Sparse Linear Models and l1Regularized 2SLS with High-Dimensional Endogenous Regressors and Instruments

Zhu, Ying

University of California, Berkeley

20 July 2015

Sparse Linear Models and l_1-Regularized 2SLS with High-Dimensional Endogenous Regressors and Instruments*

Ying Zhu
This revision: July, 2015 (First draft: August, 2013)

yingzhu@berkeley.edu. Tel: 406-465-0498.

Abstract

We explore the validity of the 2-stage least squares estimator with l_1-regularization in both stages, for linear models where the numbers of endogenous regressors in the main equation and instruments in the first-stage equations can exceed the sample size, and the regression coefficients belong to l_q- "balls" for $q \in [0, 1]$, covering both exact and approximate sparsity cases. Standard high-level assumptions on the Gram matrix for l_2-consistency require careful verifications in the two-stage procedure, for which we provide detailed analysis. We establish finite-sample bounds and conditions for our estimator to achieve l_2-consistency and variable-selection consistency. Practical guidance for choosing the regularization parameters is provided.

JEL Classification: C13, C31, C36
Keywords: High-dimensional statistics; Lasso; sparse linear models; endogeneity; two-stage estimation

1 Introduction

The objective of this paper is consistent estimation and selection of regression coefficients in models with a large number of endogenous regressors. We consider the linear model

$$Y_i = X_i^T \beta^* + \epsilon_i = \sum_{j=1}^{p} X_{ij}\beta_j^* + \epsilon_i, \quad i = 1, ..., n \tag{1}$$

where ϵ_i is a zero-mean random error possibly correlated with X_i and β^* is an unknown vector of parameters of our main interests. The j^{th} component of β^* is denoted by β_j^*. The j^{th} component of X_i is *endogenous* if $\mathbb{E}(X_{ij}\epsilon_i) \neq 0$ and *exogenous* if $\mathbb{E}(X_{ij}\epsilon_i) = 0$. Without loss of generality, we will assume all regressors are endogenous throughout the rest of this paper for notational convenience (a modification to allow mix of endogenous and exogenous regressors is straightforward.). When

*I thank James Powell, Martin Wainwright, and Demian Pouzo for useful discussions and comments. I am also grateful to the editor Jianqing Fan, the AE, and the anonymous referees for detailed feedback and suggested improvement on this paper. All errors are my own. This work was financially supported by Haas School of Business at UC Berkeley.

endogenous regressors are present, the classical least squares estimator will be inconsistent for β^* (i.e., $\hat{\beta}_{OLS} \overset{p}{\nrightarrow} \beta^*$) even when the dimension p of β^* is small relative to the sample size n. The two-stage least squares (2SLS) estimation plays an important role in accounting for endogeneity that comes from individual choice or market equilibrium (e.g., Wooldridge, 2002), and is based on the following "first-stage" equations for the components of X_i,

$$X_{ij} = Z_{ij}^T \pi_j^* + \eta_{ij} = \sum_{l=1}^{d_j} Z_{ijl} \pi_{jl}^* + \eta_{ij}, \qquad i = 1, ..., n, \ j = 1, ..., p. \tag{2}$$

For each $j = 1, ..., p$, Z_{ij} is a $d_j \times 1$ vector of instrumental variables, and η_{ij} a zero-mean random error which is uncorrelated with Z_{ij}, and π_j^* is an unknown vector of nuisance parameters. We will refer to the equation in (1) as the main equation (or second-stage equation) and the equations in (2) as the first-stage equations. Without loss of generality, the assumption $\mathbb{E}(Z_{ij}\epsilon_i) = \mathbb{E}(Z_{ij}\eta_{ij}) = \mathbf{0}$ for all $j = 1, ..., p$ and $\mathbb{E}(Z_{ij}\eta_{ij'}) = \mathbf{0}$ for all $j \neq j^{'}$ implies a triangular simultaneous equations model structure.

High dimensionality arises in (1) and (2) when the dimension p of β^* is large relative to the sample size n (namely, $p \asymp n$ or even $p \gg n$) or when the dimension d_j of π_j^* is large relative to the sample size n (namely, $d_j \asymp n$ or $d_j \gg n$) for at least one j. This paper concerns the case where $p \gg n$ and $d_j \ll n$, or the case where $p \gg n$ and $d_j \gg n$, and β^* and π_j^* (for $j = 1, ..., p$) are "sparse" in a way to be defined in Section 2. The analysis for the case $p \asymp n$ or $p \gg n$ is useful, for example, when we have the model $Y_i = f(X_i) + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, $\mathbb{E}(\epsilon_i|X_i) \neq 0$ for all i, and $f(\cdot)$ is an unknown function of interest and can be approximated by linear combinations of some set of basis functions, i.e., $f(X_i) = \sum_{j=1}^{p} \beta_j \phi_j(X_i)$.

An empirical example of the case $p \asymp n$ or $p \gg n$ concerns the estimation of network or community influence. For example, Manresa (2014) looks at how a firm's production output is influenced by the investment of other firms. As a future extension, she suggests an alternative model that looks at the network influence in terms of the output of the other firms rather than their investment:

$$Y_{it} = \alpha_i + X_{it}^T \theta + \sum_{j \in \{1,...,n\},\, j \neq i} \beta_{ji} Y_{jt} + \epsilon_{it}$$

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

In the literature on high-dimensional sparse linear regression models, a great deal of attention has been given to the l_1-penalized least squares. In particular, the Lasso is the most studied technique (see, e.g., Tibshirani, 1996; Candès and Tao, 2007; Bickel, Ritov, and Tsybakov, 2009; Belloni, Chernozhukov, and Wang, 2011; Belloni and Chernozhukov, 2011b; Loh and Wainwright, 2012; etc.). Variable selection when the dimension of the problem is larger than the sample size has also been studied in the likelihood method setting with penalty functions other than the l_1-norm (see, e.g., Fan and Li, 2001; Fan and Lv, 2011; Fan and Liao, 2014). Lecture notes by Koltchinskii (2011), as well as recent books by Bühlmann and van de Geer (2011) and Wainwright (2015) have given a more comprehensive introduction to high-dimensional statistics.

Recently, these l_1-penalized techniques have been applied in a number of econometrics papers. Caner (2009) studies a Lasso-type GMM estimator. Rosenbaum and Tsybakov (2010) study the high-dimensional errors-in-variables problem where the non-random regressors are observed with additive error and they present an application to hedge fund portfolio replication. Belloni, Chen, Chernozhukov, and Hansen (2012) estimate the optimal instruments using the Lasso and in an empirical example dealing with the effect of judicial eminent domain decisions on economic outcomes, they find the Lasso-based instrumental variable estimator outperforms an intuitive benchmark. Fan, Lv, and Li (2011) review the literature on sparse high-dimensional econometric models and also cover other regularization methods for several models including the vector autoregressive model for measuring the effects of monetary policy, panel data model for forecasting home price, and volatility matrix estimation in finance.

For the triangular simultaneous equations structure (1) and (2), the case where $d_j \gg n$ for at least one j but $p \ll n$ has been 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. In the case where $p \gg n$ and $d_j \ll n$ for all j, we can obtain the fitted regressors by a standard least squares estimation on each of the first-stage equations separately as usual and then apply the Lasso using these fitted regressors in the second-stage regression. Similarly, in the case where $p \gg n$ and $d_j \gg n$ for all j, we can obtain the fitted regressors by performing a regression with the Lasso on each of the first-stage equations separately and then apply another Lasso estimation using these fitted regressors in the second-stage.

Compared to existing 2SLS techniques which either limit the number of regressors entering the first-stage equations or the second-stage equation or both, our two-stage estimation procedures with l_1-regularization in both stages are more flexible and particularly powerful for applications in which the vector of parameters of interests is sparse and there is lack of information about the relevant explanatory variables and instruments. In terms of implementations, our high-dimensional 2SLS procedures are intuitive and can be easily implemented using built-in routines in software packages (e.g., *matlab* and *R*) for the standard Lasso estimation of linear models without endogeneity. We also provide practical guidance for choosing the regularization parameters. As we will see in Section 3, the complex structure of (1) and (2) and the nature of our regularized 2-stage least squares type estimation render existing adaptive methods (e.g., Antoniadis, 2010; Sun and Zhang, 2010, 2012; Belloni, et al., 2011; Gautier and Tsybakov, 2014; etc.) for setting the second-stage regularization parameter less useful. Instead, we recommend the model-free ESCV ("Estimation Stability and Cross Validation") criterion proposed by Lim and Yu (2013) and applied in Yu (2013). Using the estimates from the ESCV procedure, we also propose an alternative "plug-in" method for choosing the second-stage regularization parameter, which in practice may be compared with the optimal regularization parameter chosen by the ESCV criterion to determine whether the amount of penalty is sufficient.

In terms of analyzing the statistical properties, the extension from models with a few endogenous regressors to models with many endogenous regressors ($p \gg n$) in the context of triangular simultaneous equations (1) and (2) for the two-stage estimation is not obvious. This paper aims to explore the validity of these two-step estimators in the high-dimensional sparse setting. Another contribution of this paper is to introduce analysis that is suitable for showing estimation consistency of the two-step type high-dimensional estimators. When endogeneity is absent from model (1), there is a well-developed theory on what conditions on the design matrix $X \in \mathbb{R}^{n\times p}$ are sufficient for an l_1-regularized estimator to consistently estimate β^*. In some situations one can impose these conditions directly as an assumption on the underlying design matrix. However, when employing a regularized 2SLS estimator in the context of triangular simultaneous linear equation models in the high-dimensional setting, namely, (1) and (2), there is no guarantee that the random matrix $\frac{\hat{X}^T\hat{X}}{n}$ (with $\hat{X}$ obtained from regressing X on the instrumental variables) would automatically satisfy these previously established conditions for estimation consistency. To the best of our knowledge, previous literature has not dealt with this issue. This paper explicitly shows that these conditions indeed hold for $\frac{\hat{X}^T\hat{X}}{n}$ with high probability under appropriate conditions. It also establishes the sample size required for $\frac{\hat{X}^T\hat{X}}{n}$ to satisfy these conditions.

We begin in Section 2 with model assumptions imposed on (1) and (2). The high-dimensional 2SLS procedure and its theoretical properties are established in Section 3, where practical guidance for choosing the regularization parameter is also provided. Section 4 presents simulation results and compares the various practical choices of the regularization parameters. Section 5 concludes this paper and discusses future extensions. The main proofs are collected in Appendices A and B. Additional supplementary materials are included in:
https://sites.google.com/site/yingzhu1215/home/HD2SLS_Supplement.pdf.

Notation. For the convenience of the reader, we 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 $l_\infty-$norm of A. The l_2-operator norm, or spectral norm of the matrix A corresponds to its maximum singular value: it is defined as $||A||_2 := \sup_{v\in S^{m-1}} |Av|_2$, where $S^{m-1} = \{v \in \mathbb{R}^m \,|\, |v|_2 = 1\}$. The l_∞ matrix norm (maximum absolute row sum) of A is denoted by $||A||_\infty := \max_i \sum_j |a_{ij}|$ (note the difference between $|A|_\infty$ and $||A||_\infty$). For a square matrix A, denote its minimum eigenvalue and maximum eigenvalue by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$, respectively. For functions $f(n)$ and $g(n)$, write $f(n) \succsim g(n)$ to mean that $f(n) \geq cg(n)$ for a universal constant $c \in (0, \infty)$ and similarly, $f(n) \precsim g(n)$ to mean that $f(n) \leq c^{'}g(n)$ for a universal constant $c^{'} \in (0, \infty)$; $f(n) \asymp g(n)$ when $f(n) \succsim g(n)$ and $f(n) \precsim g(n)$ hold simultaneously. For some integer $s \in \{1, 2, ..., m\}$, the l_0-ball of "radius" s is given by $\mathbb{B}_0^m(s) := \{v \in \mathbb{R}^m \,|\, |v|_0 \leq s\}$ where $|v|_0 := \sum_{i=1}^m 1\{v_i \neq 0\}$. Similarly, the l_2-ball of radius r is given by $\mathbb{B}_2^m(r) := \{v \in \mathbb{R}^m \,|\, |v|_2 \leq r\}$. Also, write $\mathbb{K}(s, m, r) := \mathbb{B}_0^m(s) \cap \mathbb{B}_2^m(r)$ and $\mathbb{K}^2(s, m, r) := \mathbb{K}(s, m, r) \times \mathbb{K}(s, m, r)$. 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 $|J|$. Denote $\max\{a, b\}$ by $a \vee b$ and $\min\{a, b\}$ by $a \wedge b$. As a general rule for the proofs, c constants denote generic positive constants that are independent of n, p, d, R_{q_2}, R_{q_1}, and may change from place to place.

2 Model assumptions

Throughout the rest of this paper, the following assumptions are imposed on the model (1) and (2).

Assumption 2.1: The data $\{Y_i,\, X_i,\, Z_i\}_{i=1}^n$ are independent with finite second moments; for all $j = 1, ..., p$ and $i = 1, ..., n$, $\mathbb{E}(Z_{ij}\epsilon_i) = \mathbb{E}(Z_{ij}\eta_{ij}) = \mathbf{0}$ and $\mathbb{E}(Z_{ij}\eta_{ij'}) = \mathbf{0}$ for all $j \neq j'$.

Assumption 2.2 (Sparsity): The coefficient vector $\beta^* \in \mathbb{R}^p$ belongs to the $l_{q_2}-$"balls" $\mathcal{B}^p_{q_2}(R_{q_2})$ for a "radius" of R_{q_2} and some $q_2 \in [0,\, 1]$, where the l_q-"balls" of "radius" R for $q \in [0,\, 1]$ are defined by

$$\begin{aligned} \mathcal{B}^p_q(R) &:= \left\{ \beta \in \mathbb{R}^p \mid |\beta|_q^q = \sum_{j=1}^p |\beta_j|^q \leq R \right\} \text{ for } q \in (0,\, 1], \\ \mathcal{B}^p_0(R) &:= \left\{ \beta \in \mathbb{R}^p \mid |\beta|_0 = \sum_{j=1}^p 1\{\beta_j \neq 0\} \leq R \right\} \text{ for } q = 0. \end{aligned}$$

For $j = 1, ..., p$, the coefficient vector $\pi_j^* \in \mathbb{R}^{d_j}$ belongs to the $l_{q_{1j}}-$"balls" $\mathcal{B}^{d_j}_{q_{1j}}(R_{q_{1j}})$ for a "radius" of $R_{q_{1j}}$ and some $q_{1j} \in [0,\, 1]$, where $\mathcal{B}^{d_j}_{q_{1j}}(R_{q_{1j}})$ is defined in a similar fashion as above. For notational simplicity, $d_j = d$, $q_{1j} = q_1$, and $R_{q_{1j}} = R_{q_1}$ for all $j = 1, ..., p$.

Remark. Assumption 2.2 requires the coefficient vectors to be "sparse" and formalizes the sparsity condition by considering the l_q-"balls" $\mathcal{B}^p_q(R_q)$ of "radius" R_q where $q \in [0,\, 1]$ (see, e.g., Ye and Zhang, 2010; Raskutti, Wainwright, and Yu, 2011; Negahban, Ravikumar, Wainwright, and Yu, 2012; this notion is also used for the Bridge estimator considered in Huang, Horowitz, and Ma, 2008). For example, the exact sparsity on β^* corresponds to the case of $q = q_2 = 0$ with $R_{q_2} = k_2$, which says that β^* has at most k_2 non-zero components. In the more general setting $q_2 \in (0,\, 1]$, membership in $\mathcal{B}^p_{q_2}(R_{q_2})$ has various interpretations and one of them involves how quickly the ordered coefficients decay according to the hyperharmonic series. When $q_2 \in [0,\, 1)$, the set $\mathcal{B}^p_{q_2}(R_{q_2})$ is non-convex and the l_1-ball is the *closest convex* approximation of these non-convex sets. In terms of estimation procedure design, the idea of approximating non-convex problems with their closest convex member (so called "convex relaxation") as in the Lasso provides a tremendous computational advantage. In the rest of our analysis, we set the "radius" $R_{q_2} = \sum_{j=1}^p \left|\beta_j^*\right|^{q_2}$ when $q_2 \in (0,\, 1]$ and $R_{q_2} = k_2$ when $q_2 = 0$. The growth conditions on $(n,\, d,\, p,\, R_{q_1},\, R_{q_2})$ will be specified in Sections 3.1 and 3.2 when theoretical results are presented.

Assumption 2.3 (Restricted Identifiability): For a subset $S \subseteq \{1,\, 2, ..., p\}$ and all non-zero $\Delta \in \mathbb{C}(S;\, q_2,\, c^*) \cap \mathbb{S}_\delta$ where

$$\mathbb{C}(S;\, q_2,\, c^*) := \left\{\Delta \in \mathbb{R}^p \,:\, |\Delta_{S^c}|_1 \leq c^*|\Delta_S|_1 + (c^*+1)|\beta^*_{S^c}|_1\right\},$$

for some universal constant $c^* > 1$ (with Δ_S denoting the vector in $\mathbb{R}^p$ that has the same coordinates as Δ on S and zero coordinates on the complement S^c of S), and

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

the matrix $\Sigma_{X^*} = \mathbb{E}\left[\frac{X^{*T}X^*}{n}\right]$ satisfies

$$\frac{\Delta^T \Sigma_{X^*} \Delta}{|\Delta|_2^2} \geq \underline{\kappa}_2 > 0,$$

with parameters $(q_2,\, \delta,\, \underline{\kappa}_2)$, where $X^* := (Z_1\pi_1^*, ..., Z_j\pi_j^*, ..., Z_p\pi_p^*)$. For $j = 1, ..., p$, the matrix $\Sigma_{Z_j} = \mathbb{E}\left[\frac{Z_j^T Z_j}{n}\right]$ satisfies a similar restricted eigenvalue condition with parameters $(q_1,\, \delta_j,\, \underline{\kappa}_1)$ for a subset $S_j \subseteq \{1,\, 2, ..., d\}$. The choices of δ, δ_j, and S, S_j will be specified in Section 3.1.

Remarks. The following discussion is in regard to the RE condition on $\mathbb{E}\left[\frac{X^{*T}X^*}{n}\right]$ imposed by Assumption 2.3 (similar argument can be made for $\mathbb{E}\left[\frac{Z_j^T Z_j}{n}\right]$). When β^* is exactly sparse (namely, $q_2 = 0$), we can take $\delta = 0$ and choose $S = J(\beta^*)$ (recalling $J(\beta^*)$ denotes the support of β^*), which reduces the set $\mathbb{C}(S;\, q_2,\, c^*) \cap \mathbb{S}_\delta$ to the following cone:

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

Let us first consider a simple case where X^* is observed. The sample analog of Assumption 2.3 over the cone $\mathbb{C}(J(\beta^*);\, 0,\, c^*)$ is the so-called *restricted eigenvalue* (RE) condition on the Gram matrix $\frac{X^{*T}X^*}{n}$, studied in Bickel, et. al. (2009), Meinshausen and Yu (2009), Raskutti, et al. (2010), Bühlmann and van de Geer (2011), Loh and Wainwright (2012), Negahban, et. al. (2012), etc.

When β^* is approximately sparse (namely, $q_2 \in (0,\, 1]$), in sharp contrast to the exact sparsity case, the set $\mathbb{C}(S;\, q_2,\, c^*)$ is no longer a cone but rather contains a ball centered at the origin. Consequently, it is never possible to ensure that $\frac{|X^*\Delta|_2^2}{n}$ is bounded from below for all vectors Δ in the set $\mathbb{C}(S;\, q_2,\, c^*)$ (see Negahban, et. al., 2012 for a geometric illustration of this issue). Therefore, in order to obtain a general applicable theory, it is crucial to further restrict the set $\mathbb{C}(S;\, q_2,\, c^*)$ for $q_2 \in (0,\, 1]$ by intersecting it with the set $\mathbb{S}_\delta := \{\Delta \in \mathbb{R}^p \, : \, |\Delta|_2 \geq \delta\}$. Provided the parameter δ and the set S are properly defined, the intersection $\mathbb{C}(S;\, q_2,\, c^*) \cap \mathbb{S}_\delta$ excludes many "flat" directions (with eigenvalues of 0) in the space for the case of $q_2 \in (0,\, 1]$. To the best of our knowledge, the necessity of this additional set $\mathbb{S}_\delta$, essential for the approximately sparse case of $q_2 \in (0,\, 1]$, is first recognized explicitly in Negahban, et. al. (2012). We use this idea to derive a general upper bound on the l_2-error of the high-dimensional 2SLS estimator when β^* and π_j^* $(j = 1, ..., p)$ satisfy Assumption 2.2, which covers a spectrum of sparsity cases (exact and approximate).

