|_\infty &\leq \frac{\phi}{48(2-\frac{\phi}{4})}, \\ |\frac{1}{n}(X^* - \hat{X})^T\epsilon|_\infty &\lesssim \sqrt{\frac{\log d}{n}} = o(1) \leq \frac{\phi}{48(2-\frac{\phi}{4})}, \end{aligned}$$

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

Putting everything together, 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 $1 - c_1 \exp(-c_2 \log \max(p,\, d))$.

The proof for Part (ii) of Lemma 6.13 follows from the similar argument for proving Part (i) except that we bound the terms $|\frac{1}{n\lambda_n} X^{*T}(\hat{X} - X^*)\beta^*|_\infty$ and $|\frac{1}{n\lambda_n}(\hat{X} - X^*)^T(\hat{X} - X^*)\beta^*|_\infty$ using the discretization type of argument as in the proof for Lemma 6.4. $\square$

References

[1] Ackerberg, D. A., and G. S. Crawford (2009). "Estimating Price Elasticities in Differentiated Product Demand Models with Endogenous Characteristics". Working Paper.

[2] Amemiya, T. (1974). "The Non-Linear Two-Stage Least Squares Estimator". *Journal of Econometrics*, 2, 105-110.

[3] Angrist, J. D., and A. B. Krueger (1991). "Does Compulsory School Attendance Affect Schooling and Earnings?". *Quarterly Journal of Economics*, 106, 979-1014.

[4] 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.

[5] Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2010). "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain". Preprint: arXiv:1010.4345.

[6] Belloni, A., and V. Chernozhukov (2010). "Post L1-Penalized Estimators in High-Dimensional Linear Regression Models". Preprint: arXiv:1001.0188v2.

[7] Belloni, A., V. Chernozhukov, and L. Wang (2010). "Square-Root Lasso: Pivotal Recovery of Sparse Signals Via Conic Programming". Preprint: arXiv:1009.5689.

[8] Belloni, A., and V. Chernozhukov (2011a). "L1-Penalized Quantile Regression in High-Dimensional Sparse Models". *The Annals of Statistics*, 39, 82-130.

[9] 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.

[10] Berry, S. T., J. A. Levinsohn, and A. Pakes (1995). "Automobile Prices in Market Equilibrium". *Econometrica*, 63, 841-890.

[11] Bickel, P., J. Y. Ritov, and A. B. Tsybakov (2009). "Simultaneous Analysis of Lasso and Dantzig Selector". *The Annals of Statistics*, 37, 1705-1732.

[12] Bühlmann, P., and S. A. van de Geer (2011). *Statistics for High-Dimensional Data*. Springer, New-York.

[13] Caner, M. (2009). "LASSO Type GMM Estimator". *Econometric Theory*, 25, 1-23.

[14] 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.

[15] Carrasco, M., and J. P. Florens (2000). "Generalization of GMM to a Continuum of Moment Conditions". *Econometric Theory*, 16, 797-834.

[16] Carrasco, M. (2008). "A Regularization Approach to the Many Instruments Problem". Working Paper.

[17] 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.

[18] 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.

[19] Fan, J., and Y. Liao (2011). "Ultra High Dimensional Variable Selection with Endogenous Covariates". Working Paper.

[20] Gautier, E., and A. B. Tsybakov (2011). "High-dimensional Instrumental Variables Regression and Confidence Sets". Preprint: arXiv:1105.2454v3.

[21] Hansen, C., J. Hausman, and W. K. Newey (2008). "Estimation with Many Instrumental Variables". *Journal of Business and Economic Statistics*, 26, 398-422.

[22] Koltchinskii, V. (2009). "The Dantzig Selector and Sparsity Oracle Inequalities". *Bernoulli*, 15, 799-828.

[23] Koltchinskii, V. (2011). "Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems". Forthcoming in *Lecture Notes in Mathematics*, Springer, Berlin.

[24] Ledoux, M. (2001). *The concentration of measure phenomenon*. Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI.

[25] Ledoux, M., and M. Talagrand (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, New York, NY.

[26] Lin, Y., and H. H. Zhang (2006). "Component Selection and Smoothing in Multivariate Nonparametric Regression". *The Annals of Statistics*, 34(5): 2272-2297.

[27] Loh, P., and M. Wainwright (2012). "High-dimensional Regression with Noisy and Missing data: Provable Guarantees with Non-convexity". Preprint: arXiv:1109.3714v1.

[28] Lounici, K. (2008). "Sup-Norm Convergence Rate and Sign Concentration Property of the Lasso and Dantzig Selector". *Electronic Journal of Statistics*, 2, 90-102.

[29] Meinshausen, N., and B. Yu (2009). "Lasso-type Recovery of Sparse Representations for High-dimensional Data". *The Annals of Statistics*, 37(1): 246-270.

[30] Negahban, S., P. Ravikumar, M. J. Wainwright, and B. Yu (2009). "A Unified Framework for the Analysis of Regularized M-estimators". *In Advances in Neural Information Processing Systems.*

[31] Nevo, A. (2001). "Measuring Market Power in the Ready-to-Eat Cereal Industry". *Econometrica*, 69, 307-342.

[32] Ravikumar, P., H. Liu, J. Lafferty, and L. Wasserman (2010). "SpAM: Sparse Additive Models". *Journal of the Royal Statistical Society, Series B.* To appear.

[33] 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(3): 1287-1319.

[34] Raskutti, G., M. J. Wainwright, and B. Yu (2010). "Restricted Eigenvalue Conditions for Correlated Gaussian Designs". *Journal of Machine Learning Research*, 11: 2241-2259.

[35] 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(10): 6976-6994.

[36] Rigollet, P., and A. B. Tsybakov (2011). "Exponential Screening and Optimal Rates of Sparse Estimation". *The Annals of Statistics*, 35, 731-771.

[37] Rosenbaum, M., and A. B. Tsybakov (2010). "Sparse Recovery Under Matrix Uncertainty". *The Annals of Statistics*, 38, 2620-2651.

[38] M. Rudelson, and S. Zhou. "Reconstruction from Anisotropic Random Measurements". Technical report, University of Michigan, July 2011.

[39] Sala-i-Martin, X. (1997). "I Just Ran Two Million Regressions". *The American Economic Review*, 87, 178-183.

[40] Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso". *Journal of the Royal Statistical Society*, Series B, 58(1): 267-288.

[41] Vershynin, R. "Introduction to the Non-asymptotic Analysis of Random Matrices". *Compressed Sensing: Theory and Applications*, To appear. Available at http:// www-personal.umich.edu/~romanv/papers/non-asymptotic-rmt-plain.pdf.

[42] Wainwright. 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.

[43] Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge.

[44] 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.

[45] Zhao, P., and Yu, B. (2007). On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7, 2541-2567.