Ill-posed Estimation in High-Dimensional Models with Instrumental Variables*

Christoph Breunig† Enno Mammen‡

Emory University *Universität Heidelberg*

Anna Simoni§

CREST, CNRS

August 3, 2020

This paper is concerned with inference about low-dimensional components of a high-dimensional parameter vector β^0 which is identified through instrumental variables. We allow for eigenvalues of the expected outer product of included and excluded covariates, denoted by M, to shrink to zero as the sample size increases. We propose a novel estimator based on desparsification of an instrumental variable Lasso estimator, which is a regularized version of 2SLS with an additional correction term. This estimator converges to β^0 at a rate depending on the mapping properties of M. Linear combinations of our estimator of β^0 are shown to be asymptotically normally distributed. Based on consistent covariance estimation, our method allows for constructing confidence intervals and statistical tests for single or low-dimensional components of β^0. In Monte-Carlo simulations we analyze the finite sample behavior of our estimator. We apply our method to estimate a logit model of demand for automobiles using real market share data.

Keywords: Instrumental Variables, sparsity, central limit theorem, lasso, linear model, desparsification, ill-posed estimation problem.

*The authors gratefully thank the Co-Editor Oliver Linton, an Associate Editor, and three anonymous referees for their many constructive comments on the previous version of the paper. Financial support by ANR-11-LABEX-0047 and Deutsche Forschungsgemeinschaft through CRC TRR 190 is gratefully acknowledged.

†Department of Economics, Emory University, Rich Memorial Building, Atlanta, GA 30322, USA, e-mail: christoph.breunig@emory.edu

‡Institute for Applied Mathematics, Universität Heidelberg, Im Neuenheimer Feld 205, 69120 Heidelberg, Germany, e-mail: mammen@math.uni-heidelberg.de

§CNRS, CREST - ENSAE - École Polytechnique, 5, Avenue Henry Le Chatelier, 91120 Palaiseau, France, e-mail: anna.simoni@ensae.fr

JEL classification: C18, C26, C55

1. Introduction

In econometric applications, we may want to include a large number of regressors to account for heterogeneity of individuals or simply because economic theory is not explicit about which regressors to include in the model. These settings often lead to high-dimensional models where the number of parameters to be estimated is close to the sample size or even larger.

In this paper, we consider an instrumental variable (IV) model where the vector of parameters β^0 is identified through

$$Y = X^T\beta^0 + U, \quad \text{where } \mathbb{E}[UZ] = 0, \tag{1.1}$$

for a scalar dependent variable Y, a possibly endogenous vector of covariates X, and a vector of instrumental variables and exogenous covariates Z. Our setup is high-dimensional in the sense that the dimension of β^0 may be larger than the sample size n.

This paper is concerned with inference on inner products of β^0 of the type $a^T\beta^0$ for some vector a. In this sense, our model has a semi-parametric interpretation. When a low-dimensional subvector of β^0 is the parameter of interest and the remaining components of β^0 are considered as nuisance parameters, then inference on $a^T\beta^0$ implies inference on this low-dimensional subvector of β^0 for an appropriate choice of the vector a. We also allow the subvector of β^0 of interest to increase slowly with the sample size and provide inference for it. Our main example is when the low-dimensional subvector of β^0 is associated with endogenous regressors.

As the number of regressors in X may increase with the sample size n, also the singular values of the matrix M defined as

$$M := \mathbb{E}\left(ZX^T\right),$$

depend on n. In particular, including additional control variables in the model might affect the dependence between endogenous regressors and instruments and hence the cross second moment. This leads to situations where the singular values of M decrease with n and the vector β^0 is thus not strongly identified, following the terminology in Andrews and Cheng [2012]. Also, when the number of endogenous regressors increases with n it is well known that the singular values of M converge to zero in general and might even have an exponential decay. In the high-dimensional case, we then require some form of sparsity of the matrix M, i.e., that many entries of M are zero or sufficiently small.

A crucial insight of this paper is to show how the mapping properties of the matrix M affect the asymptotic behavior of our estimator. For instance, we see that the minimal eigenvalue of M slows down the rate of convergence and enlarges the asymptotic variance of our estimator. In addition, the sparsity pattern of M and the sparsity pattern of the parameter vector β^0 are shown to be related: less sparsity of M requires a higher degree of sparsity of β^0 and vice versa. This can be interpreted as an ℓ_1 analog of the so-called source condition used in the inverse problems literature which links the smoothness of

the unknown function to the smoothing properties of the operator that characterizes the inverse problem.

This paper proposes a novel estimation procedure based on a Lasso type estimator, suitably modified to have a tractable limiting distribution for inner products of β^0. While the Lasso estimator makes use of the underlying sparsity constraints, it is well known that it does not have a tractable limiting distribution. In this paper, we use the methodology of desparsification to make up for this drawback. Our desparsified IV Lasso estimator for β^0 corrects the high-dimensional two stage least squares (2SLS) estimator by subtracting a regularization bias. In the case of low dimensions, i.e. under a known sparsity structure, the resulting estimator coincides with the ordinary 2SLS estimator.

We establish the rate of convergence of inner products of our estimator, and show that the rate is affected by the minimum singular value of M (opportunely normalized). In particular, we can show an analog to the nonparametric IV case, where slow rates of convergence are common. Moreover, inner products of our estimator for β^0 are shown to be asymptotically normal. The normalization factor for the estimator is shown to be driven by the minimal singular value of M. We derive confidence intervals and hypothesis testing procedures for inner products of β^0. As discussed above, inference results on inner products of β^0 imply inference results on low-dimensional subvectors of β^0 or even on subvectors of β^0 slowly increasing with the sample size. In Monte Carlo simulations, we show that the proposed confidence intervals have accurate size.

It is interesting to note that having the rate of our estimator affected by the minimum singular value of M is similar to what happens for sieve estimation in the nonparametric IV (NPIV) literature. In NPIV literature the rate of convergence is derived under smoothness assumptions of the underlying IV regression functions instead of under sparsity constraints of the IV regression coefficients as in this paper. In particular, model (1.1) can be also seen as an approximation of the true relationship between Y and a vector of endogenous covariates based on a dictionary X of transformations of the endogenous covariates. Hence, the two types of assumptions (smoothness and sparsity) provide two alternative frameworks to deal with high-dimension in nonparametric IV regression models.

Related Literature. Our paper contributes to the growing literature on inference for structural parameters in sparse high-dimensional IV settings. Much work in this setting focuses on the case where the dimension of the endogenous variable is small but where there is a large number of available instruments, see Ng and Bai [2009], Belloni et al. [2012], and Belloni et al. [2011]. In this context, Chao and Swanson [2005] and Hansen et al. [2008], among others, propose methods to account for weak identification when the number of instruments is allowed to be large but not larger than n. Hansen and Kozbur [2014] and Carrasco and Doukali [2017] extend this literature by considering cases where the number of weak instruments is larger than n and propose a Ridge regularized jackknife instrumental variable estimation. When the number of endogenous regressors in model (1.1) is fixed and there are high-dimensional control variables, Chernozhukov et al. [2015] propose a three step estimator where high-dimensional sparse linear models with only exogenous variables are fitted. In particular, Lasso is only used for the fit of nuisance parameters and the use of the Lasso estimates follows standard lines. For the fit of the parameters of the endogenous covariates, the criterion function is orthogonalized such that errors in the estimation of the other parameters (i.e. of the nuisance parameters) enter into the model only quadratically. For this reason the clas-

sical bounds for the errors of the Lasso estimates of the nuisance parameters suffice. In particular, no debiasing/desparsification of Lasso estimates is needed at any point of the procedure. Belloni et al. [2017a] consider estimation of treatment effects in IV models with binary instrument and endogenous variable in the presence of a high-dimensional set of control variables.

Also relevant to this paper is the literature concerning choice of valid instruments. In the context of a scalar endogenous variable, Guo et al. [2018] propose a method to select valid instruments based on hard thresholding in setups where the number of instruments and of exogenous variables may tend to infinity. Their proposal is related to LASSO approaches for the selection of valid instruments in finite dimensional setups. Kang et al. [2016] use Lasso to instrumental variable selection in the context of invalid instruments. Based on an initial median estimator, Windmeijer et al. [2019] use adaptive Lasso for instrument selection and establish consistency of their procedure.

In model (1.1) which allows for increasing dimension of endogenous regressors, Gautier et al. [2011] establish a novel estimation procedure based on novel sensitivity characteristics of the empirical counterpart of M to obtain confidence sets with length depending on the strength of instruments. Gautier et al. [2011] also establish confidence bands after bias correction. Belloni et al. [2017b] use such sensitivity to construct simultaneously valid confidence regions and have proposed a multiplier bootstrap procedure to compute critical values and establish its validity. Their approach is based on orthogonality restrictions when considering linear combinations of the original instruments.

Our approach is essentially different from the previous ones as our sparsity condition is based on the population matrix M and not on its empirical counterpart. This allows us to provide a novel link between high-dimensional and NPIV estimation where the first is based on assuming sparsity while the latter is based on assuming smoothness in the underlying model, see *e.g.* Ai and Chen [2003], Newey and Powell [2003], Darolles et al. [2011], Chen and Pouzo [2012], and references therein for NPIV estimation. Fan and Liao [2014] propose a modified Lasso approach for estimation in high-dimensional instrumental variables models. Our paper is also related to Guo et al. [2018] and Gold et al. [2018] that, as we propose in our paper, use two-step estimators using a threshold procedure and Lasso estimation, respectively, in the first step and desparsification in the second step. However, Gold et al. [2018] make assumptions about sparsity that differ from ours and their settings exclude cases where the estimator of components of β does not achieve a parametric $\sqrt{n}$-rate. On the other hand, we do allow for singular values of M to tend to zero which yields slower rates and provide novel inference results for inner products of the estimator of β of increasing dimension. This is an important feature of our paper as we are thus able to provide an interpretation that is close to the nonparametric IV estimation. Recently, Neykov et al. [2018] considered univariate confidence set estimation in a high-dimensional setting but require that the number of instruments coincides with the number of endogenous variables. In contrast, our approach is efficient under sparsity constraints and moreover, it is robust against violations of strong identification, which, as far as we know, has not been addressed in the related literature.

Our paper is also related to the rich statistical literature on high-dimensional statistical models that contain only exogenous variables and where endogeneity and instrumental variables are not considered, see, Zhang and Zhang [2014], Javanmard and Montanari [2014a,b] and van de Geer et al. [2014]. An alternative approach to our desparsified Lasso estimator is ridge regression where an ℓ_2 penalty is used and the asymptotic distribution results can be readily obtained. This approach in high-dimensional Gaussian regression

is considered by Bühlmann [2013]. In an extensive simulation study, however, Javanmard and Montanari [2014b] show that the ridge regression approach is overly conservative, which is in line with the theoretical results. This is why we also pursue to desparsify the Lasso estimator rather than using the ridge regression.

The remainder of the paper is organized as follows. In Section 2, we describe the model, motivate the desparsification procedure and discuss sparsity requirements. Section 3 contains the rates of convergence and the asymptotic normality results of our estimator. Section 4 is concerned with the finite sample performance of our estimator. Simulations and numerical implementation of our inference procedure are in Section 4. In Section 5 we present an application to demand estimation for automobiles using market share data. All proofs can be found in the appendix.

Notation. The ℓ_p norm of a vector a is denoted by $\|a\|_p$, $1 \leq p \leq \infty$. For a set S, the cardinality of S is denoted by $|S|$. For a vector a, and S a set of indices, a_S denotes the restriction of a to indices in S: $a_S := \{a_j; j \in S\}$. Further, for a matrix A we use the notation

$$\|A\|_\infty := \max_{j,k} |A_{jk}|$$

for the element-wise sup-norm,

$$\|A\|_{op,\infty} := \max_j \sum_k |A_{jk}|$$

for the operator norm, and

$$\|A\|_1 := \max_k \sum_j |A_{jk}|$$

for the ℓ_1 norm. For vectors a we have $\|a\|_{op,\infty} = \|a\|_\infty$ and for a matrix A it holds $\|A\|_{op,\infty} = \|A^T\|_1$. The smallest and largest eigenvalue of A are denoted by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$, respectively. We denote by A_j the j-th column of the matrix A and by A_{-j} the matrix A without the j-th column. We denote by e_j the j-th unit column vector. For two positive sequences a_n, b_n we use the notation $a_n \sim b_n$ to mean that there are two positive and universal constants C_1, C_2 such that $C_1 \leq a_n/b_n \leq C_2$. We abbreviate "with probability approaching one" to "wpa1", and say that a sequence of events $\{B_n\}$ holds wpa1 if $\mathbb{P}(B_n^c) = o(1)$ as $n \to \infty$.

2. Model and Methodology

Consider again model (1.1), the high-dimensional instrumental variable model is given by

$$Y = X^T\beta^0 + U \text{ where } \mathbb{E}[UZ] = 0, \tag{1.1}$$

where β^0 is the p–dimensional, unknown parameter of interest. Some of the covariates in X are possibly endogenous in the sense that they are related to the unobservables U, i.e., $\mathbb{E}[UX]$ does not vanish. Here, Y is a scalar dependent variable, X is a p–dimensional vector of endogenous and exogenous covariates, Z is a q–dimensional vector

of instrumental variables and exogenous covariates. So, the vectors Z and X may have elements in common if X contains exogenous covariates.