In our problem, X^* is unknown and needs to be estimated. When applying the l_1-regularized 2SLS procedure to estimate β^*, there is no guarantee that the random matrix $\frac{\hat{X}^T\hat{X}}{n}$ (where $\hat{X}$ is the estimate of $X^* = \left[Z_1\pi_1^*,\, ...,\, Z_p\pi_p^*\right]$) would automatically satisfy these previously established conditions for estimation consistency. This paper provides results that imply the RE condition holds for $\frac{\hat{X}^T\hat{X}}{n}$ with high probability provided Assumption 2.3 is satisfied for a sub-Gaussian matrix X^*. Verifications of the RE condition provide finite-sample guarantees of Assumption 2.3 when the unknown X^* is replaced with its estimate $\hat{X}$ and the expectation is replaced with a sample average.

3 High-dimensional 2SLS estimation

For notational simplicity, in the main theoretical results presented below, we assume the regime of interest is $p \geq n$. The modification to allow $p < n$ is trivial. For the first-stage regression, we consider the following procedure:

$$\hat{\pi}_j \in \operatorname{argmin}_{\pi_j \in \mathbb{R}^d} \frac{1}{2n} |X_j - Z_j \pi_j|_2^2 + \lambda_{n,j} \sum_{l=1}^{d} \hat{\sigma}_{Z_{jl}} |\pi_{jl}| \tag{3}$$

for $j = 1, ..., p$ and $l = 1, ..., d$, where $\hat{\sigma}_{Z_{jl}} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} Z_{ijl}^2}$. Denote the fitted regressors using the first-stage estimates by $\hat{X}_j := Z_j \hat{\pi}_j$ for $j = 1, ..., p$, and $\hat{X} = \left(\hat{X}_1, ..., \hat{X}_p\right)$. For the second-stage regression, we consider

$$\hat{\beta}_{H2SLS} \in \operatorname{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{2n} |Y - \hat{X}\beta|_2^2 + \lambda_n \sum_{j=1}^{p} \hat{\sigma}_{X_j^*} |\beta_j|, \tag{4}$$

where $\hat{\sigma}_{X_j^*} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \hat{X}_{ij}^2}$ for $j = 1, ..., p$.

Remark. Upon solving (3), post-Lasso strategies such as thresholding or post-OLS-Lasso (which performs an OLS with the regressors in the estimated support set $J(\hat{\pi}_j)$ to obtain $\hat{\pi}_j^{OLS}$ for $j = 1, ..., p$) may be used before (4). In the third step, we apply the Lasso to estimate the main equation parameters with these fitted regressors based on the second-stage post-Lasso estimates. This type of procedure is in the similar spirit as the those in literature (see, e.g., Candès and Tao, 2007; Belloni and Chernozhukov, 2013).

We begin with Sections 3.1 and 3.2 by emphasizing the theoretical guarantees on parameter estimation and variable selection of $\hat{\beta}_{H2SLS}$, respectively. Note that these two sections do not deal with practical guidance for choosing the regularization parameters, which is the focus of Section 3.3, where we discuss two existing model-free criteria in literature for regularized estimation and then propose feasible counterparts of the theoretical choices of the regularization parameters from Section 3.1. In the simulation experiments (Section 4), we compare the various practical choices of the regularization parameters provided in Section 3.3.

3.1 Theoretical guarantees on the estimation of parameters

The first main result (Theorem 3.1) exhibits the non-asymptotic bound for $|\hat{\beta}_{H2SLS} - \beta^*|_2$, which establishes sufficient conditions for l_2-consistency of $\hat{\beta}_{H2SLS}$. This result requires some regularity conditions, which use the following definition of sub-Gaussian matrices based on Vershynin (2012) and similar to Loh and Wainwright (2012).

Definition 3.1 (Sub-Gaussian variables and matrices). A random variable X with mean $\mu = \mathbb{E}[X]$ is sub-Gaussian if there is a positive number ρ such that $\sup_{\gamma \geq 1} \gamma^{-\frac{1}{2}} \left(\mathbb{E}|X|^{\gamma}\right)^{\frac{1}{\gamma}} \leq \rho$; a random matrix $A \in \mathbb{R}^{n\times p}$ is sub-Gaussian with parameters $(\Sigma_A,\, \rho_A^2)$ where $\Sigma_A = \mathbb{E}\left[\frac{A^TA}{n}\right]$, if each row $A_i \in \mathbb{R}^p$ is sampled independently from a distribution, and for any unit vector $u \in \mathbb{R}^p$, the random variable $u^TA_i^T$ is sub-Gaussian with parameter at most ρ_A^2.

Remark. The sub-Gaussian assumption says that the variables need to be drawn from distributions with well-behaved tails like Gaussian. In contrast to the Gaussian assumption, sub-Gaussian variables constitute a more general family of distributions. In particular, one can show that $\rho = C\sigma = C\sqrt{\mathbb{E}[X^2]}$ when X is a zero-mean Gaussian random variable, and $\rho = C\frac{\overline{a}-\underline{a}}{2}$ when X is a zero-mean random variable supported on some interval $[\underline{a},\, \overline{a}]$, where $C > 0$ is a universal constant (see, e.g., Wainwright, 2015).

Assumption 3.1: The error terms ϵ and η_j for $j = 1, ..., p$ are zero-mean sub-Gaussian vectors with parameters ρ_ϵ^2 and $\rho_{\eta_j}^2$, respectively; $\rho_\eta^2 = \max_j \rho_{\eta_j}^2$. The random matrix $Z_j \in \mathbb{R}^{n\times d}$ is sub-Gaussian with parameters $(\Sigma_{Z_j},\, \rho_Z^2)$ for $j = 1, ..., p$.

Assumption 3.2: For every $j = 1, ..., p$, $X_j^* := Z_j\pi_j^*$. The matrix $X^* \in \mathbb{R}^{n\times p}$ is sub-Gaussian with parameters $(\Sigma_{X^*},\, \rho_{X^*}^2)$ where the jth column of X^* is X_j^*.

Remark. Assumptions 3.1 and 3.2 are common in the literature (see, e.g., Loh and Wainwright, 2012; Negahban, et. al 2012; Rosenbaum and Tsybakov, 2013). In fact, the second part of Assumption 3.1 on $Z_j \in \mathbb{R}^{n\times d}$ being sub-Gaussian for all j implies that $Z_j\pi_j^* = X_j^*$ is also sub-Gaussian. Therefore, the conditions that $X^* \in \mathbb{R}^{n\times p}$ is a sub-Gaussian matrix with parameters $(\Sigma_{X^*},\, \rho_{X^*}^2)$ where the jth column of X^* is X_j^*(Assumption 3.2) is a mild extension.

To state the following results, we need to introduce some definitions. First, Let

$$\mathcal{T}_0 = \max\left\{|\beta^*|_1\mathcal{T}_1,\; \rho_{X^*}\rho_\eta|\beta^*|_1\mathcal{T}_2,\; \rho_{X^*}\rho_\epsilon\mathcal{T}_2\right\}, \tag{5}$$

$$\mathcal{T}_1 = c_1\frac{\bar{\kappa}_1^{\frac{1}{2}}R_{q_1}^{\frac{1}{2}}}{\underline{\kappa}_1^{1-\frac{q_1}{2}}}\left(\sqrt{\rho_Z^2\rho_\eta^2\frac{\log(d\vee p)}{n}}\right)^{1-\frac{q_1}{2}}, \tag{6}$$

$$\mathcal{T}_2 = c_2\sqrt{\frac{\log p}{n}}. \tag{7}$$

We postpone the discussion of a practical procedure for setting the unknown parameters and constants in $\mathcal{T}_0$ until Section 3.3.

Also, recall in Section 2 the sets we introduced,

$$\begin{aligned}\mathbb{C}(S;\, q_2,\, c^*) &:= \left\{\Delta \in \mathbb{R}^p \,:\, |\Delta_{S^c}|_1 \le c^*|\Delta_S|_1 + (c^*+1)|\beta_{S^c}^*|_1\right\},\\ \mathbb{C}(S_j;\, q_1,\, c^*) &:= \left\{\Delta \in \mathbb{R}^d \,:\, |\Delta_{S_j^c}|_1 \le c^*|\Delta_{S_j}|_1 + (c^*+1)|\pi_{j,S_j^c}^*|_1\right\},\end{aligned}$$

for $j = 1, ..., p$, and some universal constant $c^* > 1$, and the spherical sets

$$\begin{aligned}\mathbb{S}_\delta : &= \left\{\Delta \in \mathbb{R}^p \,:\, |\Delta|_2 \ge \delta\right\},\\ \mathbb{S}_{\delta_j} : &= \left\{\Delta \in \mathbb{R}^d \,:\, |\Delta|_2 \ge \delta_j\right\},\end{aligned}$$

and the intersections $\mathbb{C}(S;\,q_2,\,c^*)\cap\mathbb{S}_\delta$ and $\mathbb{C}(S_j;\,q_1,\,c^*)\cap\mathbb{S}_{\delta_j}$. When β^* and π_j^* are approximately sparse (namely, $q_2,\,q_1\in(0,\,1]$), we choose S in $\mathbb{C}(S;\,q_2,\,c^*)$ and S_j in $\mathbb{C}(S_j;\,q_1,\,c^*)$ to be the following subsets

$$\begin{aligned} S_{\underline{\tau}} &:= \left\{j\in\{1,\,2,\,...,\,p\}\,:\,\left|\beta_j^*\right|>\underline{\tau}\right\}, \\ S_{\underline{\tau}_j} &:= \left\{l\in\{1,\,2,\,...,\,d\}\,:\,\left|\pi_{jl}^*\right|>\underline{\tau}_j\right\}, \end{aligned}$$

with the parameter $\underline{\tau}=\frac{c^*+1}{c^*-1}\frac{\mathcal{T}_0}{\underline{\kappa}_2}$ and $\underline{\tau}_j=c_0\frac{\sqrt{\rho_Z^2\rho_\eta^2\frac{\log(d\vee p)}{n}}}{\underline{\kappa}_1}$, respectively (recall the parameter $\underline{\kappa}_1$ and $\underline{\kappa}_2$ defined in Assumption 2.3, Section 2). When β^* and π_j^* are exactly sparse (namely, $q_2,\,q_1=0$), we set $\delta=\delta_j=\underline{\tau}=\underline{\tau}_j=0$ and choose $S=J(\beta^*)$, $S_j=J(\pi_j^*)$, which reduces the sets $\mathbb{C}(S;\,q_2,\,c^*)\cap\mathbb{S}_\delta$ and $\mathbb{C}(S_j;\,q_1,\,c^*)\cap\mathbb{S}_{\delta_j}$, respectively, to the following cones:

$$\begin{aligned} \mathbb{C}(J(\beta^*);\,0,\,c^*) &:= \left\{\Delta\in\mathbb{R}^p\,:\,|\Delta_{J(\beta^*)^c}|_1\le c^*|\Delta_{J(\beta^*)}|_1\right\}, \\ \mathbb{C}(J(\pi_j^*);\,0,\,c^*) &:= \left\{\Delta\in\mathbb{R}^d\,:\,|\Delta_{J(\pi_j^*)^c}|_1\le c^*|\Delta_{J(\pi_j^*)}|_1\right\}. \end{aligned}$$

The first main theorem provides an upper bound on $\left|\hat{\beta}_{H2SLS}-\beta^*\right|_2$ when the first- and second-stage estimations concern the programs in (3) and (4), respectively. This result concerns the case where $p\ge n$, $d\ge n$, and β^* and π_j^* (for $j=1,...,p$) satisfy Assumption 2.2. Before presenting the main theorem, we provide the following lemma to ensure that the regressors are well-behaved.

Lemma 3.1: If $\{Z_i\}_{i=1}^n$ are independent with finite second moment $\sigma_{Z_{jl}}^2=\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n Z_{ijl}^2\right)$ for $j=1,...,p$ and $l=1,...,d$, then,

$$\mathbb{P}\left(\max_{j,l}\left|\hat{\sigma}_{Z_{jl}}-\sigma_{Z_{jl}}\right|\le\frac{1}{2}\sigma_Z\right)\ge 1-O\left(\exp(-n)\right),$$

where $\hat{\sigma}_{Z_{jl}}^2=\frac{1}{n}\sum_{i=1}^n Z_{ijl}^2$ and $\sigma_Z^2=\max_{j,l}\sigma_{Z_{jl}}^2$. Furthermore, suppose Assumptions 2.1, 3.1, 3.2, and the part related to the first-stage equations in Assumption 2.2 hold. For $j=1,...,p$ and some universal constant $c^*>1$, let Assumption 2.3 hold over the restricted sets $\mathbb{C}(J(\pi_j^*);\,0,\,c^*)$ for the exact sparsity case $q_1=0$ with $R_{q_1}=k_1$, and over $\mathbb{C}(S_{\underline{\tau}_j};\,q_1,\,c^*)\cap\mathbb{S}_{\delta_j}$, where $\delta_j=c^{'}\underline{\kappa}_1^{-1+\frac{q_1}{2}}R_{q_1}^{\frac{1}{2}}\left(\sqrt{\frac{\rho_Z^2\rho_\eta^2\log(d\vee p)}{n}}\right)^{1-\frac{q_1}{2}}$ (for a sufficiently small constant $c^{'}>0$) and $\underline{\tau}_j=c_0\frac{\sqrt{\rho_Z^2\rho_\eta^2\frac{\log(d\vee p)}{n}}}{\underline{\kappa}_1}$, for the approximate sparsity case ($q_1\in(0,\,1]$). Also, for all vectors Δ in these restricted sets, $\frac{\Delta^T\Sigma_{Z_j}\Delta}{|\Delta|_2^2}\le\bar{\kappa}_1$, for $j=1,...,p$. If $n\ge c^{''}R_{q_1}^{\frac{2}{2-q_1}}\log(d\vee p)$ for some sufficiently large constant $c^{''}>0$ that depends on $\underline{\kappa}_1$, and the first-stage regularization parameters $\lambda_{n,j}$ satisfy

$$\lambda_{n,j}=c_0\sqrt{\rho_Z^2\rho_\eta^2\frac{\log(d\vee p)}{n}},\tag{8}$$

for all $j=1,...,p$, then,

$$\mathbb{P}\left(\max_{j=1,...,p}\left|\hat{\sigma}_{X_j^*}^2-\sigma_{X_j^*}^2\right|\le\sigma_{X^*}\mathcal{T}_1\right)\ge 1-O\left(\frac{1}{d\vee p}\right),$$

where $\hat{\sigma}^2_{X^*_j} = \frac{1}{n}\sum_{i=1}^n \hat{X}^2_{ij}$, $\sigma^2_{X^*} = \max_j \sigma^2_{X^*_j}$, and $\sigma^2_{X^*_j} = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n X^{*2}_{ij}\right)$.

Remark. The first part of Lemma 3.1 is implied by Lemma B.1 and the second part is proved in Section A.2. We assume in the following that the regressors Z_j are normalized such that $\hat{\sigma}_{Z_{jl}} \leq 1$ ($j = 1, ..., p$ and $l = 1, ..., d$), $\sigma_Z = 1$, and $\hat{X}_j$ are normalized such that $\hat{\sigma}_{X^*} := \max_{j=1,...,p}\sqrt{\frac{1}{n}\sum_{i=1}^n \hat{X}^2_{ij}} \leq 1$, $\sigma_{X^*} = 1$, in Lemma 3.1.

Theorem 3.1: Let the first-stage regularization parameters $\lambda_{n,j}$ satisfy (8) for $j = 1, ..., p$, and the second-stage regularization parameter λ_n satisfies

$$\lambda_n = \frac{c^* + 1}{c^* - 1}\mathcal{T}_0 \tag{9}$$

for some universal constant $c^* > 1$, with $\mathcal{T}_0$ defined in (5). Suppose: (i) Assumptions 2.1, 2.2, 3.1, and 3.2 hold; (ii) Assumption 2.3 holds over the restricted sets $\mathbb{C}(J(\beta^*);\, 0,\, c^*)$, for the exact sparsity case $q_2 = 0$ with $R_{q_2} = k_2$, and over $\mathbb{C}(S_{\underline{\tau}};\, q_2,\, c^*) \cap \mathbb{S}_\delta$ where $\delta = c_3\underline{\kappa}_2^{-1+\frac{q_2}{2}} R_{q_2}^{\frac{1}{2}}\mathcal{T}_0^{1-\frac{q_2}{2}}$ and $\underline{\tau} = \frac{c^*+1}{c^*-1}\frac{\mathcal{T}_0}{\underline{\kappa}_2}$, for the approximate sparsity case ($q_2 \in (0, 1]$); (iii) Assumption 2.3 concerning the first-stage matrices $\Sigma_{Z_j} = \mathbb{E}\left[\frac{Z_j^T Z_j}{n}\right]$ for $j = 1, ..., p$ holds according to the specifications in Lemma 3.1; (iv) for all vectors Δ in the restricted sets subject to those defined in Lemma 3.1, $\frac{\Delta^T \Sigma_{Z_j}\Delta}{|\Delta|_2^2} \leq \bar{\kappa}_1$, for $j = 1, ..., p$; (v) for some constant $c_4 > 0$ that depends on $\underline{\kappa}_2$, the condition

$$c_4 R_{q_2}\underline{\tau}^{-q_2}\left(\frac{\log p}{n} \vee \mathcal{T}_1\right) \leq 1 \tag{10}$$

holds with $\mathcal{T}_1$ defined in (6). Then,

$$|\hat{\beta}_{H2SLS} - \beta^*|_2 \leq \frac{cR_{q_2}^{\frac{1}{2}}}{\underline{\kappa}_2^{1-\frac{q_2}{2}}}\mathcal{T}_0^{1-\frac{q_2}{2}} \tag{11}$$

with probability at least $1 - O\left(\frac{1}{p}\right)$, where $c > c_3 > 0$ are some universal constants.

Remarks

The proof for Theorem 3.1 is provided in Sections A.1-A.3. If $\frac{R_{q_2}^{\frac{1}{2}}}{\underline{\kappa}_2^{1-\frac{q_2}{2}}}\mathcal{T}_0^{1-\frac{q_2}{2}} \to 0$ as $n \to \infty$, then $\hat{\beta}_{H2SLS}$ is l_2-consistent for β^*. If η_{ij}'s, ϵ_i's, Z_{ijl}'s, and X^*_{ij}'s are independent Gaussian random variables, then $\rho_\eta = C\sigma_\eta = C\max_j\sqrt{\mathbb{E}[\frac{1}{n}\sum_{i=1}^n \eta^2_{ij}]}$, $\rho_\epsilon = C\sigma_\epsilon = C\sqrt{\mathbb{E}[\frac{1}{n}\sum_{i=1}^n \epsilon^2_i]}$, $\rho_Z = C\sigma_Z = C$, and $\rho_{X^*} = C\sigma_{X^*} = C$, where $C > 0$ is a universal constant. The term $\sqrt{\rho_Z^2\rho_\eta^2\frac{\log(d\vee p)}{n}}$ in (6), $\mathcal{T}_1$, as well as in (8), the condition for $\lambda_{n,j}$ (which contrasts with $\sqrt{\rho_Z^2\rho_\eta^2\frac{\log d}{n}}$ for the Lasso estimation in a single equation problem) comes from the union bound

$$\mathbb{P}\left(\max_{j=1,...,p}\left|\frac{1}{n}Z_j^T\eta_j\right|_\infty \leq t\right) \geq 1 - O\left(\exp\left(\left(\frac{-nt^2}{\rho_Z^2\rho_\eta^2} \wedge \frac{-nt}{\rho_Z\rho_\eta}\right) + \log d + \log p\right)\right),$$

by setting $t \asymp \sqrt{\frac{\rho_Z^2\rho_\eta^2\log(d\vee p)}{n}}$ to ensure the tail probability of the order $O\left(\frac{1}{d\vee p}\right)$ (the notation $|\frac{1}{n}Z_j^T\eta_j|_\infty := \max_{l=1,...,d}|\frac{1}{n}Z_{jl}^T\eta_j|$). So, we set the first-stage regularization parameters $\lambda_{n,j} = \frac{c^*+1}{c^*-1}t = c_0\sqrt{\frac{\rho_Z^2\rho_\eta^2\log(d\vee p)}{n}}$ for all $j = 1, ..., p$ to take into account the fact that there are p endogenous regressors in the main equation and hence, p regressions to perform in the first-stage. The term $\mathcal{T}_1$ in (6) provides a sharp upper bound on the first-stage prediction error

$$\max_{j=1,...,p}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Z_{ij}\hat{\pi}_j - Z_{ij}\pi_j^*)^2}$$

when π_j^* (for all $j = 1, ..., p$) satisfies a sparsity condition as in Assumption 2.2.