To ensure identification of the parameter β^0, we assume throughout the paper that $q \geq p$. This condition can be met in at least three situations: (1) the case where one has low dimensional endogenous variables and high-dimensional exogenous controls and the interest is in inference on the coefficients of the low dimensional endogenous variables (examples are: [a] the case where one includes many exogenous controls to account for complex heterogeneity, or [b] the case where one includes many exogenous variables to account for nonlinearities to approximate a partial linear structure); (2) more generally, the case where a high-dimensional linear sieve approach is used to approximate nonlinear and nonparametric instrumental variables models; (3) models that have a rich information in exogenous variation, corresponding to high-dimensional instrumental variables, as for instance in Angrist and Krueger [1991].

We also assume throughout the paper that the matrices $M := \mathbb{E}\left(ZX^T\right)$ and $\Sigma := \mathbb{E}\left(ZZ^T\right)$ are of full column rank. Thus, the parameter vector β^0 is identified through

$$\beta^0 = (M^T\Sigma^{-1}M)^{-1}M^T\Sigma^{-1}\mathbb{E}[ZY]. \tag{2.1}$$

Estimating β^0 by simply replacing the matrices on the right hand side by their empirical counterparts fails for two reasons. First, the empirical counterparts of M and Σ are in general not of full rank in the high-dimensional case. Second, it is well known that, for large matrices, estimators simply based on the sample mean do not provide satisfactory performance. In this paper, we address these challenges by using regularization procedures.

A common assumption to obtain consistent estimation results in the high-dimensional setting is a sparsity restriction: most of the parameters of β^0 are zero (exact sparsity) or sufficiently small (approximate sparsity) which implies that a relatively small number of regressors in X is sufficient in describing the dependent variable Y.

Note that the model is not identified if the minimal eigenvalue of $M^T\Sigma^{-1}M$ is zero, which we rule out throughout the paper (see Assumption 1). We thus introduce

$$\omega := 1/\lambda_{\min}\left(M^T\Sigma^{-1}M\right)$$

which satisfies $\omega < \infty$ for each $n \geq 1$ under Assumption 1. Below, we also assume that the maximal eigenvalue of $M^T\Sigma^{-1}M$ is bounded from above uniformly in $n \geq 1$ and hence, ω is strictly positive for all $n \geq 1$. On the other hand, in many cases we expect that ω might increase with the sample size n either because the model requires a large number of functions to account for nonlinearity in the endogenous covariates or because the instruments are weak and thus the model is not strongly identified. In the first case, $X^T\beta$ is an approximation of the true nonlinear instrumental regression through approximating functions stored in X whose number increases with n. In the second case, weakness of the instruments is captured by close to zero elements in the matrices M and $\Sigma^{-1/2}M$.

Similar to Andrews and Cheng [2012], we consider the *strongly identified* case where ω is uniformly bounded above, and the *semi-strongly identified* case where ω is unbounded but satisfies $n/\omega \to \infty$. We show below that ω slows down the rate of convergence of our estimator. In the semi-strong case, the size of the confidence sets increases relative to ω. There is also a third case which is the *weak identified* case where $n/\omega = O(1)$ but the results of our paper do not apply to it. In this paper, we show that under appropriate assumptions the rate of convergence of our estimator for each component of β^0 is $\sqrt{\omega/n}$.

Throughout the paper, we assume that a sample (Y_i, X_i, Z_i), $1 \leq i \leq n$ of independent and identically distributed copies of (Y, X, Z) is available. We write the vector and matrices of observations as $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$ with $\mathbf{X}_j = (X_{1,j}, \ldots, X_{n,j})^T$ for $1 \leq j \leq p$, and $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_q)$ with $\mathbf{Z}_j = (Z_{1,j}, \ldots, Z_{n,j})^T$ for $1 \leq j \leq q$. Moreover, the n-vector of unobservables is denoted by $\mathbf{U} = (U_1, \ldots, U_n)^T$. The $(d \times d)$–dimensional identity matrix is denoted by I_d and its j–th column by e_j.

2.1. The Desparsified IV Lasso Estimator

In this section, we introduce our estimation procedure which is based on desparsifying a Lasso estimator. The methodology is based on regularized estimators of the matrices $\Theta := \Sigma^{-1}$, $M := \mathbb{E}[ZX^T]$, and $\Theta^M := (M^T \Theta M)^{-1}$ denoted by $\widehat{\Theta}$, $\widehat{M}$, and $\widehat{\Theta}^M$, respectively, which are defined in Subsection 2.3. We propose the following *desparsified IV Lasso estimator* of β^0 given by

$$\widehat{\beta} = \widehat{\Theta}^M \widehat{M}^T \widehat{\Theta} \mathbf{Z}^T \mathbf{Y}/n - \big(\widehat{\Theta}^M \widehat{M}^T \widehat{\Theta} \mathbf{Z}^T \mathbf{X}/n - I_p\big)\widetilde{\beta}, \tag{2.2}$$

where $\widetilde{\beta}$ is a consistent estimator of β^0 that makes use of the underlying sparsity assumption. The first summand on the right hand side of (2.2) corresponds to a regularized empirical analog of β^0 as in (2.1). The second summand on the right hand side of (2.2) accounts for the regularization bias of our matrix estimators.

The proposed estimator naturally extends the 2SLS estimator to the high-dimensional case. Consider the situation of a known sparsity structure where regularization is not required and so $\widehat{\Theta}$, $\widehat{M}$, and $\widehat{\Theta}^M$ are the usual empirical counterparts of Θ, M and Θ^M. In this case it holds $\widehat{\Theta}^M \widehat{M}^T \widehat{\Theta} \mathbf{Z}^T \mathbf{X}/n = I_p$ and moreover, $\widehat{\beta}$ coincides with the 2SLS estimator.

The choice of $\widetilde{\beta}$ is motivated by our sparsity assumption given below and by the asymptotic properties for $\widehat{\beta}$ that we want to obtain. To derive the asymptotic results of our desparsified IV Lasso estimator we make use of the following key decomposition

$$\sqrt{n}(\widehat{\beta} - \beta^0) = \widehat{\Theta}^M \widehat{M}^T \widehat{\Theta}\, \mathbf{Z}^T \mathbf{U}/\sqrt{n} - \Delta, \tag{2.3}$$

for a remainder term Δ which is given by

$$\Delta := \sqrt{n}\big(\widehat{\Theta}^M \widehat{M}^T \widehat{\Theta} \mathbf{Z}^T \mathbf{X}/n - I_p\big)(\widetilde{\beta} - \beta^0).$$

Then, we have to show that $\|\Delta\|_\infty$ is asymptotically negligible under regularity assumptions. In particular, to show this we require that $\|\widetilde{\beta} - \beta^0\|_1$ is sufficiently small. This property is satisfied by the Lasso estimator and thus we choose $\widetilde{\beta}$ in equation (2.2) to be the Lasso estimator which makes use of the underlying sparsity structure imposed on β^0. Therefore, our estimation procedure is based on the IV Lasso estimator of β^0 given by

$$\widetilde{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \Big\{ (\mathbf{Z}^T \mathbf{Y}/n - \widehat{M}\beta)^T \, \widehat{\Theta} \, (\mathbf{Z}^T \mathbf{Y}/n - \widehat{M}\beta) + 2\lambda \, \|\beta\|_1 \Big\} \tag{2.4}$$

for some tuning parameter $\lambda > 0$. Then, we replace $\widetilde{\beta}$ in equation (2.2) to obtain $\widehat{\beta}$.

2.2. Sparsity Constraints

In this section we introduce some notations and assumptions about sparsity that we tacitly maintain all along the paper. Let s_0 denote the cardinality of the set S_0, i.e., $s_0 := |S_0|$, where S_0 is a set such that $\|\beta^0_{S_0^c}\|_1$ is sufficiently small. That is, we assume that the set S_0 is rich enough such that the parameter vector β^0 satisfies

$$\|\beta^0_{S_0^c}\|_1 = \sum_{j \notin S_0} |\beta^0_j| \leq C s_0 \sqrt{\log(p)/n} \tag{2.5}$$

for some constant $C > 0$. Inequality (2.5) imposes approximate sparsity on β^0: The absolute value of the parameters outside the sparsity set S_0 is bounded by some value which tends to zero as the sample size tends to infinity.

Hereafter, we assume that Θ and Θ^M exist and assume sparsity with respect to rows of $\Theta := \Sigma^{-1}$. To this purpose we define

$$s_j := |\{k \neq j : \Theta_{jk} \neq 0\}| \qquad \text{and} \qquad s_{\max} := \max_{1 \leq j \leq q} s_j.$$

The sparsity restriction on Θ has the following interpretation: if the (jk)–th component of Θ is zero, then the variables Z_j and Z_k are partially uncorrelated, given the other variables. In particular if Z is jointly normal then we have that the variables Z_j and Z_k are conditionally independent, given the other variables. This also motivates the use of an ℓ_1–penalty for the estimation of Σ^{-1}, as we do in Section 2.3.1 and which was proposed by Meinshausen and Bühlmann [2006]. Note that it is possible to relax the sparsity constraints but this would lead to a less efficient estimator.

We need to assume some sparsity pattern on M, that is, most of the elements in each row or column of M are zero. We conjecture that it would suffice to assume only approximate sparsity for M and Θ but at the cost of much more technical proofs and notation. For the sparsity of M we introduce the notation

$$s_M := \max_{1 \leq k \leq p} |\{j : M_{jk} \neq 0\}|.$$

Hence, $\|M\|_1 \leq s_M \|M\|_\infty$. For $j = 1, \ldots, p$, we denote $\gamma_j := \operatorname{argmin}_{\gamma \in \mathbb{R}^{p-1}} \|(\Theta^{1/2}M)_j - (\Theta^{1/2}M)_{-j}\gamma\|_2^2$ and impose the following approximate sparsity condition on γ_j: we assume there exists a set S_j such that

$$\|\gamma_{j,S_j^c}\|_1 = \sum_{l \notin S_j} |\gamma_{jl}| \leq C \log(q)/\sqrt{n} \tag{2.6}$$

for some constant $C > 0$. Below we also denote $s_j^M := |S_j|$ and for convenience we use the notation $s^M_{\max} := \max_{1 \leq j \leq q} s_j^M$.

2.3. Regularized Matrix Estimators

In this section, we provide the regularization schemes to construct the approximate inverses $\widehat{\Theta}$ and $\widehat{\Theta}^M$ as well as the regularized estimator $\widehat{M}$. Asymptotic properties of these estimator will be studied in Section 3.

2.3.1. Construction of $\widehat{\Theta}$

Here we construct a regularized estimator of the inverse of Σ denoted by $\widehat{\Theta}$. The basic idea to construct such an estimator is to relate the inversion of a $q \times q$ matrix to q regression problems of $\mathbf{Z}_j$ over $\mathbf{Z}_{-j}$, where for $1 \leq j \leq q$, $\mathbf{Z}_j = (Z_{1,j}, \ldots, Z_{n,j})^T$ is the j-th column vector of the matrix $\mathbf{Z}$ and $\mathbf{Z}_{-j} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_{j-1}, \mathbf{Z}_{j+1}, \ldots, \mathbf{Z}_q)$. This approach was introduced by Meinshausen and Bühlmann [2006] and consists in using the Lasso estimator for nodewise regression. For every $j = 1, \ldots, q$ we consider the Lasso estimator

$$\widehat{\xi}_j = \underset{\xi \in \mathbb{R}^{q-1}}{\operatorname{argmin}} \left\{ \|\mathbf{Z}_j - \mathbf{Z}_{-j}\xi\|_2^2/n + 2\lambda_j^{\Theta}\|\xi\|_1 \right\} \tag{2.7}$$

for some tuning parameter $\lambda_j^{\Theta} > 0$ that will be let to tend to zero as the sample size increases to get asymptotic results. We introduce the q-column vector $\widehat{\Gamma}_j = (\widehat{\Gamma}_{kj})_{k=1}^q$ such that

$$\widehat{\Gamma}_{kj} = \begin{cases} 1 & \text{for } k = j \\ -\widehat{\xi}_{jk} & \text{for } k \neq j \end{cases} \tag{2.8}$$

with $\widehat{\xi}_j = (\widehat{\xi}_{jk})_{k \in \{1,\ldots,q\} \setminus \{j\}}$. By the definition of $\widehat{\Gamma}_j$, it holds $\mathbf{Z}_j - \mathbf{Z}_{-j}\widehat{\xi}_j = \mathbf{Z}\widehat{\Gamma}_j$. Then, the matrix $\widehat{\Theta} = \left(\widehat{\Theta}_1, \ldots, \widehat{\Theta}_q\right)^T$ is constructed as

$$\widehat{\Theta}_j = \widehat{\tau}_j^{-2}\widehat{\Gamma}_j \quad \text{where} \qquad \widehat{\tau}_j^2 = \|\mathbf{Z}\widehat{\Gamma}_j\|_2^2/n + \lambda_j^{\Theta}\|\widehat{\xi}_j\|_1. \tag{2.9}$$

Note that while the population counterpart Θ is symmetric, its estimator $\widehat{\Theta}$ does not need to be so. For more details on this procedure, we refer to Meinshausen and Bühlmann [2006].

2.3.2. Construction of $\widehat{M}$

A standard sample matrix estimator for the matrix M does not have good performance in the high-dimensional case and regularization is needed. Hence, we propose a thresholding estimator of M. Intuitively, we want to eliminate those values of the empirical matrix $\widetilde{M} := \mathbf{Z}^T\mathbf{X}/n$ that lie below some specified threshold. More precisely, we propose to use the thresholding estimator $\widehat{M} = (\widehat{M}_{jk})$ where

$$\widehat{M}_{jk} := \widetilde{M}_{jk}\, \mathbb{1}\left\{ |\widetilde{M}_{jk}| \geq C_0 \sqrt{\frac{\log q}{n}} \right\} \tag{2.10}$$

and $C_0 > 0$ is a constant defined as in Proposition 3.2 and is related to the constant appearing in the large deviation inequality for the components of $\widetilde{M}$. For symmetric matrices such a regularization scheme has been considered in Bickel and Levina [2008] and Cai and Zhou [2012] among others, and we refer to these papers for a discussion on this constant. For practical implementation, we choose $c_n := C_0\sqrt{\log(q)/n}$ by cross-validation, see Section 4 for more details.

2.3.3. Construction of $\widehat{\Theta}^M$

In this section, we construct the estimator $\widehat{\Theta}^M$ which is an approximate inverse of $\widehat{M}^T\widehat{\Theta}\widehat{M}$. This estimator involves the regularized estimators $\widehat{\Theta}$ and $\widehat{M}$ obtained in Sections 2.3.1 and 2.3.2 and the square root of $\widehat{\Theta}$, denoted by $\widehat{\Theta}^{1/2}$. Note that we can make

use of the Schur decomposition of $\widehat{\Theta}$ to compute its square root given that $\widehat{\Theta}$ is not necessarily symmetric.

Let $(\widehat{\Theta}^{1/2}\widehat{M})_j$ denote the j-th column vector of the matrix $\widehat{\Theta}^{1/2}\widehat{M}$ and

$$(\widehat{\Theta}^{1/2}\widehat{M})_{-j} := ((\widehat{\Theta}^{1/2}\widehat{M})_1, \ldots, (\widehat{\Theta}^{1/2}\widehat{M})_{j-1}, (\widehat{\Theta}^{1/2}\widehat{M})_{j+1}, \ldots, (\widehat{\Theta}^{1/2}\widehat{M})_p).$$

Remark that $\widehat{\Theta}^{1/2}\widehat{M}$ is the empirical cross moment of X and the (approximately) orthonormalized Z. The approximate orthonormalization of $\mathbf{Z}$ is performed by premultiplication by $\widehat{\Theta}^{1/2}$. As for the construction of $\widehat{\Theta}$, we relate the regularized inversion of a $p \times p$ matrix to p regression problems of $(\widehat{\Theta}^{1/2}\widehat{M})_j$ on $(\widehat{\Theta}^{1/2}\widehat{M})_{-j}$. To do that, for every $j = 1, \ldots, p$ we consider the Lasso estimator:

$$\widetilde{\gamma}_j = \underset{\gamma \in \mathbb{R}^{p-1}}{\operatorname{argmin}} \left\{ \|(\widehat{\Theta}^{1/2}\widehat{M})_j - (\widehat{\Theta}^{1/2}\widehat{M})_{-j}\gamma\|_2^2 + 2\lambda_j^M \|\gamma\|_1 \right\}, \tag{2.11}$$

for some tuning parameter $\lambda_j^M > 0$ that will be let to tend to zero as the sample size increases to get asymptotic results. Let $\widetilde{\Gamma}_j = (\widetilde{\Gamma}_{kj})_{k=1}^p$ be the p-column vector determined by

$$\widetilde{\Gamma}_{kj} = \begin{cases} 1 & \text{for } k = j \\ -\widetilde{\gamma}_{jk} & \text{for } k \neq j \end{cases}$$

with $\widetilde{\gamma}_j = (\widetilde{\gamma}_{jk})_{k \in \{1,\ldots,p\} \setminus \{j\}}$. The matrix $\widehat{\Theta}^M$ is then set equal to $\widehat{\Theta}^M = (\widehat{\Theta}_1^M, \ldots, \widehat{\Theta}_p^M)^T$ where

$$\widehat{\Theta}_j^M = \widetilde{\tau}_j^{-2}\widetilde{\Gamma}_j \qquad \widetilde{\tau}_j^2 = \|\widehat{\Theta}^{1/2}\widehat{M}\widetilde{\Gamma}_j\|_2^2 + \lambda_j^M \|\widetilde{\gamma}_j\|_1, \qquad 1 \leq j \leq p.$$

As already stressed in van de Geer et al. [2014], other regularization methods to obtain the approximate inverses of $\widehat{\Sigma}$ and $(\widehat{M}^T\widehat{\Theta}\widehat{M})$ that do not deliver a bound for $\left\|\widehat{M}^T\widehat{\Theta}\widehat{M}\,\widehat{\Theta}_j^M - e_j\right\|_\infty$, like the ridge regularization, may not be optimal because without this bound we cannot directly obtain asymptotic distribution results for components of β^0. The regularization methods that we use to construct $\widehat{\Theta}$ and $\widehat{\Theta}^M$ automatically include this bound in the optimization problem.

3. Inference

In this section, we derive the asymptotic distribution of the desparsified IV Lasso estimator $\widehat{\beta}$ given in (2.2). To obtain asymptotic results on which our inference will be based we have to show that the remainder term Δ in the key decomposition (2.3) is asymptotically negligible. We start by providing all the assumptions that we need to obtain our asymptotic results. After that, we first provide results about rates of convergence for the estimated matrices and for $\widehat{\beta}$, and then asymptotic normality will be established.

3.1. Assumptions

In this section we gather assumptions which we require to establish our inference results. Below, a random vector $W \in \mathbb{R}^d$ is called sub-Gaussian if $\mathbb{E}\exp\left(|v^T W|^2/C\right) = O(1)$ for all $v \in \mathbb{R}^d$ such that $\|v\|_2 \leq 1$ and some sufficiently large constant $C > 0$.

Assumption 1. *(i) We observe independent and identically distributed (i.i.d.) copies* $(Y_1, X_1, Z_1), \ldots, (Y_n, X_n, Z_n)$ *of* (Y, X, Z) *satisfying model (1.1). (ii) The vectors* X *and* Z *are sub-Gaussian. (iii) The eigenvalues of* Σ *are uniformly bounded away from zero and from infinity. (iv) The smallest eigenvalue* $\lambda_{\min}(M^T\Sigma^{-1}M)$ *is bounded from below for each* $n \geq 1$ *and the largest eigenvalue* $\lambda_{\max}(M^T\Sigma^{-1}M)$ *is bounded from above uniformly in* $n \geq 1$.

Sub-Gaussianity, as imposed in Assumption 1 (*ii*), is satisfied, for instance, if the random vectors have bounded support. Assumption 1 (*iii*) implies that $\Sigma_{jj} = O(1)$ uniformly in j since $\Sigma_{jj} \leq \lambda_{\max}(\Sigma)$. Similarly, it also implies that $\|\Theta_j\|_2 \leq \lambda_{\min}(\Sigma) = O(1)$ uniformly in j and consequently, $\|\Theta\|_1 = O\big(\sqrt{s_{\max}}\big)$ which we use below.We also make use of the notation $\mathcal{B} := \{\beta : \|\beta_{S_0^c}\|_1 \leq 3\|\beta_{S_0}\|_1\}$.

Lemma 3.1. *Let Assumption 1 be satisfied. If* $\log(q)/n = o(1)$ *then it holds for all* $\beta \in \mathcal{B}$ *that*

$$\|\beta_{S_0}\|_1^2 \leq s_0\, \beta^T M^T \Sigma^{-1} \widehat{\Sigma} \Sigma^{-1} M \beta / c^2 \tag{3.1}$$

wpa1 for some constant $c > 0$ *and where* $\widehat{\Sigma} = \mathbf{Z}^T\mathbf{Z}/n$.

The previous result shows that a modified version of the so called *compatibility condition*, see *e.g.* Bühlmann and Van De Geer [2011], is satisfied with high probability. Note that such conditions are required in the high-dimensional estimation context in order to relax the requirement of non-zero eigenvalues of associated estimated matrices. The following assumption provides more details about the choice of regularization parameters and imposes conditions on the underlying sparsity.

Assumption 2. *(i) It holds* $\lambda \sim \log(q)/\sqrt{n}$, $\lambda_j^\Theta \sim \sqrt{\log(q)/n}$, *and* $\lambda_j^M \sim \log(q)/\sqrt{n}$ *uniformly in* j. *(ii) It holds* $\|M\|_\infty = O(1)$, $\mathbb{E}\big[\max(1, |X^T\beta^0|^2)\|M^T\Sigma^{-1}Z\|_\infty^2\big] = O(\log(p))$ *and* $\mathbb{E}[U^2|Z] \leq \sigma^2 < \infty$ *for a constant* $\sigma > 0$. *(iii) Assume* $s_M\sqrt{s_{\max}}\max(s_{\max}^M, \|\beta^0\|_1) = O\big(\sqrt{\log(q)}\big)$ *and*

$$s_0 s_M \sqrt{s_{\max}} \max\big(\sqrt{s_M}, \sqrt{s_{\max}}\big)\sqrt{\log(p)\log(q)} + \omega^2 s_{\max}^M = o\big(\sqrt{n/\log(q)}\big). \tag{3.2}$$

Assumption 2 (*i*) specifies the rate of the tuning parameters λ used for the plug-in Lasso and λ_j^Θ, λ_j^M used for the nodewise Lasso estimators. The rate of the regularization parameters λ and λ_j^M is larger by $\sqrt{\log(q)}$ than the common choices of it, which is due to the additional estimation step that is involved for our initial IV Lasso estimator. Assumption 2 (*ii*) imposes upper bounds on the maximal element (in absolute value) of M and $M^T\Sigma^{-1}M$, and the conditional variance of U given Z, which is standard in the literature and is a mild restriction on the heteroscedasticity of the model. Assumption 2 (*iii*) imposes sparsity restrictions which we require in order to obtain our inference results. Specifically, this assumption restricts the sparsity of β_0 (captured by s_0) in relation to the sparsity of M (captured by s_M). Finally, condition (3.2) implies $\log(p)/\sqrt{n} = o(1)$.

For the next assumption, recall that $\gamma_j := \operatorname{argmin}_{\gamma\in\mathbb{R}^{p-1}} \|(\Theta^{1/2}M)_j - (\Theta^{1/2}M)_{-j}\gamma\|_2^2$ for $1 \leq j \leq p$. Introduce a vector $\Gamma_j := (\Gamma_{kj})_{k=1}^p$ with $\Gamma_{kj} = -\gamma_{kj}$ for $k \neq j$ and 1 otherwise, where γ_{kj} is the k–th entry of γ_j.

Assumption 3. *(i)* $\mathbb{E}\max_{1\leq j\leq p}|(\Theta M\Gamma_j)^T ZX^T\Gamma_j|^2 = O(\log(p))$ *and* $\mathbb{E}\|M^T\Theta ZX^T\Gamma_j\|_\infty^2 = O(\log(p))$ *for all* $1 \leq j \leq p$. *(ii) It holds* $\mathbb{E}\max_{1\leq j\leq p}\|(\Theta M\Gamma_j)^T Z\|_2^4 = O(\log(p)^2)$ *and further,* $\mathbb{E}\|\Gamma_j^T M^T\Theta ZZ^T\Theta M\|_\infty^2 = O(\log(p))$ *for all* $1 \leq j \leq p$.

Assumption 3 (i) imposes upper bounds on moments associated to ZX^T while Assumption 3 (ii) imposes mild rate conditions on moments of ZZ^T. Note that the logarithmic rates in Assumption 3 can be replaced by other powers of logarithms to allow for more heavy tailed variables. This would require slight changes in our constraints on the growth of dimension parameters p and q and somewhat more restrictive sparsity constraints.

3.2. Convergence Rates of estimated Matrices

In this section we provide rates of convergence for the regularized matrices used to construct our estimator $\widehat{\beta}$. These results are then used to establish asymptotic normality results in the next section.

In the following result, we derive a rate of convergence for $\widehat{M}$ in the ℓ_1 norm. The first part of the theorem provides a large deviation inequality for the components of $\widetilde{M}$ and it is derived by exploiting sub-Gaussianity of the rows of $\mathbf{X}$ and $\mathbf{Z}$ and a Bernstein-type inequality for sub-exponential random variables.

Proposition 3.2. *Let Assumption 1 hold. Then, there exists a constant $c > 0$ such that*

$$\mathbb{P}\big(\big|\widetilde{M}_{jk} - M_{jk}\big| \geq v\big) \leq 4\exp\big(-c\,v^2 n\big) \tag{3.3}$$

for $0 \leq v < 1$. Moreover, let $\widehat{M}$ be the thresholding estimator defined in (2.10) with $C_0 = \sqrt{8/c}$. If in addition Assumption 2 (i) and (ii) holds, then we have

$$\big\|\widehat{M} - M\big\|_1 = O_p\big(s_M\sqrt{\log(q)/n}\big).$$

The constant c in inequality (3.3) depends on the second order moments and cross moments of the elements in Z and X as well as on their sub-Gaussian norms. Its expression can be deduced from the proof of the proposition given in the appendix.

The next result gives a key upper bound for the approximation error of the relaxed inverses $\widehat{\Theta}_j$ and $\widehat{\Theta}_j^M$. These upper bounds depend on the regularization parameters and the values $\widehat{\tau}_j$ or $\widetilde{\tau}_j$. For the inference on the structural parameter, we thus have to control the asymptotic behavior of $\widehat{\tau}_j$ and $\widetilde{\tau}_j$.