The factor $|\beta^*|_1$ that appears in the first two terms of (5) and therefore the choice of λ_n in (9), as well as the upper bound on $|\hat{\beta}_{H2SLS} - \beta^*|_2$, is related to the fact that the second-stage procedure (4) plugs in the first-stage estimates $\hat{X}_j = Z_j\hat{\pi}_j$ as the surrogate of the unknown $X_j^* = Z_j\pi_j^*$. Indeed, our simulation results suggest that the amount of regularization needed for (4) to perform well in both estimation and selection increases with $|\beta^*|_1$. Other surrogate-type Lasso estimators such as the ones in Rosenbaum and Tsybakov (2013) and Zhu (2014) also involve the factor $|\beta^*|_1$.

For the case of approximately sparse β^* with $q_2 \in (0, 1]$, the rate $\frac{cR_{q_2}^{\frac{1}{2}}}{\underline{\kappa}_2^{1-\frac{q_2}{2}}}\mathcal{T}_0^{1-\frac{q_2}{2}}$ in (11) can be interpreted as follows. Suppose only the top s_2 components of β^* in absolute values are estimated. The fast decay imposed by the $l_{q_2}-$ "balls" assumption on β^* implies that the remaining $p - s_2$ components have relatively smaller effects, so we can view the rate for $q_2 \in (0, 1]$ intuitively as one that would be achieved if we were to choose $k_2 = s_2 = \frac{R_{q_2}}{\underline{\kappa}_2^{-q_2}}\mathcal{T}_0^{-q_2}$ for an exactly sparse problem with $q_2 = 0$, which would yield the rate $\frac{c\sqrt{s_2}}{\underline{\kappa}_2}\mathcal{T}_0 = \frac{cR_{q_2}^{\frac{1}{2}}}{\underline{\kappa}_2^{1-\frac{q_2}{2}}}\mathcal{T}_0^{1-\frac{q_2}{2}}$.

With the conditions (in Theorem 3.1) imposed on the triangular structure (1) and (2), the upper bound (11) on $|\hat{\beta}_{H2SLS} - \beta^*|_2$ and the growth requirement (10) on $(n, d, p, R_{q_1}, R_{q_2})$ are sharp. Let us consider some simpler cases of Theorem 3.1. First, suppose $\rho_\eta = 0$ so the upper bound in Theorem 3.1 reduces to $|\hat{\beta}_{H2SLS} - \beta^*|_2 \leq \frac{cR_{q_2}^{\frac{1}{2}}}{\underline{\kappa}_2^{1-\frac{q_2}{2}}}\left(\sqrt{\frac{\rho_{X^*}^2\rho_\epsilon^2\log p}{n}}\right)^{1-\frac{q_2}{2}}$, which is the minimax-optimal rate of the Lasso for the usual high-dimensional linear regression model (1) with $\mathbb{E}(X_i\epsilon_i) = \mathbf{0}$ and β^* satisfies a sparsity condition as in Assumption 2.2 (see, Raskutti, Wainwright, and Yu, 2011). Moreover, if β^* is exactly sparse ($q_2 = 0$), then $|\hat{\beta}_{H2SLS} - \beta^*|_2 \leq \frac{c}{\underline{\kappa}_2}\left(\sqrt{\frac{\rho_{X^*}^2\rho_\epsilon^2 k_2\log p}{n}}\right)$, the well-known optimal rate of the Lasso for the usual exactly sparse high-dimensional linear regression model (1) with $\mathbb{E}(X_i\epsilon_i) = \mathbf{0}$.

Now, suppose $\rho_\eta \neq 0$, and β^*, π_j^* ($j = 1, ..., p$) are exactly sparse ($q_2 = q_1 = 0$). Theorem 3.1 implies that, if the second-stage regularization parameter λ_n satisfies $\lambda_n = \frac{c^*+1}{c^*-1}\mathcal{T}_0$ with $\mathcal{T}_0$ in (5) taking the following form

$$\mathcal{T}_0 = \max\left\{c_1|\beta^*|_1\frac{\bar{\kappa}_1^{\frac{1}{2}}}{\underline{\kappa}_1}\sqrt{\rho_Z^2\rho_\eta^2\frac{k_1\log(d\vee p)}{n}},\ c_2|\beta^*|_1\sqrt{\frac{\rho_{X^*}^2\rho_\eta^2\log p}{n}},\ c_2\sqrt{\frac{\rho_{X^*}^2\rho_\epsilon^2\log p}{n}}\right\}, \tag{12}$$

then, we have

$$|\hat{\beta}_{H2SLS}-\beta^*|_2 \leq \frac{c\sqrt{k_2}}{\underline{\kappa}_2}\mathcal{T}_0 \tag{13}$$

with probability at least $1-O\left(\frac{1}{p}\right)$. If $\rho_\eta \neq 0$, $d \geq p$, $k_1 \geq 1$, and $|\beta^*|_1 = O(1)$, then aside from factors involving ρ_Z, ρ_η, $\bar{\kappa}_1$, $\underline{\kappa}_1$, and $\underline{\kappa}_2$, (13) is of the order $O\left(\sqrt{k_2}\left[|\beta^*|_1\sqrt{\frac{k_1\log d}{n}}\right]\right)$, which differs from the optimal first-stage Lasso rate $\sqrt{\frac{k_1\log d}{n}}$ by $\sqrt{k_2}\,|\beta^*|_1$. Just as the role $\sqrt{k_2}$ plays in the typical rate $\sqrt{\frac{k_2\log p}{n}} \asymp \sqrt{k_2}\lambda_n = c^{'}\sqrt{k_2}t$ (where $\left|\frac{X^T\epsilon}{n}\right|_\infty = O(t)$) for the usual exactly sparse high-dimensional linear regression model (1) with $\mathbb{E}(X_i\epsilon_i)=\mathbf{0}$, the factor $\sqrt{k_2}$ appears in the rate for $|\hat{\beta}_{H2SLS}-\beta^*|_2$. The presence of the factor $|\beta^*|_1$ is explained above.

Condition (10) in Assumption (v) of Theorem 3.1 ensures that with high probability, $\frac{\hat{X}^T\hat{X}}{n}$ satisfies the RE condition over the restricted sets subject to those in Theorem 3.1. This result is formalized in the following corollary.

Corollary 3.1: If $\lambda_{n,j}$ ($j=1,...,p$) satisfy (8) and λ_n satisfies (9), under Assumptions (i)-(v) in Theorem 3.1, for some universal constant $c^{'}>0$,

$$\frac{\Delta^T\hat{X}^T\hat{X}\Delta}{n\,|\Delta|_2^2} \geq c^{'}\underline{\kappa}_2$$

with probability at least $1-O\left(\frac{1}{p\vee d}\right)$ for all non-zero Δ in the restricted sets subject to those in Theorem 3.1.

Remark. When β^* and π_j^* ($j=1,...,p$) are exactly sparse, condition (10) implies that $n \gtrsim k_1k_2^2\log(d\vee p)$. When $\left|\hat{\pi}_j-\pi_j^*\right|_2$ is of the same order $O(\sqrt{\frac{k_1\log(d\vee p)}{n}})$ for all $j=1,...,p$, the scaling $O\left(k_1k_2^2\log(d\vee p)\right)$ on n required for $\frac{\hat{X}^T\hat{X}}{n}$ to satisfy the RE condition for the case of exactly sparse β^* and π_j^* ($j=1,...,p$) is attained and cannot be improved under the conditions of Theorem 3.1. Note that, if $\left|\hat{\pi}_j-\pi_j^*\right|_2=0$ for "most" j's (which is possible if the number of coefficients with values 0 included in $\hat{\pi}_j$ is "small"), then it is possible to reduce the scaling $O(k_1k_2^2\log(d\vee p))$ to $O(k_1k_2\log(d\vee p))$ in condition (10) for the case of exactly sparse β^* and π_j^* ($j=1,...,p$). This result is stated in the following Theorem (Theorem 3.2), which requires additional assumptions as below.

Assumption 3.3: For every $j=1,...,p$, $W_j := Z_jv_j$ where $v_j \in \mathbb{K}(c^0k_1,\,d,\,R) := \mathbb{B}_0^d(c^0k_1)\cap\mathbb{B}_2^d(R)$ and $R = 2\max_{j=1,...,p}\left|\pi_j^*\right|_2$. The matrix $W\in\mathbb{R}^{n\times p}$ is sub-Gaussian with parameters $(\Sigma_W,\,\rho_W^2)$ where the jth column of W is W_j. For all such W's, the matrix $\mathbb{E}\left[\frac{W^TW}{n}\right]$ satisfies $\frac{\Delta^T\mathbb{E}\left[\frac{W^TW}{n}\right]\Delta}{|\Delta|_2^2} \geq \underline{\kappa}_W > 0$ for all non-zero $\Delta\in\mathbb{C}(J(\beta^*);\,0,\,c^*)$ (the constant c^0 is defined in the following assumption.).

Assumption 3.4: For every $j=1,...,p$, $|J(\hat{\pi}_j)| \leq c^0k_1$ with probability at least $1-O\left(\frac{1}{d\vee p}\right)$, where $c^0>0$ is some universal constant and $|J(\hat{\pi}_j)|$ denotes the cardinality of the support of $\hat{\pi}_j$.

Remark. Assumption 3.4 can be interpreted as an *exact sparsity* constraint on the first-stage estimate $\hat{\pi}_j$ for $j = 1, ..., p$, in terms of the l_0- "ball",

$$\mathbb{B}_0^d(c^0 k_1) := \left\{ \hat{\pi}_j \in \mathbb{R}^d \mid \sum_{l=1}^{d} 1\{\hat{\pi}_{jl} \neq 0\} \leq c^0 k_1 \right\}$$

for $j = 1, ..., p$. In the simplest case where the dimension of π_j^* is fixed and small relative to n for all $j = 1, ..., p$ (e.g., in the empirical example discussed in Section 1, each endogenous regressor, firm j's output, is instrumented with an exogenous variable, firm j's investment), Assumption 3.4 is satisfied trivially. For $d \geq n$, it holds under the bounded "sparse eigenvalue condition" (e.g., Bickel, et. al, 2009; Belloni and Chernozhukov, 2013), which is sufficient for the sparsity of $\hat{\pi}_j$ to be of the order k_1 (the sparsity of π_j^* when it is exactly sparse). With sufficient "separation" requirement on $\min_{l \in J(\pi_j^*)} |\pi_{jl}^*|$, Assumption 3.4 also holds for the thresholded $\hat{\pi}_j$ which removes false inclusions of elements that are outside the support of π_j^*. The term $O\left(\frac{1}{d \vee p}\right)$ in the probability guarantee again comes from the application of a union bound which takes into account the fact that there are p endogenous regressors in the main equation and hence, p regressions to perform in the first-stage.

Theorem 3.2: Suppose Assumptions 2.1, 3.1, 3.3, and 3.4 hold. Also, assume: (i) β^* and π_j^* $(j = 1, ..., p)$ are exactly sparse with at most k_2 and k_1 non-zero coefficients, respectively; (ii) Assumption 2.3 holds over the restricted sets $\mathbb{C}(J(\beta^*);\, 0,\, c^*)$ and $\mathbb{C}(J(\pi_j^*);\, 0,\, c^*)$ $(j = 1, ..., p)$, respectively, for the exact sparsity case $q_2 = 0$ with $R_{q_2} = k_2$ and $q_1 = 0$ with $R_{q_1} = k_1$. If $n \geq c_0 k_1 k_2 \log(p \vee d)$ for some sufficiently large constant $c_0 > 0$, then, $\frac{\Delta^T \hat{X}^T \hat{X} \Delta}{n |\Delta|_2^2} \geq c' \underline{\kappa}_2$ with probability at least $1 - O\left(\frac{1}{p \vee d}\right)$, for a constant $c' > 0$ and all non-zero Δ in $\mathbb{C}(J(\beta^*);\, 0,\, c^*)$. Consequently, if $\lambda_{n,j}$ satisfies (8) and $\lambda_n = \frac{c^*+1}{c^*-1} \mathcal{T}_0$ for $\mathcal{T}_0$ defined in (12), and for all vectors Δ in $\mathbb{C}(J(\pi_j^*);\, 0,\, c^*)$, $\frac{\Delta^T \Sigma_{Z_j} \Delta}{|\Delta|_2^2} \leq \bar{\kappa}_1$, $j = 1, ..., p$, then, with probability at least $1 - O\left(\frac{1}{p}\right)$, (13) with $\underline{\kappa}_2$ replaced by $\underline{\kappa}_W$ holds.

Remark. The proof for Theorem 3.2 is provided in Section A.4. Under Assumption 3.4, for the case of exactly sparse β^* and π_j^* $(j = 1, ..., p)$, Theorem 3.2 requires $\frac{k_1 k_2 \log d}{n} = O(1)$ (in contrast with $\frac{k_1 k_2^2 \log d}{n} = O(1)$ required by Theorem 3.1) to ensure that $\frac{\hat{X}^T \hat{X}}{n}$ satisfies the RE condition over $\mathbb{C}(J(\beta^*);\, 0,\, c^*)$ with high probability.

3.2 Variable-selection for exactly sparse β^*

This section addresses the question of variable selection when β^* is exactly sparse ($q_2 = 0$). The property $\mathbb{P}[J(\hat{\beta}_{H2SLS}) = J(\beta^*)] \to 1$ is referred to as *variable-selection consistency.* We present two results regarding achievability of this property in the following, where the first one is based on thresholding and the second one based on the "incoherence condition".

3.2.1 Variable-selection consistency with thresholding

Theorem 3.3: Suppose the assumptions in Lemma 3.1 hold and $c^{'}k_2\left(\frac{\log p}{n} \vee \mathcal{T}_1\right) \leq 1$ for some sufficiently large constant $c^{'} > 0$. Assume: (i) β^* is exactly sparse with at most k_2 non-zero coefficients; (ii) Assumption 2.3 holds over the restricted sets $\mathbb{C}(J(\beta^*);\, 0,\, c^*)$ for the exact sparsity case $q_2 = 0$ with $R_{q_2} = k_2$. If the regularization parameters $\lambda_{n,j}$ $(j = 1, ..., p)$ satisfy (8), λ_n satisfies (9), and $\min_{j \in J(\beta^*)} |\beta^*_j| > \frac{c\sqrt{k_2}}{\underline{\kappa}_2}\lambda_n = B$, then, $J(\hat{\beta}_{H2SLS}) \supseteq J(\beta^*)$ with probability at least $1 - O\left(\frac{1}{p}\right)$. Moreover, let the thresholded estimator $\bar{\beta}_j = \hat{\beta}_{j,H2SLS}1\left\{\left|\hat{\beta}_{j,H2SLS}\right| > B_1\right\}$ for $j = 1, ..., p$ and $B_1 > B$. If $\min_{j \in J(\beta^*)} |\beta^*_j| > B_1$, then, $J(\bar{\beta}) \subseteq J(\beta^*)$.

Remark. The proof for Theorem 3.3 is provided in Section A.5. Theorem 3.3 is analogous to results in literature (e.g., Meinshausen and Yu, 2009; Belloni and Chernozhukov, 2011a). The first claim says as long as the minimum value of $|\beta^*_j|$ over $j \in J(\beta^*)$ is not too small, then the two-stage Lasso does not falsely exclude elements that are in the support of β^* with high probability. The second claim says that with a stronger condition on $\min_{j \in J(\beta^*)} |\beta^*_j|$, additional thresholding can remove false inclusions of elements that are outside the support of β^*.

3.2.2 Variable-selection consistency with "incoherence condition"

Under additional assumptions, it is possible for $\hat{\beta}_{H2SLS}$ to achieve perfect selection without thresholding, as we will see in the following result.

Theorem 3.4: Suppose the assumptions in Lemma 3.1 hold and $c^{'}k_2\mathcal{T}_1 \leq 1$, $n \geq c^{''}k_2^3 \log p$, for some sufficiently large constant $c^{'}$, $c^{''} > 0$. Assume: (i) β^* is exactly sparse with at most k_2 non-zero coefficients; (ii)

$$\left\|\mathbb{E}\left[X^{*T}_{J(\beta^*)^c}X^*_{J(\beta^*)}\right]\left[\mathbb{E}(X^{*T}_{J(\beta^*)}X^*_{J(\beta^*)})\right]^{-1}\right\|_\infty = 1 - \phi \tag{14}$$

for some $\phi \in (0,\, 1]$. If the regularization parameters $\lambda_{n,j}$ satisfies (8) and

$$\lambda_n = \frac{\left(2 - \frac{(\bar{c}-2)\phi}{(\bar{c}-1)}\right)(\bar{c} - 1)}{(\bar{c} - 2 - \varsigma)\phi}\mathcal{T}_0 \tag{15}$$

for some universal constant $\bar{c} > 2$ and any small number $\varsigma > 0$, with $\mathcal{T}_0$ defined in (5), then, with probability at least $1 - O\left(\frac{1}{p}\right)$: (a) program (4) has a unique optimal solution $\hat{\beta}_{H2SLS}$; (b) $J(\hat{\beta}_{H2SLS}) \subseteq J(\beta^*)$; (c)

$$|\hat{\beta}_{H2SLS,J(\beta^*)} - \beta^*_{H2SLS,J(\beta^*)}|_\infty \leq \lambda_n\left[\frac{(\bar{c} - 2 - \varsigma)\phi}{\left(2 - \frac{(\bar{c}-2)\phi}{(\bar{c}-1)}\right)(\bar{c} - 1)} + 1\right]\left\|\left(\frac{\hat{X}^T_{J(\beta^*)}\hat{X}_{J(\beta^*)}}{n}\right)^{-1}\right\|_\infty = B_2,$$

where, for some constant $c_0 > 1$,

$$\left\| \left(\frac{\hat{X}^T_{J(\beta^*)} \hat{X}_{J(\beta^*)}}{n} \right)^{-1} \right\|_\infty \leq \frac{c_0 \sqrt{k_2}}{\lambda_{\min} \left(\mathbb{E} \left[\frac{1}{n} X^{*T}_{J(\beta^*)} X^*_{J(\beta^*)} \right] \right)}; \tag{16}$$

(d) if $\min_{j \in J(\beta^*)} |\beta^*_j| > B_2$, then, $J(\hat{\beta}_{H2SLS}) \supseteq J(\beta^*)$. As a consequence, $J(\hat{\beta}_{H2SLS}) = J(\beta^*)$.

Remark. The main proof for Theorem 3.4 is provided in Section A.6. Theorem 3.4 shows that under a population "incoherence condition" (14) similar to Wainwright (2009), we have $J(\hat{\beta}_{H2SLS}) \subseteq J(\beta^*)$ with high probability. The "incoherence condition" is a refined version of the "irrepresentable condition" by Zhao and Yu (2006) and the "neighborhood stability condition" by Meinshausen and Bühlmann (2006). Bühlmann and van de Geer (2011) shows this type of conditions is sufficient and "essentially necessary" for the Lasso to correctly excludes elements that are outside the support of β^* with high probability. If each row of $X^* \in \mathbb{R}^{n \times p}$ is sampled independently from $\mathcal{N}(\mathbf{0}, \Sigma_{X^*})$ with the Toeplitz covariance matrix

$$\Sigma_{X^*} = \begin{bmatrix} 1 & \varrho_{X^*} & \varrho^2_{X^*} & \cdots & \varrho^{p-1}_{X^*} \\ \varrho_{X^*} & 1 & \varrho_{X^*} & \cdots & \varrho^{p-2}_{X^*} \\ \varrho^2_{X^*} & \varrho_{X^*} & 1 & \cdots & \varrho^{p-3}_{X^*} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \varrho^{p-1}_{X^*} & \varrho^{p-2}_{X^*} & \cdots & \varrho_{X^*} & 1 \end{bmatrix},$$

condition (14) is satisfied (see, e.g., Wainwright, 2009); moreover, evidence from our numerical integration suggests that $\phi = 1 - \varrho_{X^*}$. The correlations between explanatory variables of agents of various proximity in a network or community can be naturally interpreted by the Toeplitz structure. For example, in the empirical example discussed in Section 1, firms that are "closer" might share more similarities in terms of production levels and the correlation between two firms' production levels decays geometrically in the degree of their "closeness". Note that the second-stage regularization parameter λ_n in (15) increases as the parameter ϕ decreases. Higher dependence between the components X^*_{ij} with $j \in J(\beta^*)$ and $X^*_{ij'}$ with $j' \in J(\beta^*)^c$ leads to higher penalty level in (15); consequently, in order to ensure variable-selection consistency, the choice in (15) is generally greater than the choice in (9), which concerns parameter estimation and does not need to account for the correlation between the regressors. However, when the components of X^*_i are independent of each other so that $\phi = 1$, and as long as $\bar{c} > 2$ ($\varsigma > 0$) in (15) is sufficiently large (respectively, sufficiently small) and $c^* > 1$ in (9) is sufficiently large, then (15) and (9) are approximately equal.