Lemma 3.3. *We have*

$$\big\|\widehat{\Sigma}\widehat{\Theta}_j - e_j\big\|_\infty \leq \lambda_j^\Theta/\widehat{\tau}_j^2, \tag{3.4}$$

and

$$\big\|\widehat{M}^T\widehat{\Theta}\widehat{M}\,\widehat{\Theta}_j^M - e_j\big\|_\infty \leq \lambda_j^M/\widetilde{\tau}_j^2. \tag{3.5}$$

We now establish the rate of convergence of the regularized estimators $\widehat{\Theta}$ and $\widehat{\Theta}^M$. The first result in the next proposition was established by van de Geer et al. [2014], and hence the proof is omitted.

Proposition 3.4. *Suppose Assumption 1 is satisfied. If $s_0 = o(\sqrt{n/\log(q)})$, then we have*

$$\|\widehat{\Theta} - \Theta\|_{op,\infty} = O_p\big(s_{\max}\sqrt{\log(q)/n}\big).$$

If, in addition, Assumptions 2 and 3 are satisfied then

$$\big\|\widehat{\Theta}^M - \Theta^M\big\|_{op,\infty} = O_p\big(\omega^2 s_{\max}^M\log(q)/\sqrt{n}\big).$$

3.3. Rate of Convergence

In this subsection, we derive the rate of convergence of the desparsified IV Lasso estimator $\widehat{\beta}$. The next theorem provides an asymptotic upper bound of the bias term Δ, which is key to derive further inference results.

Theorem 3.5. *Let Assumptions 1–3 be satisfied. Then, we have*

$$\sqrt{n}(\widehat{\beta} - \beta^0) = \omega V + \Delta$$

where

$$V = \widehat{\Theta}^M \widehat{M}^T \widehat{\Theta} \mathbf{Z}^T \mathbf{U} / (\sqrt{n}\, \omega)$$

and Δ satisfies

$$\|\Delta\|_\infty = O_p\big(s_0 \max(\omega, \|\Theta M \Theta^M\|_1)(\log q)^2 / \sqrt{n}\big).$$

From Theorem 3.5 we see that the rate of convergence of the desparsified Lasso estimator $\widehat{\beta}$ is affected by the possibly increasing parameter ω. In the next result, we show that the bias term Δ is indeed asymptotically negligible under additional rate requirements. We also see below that the rate of convergence of our estimator is given by $\sqrt{\omega/n}$ under a mild assumption.

Corollary 3.6. *Let Assumptions 1–3 be satisfied. In addition, we assume*

$$s_0 \, (\log q)^2 \, \max\big(\sqrt{\omega}, \|\Theta M \Theta^M\|_1 / \sqrt{\omega}\big) = o(\sqrt{n}). \tag{3.6}$$

Then, we have

$$\sqrt{n/\omega}(\widehat{\beta} - \beta^0) = \sqrt{\omega}\, V + o_p(1).$$

We will see in the next section that V after standardization converges to the standard normal distribution and, in particular, that $\sqrt{\omega}\, V$ is stochastically bounded. We also see that ω enters the sparsity condition in equation (3.6). In the strong identified case, the components of $\widehat{\beta}$ are $\sqrt{n}$ consistent. In the semi-strongly identified case, this rate of convergence may slow down depending on the asymptotic behavior of ω.

Also the next result is an immediate consequence of Corollary 3.6 and provides a bound for linear functionals of $\widehat{\beta} - \beta^0$ uniformly over representers $a \in \mathbb{R}^p$ with ℓ_1 norm which might increase at a rate $K := K(n)$. For some constant $C > 0$, we define $\mathcal{A}_K = \{a \in \mathbb{R}^p : \|a\|_1^2 / K \leq C\}$.

Corollary 3.7. *Let Assumptions 1–3 be satisfied. In addition, we assume*

$$s_0 \, (\log q)^2 \max\big(\sqrt{\omega}, \|\Theta M \Theta^M\|_1 / \sqrt{\omega}\big) = o\big(\sqrt{n/K}\big). \tag{3.7}$$

Then, we have

$$\sup_{a \in \mathcal{A}_K} \Big| \sqrt{n/\omega}\, a^T (\widehat{\beta} - \beta^0) - \sqrt{\omega}\, a^T V \Big| = o_p\big(\sqrt{K}\big).$$

The sparsity restriction (3.7) becomes more restrictive for large values of K. Two examples of linear functionals for which Corollary 3.7 holds are given by vectors a selecting one component of β and vectors a selecting linear combinations of a finite number of components of β, for which $K = 1$ and K is bounded, respectively.

Example 3.1 (Series Approximation). *Let $\phi^K(\cdot)$ be a K–dimensional vector of basis functions used to approximate a nonlinear relationship between Y and a vector of endogenous variables X_{end} which we assume, in this example, to have bounded support. We assume that model (1.1) holds with $X = \phi^K(X_{\text{end}})$. As basis functions, we consider in this example the Cohen-Daubechies-Vial (CDV) wavelet basis. Let us denote by $supp(X_{\text{end}})$ the (bounded) support of X_{end}, then $\sup_{s\in supp(X_{\text{end}})} \|\phi^K(s)\|_1 = O(\sqrt{K})$ for CDV wavelets, see Chen and Christensen [2018, Appendix E], which guarantees that $\phi^K(x_{end}) \in \mathcal{A}_K$, for all $x_{end} \in supp(X_{\text{end}})$. If the assumptions of Corollary 3.7 and the rate restriction (3.7) are satisfied, then Corollary 3.7 yields that*

$$\sup_{s\in supp(X_{\text{end}})} \left| \sqrt{n/\omega}\, \phi^K(s)^T(\widehat{\beta}_{\text{end}} - \beta^0_{\text{end}}) - \sqrt{\omega}\, \phi^K(s)^T V \right| = o_p\big(\sqrt{K}\big).$$

Consequently, for $\phi^K(s)^T(\widehat{\beta}_{\text{end}} - \beta^0_{\text{end}})$ we obtain the rate of convergence $\sqrt{K\omega/n}$, provided that $\sqrt{\omega}\, \phi^K(s)^T V = O_p(\sqrt{K})$ which we establish in the next subsection. This corresponds to the usual variance term in nonparametric IV estimation, see Blundell et al. [2007] or Chen and Pouzo [2012] and Breunig and Johannes [2016] for pointwise rates. In contrast to the sup-norm convergence results of Chen and Christensen [2018, Lemma 3.1] we do not obtain a $\log(K)$ term since we may exploit sparsity constraints on unknown matrices.

3.4. Asymptotic Normality

In this subsection, we establish asymptotic normality of inner products of the desparsified Lasso estimator $\widehat{\beta}$. We also see that asymptotic normality of components of $\widehat{\beta}$ immediately follows.

To achieve the asymptotic distribution of our estimator $\widehat{\beta}$ we consider a normalization factor to standardize the estimator $\widehat{\beta}$. This normalization factor involves the empirical counterpart of the covariance matrix of the 2SLS estimator which is given by

$$\Omega = \Theta^M M^T \Theta \mathbb{E}[U^2 Z Z^T] \Theta M \Theta^M.$$

We then require the following assumption on this covariance matrix Ω. We introduce the set $\mathcal{A} = \{a \in \mathbb{R}^p : a \in \ell_2 \text{ and } \|a\|_1 \le C\|a\|_2\}$ for some constant $C > 0$.

Assumption 4. *There exists a constant $\underline{\sigma} > 0$ such that $\sqrt{a^T \Omega a/\omega} \ge \underline{\sigma}\, \|a\|_2$ for all $a \in \mathcal{A}$.*

Assumption 4 can be easily verified under mild regularity assumptions, such as, the lower bound $\sqrt{\mathbb{E}[U^2|Z]} \ge \underline{\sigma}$, which is a common condition to derive asymptotic distribution results. Indeed, the condition $\sqrt{\mathbb{E}[U^2|Z]} \ge \underline{\sigma}$ implies $a^T \Omega\, a \ge \underline{\sigma}^2 a^T \Theta^M\, a$ and hence, Assumption 4 holds, for instance, if the eigenvalues of Θ^M have a polynomial or exponential decay.

We now propose a heteroscedasticity robust covariance estimator. To obtain the empirical counterpart of Ω, denoted by $\widehat{\Omega}$, we replace the matrices Θ^M, M, and Θ by their regularized empirical counterparts defined in Section 2.3:

$$\widehat{\Omega} = n^{-1} \widehat{\Theta}^M \widehat{M}^T \widehat{\Theta} \mathbf{Z}^T \operatorname{diag}(\widehat{\mathbf{U}})^2 \mathbf{Z} \widehat{\Theta}^T \widehat{M} \left(\widehat{\Theta}^M\right)^T, \tag{3.8}$$

for the vector of Lasso residuals $\widehat{\mathbf{U}} = \big(Y_1 - X_1^T\widetilde{\beta}, \ldots, Y_n - X_n^T\widetilde{\beta}\big)$ and $\widetilde{\beta}$ is the IV Lasso estimator given in (2.4). We now establish asymptotic normality of linear combinations of the components of $\widehat{\beta}$.

Theorem 3.8. *Let Assumption 4 and the conditions of Corollary 3.6 be satisfied. Further, assume that* $\max(\mathbb{E}\|XX^T\|_\infty^2, \mathbb{E}\|ZZ^T\|_\infty^2) = O(1)$. *Then, for all* $a \in \mathcal{A}$ *satisfying*

$$\omega^{3/2} s_{\max}^M \sqrt{\log(q)} + \sqrt{s_M s_{\max}} \max\left(\sqrt{s_M}, \sqrt{s_{\max}}\right) \|\Theta^M\|_1 / \sqrt{\omega} = o\Big(\frac{\sqrt{n}}{\log(q)}\Big) \tag{3.9}$$

we have

$$\sqrt{n/(a^T \widehat{\Omega}\, a)}\, a^T \big(\widehat{\beta} - \beta^0\big) \xrightarrow{d} \mathcal{N}(0,1).$$

Below, we provide some implications of Theorem 3.8. An immediate consequence of Theorem 3.8 is componentwise asymptotic normality, in which case $a = e_j$ for some $1 \leq j \leq p$ where e_j is a p-vector of zeros but for the j-th component that is equal to 1. Another consequence of Theorem 3.8 is asymptotic normality of linear combinations of a finite number of components of $\widehat{\beta}$. In both cases, the restriction imposed in $\mathcal{A}$ is satisfied. But even if the dimension of the low-dimensional subvector of interest increases, the condition $\|a\|_1/\|a\|_2 \leq const.$ can be justified as the following example illustrates.

Example 3.2 (Series Approximation (cont'd))**.** *Let us assume that* $\|\phi^K(X_{end})\|_2 \sim \sqrt{K}$ *almost surely. When a CDV wavelet basis is used, recall* $\sup_{x \in supp(X_{end})} \|\phi^K(x)\|_1 = O(\sqrt{K})$. *For any* x_{end} *in the support of* X_{end}*, we may hence assume that the ratio* $\|\phi^K(x_{end})\|_1/\|\phi^K(x_{end})\|_2$ *is bounded from above uniformly in* n*. The corresponding sieve variance* $\phi^K(x_{end})'\Omega\phi^K(x_{end})$ *increases relative to the associated parameter* ω *which is thus related to Chen and Pouzo [2015] or Chen and Christensen [2018].*

The next theorem establishes asymptotically valid confidence intervals and testing procedures for inner products of β^0. The following two corollaries are direct implications of Theorem 3.8 and hence, their proofs are omitted. Below, Φ denotes the cumulative distribution function of the standard normal distribution.

Corollary 3.9. *Let the assumptions of Theorem 3.8 hold. Then, for all* $a \in \mathbb{R}^p$ *satisfying condition (3.9) we have that for any* $\alpha \in (0,1)$

$$\mathbb{P}\left(a^T\beta^0 \in \left[a^T\widehat{\beta} \pm \Phi^{-1}(1-\alpha/2)\,(a^T\widehat{\Omega}a)^{1/2}/\sqrt{n}\right]\right) = 1 - \alpha + o(1).$$

The following examples illustrate the previous theorem for the componentwise case where $a = e_j$.

Example 3.3 (Componentwise Confidence Intervals)**.** *An asymptotically valid confidence interval for* β_j^0 *at nominal level* α *is given by*

$$\left[\widehat{\beta}_j - \Phi^{-1}(1-\alpha/2)\,\widehat{\Omega}_{jj}^{1/2}/\sqrt{n}, \quad \widehat{\beta}_j + \Phi^{-1}(1-\alpha/2)\,\widehat{\Omega}_{jj}^{1/2}/\sqrt{n}\right].$$

The length of the confidence interval is given by

$$2\Phi^{-1}(1-\alpha/2)\,\widehat{\Omega}_{jj}^{1/2}/\sqrt{n}.$$

We thus see that the length of the confidence interval increases relative to the ratio $\sqrt{\omega/n}$*. This implies that in the strongly identified case the length of the interval is smaller than in the semi-strongly identified case. If the model is close to be weakly identified then the confidence interval is close to have infinite volume. This is in line with the findings of Gautier et al. [2011] who showed that in case of weak instruments, confidence sets can be arbitrarily large.*

Another direct implication of Theorem 3.8 concerns hypothesis testing. For some $a \in \mathbb{R}^p$ (satisfying condition (3.9)) consider the null hypothesis $H_{a,0} : a^T\beta^0 = a^T\beta^H$ for a given vector $\beta^H \in \mathbb{R}^p$.

Corollary 3.10. *Let the assumptions of Theorem 3.8 hold. Then under null hypothesis $H_{a,0}$ we have for any $\alpha \in (0,1)$*

$$\mathbb{P}\left(\frac{\sqrt{n}\left|a^T(\beta^0 - \beta^H)\right|}{\sqrt{a^T\,\widehat{\Omega}\,a}} \geq \Phi^{-1}(1-\alpha/2)\right) = \alpha + o(1).$$

4. Numerical Implementation

This section presents Monte Carlo experiments to analyze the finite sample properties of our estimator. We consider the situation where we have a linear reduced form equation but allow for approximate sparsity. We consider three cases: the case where the true structural relationship is linear and we have homoscedasticity (Section 4.1), the case where the true structural relationship is linear and we have heteroscedasticity (Section 4.2), and finally the homoscedastic case where the true structural relationship is non-linear in the endogenous variable and we use a series approximation (Section 4.3). All experiments are based on 1000 Monte Carlo iterations. The choice of tuning parameters is based on 10-fold cross-validation where we make use of the `R` function `cv.glmnet` of the `glmnet` package (see Appendix C for a description of the cross-validation procedure).

4.1. Linear structural relationship and homoscedasticity

We generate i.i.d. data from the following model

$$\begin{aligned} Y &= \beta_1 X_1 + \beta_{-1}^T X_{-1} + U, \qquad X = (X_1, X_{-1}^T)^T, \qquad \beta^0 = (\beta_1, \beta_{-1}^T)^T, \\ X_1 &= \alpha_1 Z_1 + \alpha_{-1}^T X_{-1} + \sqrt{1-\alpha_1^2}\,V, \qquad Z = (Z_1, X_{-1}^T)^T, \qquad \alpha^0 = (\alpha_1, \alpha_{-1}) \end{aligned} \tag{4.1}$$

with

$$\begin{pmatrix} U \\ V \\ Z \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & \rho & 0 \\ \rho & 1 & 0 \\ 0 & 0 & \Sigma \end{pmatrix}\right) \tag{4.2}$$

where $\Sigma = \left((0.5)^{|j-k|}\right)_{jk}$ is a $q \times q$ matrix. The parameter ρ captures the degree of endogeneity and is varied in the experiments below. The parameters are set in the following way: $\beta_1 = 2$, $\beta_{-1,j} = 1+(j-1)*c$ for $1 \leq j \leq 50$, where c is a constant such that the parameters $\beta_{-1,j}$ are equispaced between 1 and 3, $\beta_{-1,j} = 0$ for $51 \leq j \leq (p-1)$, and $\alpha_{-1,j} = 1/(2j^3)$ for $1 \leq j \leq (q-1)$. The parameter α_1 accounts for the strength of the instrument Z_1 and is varied in the experiments, i.e., we consider $\alpha_1 \in \{1, 0.75, 0.5, 0.25\}$. Note that we multiply the error term in the second equation by $\sqrt{1-\alpha_1^2}$, to ensure that the variance of X_1 does not depend on the value α_1. Since $\mathbb{E}[U^2|Z] = 1$ we are in the homoscedastic case where the covariance matrix simplifies to $\Omega = \mathbb{E}[U^2]\Theta^M$.

The desparsified IV Lasso estimator $\widehat{\beta}$ is computed as in Subsection 2.1. It is based on the initial IV Lasso $\widetilde{\beta}$ given in (2.3) where the tuning parameter λ is chosen via 10-fold cross-validation. The regularized estimators $\widehat{\Theta}$, $\widehat{M}$, and $\widehat{\Theta}^M$ are implemented

as described in Subsection 2.3 with the tuning parameters λ_j^{Θ}, $j = 1, \ldots, q$, and λ_j^M, $j = 1, \ldots, p$, and $c_n = C_0\sqrt{\log(q)/n}$, chosen by 10-fold cross-validation. We emphasize that our implementation of the estimators for high dimensional matrices follows standard procedures in the related literature see, for instance, Meinshausen and Bühlmann [2006]. Alternatively, one could use the procedure proposed in van de Geer et al. [2014] and choose the same tuning parameter, say $\lambda_j^{\Theta} = \lambda_{\Theta}$ (resp. $\lambda_j^M = \lambda_M$), by 10-fold cross-validation among all the q (resp. p) nodewise regressions. We examined this procedure but it slows down the computational time and the results were not better. For large choices of q and p we made use of parallel computing (which is straightforward in `R` given the `parallel` package).

To estimate the covariance matrix Ω we adapt to the instrumental variable setting the idea proposed by Sun and Zhang [2012], which consists in replacing the variance of U by the error variance estimator obtained with the initial IV Lasso $\widetilde{\beta}$, $\widetilde{\sigma}^2 := \sum_{i=1}^n (Y_i - \widetilde{\beta}_1 X_1 - \widetilde{\beta}_{-1}^T X_{-1})^2$. Then, given the estimator $\widehat{\Omega} = \widetilde{\sigma}^2(\hat{\Theta}^M)^T$ we compute the confidence interval for the structural parameter β_1 by following Example 3.3.

We first study the effect of ρ and α on the results of our inference procedure. Here, we take $p = 100$ with one endogenous variable and $q = 100$ exogenous variables (included and excluded covariates). Then, we look at the effect of α when $p = q = 200$. The sample size is fixed to $n = 100$. The results are in Table 1. Here, we report the absolute values of the mean bias for the desparsified IV Lasso estimator $\widehat{\beta}_1$ and for the IV Lasso estimator $\widetilde{\beta}_1$, for different values of the parameters ρ and α_1. The absolute mean is computed over the 1000 Monte Carlo replications. We also report the coverage of our confidence interval for β_1 at the nominal level 95%. Table 1 also reports the average coverage of the intervals for individual coefficients corresponding to variables in either S_0 or S_0^c computed as follows: $AvgCov_\alpha(S_0) = s_0^{-1} \sum_{j \in S_0} \widehat{\mathbb{P}}(\beta_j^0 \in CI_j(\alpha))$ and $AvgCov_\alpha(S_0^c) = (p - s_0)^{-1} \sum_{j \in S_0^c} \widehat{\mathbb{P}}(\beta_j^0 \in CI_j(\alpha))$, where $CI_j(\alpha) = [\widehat{\beta}_j \pm \Phi^{-1}(1-\alpha/2)\, \widehat{\Omega}_{jj}^{1/2}/n^{1/2}]$ according to Corollary 3.9 and $\widehat{\mathbb{P}}$ is obtained as an average over 1000 Monte Carlo iterations.

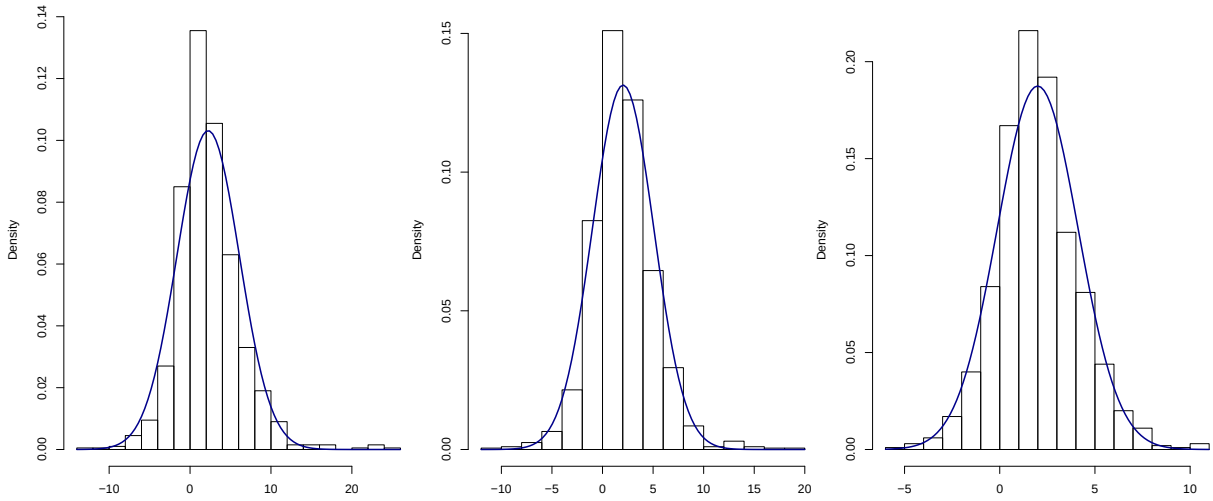
From Table 1 we see that the absolute mean bias of the desparsified IV Lasso estimator $\widehat{\beta}_1$ is considerably smaller than the absolute mean bias of the IV Lasso estimator $\widetilde{\beta}_1$ for each value of ρ and α_1, and also as $p = q$ increases. As α_1 decreases, i.e., the strength of instruments declines, we see that the values of the absolute mean bias of both the desparsified IV Lasso and of the initial lasso estimator $\widetilde{\beta}_1$ become larger when $p = q = 100$. When $p = q = 200$ we see that the effect on the instrument strength on mean the bias of the IV Lasso estimator $\widetilde{\beta}_1$ and our desparsified estimator $\widehat{\beta}_1$ is mixed. From the third column of Table 1 we see that the empirical coverage for β_1 is close to the nominal level of 95%. Concerning the coefficients corresponding to variables in S_0, we have some undercoverage (see the fourth column of Table 1), which is yet less severe when $p = q = 200$. Undercoverage for coefficients in S_0 has been shown for the desparsified Lasso in reduced from regression by van de Geer et al. [2014] in different simulation designs. On the other hand, the coverage for the coefficients corresponding to variables in S_0^c is accurate and even somewhat larger than the nominal coverage probability when $p = q = 100$.

Figure 1 shows the histograms approximating the sampling distribution of our estimator $\widehat{\beta}_1$ for different values of α_1 when $\rho = 0.5$. From this figure we see that there is a perfect fit and that for α_1 small the distribution is slightly right skewed. We have superposed the probability density function of a standard normal, which corresponds to the asymptotic distribution of the estimator.

ρ	α_1	Absolute mean bias($\widehat{\beta}_1$)	Absolute mean bias($\widetilde{\beta}_1$)	Coverage for β_1	Coverage for S_0-coefficients	Coverage for S_0^c-coefficients
$p = q = 100$						
0.7	0.75	0.002	1.750	0.945	0.897	0.978
	0.5	0.031	1.825	0.961	0.898	0.978
	0.25	0.189	1.843	0.948	0.886	0.974
0.5	0.75	0.001	1.757	0.946	0.897	0.978
	0.5	0.039	1.832	0.958	0.898	0.978
	0.25	0.220	1.863	0.944	0.884	0.974
0.3	0.75	0.004	1.758	0.947	0.897	0.978
	0.5	0.018	1.832	0.958	0.898	0.978
	0.25	0.205	1.836	0.945	0.883	0.974
$p = q = 200$						
0.7	0.75	0.079	1.910	0.962	0.923	0.984
	0.5	0.008	1.924	0.969	0.922	0.984
	0.25	0.301	1.806	0.943	0.896	0.978
0.5	0.75	0.082	1.913	0.961	0.923	0.984
	0.5	0.014	1.917	0.970	0.922	0.984
	0.25	0.346	1.819	0.944	0.896	0.978
0.3	0.75	0.087	1.913	0.962	0.923	0.984
	0.5	0.026	1.917	0.969	0.922	0.984
	0.25	0.437	1.856	0.947	0.897	0.978

Table 1: The simulation design is (4.1) with $n = 100$ and varying parameters ρ and α. Absolute mean of the bias for the desparsified IV estimator $\widehat{\beta}_1$ and the initial IV Lasso estimator $\widetilde{\beta}_1$. The last three columns provide coverages of our 95%-confidence interval for β_1, and for coefficients corresponding to variables in either S_0 or S_0^c.

Random support of β^0. As a further exercise, we have analyzed the situation where the support of β^0 is randomly selected. We fix the cardinality of the active set of β^0 equal to $s_0 = 15$ and then the support S_0 of β^0 is obtained as $S_0 = \{u_1, \ldots, u_{15}\}$ where $u_1, \ldots, u_{15}$ is a realization of 15 draws without replacement from $\{1, \ldots, p\}$. The simulation design is as in (4.1) but where we present here only the result for $\alpha_1 = 0.05$. In Table 2 we report the absolute value of the estimated bias (again computed as the difference between the average over the Monte Carlo iterations and the true value of β_1) for our desparsified IV Lasso estimator $\widehat{\beta}_1$ and for the IV Lasso estimator $\widetilde{\beta}_1$. We see that the coverage of our confidence interval for β_1 is close to the nominal coverage of 95%. Table 2 also reports the average coverage of the intervals for individual coefficients corresponding to variables in either S_0 or S_0^c. Again there is some undercoverage for the S_0 coefficients.



(a) $\rho = 0.5$, $\alpha = 0.25$ (b) $\rho = 0.5$, $\alpha = 0.5$ (c) $\rho = 0.5$, $\alpha = 0.75$

Figure 1: Histograms approximating the sampling distribution of $\widehat{\beta}_1$ for the simulation design (4.1).

	Absolute mean bias($\widehat{\beta}_1$)	Absolute mean bias($\widetilde{\beta}_1$)	Coverage for β_1	Coverage for S_0-coefficients	Coverage for S_0^c-coefficients
$p = q = 100$	0.489	1.802	0.941	0.7741	0.9753
$p = q = 150$	0.401	1.869	0.942	0.691	0.979
$p = q = 200$	0.505	1.928	0.936	0.603	0.981

Table 2: Random support for β^0 for varying p and q. Absolute mean of the bias for the desparsified IV estimator $\widehat{\beta}_1$ and the initial IV Lasso estimator $\widetilde{\beta}_1$. The last three columns provide coverages of our 95%-confidence interval for β_1, and for coefficients corresponding to variables in either S_0 or S_0^c when $n = 100$.