Imperfect variable selection and post-penalized procedures

The variable selection consistency of $\hat{\beta}_{H2SLS}$ is a desirable property; not only it guarantees the sparsity of $\hat{\beta}_{H2SLS}$ to be the same as the sparsity of β^*, most importantly it allows us to conduct post-selection inference by performing low-dimensional procedures on the selected model. However, we recognize that the conditions required in Theorem 3.3 or Theorem 3.4 are strong and perfect variable selection might be hard to achieve in practice. We briefly discuss a few solutions to the issue of imperfect variable selection in the following.

If the interest is only the sparsity of $\hat{\beta}_{H2SLS}$, the bounded "sparse eigenvalue condition" (e.g., Bickel, et. al, 2009; Belloni and Chernozhukov 2011a, 2013) is sufficient for the number of additional unnecessary components selected by $\hat{\beta}_{H2SLS}$ to be of the order k_2. "Sparse eigenvalue conditions" are also useful for analyzing a post $\hat{\beta}_{H2SLS}$ estimator similar to Belloni and Chernozhukov (2011a, 2013), which may attain a rate no slower than $\hat{\beta}_{H2SLS}$. If the interest is post-selection inference, it is possible to build another type of post procedure which uses $\hat{\beta}_{H2SLS}$ as an initial estimate to construct confidence intervals for individual coefficients and linear combinations of several of them (similar to Zhang and Zhang, 2013). Given that our focus here is the validity of the traditional 2SLS estimator with the l_1-regularization in both stages under high-dimensional scenarios, these aforementioned post strategies are beyond the scope of this paper but they are definitely worthwhile exploring in future research.

3.3 Choosing the regularization parameters

Because of the complex structure of model (1) and (2) and the nature of our two-stage estimation, existing adaptive methods (e.g., Antoniadis, 2010; Sun and Zhang, 2010, 2012; Belloni, et al., 2011; Gautier and Tsybakov, 2014; etc.) for setting the second-stage regularization parameter λ_n are less useful as they only have to deal with one unknown parameter related to the size of noise in a single linear regression model. As we have seen in (9), the choice of our λ_n depends on several unknown parameters: ρ_{X^*}, ρ_ϵ, $|\beta^*|_1$, ρ_Z, ρ_η, $\bar{\kappa}_1$, $\underline{\kappa}_1$, and R_{q_1}. Data-driven regularization parameter selection with theoretical guarantee turns out to be a particular challenge for the problem of our interest. In the following, we discuss two model-free criteria for choosing the regularization parameters in literature and also propose a feasible counterpart of the theoretical choice of the regularization parameter in (9). We then compare in our simulation experiments (Section 4) the amount of regularization imposed by these model-free criteria with the feasible counterpart of the theoretical choice.

When the Lasso is applied to estimate the standard high-dimensional sparse linear regression model (1) with exogenous X, Cross-Validation (CV) is the most popular approach for choosing data-driven regularization parameters (Allen 1974; Stone 1974). When facilitated by data resampling and parallel computing, CV finds a regularization parameter that locally minimizes the prediction error at a feasible computational cost (Breiman 1995, 1996, 2001; Hastie et al. 2002). However, Lasso+CV tends to overfit the model and perform poorly in parameter estimation especially when the regressors are correlated (see e.g., Bach, 2008; Meinshausen and Bühlmann, 2010; Lim and Yu, 2013; Yu, 2013). By combining a new metric, "Estimation Stability" (ES), with the CV, Lim and Yu (2013) propose an alternative model-free criterion ESCV, which yields a smaller-size model but similar performance in prediction relative to the CV choice. According to Lim and Yu (2013) as well as Yu (2013), the ESCV outperforms the CV in variable selection and substantially reduces false positive rates for exactly sparse models, and also outperforms the CV in parameter estimation for models with correlated regressors. To define the ES criterion, they adopt the idea of cross-validation data perturbation where n observations are randomly assigned into T subsamples of size $(n-L)$ with $L = \lfloor \frac{n}{T} \rfloor$. Given a regularization parameter λ^m and the subsample t, the Lasso is performed to obtain $\hat{\beta}_t(\lambda^m)$ and $\hat{Y}_t(\lambda^m) = X\hat{\beta}_t(\lambda^m)$. For $m = 1, ..., M$, Lim and Yu then form

$$\text{ES}(\lambda^m) := \frac{\widehat{\text{Var}}(\hat{Y}(\lambda^m))}{\left|\bar{\hat{Y}}(\lambda^m)\right|_n^2} = \frac{L}{n-L}\frac{1}{\mathcal{Z}^2(\lambda^m)} \tag{17}$$

where

$$\begin{aligned} \widehat{\text{Var}}(\hat{Y}(\lambda^m)) &:= \frac{1}{T}\sum_{t=1}^{T}\left|\hat{Y}_t(\lambda^m) - \bar{\hat{Y}}(\lambda^m)\right|_n^2, \\ \mathcal{Z}^2(\lambda^m) &:= \frac{\bar{\hat{Y}}(\lambda^m)}{\sqrt{\frac{n-L}{L}\widehat{\text{Var}}(\hat{Y}(\lambda^m))}} \end{aligned} \tag{18}$$

$|w|_n^2 := \frac{1}{n}\sum_{i=1}^{n} w_i^2$, $\bar{\hat{Y}}(\lambda^m) := \frac{1}{T}\sum_{t=1}^{T}\hat{Y}_t(\lambda^m)$. Note that (18) is proportional to the average pairwise squared Euclidean distance:

$$A(\lambda^m) := \frac{1}{\binom{T}{2}}\sum_{t\neq t'}\left|\hat{Y}_t(\lambda^m) - \hat{Y}_{t'}(\lambda^m)\right|_n^2. \tag{19}$$

They further point out that ES (17) is in fact the reciprocal of a test statistic for testing $H_0 : X\beta^* = 0$. To deal with the high noise situation where ES may not have a well-defined minimum, Lim and Yu suggest the combined ESCV criterion: Choose λ^m such that it minimizes $\text{ES}(\lambda^m)$ over all m and $\sum_{j=1}^{p}\hat{\sigma}_{X_j}\left|\hat{\beta}_j(\lambda^m)\right|$ ($\hat{\sigma}_{X_j} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}X_{ij}^2}$ and $\hat{\beta}_j(\lambda^m)$ is the Lasso estimate based on λ^m using the entire sample) is no greater than the one resulting from the optimal CV choice. They recommend a grid-search algorithm to find a local minimum of ES as often done for CV. Consequently, the ESCV enjoys a similar computational advantage to that of the CV and they both work well in the parallel computing paradigm.

To test the applicability of the model-free criteria discussed above in our problem, we simulate data sets with various model structures in Section 4 and apply either the Lasso+CV or the Lasso+ESCV in both (3) and (4). An estimate $\hat{\beta}$ of β^* is a function of $\left(\lambda_{n,j}^{m_j}\right)_{j=1}^{p}$ and λ_n^m where $m_j = 1, ..., M$ for $j = 1, ..., p$, and $m = 1, ..., M$. Ideally, the best λ_n^m should be selected as the optimum that minimizes the CV or the ESCV criterion over all combinations $\left[\lambda_n^m,\, (\lambda_{n,j}^{m_j})_{j=1}^{p}\right]$. This procedure, however, is computationally expensive when p is large as the number of combinations scales as M^p. Instead, we use the heuristic which selects λ_n^m only as the optimum that minimizes the CV or the ESCV criterion over combinations $\left[\lambda_n^m,\, (\lambda_{n,j}^{m_j^*})_{j=1}^{p}\right]$ where $\lambda_{n,j}^{m_j^*}$ is the optimum choice for estimating the jth equation in the first-stage. We then compare such $\lambda_n^m := \lambda_n^{m^*}$ with the feasible (plug-in) counterpart of the theoretical choice in (9).

To construct the feasible (plug-in) counterpart of (9), instead of trying to deal with all the unknown parameters and constant c_1 in $\mathcal{T}_1$ (6) (which bounds the first-stage prediction error $\max_{j=1,...,p}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Z_{ij}\hat{\pi}_j - Z_{ij}\pi_j^*)^2}$ from above), we suggest estimating $\frac{1}{n}\sum_{i=1}^{n}(Z_{ij}\hat{\pi}_j - Z_{ij}\pi_j^*)^2$

directly by the formula as in (19):

$$\hat{\mathcal{T}}_{1,j} := \frac{1}{\begin{pmatrix} T \\ 2 \end{pmatrix}} \sum_{t \neq t'} \left| Z_j \hat{\pi}_{jt}(\lambda_{n,j}^{m_j^*}) - Z_j \hat{\pi}_{jt'}(\lambda_{n,j}^{m_j^*}) \right|_n^2 \tag{20}$$

using the optimal first-stage regularization parameters $\lambda_{n,j}^{m_j^*}$, $j = 1, ..., p$ according to either the CV or the ESCV criterion. For the second-stage regularization parameter selection, when either the ES criterion (17) or the feasible plug-in method is used, it adjusts the amount of regularization to account for the noise from the first-stage estimates $\hat{X}_j$ as the surrogate of the unknown $X_j^* = Z_j \pi_j^*$ in the second-stage estimation (4).

Apart from the first-stage prediction error, the second-stage regularization parameter λ_n in (9) also depends on β^*, ρ_η, ρ_ϵ, and ρ_{X^*}. Upon the Lasso+CV or Lasso+ESCV estimates $\hat{\pi}_j = \hat{\pi}_j(\lambda_{n,j}^{m_j^*})$ of π_j^* from (3) for all $j = 1, ..., p$ and $\hat{\beta} = \hat{\beta}(\Lambda_n)$ ($\Lambda_n = \left[\lambda_n^{m^*}, (\lambda_{n,j}^{m_j^*})_{j=1}^p \right]$) of β^* from (4), we can estimate the unknown parameters β^* by $\hat{\beta}$, ρ_η by $\hat{\rho}_\eta = \max_j \sup_{\gamma \geq 1} \gamma^{-\frac{1}{2}} \left(\frac{1}{n} \sum_{i=1}^n |X_{ij} - Z_{ij} \hat{\pi}_j|^\gamma \right)^{\frac{1}{\gamma}}$, ρ_ϵ by $\hat{\rho}_\epsilon = \sup_{\gamma \geq 1} \gamma^{-\frac{1}{2}} \left(\frac{1}{n} \sum_{i=1}^n \left| Y_i - X_i \hat{\beta} \right|^\gamma \right)^{\frac{1}{\gamma}}$, and ρ_{X^*} by $\hat{\rho}_{X^*} = \max_j \sup_{\gamma \geq 1} \gamma^{-\frac{1}{2}} \left(\frac{1}{n} \sum_{i=1}^n \left| \hat{X}_{ij} \right|^\gamma \right)^{\frac{1}{\gamma}}$. The computation of the "sup" part in $\hat{\rho}_\eta$, $\hat{\rho}_\epsilon$, and $\hat{\rho}_{X^*}$can be carried out numerically for a sufficiently wide range of $\gamma \geq 1$. With all the estimated pieces from above in hand, the feasible plug-in counterpart λ_n^f of the theoretical choice in (9) can be formed by

$$\lambda_n^f = \frac{c^* + 1}{c^* - 1} \max_{r=1,2,3} \hat{Q}_r, \tag{21}$$

where $\hat{Q}_1 = \left| \hat{\beta} \right|_1 \max_{j=1,...,p} \hat{\mathcal{T}}_{1,j}$, $\hat{Q}_2 = c' \hat{\rho}_{X^*} \hat{\rho}_\eta \left| \hat{\beta} \right|_1 \sqrt{\frac{\log p}{n}}$, and $\hat{Q}_3 = c' \hat{\rho}_{X^*} \hat{\rho}_\epsilon \sqrt{\frac{\log p}{n}}$. In practice, one may "standardize" the choice of the constant c' in $\hat{Q}_2$ and $\hat{Q}_3$ according to some convenient distributions of X_{ij}^*, η_{ij} ($j = 1, ..., p$), and ϵ_i; for example, $c' = \sqrt{2} + \varsigma_0$ for any small number $\varsigma_0 > 0$ if X_{ij}^*'s, η_{ij}'s, ϵ_i's are independent Gaussian random variables, $\frac{1}{\sqrt{n}} |X_j^*|_2 \leq 1$, and $\mathbb{E}(\eta_{ij} | X_{ij}^*) = \mathbb{E}(\epsilon_i | X_{ij}^*) = 0$ for all $i = 1, ..., n$ and $j = 1, ..., p$; under such "standardization", we can replace $\hat{\rho}_\eta$ by $\hat{\sigma}_\eta = \max_j \sqrt{\frac{1}{n} \sum_{i=1}^n (X_{ij} - Z_{ij} \hat{\pi}_j)^2}$, $\hat{\rho}_\epsilon$ by $\hat{\sigma}_\epsilon = \sqrt{\frac{1}{n} \sum_{i=1}^n \left(Y_i - X_i \hat{\beta} \right)^2}$, and $\hat{\rho}_{X^*}$ by 1 ($\hat{\sigma}_{X^*} \leq 1$ for normalized $\hat{X}_j$). This "standardization" is similar to the usual practice in kernel density estimation for choosing bandwidth parameters (e.g., the "Silverman rule"; see Section 3.4.2 of Silverman, 1986). In terms of the constant $\frac{c^*+1}{c^*-1} > 1$, we recommend in practice choosing $\frac{c^*+1}{c^*-1}$ so that the resulting λ_n^f is not substantially different from the regularization parameter $\lambda_n^{m^*} := \lambda_n^{ESCV}$ to obey the "data faithfulness" requirement imposed by the ESCV criterion.

4 Simulations

We now turn to the Monte-Carlo simulation experiments. The data is generated according to (1) and (2) where

$$(\epsilon_i,\, \eta_i) \sim_{i.i.d.} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\epsilon^2 & \varrho\sigma_\epsilon\sigma_\eta & \cdots & \cdots & \varrho\sigma_\epsilon\sigma_\eta \\ \varrho\sigma_\epsilon\sigma_\eta & \sigma_\eta^2 & 0 & \cdots & 0 \\ \vdots & 0 & \sigma_\eta^2 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \varrho\sigma_\epsilon\sigma_\eta & 0 & \cdots & 0 & \sigma_\eta^2 \end{pmatrix}\right).$$

The matrix Z_i^T is a $p\times d$ matrix of Gaussian random variables with identical variances $\sigma_Z = \sigma_{z_{jl}} = 1$ for all $j = 1, ..., p$, $l = 1, ..., d$, and Z_{ij}^T is independent of $(\epsilon_i,\, \eta_{i1}, ..., \eta_{ip})$ for all $j = 1, ..., p$. We set the correlation level $\varrho = 0.1$ between ϵ_i and η_{ij} for all $j = 1, ..., p$. With this setup, we simulate 100 sets of *i.i.d.* $(Y_i,\, X_i^T,\, Z_i^T,\, \epsilon_i,\, \eta_i)_{i=1}^n$ where n is the sample size in each set, and construct Monte Carlo simulation experiments with different model parameters (β^*, σ_ϵ, and σ_η) and the design of Z_i. In terms of the dimensions, we set $d = 46$, $p = 50$, $n = 45$. In the first 5 experiments, $(\pi^*_{j,1}, ..., \pi^*_{j,4}) = \mathbf{0.5}$, $(\pi^*_{j,5}, ..., \pi^*_{j,46}) = \mathbf{0}$ for all $j = 1, ..., 50$; as a result, we have $\sigma_{X^*} = \sigma_{X_j^*} = 1$ for all $j = 1, ..., 50$. In addition, we set $(\beta^*_1, ..., \beta^*_4) = \mathbf{0.5}$, $(\beta^*_5, ..., \beta^*_{50}) = \mathbf{0}$ for the first 4 experiments; and $(\beta^*_1, ..., \beta^*_4) = \mathbf{1}$, $(\beta^*_5, ..., \beta^*_{50}) = \mathbf{0}$ for Experiment 5. Experiment 2 sets the ratio $\frac{\sigma_\epsilon}{\sigma_{X^*}}$ to 1 : 2 while the rest of experiments set it to 1 : 10; Experiment 3 sets the ratio $\frac{\sigma_\eta}{\sigma_{X^*}}(= \frac{\sigma_\eta}{\sigma_Z})$ to 1 : 2 while the rest of experiments set it to 1 : 10. Experiment 4 introduces correlations between the "purged" regressors X_j^* and $X_{j'}^*$ by setting $\mathrm{Corr}(Z_{ijl},\, Z_{ij'l}) = 0.5^{|j-j'|}$ for all $l = 1, ..., 46$ and $j,\, j' = 1, ..., 50$. Table 4.1 summarizes the designs of these experiments. We include four additional experiments (Experiments 6-9) in Section S.2 of the supplementary materials (https://sites.google.com/site/yingzhu1215/home/HD2SLS_Supplement.pdf) for approximate sparsity scenarios as in Assumption 2.2.

Table 4.1: Designs of the Monte-Carlo simulation experiments

Parameters	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5
β_j^*	$(\mathbf{0.5}, \mathbf{0})$	$(\mathbf{0.5}, \mathbf{0})$	$(\mathbf{0.5}, \mathbf{0})$	$(\mathbf{0.5}, \mathbf{0})$	$(\mathbf{1}, \mathbf{0})$
π_{jl}^*	$(\mathbf{0.5}, \mathbf{0})$	$(\mathbf{0.5}, \mathbf{0})$	$(\mathbf{0.5}, \mathbf{0})$	$(\mathbf{0.5}, \mathbf{0})$	$(\mathbf{0.5}, \mathbf{0})$
$\frac{\sigma_\epsilon}{\sigma_{X^*}}$	1 : 10	1 : 2	1 : 10	1 : 10	1 : 10
$\frac{\sigma_\eta}{\sigma_{X^*}}$	1 : 10	1 : 10	1 : 2	1 : 10	1 : 10
$\mathrm{Corr}(Z_{ijl}, Z_{ij'l})$	0	0	0	$0.5^{\lvert j-j'\rvert}$	0

For each simulation run $h = 1, ..., 100$, we first apply the Lasso+CV in both (3) and (4) and also apply the Lasso+ESCV in the same way; following the methods described in Section 3.3, we then compute the quantities in (21): $\hat{Q}_r^h$ $(r = 1, ..., 3)$ with $c' = \sqrt{2} + 0.01$ in $\hat{Q}_2^h$ and $\hat{Q}_3^h$, and set $\frac{c^*+1}{c^*-1} = 1.01$. Table 4.2 displays the amount of second-stage regularization averaged over 100 simulations according the CV criterion (column "CV") and the ESCV criterion (column "ESCV") as well as the feasible plug-in choices $\overline{\lambda_n^f} := 1.01 \max_{r=1,2,3} \frac{1}{100} \sum_{h=1}^{100} \hat{Q}_r^h$ (columns

"PLUG-1" and "PLUG-2"); column "PLUG-1" (column "PLUG-2") are choices that use the CV estimates (respectively, the ESCV estimates) to form $\hat{X}$ in (4) and $\hat{\rho}_{\eta}$, $\hat{\mathcal{T}}_{1,j}$, $\hat{\rho}_{\epsilon}$, $\hat{\rho}_{X^*}$, and $\hat{\beta}_j$ in (21). Under the CV, the ESCV, and the feasible plug-in choices, respectively, Table 4.2 also displays the mean of the l_2-errors, $\frac{1}{100}\sum_{h=1}^{100}|\hat{\beta}^h-\beta^*|_2$ as well as the mean of the selection percentages, $\frac{1}{100}\sum_{h=1}^{100}\frac{1}{50}\sum_{j=1}^{p}1\{\text{sgn}(\hat{\beta}_j^h)=\text{sgn}(\beta_j^*)\}$.

Table 4.2 shows that the two-stage Lasso+ESCV outperforms the two-stage Lasso+CV in variable selection while giving similar l_2-errors; the two-stage Lasso+CV procedure overfits the models by under penalizing and selects more "irrelevant" variables (ones whose true coefficients are zero). As a consequence, when computing the plug-in quantities $\hat{Q}_r$, we noticed that $\hat{Q}_1$ and $\hat{Q}_2$ with $\hat{\beta}_j$ obtained from the CV estimates tend to be greater than those from the ESCV estimates, while $\hat{Q}_3$ with $\hat{\rho}_{\epsilon}$ obtained from the CV estimates tend to be smaller than those from the ESCV estimates. Experiment 5 shows that the amount of regularization needed for (4) to perform well in both estimation and selection increases with $|\beta^*|_1$, and the ESCV procedure appears to do better at accounting for the increasing $|\beta^*|_1$ than the CV. From Table 4.2, we see that overall, the choices which use the ESCV estimates to produce $\overline{\lambda_n^f}$ (column "PLUG-2") tend to over penalize but still give satisfactory performance in parameter estimation and variable selection; except when the ratio $\frac{\sigma_\eta}{\sigma_{X^*}}$ is sufficiently high as in Experiment 3, the "plug-in" choices result in significant reduction of true positive rates (given that the mean of the l_2-errors is greater than $\beta_j^*=0.5$ for $j=1,...,4$). Based on these simulation results, the Lasso+ESCV procedure described in Section 3.3 for (3) and (4) appears to be the most effective method in terms of both estimation and selection. In practice, one may also consider our alternative "plug-in" method (21) using the estimates from the ESCV procedure and compare it with the optimal regularization parameter chosen by the ESCV criterion to determine whether the amount of regularization is sufficient.

Table 4.2: 2nd-stage regularization level, l_2-error, and selection %

Exp	CV			ESCV			PLUG-1			PLUG-2		
#	reg	l_2-err	sel %	reg	l_2-err	sel %	reg	l_2-err	sel %	reg	l_2-err	sel %
1	0.020	0.081	89.2	0.045	0.071	97.2	0.154	0.323	99.9	0.113	0.204	99.8
2	0.078	0.345	89.9	0.120	0.337	94.3	0.168	0.414	96.9	0.198	0.444	98.1
3	0.057	0.268	87.7	0.121	0.278	94.3	0.956	0.998	92.1	0.728	1.014	92.0
4	0.024	0.073	92.2	0.056	0.063	99.1	0.155	0.162	100	0.116	0.097	99.9
5	0.028	0.113	88.9	0.070	0.098	97.2	0.305	0.642	99.9	0.230	0.416	99.9

5 Conclusion and extensions

This paper has explored the validity of the l_1-regularized 2SLS estimation for linear models where the number of endogenous regressors in the main equation and the number of instruments in the first-stage equations can exceed the sample size n, and the regression coefficients belong to l_q-"balls" for $q\in[0,\,1]$, which covers both exact and approximate sparsity cases. Standard high-level assumptions on the Gram matrix for l_2-consistency require careful verifications in the two-stage procedure, for which we provide detailed theoretical analysis. Conditions for estimation consistency in l_2-norm and variable-selection consistency of the high-dimensional two-stage estimators have been established. We also provide practical methods for choosing the regularization parameters and the effectiveness of these methods is demonstrated on simulated data sets.

In addition to the research directions already proposed in the previous sections for the future, we discuss some more extensions in the following. First, as pointed out by a reviewer, it would be ideal to test the performance of our procedure on real data sets to see the shortcoming of our estimator and the way the regularization parameters are chosen. Second, as an alternative to the l_1-regularized 2SLS procedure proposed here, a high-dimensional two-stage estimator based on the "control function" approach would be interesting to explore.

Third, it may be worthwhile to extend our analysis to allow non-sub-Gaussian errors ϵ and η in (1) and (2). There are a couple of ways to relax the sub-Gaussian condition on the error terms. For example, the square-root Lasso (as in Belloni, Chernozhukov, and Wang, 2011) and the pivotal Dantzig selector (as in Gautier and Tsybakov, 2014) whose "score" functions (the first derivative of the sample square root of the residual sum of squares loss evaluated at the true parameters) allow these authors to evoke a bound for moderate deviations of self-normalized sums of random variables (Lemma 2.11 by Jing, Shao and Wang, 2003). The bound in Jing, et al. does not require sub-Gaussian tails. However, compared to the standard Lasso, the square-root Lasso or the pivotal Dantzig selector involves a more sophisticated optimization algorithm computation-wise. Another paper by Minsker (2014) that uses a "trick" originally noted in Nemirovski and Yudin (1983) is also able to avoid imposing a sub-Gaussian condition on the error terms when deriving the nonasymptotic bounds for the standard Lasso. It is possible to apply these techniques in our problem, albeit doing so would distract the main focus of this paper; therefore, we leave these extensions to future research.

A Appendix: Main Proofs

For notational simplicity, in the following proofs, assume $d_j = d$ for all $j = 1, ..., p$; additionally, as in most high-dimensional statistics literature, we assume the regime of interest is $p \geq n$ and $d \geq n$. The modification to allow $p < n$ or $d < n$ or $d_j \neq d_{j'}$ for some j and j' is straightforward. Also, as a general rule for the proofs, c constants denote generic positive constants that are independent of n, p, d, R_{q_2}, R_{q_1}, and may change from place to place.

A.1 Lemmas A.1-A.3

Lemma A.1 (General upper bound on the l_2-error). Let $\hat{\Gamma} = \frac{1}{n}\hat{X}^T\hat{X}$, $\hat{D} = \text{diag}\left[\hat{\sigma}_{X_1^*}, ..., \hat{\sigma}_{X_p^*}\right]$, and $e = (X^* - \hat{X})\beta^* + \eta\beta^* + \epsilon$. For some universal constant $c^* > 1$, if λ_n in program (4) satisfies

$$\lambda_n \geq \frac{c^* + 1}{c^* - 1}|\hat{D}^{-1}\frac{1}{n}\hat{X}^T e|_\infty > 0,$$

and $c' R_{q_2}\underline{\tau}^{-q_2}\left(\frac{\log p}{n} \vee \mathcal{T}_1\right) \leq 1$ for some constant $c' > 0$ that depends on $\underline{\kappa}_2$, then there is a constant $c > 0$ such that under Assumption 2.2,

$$|\hat{\beta}_{H2SLS} - \beta^*|_2 \leq \frac{c}{\underline{\kappa}_2^{1-\frac{q_2}{2}}} R_{q_2}^{\frac{1}{2}} \lambda_n^{1-\frac{q_2}{2}}.$$

Proof. First, write

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

where $e := (X^* - \hat{X})\beta^* + \eta\beta^* + \epsilon$. Define the thresholded subset

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

where $\underline{\tau} = \frac{\lambda_n}{\underline{\kappa}_2}$ is the threshold parameter. For any $p-$dimensional vector v, denote $|v|_{1,n} = \sum_{j=1}^{p} \hat{\sigma}_{X_j^*} |v_j|$, the l_1-norm weighed by $\hat{\sigma}_{X_j^*}$s. Define $\hat{v}^0 = \hat{\beta}_{H2SLS} - \beta^*$ and the Lagrangian $L(\beta;\, \lambda_n) = \frac{1}{2n}|Y - \hat{X}\beta|_2^2 + \lambda_n |\beta|_{1,n}$. Since $\hat{\beta}_{H2SLS}$ is optimal, we have

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

which yields

$$\begin{aligned} 0 \leq \frac{1}{2n}|\hat{X}\hat{v}^0|_2^2 &\leq \frac{1}{n}e^T\hat{X}\hat{v}^0 + \lambda_n \left\{ |\beta^*_{S_{\underline{\tau}}}|_{1,n} + |\beta^*_{S^c_{\underline{\tau}}}|_{1,n} - |(\beta^*_{S_{\underline{\tau}}} + \hat{v}^0_{S_{\underline{\tau}}},\, \beta^*_{S^c_{\underline{\tau}}} + \hat{v}^0_{S^c_{\underline{\tau}}})|_{1,n} \right\} && (22) \\ &\leq |\hat{D}\hat{v}^0|_1 |\hat{D}^{-1}\frac{1}{n}\hat{X}^T e|_\infty + \lambda_n \left\{ |\hat{v}^0_{S_{\underline{\tau}}}|_{1,n} - |\hat{v}^0_{S^c_{\underline{\tau}}}|_{1,n} + 2|\beta^*_{S^c_{\underline{\tau}}}|_{1,n} \right\} && (23) \\ &\leq \lambda_n \frac{c^*-1}{c^*+1} \left\{ \frac{2c^*}{c^*-1}|\hat{v}^0_{S_{\underline{\tau}}}|_{1,n} - \frac{2}{c^*-1}|\hat{v}^0_{S^c_{\underline{\tau}}}|_{1,n} + \frac{2(c^*+1)}{c^*-1}|\beta^*_{S^c_{\underline{\tau}}}|_{1,n} \right\} \\ &\leq \lambda_n \frac{c^*-1}{c^*+1} \left\{ \frac{3c^*}{c^*-1}|\hat{v}^0_{S_{\underline{\tau}}}|_1 - \frac{3}{c^*-1}|\hat{v}^0_{S^c_{\underline{\tau}}}|_1 + \frac{3(c^*+1)}{c^*-1}|\beta^*_{S^c_{\underline{\tau}}}|_1 \right\} && (24) \end{aligned}$$

where the third inequality holds as long as $\lambda_n \geq \frac{c^*+1}{c^*-1}|\hat{D}^{-1}\frac{1}{n}\hat{X}^T e|_\infty$, and the last inequality follows from (37). Consequently,

$$|\hat{v}^0|_1 \leq (c^*+1)|\hat{v}^0_{S_{\underline{\tau}}}|_1 + (c^*+1)|\beta^*_{S^c_{\underline{\tau}}}|_1 \leq (c^*+1)\sqrt{|S_{\underline{\tau}}|}|\hat{v}^0|_2 + (c^*+1)|\beta^*_{S^c_{\underline{\tau}}}|_1.$$

We now upper bound the cardinality of $S_{\underline{\tau}}$ in terms of the threshold $\underline{\tau}$ and the l_q- "ball" with "radius" of R_{q_2} condition on β^*. Note that we have

$$R_{q_2} \geq \sum_{j=1}^{p} \left|\beta_j^*\right|^{q_2} \geq \sum_{j \in S_{\underline{\tau}}} \left|\beta_j^*\right|^{q_2} \geq \underline{\tau}^{q_2} \left|S_{\underline{\tau}}\right|$$

and therefore $\left|S_{\underline{\tau}}\right| \leq \underline{\tau}^{-q_2} R_{q_2}$. To upper bound the approximation error $|\beta^*_{S^c_{\underline{\tau}}}|_1$, we use the fact that $\beta^* \in \mathcal{B}^p_{q_2}(R_{q_2})$ and have

$$|\beta^*_{S^c_{\underline{\tau}}}|_1 = \sum_{j \in S^c_{\underline{\tau}}} \left|\beta_j^*\right| = \sum_{j \in S^c_{\underline{\tau}}} \left|\beta_j^*\right|^{q_2} \left|\beta_j^*\right|^{1-q_2} \leq R_{q_2}\underline{\tau}^{1-q_2}.$$

Putting the pieces together yields

$$|\hat{v}^0|_1 \leq (c^* + 1)\sqrt{\underline{\tau}^{-q_2} R_{q_2}}|\hat{v}^0|_2 + (c^* + 1) R_{q_2} \underline{\tau}^{1-q_2}. \tag{25}$$

Let us first prove the case of $q_2 \in (0, 1]$. Note that from (22), (23), and (37), we have

$$\begin{aligned}
\frac{1}{2n}|\hat{X}\hat{v}^0|_2^2 &\leq |\hat{v}^0|_{1,n}|\hat{D}^{-1}\frac{1}{n}\hat{X}^T e|_\infty + \lambda_n\left\{|\hat{v}^0_{S_{\underline{\tau}}}|_{1,n} - |\hat{v}^0_{S^c_{\underline{\tau}}}|_{1,n} + 2|\beta^*_{S^c_{\underline{\tau}}}|_{1,n}\right\} \\
&\leq 2\left[|\hat{v}^0|_1|\hat{D}^{-1}\frac{1}{n}\hat{X}^T e|_\infty + \lambda_n\left\{|\hat{v}^0_{S_{\underline{\tau}}}|_1 - |\hat{v}^0_{S^c_{\underline{\tau}}}|_1 + 2|\beta^*_{S^c_{\underline{\tau}}}|_1\right\}\right] \\
&\leq \left(c_0\sqrt{\underline{\tau}^{-q_2} R_{q_2}}|\hat{v}^0|_2 + c_1 R_{q_2}\underline{\tau}^{1-q_2}\right)\lambda_n \\
&\leq c_0\sqrt{\underline{\tau}^{-q_2} R_{q_2}}|\hat{v}^0|_2\lambda_n + c_1\underline{\delta} \qquad (26)\\
&\leq \max\left\{c_0 R_{q_2}^{\frac{1}{2}}\underline{\kappa}_2^{\frac{q_2}{2}}\lambda_n^{1-\frac{q_2}{2}}|\hat{v}^0|_2,\ c_1\underline{\delta}\right\}
\end{aligned}$$

where the third and fourth inequalities follow from our choices of $\underline{\tau} = \frac{\lambda_n}{\underline{\kappa}_2}$ and $\underline{\delta} = R_{q_2}\lambda_n\tau^{1-q_2}$. Now we proceed by cases. If

$$\max\left\{c_0 R_{q_2}^{\frac{1}{2}}\underline{\kappa}_2^{\frac{q_2}{2}}\lambda_n^{1-\frac{q_2}{2}}|\hat{v}^0|_2,\ c_1\underline{\delta}\right\} = c_0 R_{q_2}^{\frac{1}{2}}\underline{\kappa}_2^{\frac{q_2}{2}}\lambda_n^{1-\frac{q_2}{2}}|\hat{v}^0|_2,$$

and if $c^{'} R_{q_2}\underline{\tau}^{-q_2}\left(\frac{\log p}{n} \vee \mathcal{T}_1\right) \leq 1$ for some constant $c^{'} > 0$ that depends on $\underline{\kappa}_2$, we have

$$|\hat{v}^0|_2 \geq c_3\underline{\kappa}_2^{-1+\frac{q_2}{2}} R_{q_2}^{\frac{1}{2}}\lambda_n^{1-\frac{q_2}{2}} \geq \delta^* \tag{27}$$

where $\delta^* = \frac{c_2}{\underline{\kappa}_2^{\frac{1}{2}}} R_{q_2}\underline{\tau}^{1-q_2}\left(\sqrt{\mathcal{T}_1 \vee \frac{b_0 \log p}{n}}\right)$ and $b_0 = \underline{\kappa}_2\left(\frac{1}{\underline{\kappa}_2^2} \vee 1\right)$. Consequently, (24) and (27) together imply that

$$\hat{v}^0 \in \mathbb{K}(\underline{\delta},\, S_{\underline{\tau}}) := \mathbb{C}(S_{\underline{\tau}};\, q_2,\, c^*) \cap \left\{v^0 \in \mathbb{R}^p\ :\ \left|v^0\right|_2 \geq \delta^*\right\} \tag{28}$$

where

$$\mathbb{C}(S_{\underline{\tau}};\, q_2,\, c^*) = \left\{v^0 \in \mathbb{R}^p\ :\ |v^0_{S^c_{\underline{\tau}}}|_1 \leq c^*|v^0_{S_{\underline{\tau}}}|_1 + (c^* + 1)|\beta^*_{S^c_{\underline{\tau}}}|_1\right\}.$$

By Lemma A.2 and Lemma A.4, the random matrix $\hat{\Gamma} = \frac{\hat{X}^T\hat{X}}{n}$ satisfies the RE condition over

$$\mathbb{C}(S_{\underline{\tau}};\, q_2,\, c^*) \cap \left\{v^0 \in \mathbb{R}^p\ :\ \left|v^0\right|_2 \geq \delta^*\right\}, \tag{29}$$

therefore, we have

$$c^{''}\underline{\kappa}_2|\hat{v}^0|_2^2 \leq \frac{1}{2n}|\hat{X}\hat{v}^0|_2^2 \leq c_0 R_{q_2}^{\frac{1}{2}}\underline{\kappa}_2^{\frac{q_2}{2}}\lambda_n^{1-\frac{q_2}{2}}|\hat{v}^0|_2,$$

so the claim follows. It is sufficient to set δ in Assumption 2.3 to $\delta = c_3\underline{\kappa}_2^{-1+\frac{q_2}{2}} R_{q_2}^{\frac{1}{2}}\lambda_n^{1-\frac{q_2}{2}} \geq \delta^*$ where $c > c_3 > 0$. On the other hand, if

$$\max\left\{c_0 R_{q_2}^{\frac{1}{2}}\underline{\kappa}_2^{\frac{q_2}{2}}\lambda_n^{1-\frac{q_2}{2}}|\hat{v}^0|_2,\ c_1\underline{\delta}\right\} = c_1\underline{\delta},$$

then

$$|\hat{v}^0|_2 \leq c\underline{\kappa}_2^{-1+\frac{q_2}{2}} R_{q_2}^{\frac{1}{2}} \lambda_n^{1-\frac{q_2}{2}}$$

so again the claim follows.

To prove the case of $q_2 = 0$, simply choose $S_{\underline{\tau}} = J(\beta^*)$ and $\underline{\delta} = 0$ in (24) and (26), respectively, and the claim follows trivially from the above argument.□

Remark. Inequality (25) implies that $|\hat{v}^0|_1 \precsim \underline{\kappa}_2^{q_2-1} R_{q_2} \lambda_n^{1-q_2}$.

Lemma A.2: Define the thresholded subset

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

Under the assumptions in Theorem 3.1 and the choice $\underline{\tau} = \frac{\lambda_n}{\underline{\kappa}_2}$, if

$$c_0 R_{q_2} \underline{\tau}^{-q_2} \left(\frac{b_0 \log p}{n} \vee \mathcal{T}_1 \right) \leq \underline{\kappa}_2,$$

the RE condition holds for $\frac{\hat{X}^T \hat{X}}{n}$ over the set

$$\mathbb{C}(S_{\underline{\tau}};\, q_2,\, c^*) \cap \left\{ v^0 \in \mathbb{R}^p \, : \, \left|v^0\right|_2 \geq \delta^* \right\}$$

where $\delta^* = \frac{c_1}{\underline{\kappa}_2^{\frac{1}{2}}} R_{q_2} \underline{\tau}^{1-q_2} \left(\sqrt{\mathcal{T}_1 \vee \frac{b_0 \log p}{n}} \right)$ and $b_0 = \underline{\kappa}_2 \left(\frac{1}{\underline{\kappa}_2^2} \vee 1 \right)$, for some universal constant $c^* > 1$.