4.2. Linear structural relationship and heteroscedasticity

We generate i.i.d. data from the model (4.1) where

$$U = \varepsilon\sqrt{1/2 + \Phi(X_1)} \quad \text{and} \quad \begin{pmatrix} \varepsilon \\ V \\ Z \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0 \\ 0 & 0 & \Sigma \end{pmatrix}\right) \tag{4.3}$$

where Σ is a $q \times q$ matrix that can be set in two different ways. Denote $p_z = q - p_w$, $q = \dim(Z)$ and $p_w = \dim(X_{-1})$, then we have the two following designs for Σ.

Design 1: $\Sigma = \left((0.5)^{|j-k|}\right)_{jk}$;

Design 2: Uncorrelated block structure:

$$\Sigma = \begin{pmatrix} \Sigma_{Z_1 Z_1} & \Sigma_{Z_1 X_{-1}} \\ \Sigma_{Z_1 X_{-1}}^T & \Sigma_{X_{-1} X_{-1}} \end{pmatrix},$$

where $\Sigma_{Z_1X_{-1}}$ is a $p_z \times p_w$ matrix of zeros, $\Sigma_{Z_1Z_1} = \left((0.5)^{|j-k|}\right)_{1\leq j,k\leq p_z}$, and $\Sigma_{X_{-1}X_{-1}} = \left((0.5)^{|j-k|}\right)_{1\leq j,k\leq p_w}$.

In the rest of this section, we fix the degree of endogeneity and the strength of the instruments by setting $\rho = 0.5$ and $\alpha_1 = 1$. The other parameters are set as in the previous simulation with homoscedastic errors. The covariance estimator under heteroscedasticity is implemented as the matrix $\widehat{\Omega}$ in (3.8).

In Tables 3 and 4 we report the absolute value of the estimated bias – computed as the difference between the average over 1000 Monte Carlo iterations and the true value of β_1 given by 2 – for our desparsified IV Lasso estimator $\widehat{\beta}_1$ and for the IV Lasso estimator $\widetilde{\beta}_1$. Table 3 refers to *Design 1* while Table 4 refers to *Design 2*. We see that the bias of $\widehat{\beta}_1$ is again considerably smaller than the one of the initial IV Lasso estimator $\widetilde{\beta}_1$ in absolute value. Compared to the bias reported in Table 1, in presence of heteroskedasticity the bias is larger in absolute value. However the bias of our estimator $\widehat{\beta}_1$ is less affected by heteroscedasticity than the bias of the initial IV Lasso estimator $\widetilde{\beta}_1$. In addition, we report the average coverage of our confidence interval for β_1 at the confidence level of 95%. We see that the coverage increases with q. We also report the average coverage of the intervals for individual coefficients corresponding to variables in either S_0 or S_0^c. For the *Design 2* we also outline in Table 4 the effect of augmenting p. Figure 2 repots the histograms relative to *Design 2* which show that the distribution of $\widehat{\beta}_1$ is more and more concentrated around the true value of β_1 as n and q increase.

	Absolute mean bias($\widehat{\beta}_1$)	Absolute mean bias($\widetilde{\beta}_1$)	Coverage for β_1	Coverage for S_0-coefficients	Coverage for S_0^c-coefficients
$p = q = 100$	0.326	1.605	0.908	0.880	0.967
$p = q = 150$	0.387	1.735	0.908	0.894	0.969
$p = q = 200$	0.450	1.820	0.922	0.901	0.969

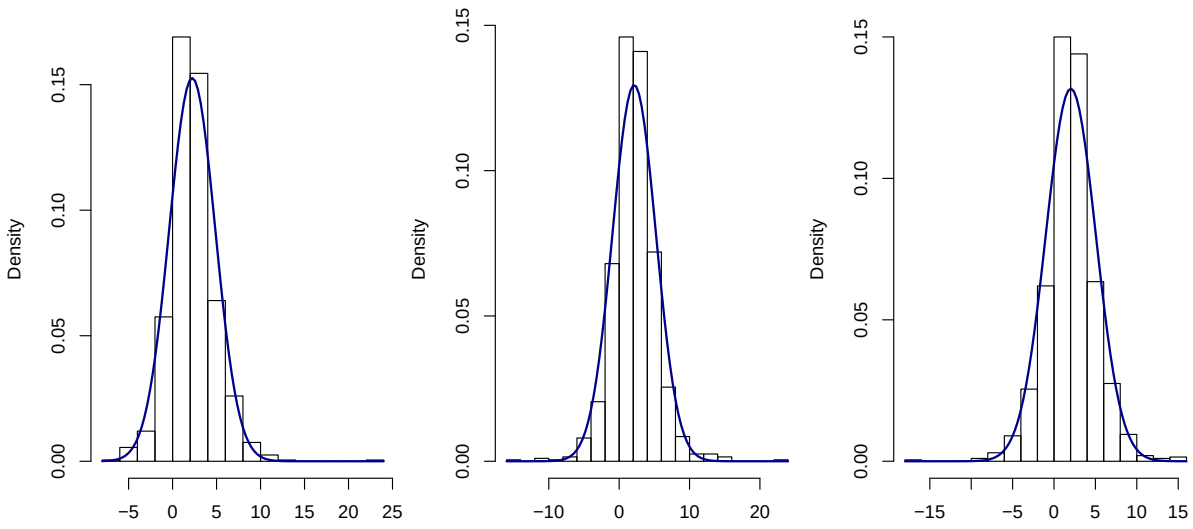
Table 3: Heteroskedastic case - Design 1 from model (4.3) with $n = 100$ and varying p and q. The same explanations as in Table 1 apply.

	Absolute mean bias($\widehat{\beta}_1$)	Absolute mean bias($\widetilde{\beta}_1$)	Coverage for β_1	Coverage for S_0-coeff.	Coverage for S_0^c-coeff.
$p = q = 100$	0.236	1.734	0.936	0.879	0.966
$p = 100$, $q = 150$	0.016	0.665	0.794	0.892	0.966
$p = q = 150$	0.138	1.800	0.936	0.894	0.969
$p = 150$, $q = 200$	0.029	0.653	0.836	0.897	0.965
$p = q = 200$	0.029	1.885	0.939	0.899	0.968

Table 4: Heteroskedastic case - Design 2 from model (4.3) with $n = 100$ and varying p and q. The same explanations as in Table 1 apply.

4.3. Increasing number of endogenous variables

This simulation corresponds to Example 3.2 about series approximation. Let $\phi^J(X_1) := (\phi_1(X_1), \ldots, \phi_J(X_1))^T$ be a J-vector of basis functions. We generate i.i.d. data from the



(a) $n = p = q = 100$ (b) $p = q = 150$, $n = 100$ (c) $p = q = 200$, $n = 100$

Figure 2: Heteroskedastic case - Design 2. Histograms approximating the sampling distribution of $\widehat{\beta}_1$ for the simulation design (4.3) based on a Monte Carlo experiment with 1000 iterations.

following model

$$\begin{aligned} Y &= \varphi(X_1) + \beta_{-1}^T X_{-1} + U, \qquad X = (\phi^J(X_1), X_{-1}^T)^T, \\ X_1 &= \alpha_1^T Z_1 + \alpha_{-1}^T X_{-1} + V/2, \qquad Z = (Z_1^T, X_{-1}^T)^T, \end{aligned} \tag{4.4}$$

with (U, V, Z) is generated from (4.2) again with $\Sigma = \left((0.5)^{|j-k|}\right)_{jk}$. Moreover, $\varphi(X_1) = X_1^2/4$ so that if $(\phi_j)_j$ are polynomials we have that $\phi^J(X_1) = (1, X_1, (X_1/2)^2, \ldots, (X_1/J)^J)$ and $\beta_1 = (0, 0, 1, 0, \ldots, 0)^T$. We take $J = 10$, $\dim(X_{-1}) = 100 - J$ and $q = 100$ exogenous variables (included and excluded covariates). The parameters are set in the following way: $\beta_{-1,j} = 1 + (j-1) * c$ for $1 \leq j \leq 50$, where c is a constant such that the parameters $\beta_{-1,j}$ are equispaced between 1 and 3, $\alpha_{1,j}$ has components equispaced between 1 and 0.5, $\beta_{-1,j} = 0$ for $51 \leq j \leq (p - J)$and $\alpha_{-1,j} = j^{-3}/2$ for $1 \leq j \leq 50$. The results of this simulation are reported in Table 5. We report the absolute mean of the bias for the desparsified IV estimator and the initial IV Lasso estimator of the parameter $\beta_{1,3} = 1$, that is, the coefficient of the second order polynomial. We see that we obtain some undercoverage for the coefficient $\beta_{1,3}$ but the coverages for S_0-coefficients and S_0^c-coefficients are close or beyond the 95% nominal level.

5. Application to the Logit Demand Estimation

In this section we apply our method to estimate the price coefficient in a logit model of demand for automobiles using market share data. This application follows the empirical illustration in Chernozhukov et al. [2015] and the aim is to estimate the price effect on

	Absolute mean bias($\widehat{\beta}_{1,3}$)	Absolute mean bias($\widetilde{\beta}_{1,3}$)	Coverage for $\beta_{1,3}$	Coverage for S_0-coefficients	Coverage for S_0^c-coefficients
$p = q = 100$	0.070	0.890	0.900	0.967	0.966
$p = q = 150$	0.033	0.800	0.907	0.955	0.977
$p = q = 200$	0.048	0.706	0.883	0.932	0.978

Table 5: The simulation design is (4.4) with $n = 100$. The same explanations as in Table 1 apply.

the market share of a particular car. We consider the following system of equations:

$$\begin{aligned} \log(s_{it}) - \log(s_{0t}) &= \beta_0^0 p_{it} + x_{it}^T \beta_1^0 + u_{it}, \\ p_{it} &= z_{it}^T \alpha_{0,0} + x_{it}^T \alpha_{0,1} + \varepsilon_{it}, \end{aligned}$$

where s_{it} is the market share of product i in market t, s_{0t} denotes the outside option, p_{it} is the price which is endogenous, x_{it} are observed product characteristics which are exogenous, and z_{it} is a set of instrumental variables.

In our application we use the same product characteristics as in Chernozhukov et al. [2015] and Berry et al. [1995], that is, x_{it} contains an air conditioning dummy, horsepower divided by weight, miles per dollar, vehicle size and a time trend. We center all the variables in order to eliminate the constant. The instruments for price are formed by using the idea developed in Berry et al. [1995] that characteristics of other products satisfy an exclusion restriction of the type $\mathbb{E}[u_{it}|x_{jt'}] = 0$ for any t' and any $j \neq i$. Therefore, any function of characteristics of other products may be used as an instrument for price. We follow Chernozhukov et al. [2015] and form instruments as

$$z_{k,it} = \left(\sum_{j \neq i, j \in \mathcal{I}_f} x_{k,jt}, \sum_{j \neq i, j \notin \mathcal{I}_f} x_{k,jt} \right), \tag{5.1}$$

where $x_{k,it}$ and $z_{k,it}$ denote the k-th element of x_{it} and z_{it}, respectively, and $\mathcal{I}_f$ denotes the set of products produced by firm f. In this way we have a set of 10 excluded instruments.

In addition, because economic theory does not specify the functional form in which the elements of x_{it} enter the regression model, we also consider first-order interaction terms of the variables in x_{it}, and quadratic and cubic transformations of the continuous variables in x_{it} for a total of 18 new variables. In this way, we have a vector of augmented controls, denoted by x_{it}^a that contains x_{it} and these new variables. The corresponding vector of augmented excluded instruments is then given by z_{it}^a where z_{it}^a is constructed as in (5.1) but with $x_{k,jt}$ replaced by $x_{k,jt}^a$. By using the generic notation in the paper: $X = (p_{it}, x_{it}^{aT})^T$, $Z = (z_{it}^{aT}, x_{it}^{aT})^T$, and $\beta^0 = (\beta_0^0, \beta_1^{0T})^T$.

In our data set, we have a total of $n = 2217$ observations, 23 augmented controls, and 71 augmented instruments Z. According to our theory, the strength of identification is measured through the parameter ω. In the non-augmented framework with 10 excluded instruments, the estimated ω is equal to $6.44 \cdot 10^{-06}$ and hence, relatively small. When we augment the number of controls and instruments the estimate of ω increases. This means that adding polynomial transformations and interactions, if on the one hand makes the model more flexible, on the other hand reduces the strength of identification, which is not surprising. This is not a problem since our approach is robust to semi strongly identified models.

In our application we estimate the covariance matrix to construct the confidence intervals by using our heteroscedastic robust estimator. In Table 6 we show the results obtained with different estimators. Together with the point estimate, we also report the lower and upper bound of the 95%-confidence interval. We first compute the OLS and 2SLS estimators obtained without augmenting the controls and the instruments. Then, we show the results obtained with augmented controls and instruments with three estimators: the OLS, the 2SLS and our desparsified IV Lasso estimator. To incorporate uncertainty induced by sample splitting for the selection of the tuning parameters, we use the finite-sample adjustments proposed by Chernozhukov et al. [2018]. Specifically, we present estimation results as the median of desparsified IV Lasso estimate for 200 different sample splits. Here, we retain estimates of our desparsified IV Lasso estimate gives a low number of products with inelastic demand. As we explained below, inelastic demand is unrealistic in a setup of firms maximizing their profit. The standard deviation is computed from the median, over the same 200 seeds, of the adjusted variances (following the variance adjustment in equation (3.14) in Chernozhukov et al. [2018]).