Proof. The argument is similar to what is used in the proof of Lemma 2 from Negahban, et. al (2010). For any $v^0 \in \mathbb{C}(S_{\underline{\tau}};\, q_2,\, c^*)$, we have

$$\begin{aligned} |v^0|_1 &\leq (c^* + 1)|v^0_{S_{\underline{\tau}}}|_1 + (c^* + 1)|\beta^*_{S^c_{\underline{\tau}}}|_1 \\ &\leq (c^* + 1)\sqrt{R_{q_2}} \underline{\tau}^{-\frac{q_2}{2}} |v^0|_2 + (c^* + 1) R_{q_2} \underline{\tau}^{1-q_2}, \end{aligned}$$

where we have used the bound in (25) from the proof of Lemma A.1. Therefore, for any vector $\Delta \in \mathbb{C}(S_{\underline{\tau}}; q_2,\, c^*)$ and the choice $\underline{\tau} = \frac{\lambda_n}{\underline{\kappa}_2}$, substituting the upper bound $(c^* + 1)\sqrt{R_{q_2}} \underline{\tau}^{-\frac{q_2}{2}} |v^0|_2 + (c^* + 1) R_{q_2} \underline{\tau}^{1-q_2}$ on $|v^0|_1$ into condition (38) from Lemma A.4 yields

$$\left| v^{0T} \frac{\hat{X}^T \hat{X}}{n} v^0 \right| \geq |v^0|_2^2 \left\{ c\underline{\kappa}_2 - c_0 R_{q_2} \underline{\tau}^{-q_2} \left(\mathcal{T}_1 \vee \frac{b_0 \log p}{n} \right) \right\} - c_0 R_{q_2}^2 \underline{\tau}^{2-2q_2} \left(\mathcal{T}_1 \vee \frac{b_0 \log p}{n} \right),$$

for some sufficiently small c_0, where $b_0 = \underline{\kappa}_2 \left(\frac{1}{\underline{\kappa}_2^2} \vee 1 \right)$. With the choice of

$$\frac{c_1}{\underline{\kappa}_2^{\frac{1}{2}}} R_{q_2} \underline{\tau}^{1-q_2} \left(\sqrt{\mathcal{T}_1 \vee \frac{b_0 \log p}{n}} \right) = \delta^*,$$

for some sufficiently small c_1, and if

$$c_0 R_{q_2} \underline{\tau}^{-q_2} \left(\frac{b_0 \log p}{n} \vee \mathcal{T}_1 \right) \leq \frac{c\underline{\kappa}_2}{2},$$

we have

$$\left| v^{0T} \frac{\hat{X}^T \hat{X}}{n} v^0 \right| \geq c^{'} \underline{\kappa}_2 \left\{ |v^0|_2^2 - \frac{|v^0|_2^2}{2} \right\} = c^{''} \underline{\kappa}_2 |v^0|_2^2$$

for any v^0 such that $\left|v^0\right|_2 \geq \delta^*$. $\square$

Lemma A.3: Suppose the assumptions in Lemma 3.1 hold. If $\hat{\pi}_j$ solves program (3) with the regularization parameter $\lambda_{n,j} \geq c_0 \rho_Z \rho_\eta \sqrt{\frac{\log(d \vee p)}{n}}$ for $j = 1, ..., p$, then,

$$\max_{j=1,\ldots,p} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[Z_{ij}^T \hat{\pi}_j - Z_{ij}^T \pi_j^* \right]^2 \right\} \leq \frac{c_1 \bar{\kappa}_1}{\underline{\kappa}_1^{2-q_1}} R_{q_1} \left(\rho_Z^2 \rho_\eta^2 \frac{\log(d \vee p)}{n} \right)^{1-\frac{q_1}{2}}$$

with probability at least $1 - O\left(\frac{1}{d \vee p}\right)$.

Proof. Applying Lemma B.1 with $t = c_0 \rho_Z \rho_\eta \sqrt{\frac{\log(d \vee p)}{n}}$ and a union bound yields

$$\mathbb{P}\left(\max_{j=1,\ldots,p} \left| \frac{1}{n} Z_j^T \eta_j \right|_\infty \leq c_0 \rho_Z \rho_\eta \sqrt{\frac{\log(d \vee p)}{n}} \right) \geq O\left(\frac{1}{p \vee d} \right).$$

We can use (40) in Lemma B.3 with $s = c_1 \frac{n}{\log(d \vee p)} \frac{\underline{\kappa}_1^{1+\frac{q_1}{2}}}{\bar{\kappa}_1}$, $U = Z_j$, and $\underline{\kappa} = \underline{\kappa}_1$ to show that

$$\frac{|Z_j v^j|_2^2}{n} \geq \frac{\underline{\kappa}_1}{2} |v^j|_2^2 - c \frac{\bar{\kappa}_1}{\underline{\kappa}_1^{\frac{q_1}{2}}} \frac{\log(d \vee p)}{n} |v^j|_1^2,$$

for any v^j in the restricted set subject to $\mathbb{C}(S_{\underline{\tau}_j}; q_1, c^*) \cap \mathbb{S}_{\delta_j}$, $j = 1, ..., p$, where $\underline{\tau}_j = \frac{\lambda_{n,j}}{\underline{\kappa}_1}$ and $\delta_j = c_2 \underline{\kappa}_1^{-1+\frac{q_1}{2}} R_{q_1}^{\frac{1}{2}} \lambda_{n,j}^{1-\frac{q_1}{2}}$ for some sufficiently small constant $c_2 > 0$. Follow the argument in Lemmas A.1 and A.2 where we set

$$\delta_j^* = O\left(\underline{\kappa}_1^{-\frac{1}{2}} R_{q_1} \underline{\tau}_j^{1-q_1} \sqrt{\frac{\log(d \vee p)}{n}} \right)$$

for all $j = 1, ..., p$ so that $\delta_j^* \leq \delta_j$. If $n \geq c^{'} R_{q_1}^{\frac{2}{2-q_1}} \log(d \vee p)$ for some sufficiently large constant $c^{'} > 0$ that depends on $\underline{\kappa}_1$, we have, for some $c_3 > c_2 > 0$,

$$\left| \hat{\pi}_j - \pi_j^* \right|_2 \leq \frac{c_3 \sqrt{\bar{\kappa}_1}}{\underline{\kappa}_1^{1-\frac{q_1}{2}}} R_{q_1}^{\frac{1}{2}} \left(\rho_Z \rho_\eta \sqrt{\frac{\log(d \vee p)}{n}} \right)^{1-\frac{q_1}{2}}, \tag{30}$$

and

$$\left|\hat{v}^j\right|_1 \leq c_4 \underline{\kappa}_1^{q_1-1} R_{q_1} \lambda_{n,j}^{1-q_1} = c_5 \underline{\kappa}_1^{\frac{q_1}{2}} R_{q_1}^{\frac{1}{2}} \left|\hat{v}^j\right|_2 \left(\sqrt{\frac{\rho_Z^2 \rho_\eta^2 \log(d \vee p)}{n}}\right)^{-\frac{q_1}{2}}, \tag{31}$$

where $\hat{v}^j = \hat{\pi}_j - \pi_j^*$ for $j = 1, ..., p$. The bound (41) in Lemma B.3 with $s = c_1 \frac{n}{\log(d \vee p)} \frac{\underline{\kappa}_1^{1+\frac{q_1}{2}}}{\bar{\kappa}_1}$ then implies

$$\begin{aligned}
\frac{\left|Z_j \hat{v}^j\right|_2^2}{n} &\leq \frac{3\bar{\kappa}_1}{2}\left|\hat{v}^j\right|_2^2 + \frac{\bar{\kappa}_1}{2c_1 \underline{\kappa}_1^{1+\frac{q_1}{2}}} \frac{\log(d \vee p)}{n}\left|\hat{v}^j\right|_1^2 \\
&\leq \frac{3\bar{\kappa}_1}{2}\left|\hat{v}^j\right|_2^2 + \frac{\bar{\kappa}_1 R_{q_1}}{2c_1 \underline{\kappa}_1^{1-\frac{q_1}{2}}} \left(\sqrt{\frac{\log(d \vee p)}{n}}\right)^{2-q_1} \left|\hat{v}^j\right|_2^2 \\
&\leq \frac{(3+\varsigma)\bar{\kappa}_1}{2}\left|\hat{v}^j\right|_2^2
\end{aligned} \tag{32}$$

for any v^j in the restricted set subject to $\mathbb{C}(S_{\underline{\tau}_j};\, q_1,\, c^*) \cap \mathbb{S}_{\delta_j}$, where the last inequality follows as long as

$$\frac{\bar{\kappa}_1 R_{q_1}}{2c_1 \underline{\kappa}_1^{1-\frac{q_1}{2}}} \left(\sqrt{\frac{\log(d \vee p)}{n}}\right)^{2-q_1} \leq \frac{\varsigma \bar{\kappa}_1}{2}$$

for any $\varsigma > 0$. Combining (32) and (30) yields the claim.

A.2 Proof for Lemma 3.1

Proof. We provide a proof for a more general result that implies Lemma 3.1. This more general result is useful for proving Theorem 3.1 later on. Note that we have

$$\begin{aligned}
\left|\frac{\hat{X}^T \hat{X} - X^{*T} X^*}{n}\right|_\infty &\leq \left|\frac{X^{*T}(\hat{X} - X^*)}{n}\right|_\infty + \left|\frac{(\hat{X} - X^*)^T \hat{X}}{n}\right|_\infty \\
&\leq \left|\frac{X^{*T}(\hat{X} - X^*)}{n}\right|_\infty + \left|\frac{(\hat{X} - X^*)^T X^*}{n}\right|_\infty + \left|\frac{(\hat{X} - X^*)^T (\hat{X} - X^*)}{n}\right|_\infty
\end{aligned} \tag{33}$$

To bound the term $\left|\frac{X^{*T}(\hat{X} - X^*)}{n}\right|_\infty$, first note that by Lemma A.3, we have

$$\max_{j=1,...,p} \sqrt{\frac{1}{n}\sum_{i=1}^n \left[Z_{ij}(\hat{\pi}_j - \pi_j^*)\right]^2} \leq c \frac{\sqrt{\bar{\kappa}_1} R_{q_1}^{\frac{1}{2}}}{\underline{\kappa}_1^{1-\frac{q_1}{2}}} \left(\sqrt{\frac{\rho_Z^2 \rho_\eta^2 \log(d \vee p)}{n}}\right)^{1-\frac{q_1}{2}}$$

with probability at least $1 - c_1 \exp(-c_2 \log(d \vee p))$. As a consequence, we apply a Cauchy-Schwarz inequality and obtain

$$
\begin{aligned}
\max_{j',j}\left|\frac{1}{n}X_{j'}^{*T}(\hat{X}_j - X_j^*)\right| &= \max_{j',j}\left|\frac{1}{n}\sum_{i=1}^{n}X_{ij'}^{*}Z_{ij}(\hat{\pi}_j - \pi_j^*)\right| \\
&\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}X_{ij'}^{*2}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[Z_{ij}(\hat{\pi}_j - \pi_j^*)\right]^2} \\
&\leq c\sigma_{X^*}\frac{\sqrt{\bar{\kappa}_1}R_{q_1}^{\frac{1}{2}}}{\underline{\kappa}_1^{1-\frac{q_1}{2}}}\left(\sqrt{\frac{\rho_Z^2\rho_\eta^2\log(d\vee p)}{n}}\right)^{1-\frac{q_1}{2}}, \qquad (34)
\end{aligned}
$$

where $\sigma_{X^*} = \max_{j=1,...,p}\sigma_{X_j^*}$. To bound the term $\left|\frac{(\hat{X}-X^*)^T(\hat{X}-X^*)}{n}\right|_\infty$, we again apply a Cauchy-Schwarz inequality and obtain

$$
\left|\frac{(\hat{X}-X^*)^T(\hat{X}-X^*)}{n}\right|_\infty \leq c'\frac{\bar{\kappa}_1 R_{q_1}}{\underline{\kappa}_1^{2-q_1}}\left(\rho_Z^2\rho_\eta^2\frac{\log(d\vee p)}{n}\right)^{1-\frac{q_1}{2}} \qquad (35)
$$

with probability at least $1-c_1\exp(-c_2\log(p\vee d))$. Putting everything together, if $n \geq c' R_{q_1}^{\frac{2}{2-q_1}}\log(d\vee p)$ for some sufficiently large constant $c' > 0$, we have

$$
\left|\frac{\hat{X}^T\hat{X} - X^{*T}X^*}{n}\right|_\infty \leq c\sigma_{X^*}\frac{\sqrt{\bar{\kappa}_1}R_{q_1}^{\frac{1}{2}}}{\underline{\kappa}_1^{1-\frac{q_1}{2}}}\left(\sqrt{\frac{\rho_Z^2\rho_\eta^2\log(d\vee p)}{n}}\right)^{1-\frac{q_1}{2}}.
$$

The bound above implies

$$
\mathbb{P}\left(\max_j\left|\frac{1}{n}\hat{X}_j^T\hat{X}_j - \sigma_{X_j^*}^2\right| \leq \sigma_{X^*}\mathcal{T}_1\right) \geq 1 - O\left(\frac{1}{d\vee p}\right), \qquad (36)
$$

as long as $n \geq c' R_{q_1}^{\frac{2}{2-q_1}}\log(d\vee p)$ for some sufficiently large constant $c' > 0$. □

Remark. In the rest of proofs, we assume the regressors $\hat{X}_j$ $(j = 1,...,p)$ are normalized such that $\sigma_{X_j^*} = 1$. So long as $\mathcal{T}_1 \leq 1$, (36) implies that

$$
\mathbb{P}\left(\max_j\left|\sqrt{\frac{1}{n}\hat{X}_j^T\hat{X}_j} - 1\right| \leq 1\right) \geq 1 - O\left(\frac{1}{d\vee p}\right). \qquad (37)
$$

A.3 Theorem 3.1

To apply Lemma A.1 to show Theorem 3.1, we need to show Lemmas A.4 and A.5.

Lemma A.4 (RE condition): Under the conditions in Lemma 3.1, we have

$$
\frac{|\hat{X}v^0|_2^2}{n} \geq \frac{\underline{\kappa}_2}{2}\left|v^0\right|_2^2 - c_0\underline{\kappa}_2\left(\frac{1}{\underline{\kappa}_2^2}\vee 1\right)\frac{\log p}{n}|v^0|_1^2 - \mathcal{T}_1\left|v^0\right|_1^2, \qquad (38)
$$

for any v^0 in the restricted set subject to (29), with probability at least $1 - c_1 \exp(-c_2 \log(p \vee d))$.

Proof. Note that

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

From (33), we have

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

Using (34) and (35), under the condition $n \geq c^{'} R_{q_1}^{\frac{2}{2-q_1}} \log(d \vee p)$ for some sufficiently large $c^{'} > 0$, and applying (40) in Lemma B.3 with $s = \frac{1}{c_0} \frac{n}{\log p} (\underline{\kappa}_2^2 \wedge 1)$, $U = X^*$, and $\underline{\kappa} = \underline{\kappa}_2$, we have

$$\begin{aligned} \left| v^{0T} \frac{\hat{X}^T \hat{X}}{n} v^0 \right| &\geq \left| v^{0T} \frac{X^{*T} X^*}{n} v^0 \right| - c^{'} \frac{\sqrt{\bar{\kappa}_1} R_{q_1}^{\frac{1}{2}}}{\underline{\kappa}_1^{1-\frac{q_1}{2}}} \left(\sqrt{\frac{\rho_Z^2 \rho_\eta^2 \log(d \vee p)}{n}} \right)^{1-\frac{q_1}{2}} \left| v^0 \right|_1^2 \\ &\geq \frac{\underline{\kappa}_2}{2} \left| v^0 \right|_2^2 - c_0 \underline{\kappa}_2 \left(\frac{1}{\underline{\kappa}_2^2} \vee 1 \right) \frac{\log p}{n} |v^0|_1^2 - c_1 \frac{\sqrt{\bar{\kappa}_1} R_{q_1}^{\frac{1}{2}}}{\underline{\kappa}_1^{1-\frac{q_1}{2}}} \left(\sqrt{\frac{\rho_Z^2 \rho_\eta^2 \log(d \vee p)}{n}} \right)^{1-\frac{q_1}{2}} \left| v^0 \right|_1^2 \end{aligned}$$

for any v^0 in the restricted set subject to (29), with probability at least $1 - c_2 \exp(-c_3 \log(p \vee d))$. Notice the last inequality can be written in the form of (38).$\square$

Lemma A.5 (Upper bound on $|\frac{1}{n} \hat{D}^{-1} \hat{X}^T e|_\infty$): Under the conditions for Lemma 3.1, we have

$$\left| \hat{D}^{-1} \frac{\hat{X}^T e}{n} \right|_\infty \leq \mathcal{T}_0,$$

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

Proof. By (36), we have $\left| \hat{D}^{-1} \frac{\hat{X}^T e}{n} \right|_\infty \leq c^{'} \left| D^{-1} \frac{\hat{X}^T e}{n} \right|_\infty$, where $D = \text{diag}\left[\sigma_{X_1^*}, \ldots, \sigma_{X_p^*} \right] = \text{diag}\left[1\right]$ and $c^{'} > 1$. Furthermore,

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

Hence,

$$|\frac{1}{n}\hat{X}^{T}e|_{\infty} \leq |\frac{1}{n}\hat{X}^{T}(\hat{X}-X^{*})\beta^{*}|_{\infty}+|\frac{1}{n}X^{*T}\eta\beta^{*}|_{\infty}+|\frac{1}{n}X^{*T}\epsilon|_{\infty} \tag{39}$$
$$+|\frac{1}{n}(\hat{X}-X^{*})^{T}\eta\beta^{*}|_{\infty}+|\frac{1}{n}(\hat{X}-X^{*})^{T}\epsilon|_{\infty}.$$

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

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

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

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

We apply Lemma A.3 and a Cauchy-Schwarz inequality to bound $\left|\frac{\hat{X}^{T}(\hat{X}-X^{*})}{n}\right|_{\infty}$ and obtain

$$\begin{aligned}\max_{j^{'},j}\left|\frac{1}{n}\hat{X}_{j'}^{T}(\hat{X}_{j}-X_{j}^{*})\right| &= \max_{j^{'},j}\left|\frac{1}{n}\sum_{i=1}^{n}\hat{X}_{ij'}Z_{ij}(\hat{\pi}_{j}-\pi_{j}^{*})\right|\\ &\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{X}_{ij'}^{2}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[Z_{ij}(\hat{\pi}_{j}-\pi_{j}^{*})\right]^{2}}\\ &\leq c_{1}\frac{\sqrt{\bar{\kappa}_{1}}R_{q_{1}}^{\frac{1}{2}}}{\underline{\kappa}_{1}^{1-\frac{q_{1}}{2}}}\left(\sqrt{\frac{\rho_{Z}^{2}\rho_{\eta}^{2}\log(d\vee p)}{n}}\right)^{1-\frac{q_{1}}{2}}.\end{aligned}$$

The last inequality follows because we normalize $\hat{X}_{ij'}$ for $j^{'}=1,...,p$ so that $\max_{j^{'}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{X}_{ij'}^{2}}\leq 1$. Consequently,

$$|\frac{1}{n}\hat{X}^{T}(\hat{X}-X^{*})\beta^{*}|_{\infty}\leq c_{1}|\beta^{*}|_{1}\frac{\sqrt{\bar{\kappa}_{1}}R_{q_{1}}^{\frac{1}{2}}}{\underline{\kappa}_{1}^{1-\frac{q_{1}}{2}}}\left(\sqrt{\frac{\rho_{Z}^{2}\rho_{\eta}^{2}\log(d\vee p)}{n}}\right)^{1-\frac{q_{1}}{2}},$$

with probability at least $1-c_{1}^{'}\exp(-c_{2}^{'}\log(p\vee d))$. For the term $|\frac{1}{n}X^{*T}\eta\beta^{*}|_{\infty}$, we have

$$\begin{aligned}|\frac{1}{n}X^{*T}\eta\beta^{*}|_{\infty} &\leq \max_{j^{'},j}|\frac{1}{n}\sum_{i=1}^{n}X_{ij'}^{*}\eta_{ij}||\beta^{*}|_{1}\\ &\leq c_{2}\rho_{X^{*}}\rho_{\eta}|\beta^{*}|_{1}\sqrt{\frac{\log p}{n}}\end{aligned}$$

with probability at least $1 - c_1^{'} \exp(-c_2^{'} \log p)$. The last inequality follows from Lemma B.1 and Assumption 2.1 that $\mathbb{E}(Z_{ij^{'}}\eta_{ij}) = \mathbf{0}$ for all $j^{'}$, j as well as Assumptions 3.1 and 3.2. For the term $|\frac{1}{n}(X^* - \hat{X})^T \eta \beta^*|_\infty$, applying (31) to bound $\max_{j^{'}} |\hat{\pi}_{j^{'}} - \pi^*_{j^{'}}|_1$ and applying Lemma B.1 to bound $\max_{j^{'},j} |\frac{1}{n}\sum_{i=1}^n Z_{ij^{'}}^T \eta_{ij}|_\infty$ by setting $t = \sqrt{\frac{\rho_Z^2 \rho_\eta^2 \log(d \vee p)}{n}}$ yields

$$\begin{aligned} |\frac{1}{n}(X^* - \hat{X})^T \eta \beta^*|_\infty &\leq \max_{j^{'}} |\hat{\pi}_{j^{'}} - \pi^*_{j^{'}}|_1 \max_{j^{'},j} |\frac{1}{n}\sum_{i=1}^n Z_{ij^{'}}^T \eta_{ij}|_\infty |\beta^*|_1 \\ &\leq c_3 |\beta^*|_1 \sqrt{\bar{\kappa}_1} \underline{\kappa}_1^{-1+q_1} R_{q_1} \left(\frac{\rho_Z^2 \rho_\eta^2 \log(d \vee p)}{n} \right)^{1-\frac{q_1}{2}}, \end{aligned}$$

with probability at least $1 - c_1^{'} \exp(-c_2^{'} \log(p \vee d))$. To bound the term $|\frac{1}{n} X^{*T} \epsilon|_\infty$, note under Assumptions 3.1 and 3.2 as well as Assumption 2.1, again by Lemma B.1,

$$|\frac{1}{n} X^{*T} \epsilon|_\infty \leq c_2 \rho_{X^*} \rho_\epsilon \sqrt{\frac{\log p}{n}},$$

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 apply similar techniques used for bounding $|\frac{1}{n}(X^* - \hat{X})^T \eta \beta^*|_\infty$ and obtain

$$|\frac{1}{n}(X^* - \hat{X})^T \epsilon|_\infty \leq c_4 \sqrt{\bar{\kappa}_1} \underline{\kappa}_1^{-1+q_1} R_{q_1} \rho_\epsilon \rho_\eta^{1-q_1} \left(\frac{\rho_Z^2 \log(d \vee p)}{n} \right)^{1-\frac{q_1}{2}}$$

with probability at least $1 - c_1^{'} \exp(-c_2^{'} \log(p \vee d))$. Putting everything together, as long as

$$\begin{aligned} c_3^{'} \underline{\kappa}_1^{\frac{q}{2}} R_{q_1}^{\frac{1}{2}} \left(\sqrt{\frac{\log(d \vee p)}{n}} \right)^{1-\frac{q_1}{2}} &\leq 1, \\ c_4^{'} |\beta^*|_1^{-1} \underline{\kappa}_1^{\frac{q}{2}} R_{q_1}^{\frac{1}{2}} \left(\sqrt{\frac{\log(d \vee p)}{n}} \right)^{1-\frac{q_1}{2}} &\leq 1, \end{aligned}$$

for some constants $c_3^{'} > 0$ and $c_4^{'} > 0$ depending on ρ_Z, ρ_η, and ρ_ϵ, the claim in Lemma A.5 follows. □

Now, by applying Lemma A.1 and setting λ_n according to (9), we obtain

$$|\hat{\beta}_{H2SLS} - \beta^*|_2 \leq \frac{c R_{q_2}^{\frac{1}{2}}}{\underline{\kappa}_2^{1-\frac{q_2}{2}}} \mathcal{T}_0^{1-\frac{q_2}{2}}$$

with probability at least $1 - O\left(\frac{1}{p}\right)$.□

A.4 Theorem 3.2

The verification of the RE condition for $\frac{\hat{X}^T \hat{X}}{n}$ in Theorem 3.2 is done via Lemma A.6.

Lemma A.6 (RE condition): Let $r \in [0, 1]$. Under Assumptions 2.1, 3.1, 3.3, 3.4, and the condition $n \geq c_0 k_1 \log(p \vee d)$ for some sufficiently large positive constant c_0, we have,

$$\frac{|\hat{X}v^0|_2^2}{n} \geq \frac{\underline{\kappa}_W}{2}|v^0|_2^2 - c\underline{\kappa}_W \frac{k_1 \log(p \vee d)}{n}|v^0|_1^2,$$

for any v^0 in the restricted set subject to $\mathbb{C}(J(\beta^*);\, 0,\, c^*)$ for some universal constant $c^* > 1$, with probability at least $1 - c_1 \exp(-c_2 \log(p \vee d))$.

Proof. Under Assumption 3.4, we have $|J(\hat{\pi}_j)| \leq c^0 k_1$ for some universal constant $c^0 > 0$. To bound $\left|v^{0T}\frac{\hat{X}^T\hat{X}}{n}v^0\right|$, I apply a discretization argument motivated by the idea in Loh and Wainwright (2012). This type of argument is often used in statistical problems requiring manipulating and controlling collections of random variables indexed by sets with an infinite number of elements. For the particular problem in this paper, I work with the space $\Omega = \mathbb{K}(2s,\, p,\, 1) \times \mathbb{K}^2(c^0 k_1,\, d_1,\, R) \times \ldots \times \mathbb{K}^2(c^0 k_1,\, d_p,\, R)$ where $d_j = d$ for all $j = 1, ..., p$. For $s \geq 1$ and $L \geq 1$, recall the notation $\mathbb{K}(s,\, L,\, R) := \{v \in \mathbb{R}^L \,|\, |v|_2 \leq R,\, |v|_0 \leq s\}$. Given $V^j \subseteq \{1, ..., d\}$ and $V^0 \subseteq \{1, ..., p\}$, define $S_{V^j} = \{v \in \mathbb{R}^d : |v|_2 \leq R,\, J(v) \subseteq V^j\}$ and $S_{V^0} = \{v \in \mathbb{R}^p : |v|_2 \leq 1,\, J(v) \subseteq V^0\}$. Note that $\mathbb{K}(c^0 k_1,\, d,\, R) = \cup_{|V^j| \leq c^0 k_1} S_{V^j}$ and $\mathbb{K}(2s,\, p,\, 1) = \cup_{|V^0| \leq 2s} S_{V^0}$. If $\mathcal{V}^j = \{t_1^j, ..., t_{m_j}^j\}$ is a $\frac{R}{9}-$cover of S_{V^j} ($\mathcal{V}^0 = \{t_1^0, ..., t_{m_0}^0\}$ is a $\frac{1}{9}-$cover of S_{V^0}), for every $v^j \in S_{V^j}$ ($v^0 \in S_{V^0}$), we can find some $t_i^j \in \mathcal{V}^j$ ($t_{i'}^0 \in \mathcal{V}^0$) such that $|\triangle v^j|_2 \leq \frac{R}{9}$ ($|\triangle v^0|_2 \leq \frac{1}{9}$), 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 $|\mathcal{V}^j| \leq 81^{c^0 k_1}$ and $|\mathcal{V}^0| \leq 81^{2s}$. Therefore, for $v^0 \in \mathbb{K}(2s,\, p,\, 1)$, there is some S_{V^0} and $t_i^0 \in \mathcal{V}^0$ such that

$$\begin{aligned} v^{0T}\frac{\hat{X}^T\hat{X}}{n}v^0 &= (t_i^0 + v^0 - t_i^0)^T\frac{\hat{X}^T\hat{X}}{n}(t_i^0 + v^0 - t_i^0) \\ &= t_i^{0T}\frac{\hat{X}^T\hat{X}}{n}t_i^0 + 2\triangle v^{0T}\frac{\hat{X}^T\hat{X}}{n}t_i^0 + \triangle v^{0T}\frac{\hat{X}^T\hat{X}}{n}\triangle v^0 \end{aligned}$$

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

$$\frac{1}{n}\hat{X}_{j'}^T\hat{X}_j = \frac{1}{n}\sum_{i=1}^n \hat{\pi}_{j'}^T Z_{ij'}^T Z_{ij}\hat{\pi}_j.$$

Notice that, under Assumption 3.4, $|J(\hat{\pi}_j)| \leq c^0 k_1$ for every $j = 1, ..., p$ and as long as $n \geq c_0 k_1 \log(p \vee d)$ for some sufficiently large constant $c_0 > 0$, (so that by (30) specialized to $q_1 = 0$, $\max_j |\hat{\pi}_j|_2 \leq R := 2\max_{j=1,...,p}\left|\pi_j^*\right|_2$), we have $\hat{\pi}_j \in \mathbb{K}(c^0 k_1,\, d,\, R) = \cup_{|V^j| \leq c^0 k_1} S_{V^j}$. Therefore, there are some S_{V^j} and $S_{V^{j'}}$ with $|V^j| \leq c^0 k_1$ and $|V^{j'}| \leq c^0 k_1$, $t_{i'}^j \in \mathcal{V}^j$ and $t_{i''}^{j'} \in \mathcal{V}^{j'}$ (where $\mathcal{V}^j = \{t_1^j, ..., t_{m_j}^j\}$ is a $\frac{R}{9}-$cover of S_{V^j} and $\mathcal{V}^{j'} = \{t_1^{j'}, ..., t_{m_{j'}}^{j'}\}$ is a $\frac{R}{9}-$cover of $S_{V^{j'}}$) such that

$$\begin{aligned} \frac{1}{n}\hat{\pi}_{j'}^T Z_{j'}^T Z_j \hat{\pi}_j^T &= (t_{i''}^{j'} + \hat{\pi}_{j'}^T - t_{i''}^{j'})^T\frac{Z_{j'}^T Z_j}{n}(t_{i'}^j + \hat{\pi}_j^T - t_{i'}^j) \\ &= t_{i''}^{j'T}\frac{Z_{j'}^T Z_j}{n}t_{i'}^j + t_{i''}^{j'T}\frac{Z_{j'}^T Z_j}{n}\triangle v^j + \triangle v^{j'T}\frac{Z_{j'}^T Z_j}{n}t_{i'}^j + \triangle v^{j'T}\frac{Z_{j'}^T Z_j}{n}\triangle v^j \end{aligned}$$

with $|\triangle v^j|_2 \le \frac{R}{9}$ and $|\triangle v^{j'}|_2 \le \frac{R}{9}$. Denote a matrix A by $\left[A_{j'j}\right]_M$, where the $(j',\, j)$ element of A is $A_{j'j}$, and let $A_{j'j} = \frac{Z_{j'}^T Z_j}{n} - \mathbb{E}\left(\frac{Z_{j'}^T Z_j}{n}\right)$ Define $v = (v^0,\, v^1,\, ...,\, v^p) \in S_V := S_{V^0} \times S^2_{V^1} \times ... \times S^2_{V^p}$. Hence,

$$\left|v^{0T}\left[\frac{\hat{X}_{j'}^T\hat{X}_j}{n} - \mathbb{E}\frac{\hat{X}_{j'}^T\hat{X}_j}{n}\right]_M v^0\right| \le \left|v^{0T}\left[v^{j'T}A_{j'j}v^j\right]_M v^0\right|$$

$$\begin{aligned}
\le\ & \max_{i'',i',i}\left|t_i^{0T}\left[t_{i''}^{j'T}A_{j'j}t_{i'}^j\right]_M t_i^0\right| + \sup_{v\in S_V}\left|t_i^{0T}\left[t_{i''}^{j'T}A_{j'j}\triangle v^j\right]_M t_i^0\right| \\
& + \sup_{v\in S_V}\left|t_i^{0T}\left[\triangle v^{j'T}A_{j'j}t_{i'}^j\right]_M t_i^0\right| + \sup_{v\in S_V}\left|t_i^{0T}\left[\triangle v^{j'T}A_{j'j}\triangle v^j\right]_M t_i^0\right| \\
& + \sup_{v\in S_V} 2\left|\triangle v^{0T}\left[t_{i''}^{j'T}A_{j'j}t_{i'}^j\right]_M t_i^0\right| + \sup_{v\in S_V} 2\left|\triangle v^{0T}\left[t_{i''}^{j'T}A_{j'j}\triangle v^j\right]_M t_i^0\right| \\
& + \sup_{v\in S_V} 2\left|\triangle v^{0T}\left[\triangle v^{j'T}A_{j'j}t_{i'}^j\right]_M t_i^0\right| + \sup_{v\in S_V} 2\left|\triangle v^{0T}\left[\triangle v^{j'T}A_{j'j}\triangle v^j\right]_M t_i^0\right| \\
& + \sup_{v\in S_V}\left|\triangle v^{0T}\left[t_{i''}^{j'T}A_{j'j}t_{i'}^j\right]_M \triangle v^0\right| + \sup_{v\in S_V}\left|\triangle v^{0T}\left[t_{i''}^{j'T}A_{j'j}\triangle v^j\right]_M \triangle v^0\right| \\
& + \sup_{v\in S_V}\left|\triangle v^{0T}\left[\triangle v^{j'T}A_{j'j}t_{i'}^j\right]_M \triangle v^0\right| + \sup_{v\in S_V}\left|\triangle v^{0T}\left[\triangle v^{j'T}A_{j'j}\triangle v^j\right]_M \triangle v^0\right| \\
\le\ & \max_{i'',i',i}\left|t_i^{0T}\left[t_{i''}^{j'T}A_{j'j}t_{i'}^j\right]_M t_i^0\right| + \frac{1}{9}\sup_{v\in S_V}\left|v^{0T}\left[v^{j'T}A_{j'j}v^j\right]_M v^0\right| \\
& + \frac{1}{9}\sup_{v\in S_V}\left|v^{0T}\left[v^{j'T}A_{j'j}v^j\right]_M v^0\right| + \frac{1}{81}\sup_{v\in S_V}\left|v^{0T}\left[v^{j'T}A_{j'j}v^j\right]_M v^0\right| \\
& + \frac{1}{9}\sup_{v\in S_V} 2\left|v^{0T}\left[v^{j'T}A_{j'j}v^j\right]_M v^0\right| + \frac{1}{81}\sup_{v\in S_V} 2\left|v^{0T}\left[v^{j'T}A_{j'j}v^j\right]_M v^0\right| \\
& + \frac{1}{81}\sup_{v\in S_V} 2\left|v^{0T}\left[v^{j'T}A_{j'j}v^j\right]_M v^0\right| + \frac{1}{729}\sup_{v\in S_V} 2\left|v^{0T}\left[v^{j'T}A_{j'j}v^j\right]_M v^0\right| \\
& + \frac{1}{81}\sup_{v\in S_V}\left|v^{0T}\left[v^{j'T}A_{j'j}v^j\right]_M v^0\right| + \frac{1}{729}\sup_{v\in S_V}\left|v^{0T}\left[v^{j'T}A_{j'j}v^j\right]_M v^0\right| \\
& + \frac{1}{729}\sup_{v\in S_V}\left|v^{0T}\left[v^{j'T}A_{j'j}v^j\right]_M v^0\right| + \frac{1}{6561}\sup_{v\in S_V}\left|v^{0T}\left[v^{j'T}A_{j'j}v^j\right]_M v^0\right|
\end{aligned}$$

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

$$\begin{aligned}
\sup_{v\in S_V}\left|v^{0T}\left[v^{j'T}A_{j'j}v^j\right]_M v^0\right| &\le \frac{6561}{3122}\max_{i'',i',i}\left|t_i^{0T}\left[t_{i''}^{j'T}A_{j'j}t_{i'}^j\right]_M t_i^0\right| \\
&\le 3\max_{i'',i',i}\left|t_i^{0T}\left[t_{i''}^{j'T}A_{j'j}t_{i'}^j\right]_M t_i^0\right|.
\end{aligned}$$

Under Assumption 3.3, $Z_j t_i^j = W_j$ is a sub-Gaussian vector with parameter at most ρ_{W^*}. An application of Lemma B.1 and a union bound yields

$$\mathbb{P}\left(\sup_{v\in S_V}\left|v^{0T}\left[v^{j'T}A_{j'j}v^j\right]_M v^0\right|\geq t\right)\leq 81^{2sc^0k_1}81^{2s}2\exp(-cn\min(\frac{t^2}{\rho_{X^*}^2\rho_W^2},\ \frac{t}{\rho_{X^*}\rho_W})),$$

where the exponent $2sc^0k_1$ in $81^{2sc^0k_1}$ 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 $\binom{d}{\lfloor c^0k_1\rfloor}\leq d^{c^0k_1}$ choices of V^j and respectively, the $\binom{p}{\lfloor 2s\rfloor}\leq p^{2s}$ choices of V^0 yields

$$\begin{aligned}\mathbb{P}\left(\sup_{v\in\Omega}\left|v^{0T}\left[v^{j'T}A_{j'j}v^j\right]_M v^0\right|\geq t\right) &\leq p^{2s}d^{2sc^0k_1}\cdot 2\exp(-cn\min(\frac{t^2}{\rho_{X^*}^2\rho_W^2},\ \frac{t}{\rho_{X^*}\rho_W}))\\ &\leq 2\exp(-cn\min(\frac{t^2}{\rho_{X^*}^2\rho_W^2},\ \frac{t}{\rho_{X^*}\rho_W})+2sc^0k_1\log d+2s\log p).\end{aligned}$$

With the choice of $s=\frac{c'n}{k_1\log(p\vee d)}(\underline{\kappa}_2^2\wedge 1)$ and $t=\frac{\underline{\kappa}_W}{54}$ for some sufficiently large universal constant $c'\geq 1$, we have

$$\left|v^{0T}\left[\frac{\hat{X}_{j'}^T\hat{X}_j}{n}-\mathbb{E}\frac{\hat{X}_{j'}^T\hat{X}_j}{n}\right]_M v^0\right|\leq\frac{\underline{\kappa}_W}{54}$$

with probability at least $1-c_1'\exp(-c_2'n)-c_1''\exp(-c_2''\log(p\vee d))=1-c_1\exp(-c_2\log(p\vee d))$ provided $n\geq c\log(p\vee d)$ for some sufficiently large constant $c>0$. Under Assumption 3.3, applying Lemma B.2 with $\Gamma=\frac{\hat{X}^T\hat{X}}{n}-\mathbb{E}\left(\frac{\hat{X}^T\hat{X}}{n}\right)$ and (40) in Lemma B.3 with the choice $s=\frac{c'n}{k_1\log(p\vee d)}(\underline{\kappa}_2^2\wedge 1)$, we have

$$\begin{aligned}v^{0T}\left[\frac{\hat{X}_{j'}^T\hat{X}_j}{n}\right]_M v^0 &\geq \frac{\underline{\kappa}_W}{2}\left|v^0\right|_2^2-\frac{\underline{\kappa}_W}{2s}\left|v^0\right|_1^2\\ &\geq \frac{\underline{\kappa}_W}{2}\left|v^0\right|_2^2-c'\underline{\kappa}_W\frac{k_1\log(p\vee d)}{2n}\left|v^0\right|_1^2\end{aligned}$$

for all $v^0\in\mathbb{C}(J(\beta^*);\,0,\,c^*)$. □

Recalling in proving Lemma A.1, for exactly sparse β^* (i.e., $q_2=0$), upon our choice λ_n, we have shown

$$\hat{v}=\hat{\beta}_{H2SLS}-\beta^*\in\mathbb{C}(J(\beta^*);\,0,\,c^*),$$

and $|\hat{v}^0|_1^2\leq c_0|\hat{v}^0_{J(\beta^*)}|_1^2\leq c_0k_2|\hat{v}^0_{J(\beta^*)}|_2^2$. Therefore, if $n\geq c_1k_1k_2\log(p\vee d)$ for some sufficiently large c_1, then,

$$\left|\hat{v}^{0T}\frac{\hat{X}^T\hat{X}}{n}\hat{v}^0\right|\geq c_2\underline{\kappa}_W\left|\hat{v}^0\right|_2^2.$$

The above inequality implies RE on $\frac{\hat{X}^T\hat{X}}{n}$. □

A.5 Proof for Theorem 3.3

Proof. Note that $|\hat{\beta}-\beta^*|_\infty \leq |\hat{\beta}-\beta^*|_2 \leq B$, which implies that $-B+\beta_j^* \leq \hat{\beta}_j \leq B+\beta_j^*$. Given $B < \min_{j\in J(\beta^*)} |\beta_j^*|$, for $j \in J(\beta^*)$, if $\beta_j^* > 0$, then the left inequality ensures that $\hat{\beta}_j > 0$ and on the other hand if $\beta_j^* < 0$, then the right inequality ensures that $\hat{\beta}_j < 0$. In either case, we must have $J(\hat{\beta}_{H2SLS}) \supseteq J(\beta^*)$. To show the correct inclusion of the thresholded estimator, note that $\max_{j\notin J(\beta^*)} |\hat{\beta}_j| \leq B < B_1$. Because the thresholded estimator $\bar{\beta}$ excludes all components smaller than B_1, we must have $J(\bar{\beta}) \subseteq J(\beta^*)$. $\square$

A.6 Main proofs for Theorem 3.4

The proof for Theorem 3.4 is based on a construction called Primal-Dual Witness (PDW) method developed by Wainwright (2009). This method constructs a pair $(\hat{\beta},\, \hat{\mu})$. When this procedure succeeds, the constructed pair is primal-dual optimal, and acts as a witness for the fact that the Lasso has a unique optimal solution with the correct signed support. The procedure is described in the following.

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 $|\cdot|_1 : \mathbb{R}^{k_2} \to \mathbb{R}$.

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.