We see that the estimated price coefficient becomes larger in absolute value when we move from the Baseline OLS (-0.0886 with a standard deviation of 0.0043) to the Baseline 2SLS (-0.1419 with a standard deviation of 0.0119), which can be interpreted as the fact that the OLS estimator is biased because of endogeneity of price. The magnitude of the OLS estimated price coefficient increases when we use augmented controls (the Augmented OLS estimate is -0.0991 with a standard deviation of 0.0046) and, when we use augmented controls and instruments, the Augmented 2SLS price coefficient estimate is -0.1273 with a standard deviation of 0.0076. The largest value in absolute value is obtained with our desparsified IV Lasso estimator which gives a price coefficient estimate equal to -0.2104 with a (finite-sample adjusted) standard deviation of 0.0306. The estimated price coefficient that we obtain with our estimator is similar to the one in Chernozhukov et al. [2015] based on the double-selection approach, which is equal to -0.221. The latter is contained in our 95%-confidence interval for β_0^0. The slight difference is mainly due to the randomness in choosing the tuning parameters.

Notice that as we move from the baseline results to the results based on augmented controls and instruments, the estimates become more plausible from an economic theory point of view. Indeed, in our setup firms maximizing their profit should face elastic demand for all products. In line with this insight, whereas the baseline OLS (resp. 2SLS) point estimates imply inelastic demand for 1502 (resp. 670) products, our desparsified IV Lasso estimate inelastic demand for only 32 products using augmented controls and variables. The number of products with inelastic demand are reported on the last column of Table 6. For our desparsified IV estimator, the number of inelastic is larger than the one based on the double-selection procedure of Chernozhukov et al. [2015] which is equal to 12. This difference is due to the fact that our estimate of the price coefficient is slightly lower than the double-selection based estimate, as discussed above.

Overall, we see that our desparsified IV estimator and inference procedure perform well in empirical applications and give plausible results. In addition, because our procedure is robust to heteroscedasticity, our inference remains valid when the regression error term is heteroscedastic.

	Price Coefficient	Lower	Upper	Number Inelastic
Baseline OLS	-0.0886	-0.0971	-0.0802	1502
Baseline 2SLS	-0.1419	-0.1651	-0.1186	670
Augmented OLS	-0.0991	-0.1081	-0.0901	1405
Augmented 2SLS	-0.1273	-0.1423	-0.1124	874
Desparsified IV	-0.2104	-0.2704	-0.1504	32

Table 6: Logit Demand Estimation. Comparison of different estimators for the price coefficient β_0^0. "Baseline OLS" refers to the OLS estimate obtained with the non augmented controls x_{it}, "Augmented OLS" refers to the OLS estimate obtained with the augmented controls x_{it}^a, "Baseline 2SLS" refers to the 2SLS estimate obtained with the non augmented controls x_{it} and the non augmented instruments z_{it}, "Augmented 2SLS" refers to the 2SLS estimate obtained with the augmented controls x_{it}^a and the augmented instruments z_{it}^a, "Desparsified IV" refers to our desparsified IV Lasso estimate obtained with the augmented controls x_{it}^a and the augmented instruments z_{it}^a. "Lower" and "Upper" denote the lower and upper bound of the 95%-confidence interval for β_0^0. Finally, "Number Inelastic" refers to the point estimate of the number of products for which demand is estimated to be inelastic.

A. Appendix: Proofs

Let Assumption 1 hold. By using the Cauchy-Schwarz inequality, the definition of s_M, and the assumption that $\lambda_{\max}(\Sigma) = O(1)$ and $\lambda_{\max}(M^T\Theta M) = O(1)$ we obtain

$$\begin{aligned}\|M\|_1 &\leq \sqrt{s_M} \max_{1\leq j\leq p} \|M_j\|_2 \\ &= O\big(\sqrt{s_M} \max_{1\leq j\leq p} \|\Theta^{1/2} M e_j\|_2\big) \\ &= O\big(\sqrt{s_M}\big)\end{aligned}$$

where M_j denotes the j–th column of the matrix M. Similarly, the sparsity constraint on Θ implies

$$\|\Theta\|_1 \leq \sqrt{s_{\max}} \max_{1\leq j\leq q} \|\Theta_j\|_2 = O(\sqrt{s_{\max}}).$$

For the next proofs, we require the following notation. For $j = 1, \ldots, p$, recall the definition $\gamma_j := \operatorname{argmin}_{\gamma\in\mathbb{R}^{p-1}} \|(\Theta^{1/2}M)_j - (\Theta^{1/2}M)_{-j}\gamma\|_2^2$. We also define $\tau_j^2 := \|(\Theta^{1/2}M)_j - (\Theta^{1/2}M)_{-j}\gamma_j\|_2^2$. Introduce a vector $\Gamma_j := (\Gamma_{kj})_{k=1}^p$ with $\Gamma_{kj} = -\gamma_{kj}$ for $k \neq j$ and otherwise 1, where γ_{kj} is the k–th entry of γ_j. Then, we have $\tau_j^2 = \Gamma_j^T M^T \Theta M \Gamma_j$ since $\Theta^{1/2}M\Gamma_j = (\Theta^{1/2}M)_j - (\Theta^{1/2}M)_{-j}\gamma_j$. It also holds $\tau_j^2 = 1/\Theta_{jj}^M$, which can be seen as follows. The first order condition for γ_j yields

$$(\Theta^{1/2}M)_{-j}^T \Theta^{1/2} M \Gamma_j = 0,$$

and thus,

$$M^T\Theta M\Gamma_j = \big((\Theta^{1/2}M)_j^T \Theta^{1/2} M\Gamma_j\big)\, e_j = \Gamma_j^T M^T \Theta M \Gamma_j e_j = \tau_j^2 e_j,$$

where we have used the fact that $\Theta^{1/2}M\Gamma_j = (\Theta^{1/2}M)_j - (\Theta^{1/2}M)_{-j}\gamma_j$ together with the first order condition for γ_j to get the second equality. Further, by premultiplying with Θ^M we obtain

$$\Gamma_j = \tau_j^2\, \Theta^M e_j$$

and since $e_j^T \Gamma_j = 1$ we obtain $\tau_j^2 = 1/\Theta_{jj}^M$. By the definition of ω we obtain the following lower bound for τ_j:

$$\tau_j^2 = 1/\Theta_{jj}^M \geq 1/\lambda_{\max}(\Theta^M) = \lambda_{\min}(M^T \Theta M) = \omega^{-1}, \tag{A.1}$$

which we will use in the following proofs. Reversely, τ_j is bounded from above by the maximal eigenvalue of $M^T \Theta M$ which we assume to be bounded. This implies that

$$\begin{aligned} \|\gamma_j\|_1^2 &\leq C\Big(s_j^M \|\gamma_j\|_2^2 + (\log(q))^2/n\Big) \\ &\leq C\Big(s_j^M + (\lambda_j^M)^2\Big). \end{aligned}$$

where we have used the upper bound $\|\gamma_j\|_1 \leq \|\gamma_{j,S_j}\|_1 + \|\gamma_{j,S_j^c}\|_1$, the Cauchy-Schwarz inequality and (2.6) to get the first inequality. Below we also use for matrices A and B the inequalities

$$\|AB\|_\infty \leq \|A\|_\infty \|B\|_1 \qquad \text{and} \qquad \|AB\|_\infty \leq \|B\|_\infty \|A^T\|_1.$$

A.1. Proofs of the Main Results

Proof of Lemma 3.1. We make use of the inequality

$$\|M^T \Sigma^{-1} v\|_2 \leq \|v\|_2 \sqrt{\lambda_{\max}(MM^T)}/\lambda_{\min}(\Sigma)$$

for all $v \in \mathbb{R}^q$ and the fact that Σ has eigenvalues uniformly bounded away from zero by Assumption 1 *(iii)*. Consequently, sub-Gaussianity of Z implies sub-Gaussianity of $\widetilde{Z} := M^T \Sigma^{-1} Z$. We make use Lemma 5.2 (and the proof of Theorem 2.4) in van de Geer et al. [2014] to the reduced form model

$$Y = \widetilde{Z}^T \beta^* + V,$$

where $\beta^* := \Sigma^{-1/2} M \beta^0$ and $V = Y - Z^T \Sigma^{-1} \mathbb{E}[YZ]$. Hence, sub-Gaussianity of $\widetilde{Z}$ and Assumption 1 *(iii)* imply

$$\|\beta_{S_0}\|_1^2 \leq C s_0 \, \beta^T M^T \Sigma^{-1} \widehat{\Sigma} \Sigma^{-1} M \beta \tag{A.2}$$

wpa1, for all $\beta \in \mathcal{B}$. □

Proof of Theorem 3.5. The proof is based on the decomposition

$$\Delta = \sqrt{n}\big(\widehat{\Theta}^M \widehat{M}^T \widehat{\Theta} \widehat{M} - I_p\big)(\widetilde{\beta} - \beta^0) - \sqrt{n}\, \widehat{\Theta}^M \widehat{M}^T \widehat{\Theta}(\widehat{M} - \widetilde{M})(\widetilde{\beta} - \beta^0).$$

We observe

$$\begin{aligned} \|\Delta\|_\infty/\sqrt{n} &\leq \|\big(\widehat{\Theta}^M \widehat{M}^T \widehat{\Theta} \widehat{M} - I_p\big)(\widetilde{\beta} - \beta^0)\|_\infty + \|\widehat{\Theta}^M \widehat{M}^T \widehat{\Theta}(\widehat{M} - \widetilde{M})(\widetilde{\beta} - \beta^0)\|_\infty \\ &\leq \|\widehat{\Theta}^M \widehat{M}^T \widehat{\Theta} \widehat{M} - I_p\|_\infty \|\widetilde{\beta} - \beta^0\|_1 + \|\widehat{\Theta}^M \widehat{M}^T \widehat{\Theta}\|_{op,\infty} \|(\widehat{M} - \widetilde{M})(\widetilde{\beta} - \beta^0)\|_\infty. \end{aligned}$$

Further, the upper bound given in (3.5) implies that

$$\|\Delta\|_\infty \leq \sqrt{n} \max_{1 \leq j \leq p} \big\{\lambda_j^M / \widetilde{\tau}_j^2\big\} \, \|\widetilde{\beta} - \beta^0\|_1 + \sqrt{n} \|\widehat{\Theta}^T \widehat{M} (\widehat{\Theta}^M)^T\|_1 \|\widehat{M} - \widetilde{M}\|_\infty \|\widetilde{\beta} - \beta^0\|_1.$$

By the definition of the regularized estimator $\widehat{M}$ given in (2.10) it holds for all j, k:

$$\begin{aligned}|\widetilde{M}_{jk} - \widehat{M}_{jk}| &= |\widetilde{M}_{jk}| \, \mathbb{1}\left\{|\widetilde{M}_{jk}| < C_0\sqrt{\log(q)/n}\right\} \\ &< C_0\sqrt{\log(q)/n},\end{aligned}$$

which implies

$$\|\widetilde{M} - \widehat{M}\|_\infty < C_0\sqrt{\log(q)/n}. \tag{A.3}$$

Thus, using that $\lambda_j^M \sim \log(q)/\sqrt{n}$ uniformly in j, by Assumption 2 (i) we obtain

$$\|\Delta\|_\infty \leq C\log(q)\left(\max_{1\leq j\leq p}\widetilde{\tau}_j^{-2} + \|\widehat{\Theta}^T\widehat{M}(\widehat{\Theta}^M)^T\|_1\right)\|\widetilde{\beta} - \beta^0\|_1.$$

In the following, we consider the events

$$\mathcal{C} := \left\{\|\beta_{S_0}\|_1^2 \leq s_0\beta^T\widehat{M}^T(\widehat{\Theta} + \widehat{\Theta}^T)\widehat{M}\beta/c^2 \text{ for all } \|\beta_{S_0^c}\|_1 \leq 3\|\beta_{S_0}\|_1\right\}$$

and $\mathcal{T} := \left\{\|\widehat{M}^T(\widehat{\Theta} + \widehat{\Theta}^T)\mathbf{Z}^T\mathbf{U}/n + \widehat{M}^T(\widehat{\Theta} + \widehat{\Theta}^T)(\widetilde{M} - \widehat{M})\beta^0\|_\infty \leq C\lambda^*\right\}$ for some sufficiently large constant C where $\lambda^* > Cc_1\lambda$ for some $c_1 > 1$ and recall $\lambda \sim \log(q)/\sqrt{n}$.