We let $J(\beta^*) := K$, $J(\beta^*)^c := K^c$, $\Sigma_{K^cK} := \mathbb{E}\left[\frac{1}{n}X_{K^c}^{*T}X_K^*\right]$, $\hat{\Sigma}_{K^cK} := \frac{1}{n}X_{K^c}^{*T}X_K^*$, and $\tilde{\Sigma}_{K^cK} := \frac{1}{n}\hat{X}_{K^c}^T\hat{X}_K$. Similarly, let $\Sigma_{KK} := \mathbb{E}\left[\frac{1}{n}X_K^{*T}X_K^*\right]$, $\hat{\Sigma}_{KK} := \frac{1}{n}X_K^{*T}X_K^*$, and $\tilde{\Sigma}_{KK} := \frac{1}{n}\hat{X}_K^T\hat{X}_K$. The proof for the first claim in Theorem 3.4 is established in Lemma A.7, which shows that $\hat{\beta}_{H2SLS} = (\hat{\beta}_K,\, \mathbf{0})$ where $\hat{\beta}_K$ is the solution obtained in step 2 of the PDW construction. The second and third claims are proved using Lemma A.8. The last claim is a consequence of the third claim (which can be shown in the similar way as the proof for the first part of Theorem 3.3).

Lemma A.7: If the PDW construction succeeds and if $\lambda_{\min}(\Sigma_{KK}) \geq C_{\min} > 0$, then the vector $(\hat{\beta}_K,\, \mathbf{0}) \in \mathbb{R}^p$ is the unique optimal solution of the Lasso.

Proof. The proof for Lemma A.7 adopts the proof for Lemma 1 from Chapter 6.4.2 of Wainwright (2015). If the PDW construction succeeds, then $\hat{\beta} = (\hat{\beta}_K,\, \mathbf{0})$ is an optimal solution with subgradient $\hat{\mu} \in \mathbb{R}^p$ and $|\hat{\mu}_{K^c}|_\infty < 1$, $\left\langle \hat{\mu},\, \hat{\beta} \right\rangle = |\hat{\beta}|_1$. Suppose $\tilde{\beta}$ is another optimal solution. Letting

$F(\beta) = \frac{1}{2n}|Y - \hat{X}\beta|_2^2$, then $F(\hat{\beta}) + \lambda_n \left\langle \hat{\mu},\, \hat{\beta} \right\rangle = F(\tilde{\beta}) + \lambda_n |\tilde{\beta}|_1$ and $F(\hat{\beta}) - \lambda_n \left\langle \hat{\mu},\, \tilde{\beta} - \hat{\beta} \right\rangle = F(\tilde{\beta}) + \lambda_n \left(|\tilde{\beta}|_1 - \left\langle \hat{\mu},\, \tilde{\beta} \right\rangle \right)$. However, by the zero-subgradient[1] optimality conditions, $\lambda_n \hat{\mu} = -\nabla F(\hat{\beta})$, so that $F(\hat{\beta}) + \left\langle \nabla F(\hat{\beta}),\, \tilde{\beta} - \hat{\beta} \right\rangle - F(\tilde{\beta}) = \lambda_n \left(|\tilde{\beta}|_1 - \left\langle \hat{\mu},\, \tilde{\beta} \right\rangle \right)$. Convexity of F ensures that the left-hand side is non-positive and consequently $|\tilde{\beta}|_1 \leq \left\langle \hat{\mu},\, \tilde{\beta} \right\rangle$. On the other hand, since $\left\langle \hat{\mu},\, \tilde{\beta} \right\rangle \leq |\hat{\mu}|_\infty |\tilde{\beta}|_1$, we must have $|\tilde{\beta}|_1 = \left\langle \hat{\mu},\, \tilde{\beta} \right\rangle$. Given $|\hat{\mu}_{K^c}|_\infty < 1$, this equality can only hold if $\tilde{\beta}_j = 0$ for all $j \in K^c$. Therefore, all optimal solutions must have the same support K and can be obtained by solving the oracle subproblem in the PDW procedure. The bound $\lambda_{\min}(\tilde{\Sigma}_{KK}) \geq c\lambda_{\min}(\hat{\Sigma}_{KK}) \geq c(1 - c')\lambda_{\min}(\Sigma_{KK})$ for some $c,\, c' \in (0,\, 1)$ (inequalities (7) and (13) of Section S.1 from the proofs for Lemma S.2 and S.3) and the condition $\lambda_{\min}(\Sigma_{KK}) \geq C_{\min} > 0$ ensures that this subproblem is strictly convex and has a unique minimizer. □

Lemma A.8: Suppose the assumptions in Theorem 3.4 hold. Then, with probability at least $1 - O\left(\frac{1}{p}\right)$: (i) $|\hat{\mu}_{K^c}|_\infty \leq 1 - \frac{\varsigma\phi}{\bar{c}-1}$ for some universal constant $\bar{c} > 2$ and any small number $\varsigma > 0$; (ii)

$$|\hat{\beta}_{H2SLS,J(\beta^*)} - \beta^*_{H2SLS,J(\beta^*)}|_\infty \leq \lambda_n \left[\frac{(\bar{c} - 2 - \varsigma)\phi}{\left(2 - \frac{(\bar{c}-2)\phi}{(\bar{c}-1)}\right)(\bar{c} - 1)} + 1 \right] \left\| \left(\frac{\hat{X}^T_{J(\beta^*)}\hat{X}_{J(\beta^*)}}{n} \right)^{-1} \right\|_\infty = B_2,$$

where, for some constant $c'' > 1$,

$$\left\| \left(\frac{\hat{X}^T_{J(\beta^*)}\hat{X}_{J(\beta^*)}}{n} \right)^{-1} \right\|_\infty \leq \frac{c''\sqrt{k_2}}{\lambda_{\min}\left(\mathbb{E}\left[\frac{1}{n} X^{*T}_{J(\beta^*)} X^*_{J(\beta^*)} \right] \right)}.$$

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

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

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

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

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

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

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

where the fact that $|\hat{\mu}_K|_\infty \leq 1$ is used in the inequality above. By Lemma S.1, $\left\|\tilde{\Sigma}_{K^cK}\tilde{\Sigma}_{KK}^{-1}\right\|_\infty \leq 1 - \frac{(\bar{c}-2)\phi}{(\bar{c}-1)}$ with probability at least $1 - O\left(\frac{1}{p}\right)$. Hence,

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

Therefore, it suffices to show that $\left(2 - \frac{(\bar{c}-2)\phi}{(\bar{c}-1)}\right)\left|\hat{X}^T\frac{e}{n\lambda_n}\right|_\infty \leq \frac{(\bar{c}-2-\varsigma)\phi}{\bar{c}-1}$ with high probability, for any small number $\varsigma > 0$. This result holds if $\lambda_n \geq \frac{\left(2-\frac{(\bar{c}-2)\phi}{(\bar{c}-1)}\right)(\bar{c}-1)}{(\bar{c}-2-\varsigma)\phi}\mathcal{T}_0$ with $\mathcal{T}_0$ defined in (5), Thus, we have $|\hat{\mu}_{K^c}|_\infty \leq 1 - \frac{\varsigma\phi}{\bar{c}-1}$ with probability at least $1 - O\left(\frac{1}{p}\right)$. It remains to establish a bound on the l_∞−norm of the error $\hat{\beta}_K - \beta_K^*$. By the triangle inequality, we have

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

Using the following bound (inequality (14) of Section S.1) from the proof for Lemma S.3:

$$\left\|\left(\frac{\hat{X}_K^T\hat{X}_K}{n}\right)^{-1}\right\|_\infty \leq \frac{c'\sqrt{k_2}}{\lambda_{\min}(\hat{\Sigma}_{KK})} \leq \frac{c''\sqrt{k_2}}{\lambda_{\min}(\Sigma_{KK})}$$

for some $c'' > c' > 1$, and putting everything together with the choice of λ_n stated in Theorem 3.4 yields claim (ii). □

B Technical lemmas

Lemma B.1: If $X \in \mathbb{R}^{n\times p_1}$ is a sub-Gaussian matrix with parameters (Σ_X, ρ_X^2) and each row is sampled independently, then for any fixed (unit) vector $v \in \mathbb{R}^{p_1}$, we have

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

Moreover, if $Y \in \mathbb{R}^{n\times p_2}$ is a sub-Gaussian matrix with parameters $(\Sigma_Y,\, \rho_Y^2)$ and each row is sampled independently, then

$$\mathbb{P}(|Y^TX - \mathbb{E}(Y^TX)|_\infty \geq nt) \leq 6p_1p_2 \exp(-cn \min\{\frac{t^2}{\rho_X^2\rho_Y^2},\, \frac{t}{\rho_X\rho_Y}\}),$$

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

$$\mathbb{P}(|\frac{Y^TX}{n} - \mathbb{E}(\frac{Y^TX}{n})|_\infty \geq c_0\rho_X\rho_Y\sqrt{\frac{\log(p_1 \vee p_2)}{n}}) \leq c_1 \exp(-c_2 \log(p_1 \vee p_2)).$$

Remark. Lemma B.1 is Lemma 14 in Loh and Wainwright (2012), based on Lemma 5.14 and Corollary 5.17 in Vershynin (2012).

Lemma B.2: For a fixed matrix $\Gamma \in \mathbb{R}^{p\times p}$, parameter $s \geq 1$, and tolerance $\tau > 0$, suppose we have the deviation condition

$$|v^T\Gamma v| \leq \tau \qquad \forall v \in \mathbb{K}(2s,\, p,\, 1).$$

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 B.2 is Lemma 12 in Loh and Wainwright (2012).

Lemma B.3: Suppose the matrix $U \in \mathbb{R}^{n\times q}$ is sub-Gaussian with parameters $(\Sigma_U,\, \rho_U^2)$ where the jth column of U is U_j, and each row is sampled independently, we have

$$v^{0T}\frac{U^TU}{n}v^0 \geq v^{0T}\Sigma_U v^0 - \frac{\kappa}{2}\left(|v^0|_2^2 + \frac{1}{s}|v^0|_1^2\right), \tag{40}$$

$$v^{0T}\frac{U^TU}{n}v^0 \leq v^{0T}\Sigma_U v^0 + \frac{\kappa}{2}\left(|v^0|_2^2 + \frac{1}{s}|v^0|_1^2\right), \tag{41}$$

for all $v \in \mathbb{R}^q$ with probability at least $1 - c_1\exp(-c_2n + 2s\log q)$.

Proof. First, we show

$$\sup_{v\in\mathbb{K}(2s,\,q,\,1)}\left|v^T\left(\frac{U^TU}{n} - \Sigma_U\right)v\right| \leq \frac{\kappa}{54}$$

with high probability, where $\Sigma_U = \mathbb{E}(\frac{U^TU}{n})$. By Lemma B.1 and a discretization argument similar to those in the proof for Lemma A.6, we have

$$\mathbb{P}\left(\sup_{v\in\mathbb{K}(2s,\,q,\,1)}\left|v^T\left(\frac{U^TU}{n} - \Sigma_U\right)v\right| \geq t\right) \leq 2\exp(-cn\min(\frac{t^2}{\rho_U^4},\, \frac{t}{\rho_U^2}) + 2s\log q),$$

for some universal constant $c > 0$. By choosing $t = \frac{\kappa}{54}$, $s \geq 1$, we obtain

$$\mathbb{P}\left(\sup_{v^0 \in \mathbb{K}(2s,\, q,\, 1)} \left| v^T \left(\frac{U^T U}{n} - \Sigma_U \right) v \right| \geq \frac{\kappa}{54} \right) \leq 2\exp(-c_2 n + 2s \log q).$$

Now, by Lemma B.2 with the following substitutions $\Gamma = \frac{U^T U}{n} - \Sigma_U$ and $\tau := \frac{\kappa}{54}$, we obtain

$$\left| v^{0T} \left(\frac{U^T U}{n} - \Sigma_U \right) v^0 \right| \leq \frac{\kappa}{2} \left(|v^0|_2^2 + \frac{1}{s} |v^0|_1^2 \right),$$

with probability at least $1 - c_1 \exp(-c_2 n + 2s \log q)$. The claims follow from the bound above. $\square$

S Supplementary materials

The supplementary materials include additional technical lemmas with proofs, as well as additional simulation results (https://sites.google.com/site/yingzhu1215/home/HD2SLS_Supplement.pdf).

References

Allen, D. M. (1974). "The relationship between variable selection and data argumentation and a method of prediction." *Technometrics*, 16, 125-127.

Amemiya, T. (1974). "The non-linear two-stage least squares estimator." *Journal of Econometrics*, 2, 105-110.

Antoniadis, A. (2010). "Comments on: l_1-penalization for mixture regression models." *Test*, 19, 257-258.

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

Belloni, A., V. Chernozhukov, and L. Wang (2011). "Square-root Lasso: pivotal recovery of sparse signals via conic programming." *Biometrika*, 98, 791-806.

Belloni, A. and V. Chernozhukov (2011a). "L1-penalized quantile regression in high-dimensional sparse models." *The Annals of Statistics*, 39, 82-130.

Belloni, A. and V. Chernozhukov (2011b). "High dimensional sparse econometric models: an introduction", in: Inverse problems and high dimensional estimation, Stats in the Château 2009, Alquier, P., E. Gautier, and G. Stoltz, Eds., *Lecture Notes in Statistics*, 203, 127-162, Springer, Berlin.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). "Sparse models and methods for instrumental regression, with an application to eminent domain." *Econometrica*, 80, 2369-2429.

Belloni, A. and V. Chernozhukov (2013). "Least squares after model selection in high-dimensional sparse models." *Bernoulli*, 19, 521-547.

Bickel, P., J. Y. Ritov, and A. B. Tsybakov (2009). "Simultaneous analysis of Lasso and Dantzig selector." *The Annals of Statistics*, 37, 1705-1732.

Breiman, L. (1995). "Better subset regression using the nonnegative garrote." *Technometrics*, 37.

Breiman, L. (1996). “Heuristics of instability and stabilization in model selection.” *The Annals of Statistics*, 24, 2350-2383.

Breiman, L. (2001). “Statistical modeling: the two cultures.” *Statistical Science*, 16, 199-231.

Bühlmann, P. and S. A. van de Geer (2011). *Statistics for high-dimensional data.* Springer, New-York.

Caner, M. (2009). “Lasso type GMM estimator.” *Econometric Theory*, 25, 1-23.

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

Carrasco, M. and J. P. Florens (2000). “Generalization of GMM to a continuum of moment conditions.” *Econometric Theory*, 16, 797-834.

Carrasco, M. (2012). “A regularization approach to the many instruments problem.” *Journal of Econometrics*, 170, 383-398.

Chen, X. H. and M. Reiss (2011). “On rate optimality for ill-posed inverse problems in econometrics.” *Econometric Theory*, 27, 497-521.

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

Fan, J. and Y. Liao (2014). “Endogeneity in ultrahigh dimension.” *The Annals of Statistics*, 42, 872-917.

Fan, J. and J. Lv (2010). “A Selective overview of variable selection in high dimensional feature space.” *Statistica Sinica*, 20, 101-148.

Fan, J. and J. Lv (2011). “Non-concave penalized likelihood with NP-dimensionality.” *IEEE Transactions on Information Theory*, 57, 5467-5484.

Fan, J., J. Lv, and L. Qi (2011). “Sparse high dimensional models in economics.” *Annual Review of Economics*, 3, 291-317.

Garen, J. (1984). “The returns to schooling: a selectivity bias approach with a continuous choice variable.” *Econometrica*, 52, 1199-1218.

Gautier, E. and A. B. Tsybakov (2014). “High-dimensional instrumental variables regression and confidence sets.” Manuscript. CREST (ENSAE).

Hastie, T., R. Tibshirani, and J. Friedman (2002). *The elements of statistical learning: data mining, inference, and prediction*, Springer.

Huang, J., J. L. Horowitz, and S. Ma (2008). “Asymptotic properties of Bridge estimators in sparse high-dimensional regression models.” *The Annals of Statistics*, 36, 587-613.

Jing, B.-Y., Q. M. Shao, and Q. Wang (2003). “Self-normalized Cramér-type large deviations for independent random variables.” *The Annals of Probability*, 31, 2167-2215.

Koltchinskii, V. (2009). “The Dantzig selector and sparsity oracle inequalities.” *Bernoulli*, 15, 799-828.

Koltchinskii, V. (2011). “Oracle inequalities in empirical risk minimization and sparse recovery problems.” Forthcoming in *Lecture Notes in Mathematics*, Springer, Berlin.

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

Ledoux, M. and M. Talagrand (1991). *Probability in Banach spaces: isoperimetry and processes.* Springer-Verlag, New York, NY.

Lim, C. and B. Yu. (2013). “Estimation stability with cross validation (ESCV).” arXiv:1303.3128.

Lin, Y. and H. H. Zhang (2006). "Component selection and smoothing in multivariate nonparametric regression." *The Annals of Statistics*, 34(5): 2272-2297.

Loh, P., and M. Wainwright (2012). "High-dimensional regression with noisy and missing data: provable guarantees with non-convexity." *The Annals of Statistics*, 40, 1637-1664.

Manresa, E. (2014). "Estimating the structure of social interactions using panel data." Working paper. CEMFI.

Meinshausen, N., and P. Bühlmann (2006). "High-dimensional graphs and variable selection with the Lasso." *The Annals of Statistics*, 34:1436-1462.

Meinshausen, N., and P. Bühlmann (2010). "Stability selection." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72, 417-473.

Meinshausen, N., and B. Yu (2009). "Lasso-type recovery of sparse representations for high-dimensional Data." *The Annals of Statistics*, 37, 246-270.

Minsker, S. (2014). "Geometric median and robust estimation in Banach spaces." arXiv:1308.1334v5.

Negahban, S., P. Ravikumar, M. J. Wainwright, and B. Yu (2012). "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers." *Statistical Science*, 27, 538-557.

Nemirovski, A., and D. Yudin (1983). *Problem complexity and method efficiency in optimization.* John Wiley and Sons Inc.

Ravikumar, P., H. Liu, J. Lafferty, and L. Wasserman (2009). "Sparse additive models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71, 1009-1030.

Ravikumar, P., M. J. Wainwright, and J. Lafferty (2010). "High-dimensional Ising model selection using l_1-regularized logistic regression." *The Annals of Statistics*, 38, 1287-1319.

Raskutti, G., M. J. Wainwright, and B. Yu (2010). "Restricted eigenvalue conditions for correlated Gaussian designs." *Journal of Machine Learning Research*, 11, 2241-2259.

Raskutti, G., M. J. Wainwright, and B. Yu (2011). "Minimax rates of estimation for high-dimensional linear regression over l_q-balls." *IEEE Trans. Information Theory*, 57, 6976-6994.

Rosenbaum, M. and A. B. Tsybakov (2010). "Sparse recovery under matrix uncertainty." *The Annals of Statistics*, 38, 2620-2651.

Rosenbaum, M. and A. B. Tsybakov (2013). "Improved matrix uncertainty selector", in: *From Probability to Statistics and Back: High-Dimensional Models and Processes* - A Festschrift in Honor of Jon A. Wellner, Banerjee, M. et al. Eds, *IMS Collections*, 9, 276-290, Institute of Mathematical Statistics.

Rudelson, M. and S. Zhou (2011). "Reconstruction from anisotropic random measurements." Technical report, University of Michigan.

Sala-i-Martin, X. (1997). "I Just ran two million regressions." *The American Economic Review*, 87, 178-183.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis. Monographs on Statistics and Applied Probability*, 26, Chapman and Hall, London.

Stone, M. (1974). "Cross-validation choice and assessment of statistical prediction." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39, 44-47.

Sun, T. and C.-H. Zhang (2010). "Comments on: l_1-penalization for mixture regression models." *Test*, 19, 270-275.

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

Tibshirani, R. (1996). “Regression shrinkage and selection via the Lasso.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58, 267-288.

Vershynin, R. (2012). “Introduction to the non-asymptotic analysis of random matrices”, in Eldar, Y. and G. Kutyniok, Eds, *Compressed Sensing: Theory and Applications*, 210-268, Cambridge.

Wainwright, J. M. (2009). “Sharp thresholds for high-dimensional and noisy sparsity recovery using l_1-constrained quadratic programming (Lasso).” *IEEE Trans. Information Theory*, 55, 2183-2202.

Wainwright, J. M. (2015). *High-dimensional statistics: A non-asymptotic viewpoint.* In preparation. University of California, Berkeley.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data.* MIT Press, Cambridge.

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.

Yu, B. (2013). “Stability.” *Bernoulli*, 19, 1484-1500.

Zhang C.-H. and S. S. Zhang (2013). “Confidence intervals for low dimensional parameters in high dimensional linear models.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76, 217-242.

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

Zhu, Y. (2014). “High-dimensional linear models with endogeneity and sparsity.” *The California Econometrics Conference*. Stanford University.

Zhu, Y. (2014). “High-dimensional semiparametric selection models: estimation theory with an application to the retail gasoline market.” Working paper. University of California, Berkeley.