On the event $\mathcal{C} \cap \mathcal{T}$ we have

$$\|\widetilde{\beta} - \beta^0\|_1 \leq Cs_0\log(q)/\sqrt{n},$$

which follows directly from van de Geer [2016, Theorem 2.2].[1] From the proof of Proposition 3.4 in Appendix A.2 we also have that $\widetilde{\tau}_j^2$ is a consistent estimator of τ_j^2. Further, Propositions 3.3 and 3.4 together with the lower bound (A.1) yield

$$\|\Delta\|_\infty \, \mathbb{1}_{\mathcal{C}\cap\mathcal{T}} = O_p\Big(s_0\log(q)^2/\sqrt{n}\,\max\left(\omega, \|\Theta M\Theta^M\|_1\right)\Big).$$

It is thus sufficient to show $\mathbb{1}_{\mathcal{C}\cap\mathcal{T}} = 1$ wpa1. We proceed in two steps and control the sets $\mathcal{T}$ and $\mathcal{C}$ separately. To handle the set $\mathcal{T}$ note that

$$\begin{aligned}&\left\|\widehat{M}^T\,\widehat{\Theta}\mathbf{Z}^T\mathbf{U}/n + \widehat{M}^T\widehat{\Theta}(\widetilde{M} - \widehat{M})\beta^0\right\|_\infty \\ &\qquad\leq \underbrace{\|\mathbf{U}^T\mathbf{Z}\,\Theta\,M/n\|_\infty}_{I} + \underbrace{\|(\widehat{M}^T\widehat{\Theta} - M^T\Theta)\big(\mathbf{U}^T\mathbf{Z}/n + (\widetilde{M} - \widehat{M})\beta^0\big)\|_\infty}_{II} \\ &\qquad\quad + \underbrace{\|M^T\Theta(\widetilde{M} - \widehat{M})\beta^0\|_\infty}_{III}.\end{aligned}$$

To bound I, we make use of Nemirovski's inequality (see, for instance, p. 509 in Bühlmann and Van De Geer [2011]) and $\mathbb{E}[U^2|Z] \leq \sigma^2$ to get

$$\begin{aligned}\mathbb{E}\Big(\max_{1\leq j\leq p}\big|(\mathbf{U}^T\mathbf{Z}\,\Theta\,M)_j/n\big|\Big)^2 &\leq 8\log(2p)\frac{1}{n^2}\sum_{i=1}^n\mathbb{E}\max_{1\leq j\leq p}|U_i(Z_i^T\,\Theta\,M)_j|^2 \\ &\leq 8\log(2p)n^{-1}\sigma^2\mathbb{E}\max_{1\leq j\leq p}|(Z^T\,\Theta\,M)_j|^2\end{aligned}$$

[1] Apply van de Geer [2016, Theorem 2.2] with, in their notation, $L = 3$, $\widehat{\phi}^2(3, \widetilde{S}_0) = c^2$, $X = (\widehat{\Theta} + \widehat{\Theta}^T)^{1/2}\widehat{M}$, $Y = (\widehat{\Theta} + \widehat{\Theta}^T)^{1/2}\mathbf{Z}^T\mathbf{Y}$, $\epsilon = (\widehat{\Theta} + \widehat{\Theta}^T)^{1/2}(\mathbf{Z}^T\mathbf{Y} - \widehat{M}\beta^0)$.

and hence, we obtain $I = O_p\big(\sqrt{\mathbb{E}\|Z^T\Theta M\|_\infty^2 \log(p)/n}\big) = O_p\left(\log(p)/\sqrt{n}\right)$ by using Assumption 2 (ii). Under Assumption 2 (i) we have that $\lambda \sim \log(q)/\sqrt{n}$ and thus, $I = O_p(\lambda)$.

Next, we consider II. We have

$$\begin{aligned} II &= \|\widehat{M}^T\widehat{\Theta} - M^T\Theta\|_{op,\infty}\|\mathbf{U}^T\mathbf{Z}/n + (\widetilde{M} - \widehat{M})\beta^0\|_\infty \\ &\leq \Big(\|\widehat{\Theta}\|_{op,\infty}\|\widehat{M} - M\|_1 + \|M\|_1\|\widehat{\Theta} - \Theta\|_{op,\infty}\Big)\big(\|\mathbf{U}^T\mathbf{Z}/n\|_\infty + \|\widetilde{M} - \widehat{M}\|_\infty\|\beta^0\|_1\big). \end{aligned}$$

Again, due to Nemirovski's inequality, we have $\|\mathbf{U}^T\mathbf{Z}/n\|_\infty = O_p(\sqrt{\log(q)/n})$ under Assumption 1 (iii) and condition $\mathbb{E}[U^2|Z] \leq \sigma^2$ imposed in Assumption 2 (ii). Furthermore, $\|\widetilde{M} - \widehat{M}\|_\infty = O\big(\sqrt{\log(q)/n}\big)$ by inequality (A.3). We also have $\|\widehat{M} - M\|_1 = O_p\big(s_M\sqrt{\log(q)/n}\big)$ and $\|\widehat{\Theta} - \Theta\|_{op,\infty} = O_p\big(s_{\max}\sqrt{\log(q)/n}\big)$ from Propositions 3.2 and 3.4. Now using that $\|\beta^0\|_1 = O(s_0)$ (since $\|\beta^0\|_1 \leq \|\beta^0_{S_0}\|_1 + Cs_0\sqrt{\log(p)/n} \leq s_0\|\beta^0\|_\infty + o(1)$ by using the fact that $\beta^0_{S_0} \in \mathcal{B}$, (2.5) and (3.2), and $\|\beta^0\|_\infty = O(1)$) we obtain

$$\begin{aligned} II &= O_p\big((1 + \|\beta^0\|_1)\big(s_M\|\Theta\|_1 + s_{\max}\|M\|_1\big)\log(q)/n\big) \\ &= O_p\big(s_0\max\big(s_M\sqrt{s_{\max}}, s_{\max}\sqrt{s_M}\big)\log(q)/n\big) \\ &= o_p(\sqrt{\log(p)/n}) \end{aligned}$$

employing (3.2) in Assumption 2 to get the last equality. Remark that to get the first equality we have used the fact that $\|\widehat{\Theta}\|_{op,\infty} \leq \|\widehat{\Theta} - \Theta\|_{op,\infty} + \|\Theta\|_1$ because Θ is symmetric, and by Proposition 3.4 $\|\widehat{\Theta} - \Theta\|_{op,\infty} = O_p(s_{\max}\sqrt{\log(q)/n})$ which is negligible with respect to the other terms under (3.2). Consider III. We have

$$III \leq \max_j \big|(\Theta M)_j^T(\widehat{M} - M)\beta^0\big| + \max_j \big|(\Theta M)_j^T(\widetilde{M} - M)\beta^0\big|, \tag{A.4}$$

where the second summand can be bounded again by using Nemirovski's inequality:

$$\begin{aligned} \mathbb{E}\big\|M^T\Theta(\widetilde{M} - M)\beta^0\big\|_\infty^2 &= \mathbb{E}\max_{1\leq j\leq p}\big|n^{-1}\sum_i (\Theta M)_j^T Z_i X_i^T\beta^0 - (\Theta M)_j^T M\beta^0\big|^2 \\ &\leq 8\log(2p)n^{-1}\mathbb{E}\max_{1\leq j\leq p}|(\Theta M)_j^T ZX^T\beta^0|^2 \\ &\leq 8\log(2p)n^{-1}\mathbb{E}\big[(X^T\beta^0)^2\|M^T\Theta Z\|_\infty^2\big] \\ &= O\big(\log(p)^2/n\big), \end{aligned}$$

where we have used Assumption 2 (ii) to get the last line. For the first summand on the right hand side of (A.4) we observe

$$\begin{aligned} \max_j \big|(\Theta M)_j^T(\widehat{M} - M)\beta^0\big| &\leq \|\Theta M\|_\infty\|\widehat{M} - M\|_1\|\beta^0\|_1 \\ &= O_p\left(s_0\sqrt{s_{\max}}s_M\sqrt{\log(q)/n}\right) \\ &= O_p\left(\log(q)/\sqrt{n}\right) \end{aligned} \tag{A.5}$$

due to Assumption 1 *(iii)* which implies $\|\Theta\|_1 = O(\sqrt{s_{\max}})$, Assumption 2 *(ii)*, the second result of Proposition 3.2 and the first rate restriction imposed in Assumption 2 *(iii)*.

It remains to control $\mathcal{C}$. By Lemma 3.1 it holds for all $\|\beta_{S_0^c}\|_1 \leq 3\|\beta_{S_0}\|_1$ that

$$\|\beta_{\widetilde{S}_0}\|_1^2 \leq s_0\,\beta^T M^T\Sigma^{-1}\widehat{\Sigma}\Sigma^{-1}M\beta/\widetilde{c}^2$$

wpa1, for some constant $\widetilde{c} > 0$. Thus, in order to prove that $\mathcal{C}$ holds wpa1 it suffices to show that for some sufficiently small constant $c^* > 0$ it holds

$$s_0 \|M^T \Theta (\widehat{\Sigma} - \Sigma) \Theta M\|_\infty \leq c^*/2 \quad \text{wpa1} \tag{A.6}$$

and

$$s_0 \|M^T \Theta M - \widehat{M}^T \widehat{\Theta} \widehat{M}\|_\infty \leq c^*/2 \quad \text{wpa1}. \tag{A.7}$$

To prove (A.6), note that $\|\widehat{\Sigma} - \Sigma\|_\infty \leq c' \sqrt{\log(q)/n}$ wpa1 for some constant $c' > 0$, see *e.g.* van de Geer [2016, Problem 14.2], and thus the result follows by

$$s_0 \|\Theta M\|_1^2 \sqrt{\frac{\log(q)}{n}} \leq c^{**}$$

for some constant c^{**} that is chosen small enough. This inequality is indeed satisfied due to $\|\Theta M\|_1^2 \leq s_{\max} s_M$ and the rate requirement imposed in Assumption 2 (*iii*).

To show (A.7) we first make the decomposition $\|M^T \Theta M - \widehat{M}^T \widehat{\Theta} \widehat{M}\|_\infty \leq \|M^T \Theta M - \widehat{M}^T \Theta \widehat{M}\|_\infty + \|\widehat{M}^T (\widehat{\Theta} - \Theta) \widehat{M}\|_\infty$. Then,

$$\begin{aligned} s_0 \|\widehat{M}\|_1^2 \|\widehat{\Theta} - \Theta\|_\infty &\leq 2 s_0 \left(\|M\|_1^2 + \|\widehat{M} - M\|_1^2 \right) \|\widehat{\Theta} - \Theta\|_\infty \\ &\leq C s_0 s_M^2 \big(1 + \log(q)/n\big) \sqrt{s_{\max}} \sqrt{\log(q)/n}, \end{aligned}$$

wpa1, where we have used Assumption 2 (*ii*) to get $\|M\|_1 \leq s_M \|M\|_\infty = O(s_M)$, the second result of Proposition 3.2 and the result $\|\widehat{\Theta} - \Theta\|_\infty = O_p(\sqrt{s_{\max} \log(q)/n})$ (see van de Geer et al. [2014]). Moreover,

$$\begin{aligned} &\|M^T \Theta M - \widehat{M}^T \Theta \widehat{M}\|_\infty \\ &\qquad \leq \|M - \widehat{M}\|_1 \|\Theta\|_1 \|M\|_\infty + \left(\|M\|_\infty + \|\widehat{M} - M\|_1 \right) \|\Theta\|_1 \|\widehat{M} - M\|_1 \\ &\qquad \leq C s_M \sqrt{\log(q)/n} \sqrt{s_{\max}} \left(1 + s_M \sqrt{\log(q)/n} \right) \end{aligned}$$

wpa1, where we have used Assumptions 1 (*iii*) and 2 (*ii*) and the second result of Proposition 3.2. Consequently, by the rate restriction $s_0 s_M \sqrt{s_{\max}} = o(\sqrt{n/\log(q)})$ in Assumption 2 (*iii*), result (A.7) holds wpa1. □

Proof of Theorem 3.8. We proceed in two steps. First, we show $\sqrt{n/(a^T \Omega\, a)}\, a^T \big(\widehat{\beta} - \beta^0\big) \xrightarrow{d} \mathcal{N}(0,1)$. We make use of the decomposition

$$\begin{aligned} \sqrt{n/(a^T \Omega\, a)}\, a^T \big(\widehat{\beta} - \beta^0\big) = &\underbrace{a^T \Theta^M M^T \Theta \mathbf{Z}^T \mathbf{U} / \sqrt{n(a^T \Omega\, a)}}_{=I} \\ &+ \underbrace{a^T \big(\widehat{\Theta}^M \widehat{M}^T \widehat{\Theta} - \Theta^M M^T \Theta\big) \mathbf{Z}^T \mathbf{U} / \sqrt{n(a^T \Omega\, a)}}_{=II} \\ &+ \Delta \|a\|_1 / \sqrt{a^T \Omega\, a}. \end{aligned}$$

Since $\sqrt{a^T \Omega\, a} \geq \underline{\sigma} \sqrt{\omega} \|a\|_2$ it holds $\|a\|_1 / \sqrt{a^T \Omega\, a} = O(\sqrt{\omega})$ for all $a \in \mathcal{A}$. By Theorem 3.5 and rate condition (3.6) we obtain $\Delta \|a\|_1 / \sqrt{a^T \Omega\, a} = o_p(1)$. We have that $I \xrightarrow{d}$

$\mathcal{N}(0,1)$ and moreover, $II = o_p(1)$ which can be seen as follows. We observe

$$\begin{aligned} II \le \sqrt{\omega/(a^T\Omega\, a)}\, \|a\|_1 \Big(& \|\widehat{\Theta}^M - \Theta^M\|_{op,\infty} \|M^T \Theta \mathbf{Z}^T \mathbf{U}\|_\infty / \sqrt{\omega n} \\ & + \|\widehat{M} - M\|_1 \|\Theta^M\|_1 \|\widehat{\Theta}\|_{op,\infty} \|\mathbf{Z}^T \mathbf{U}\|_\infty / \sqrt{\omega n} \\ & + \|\widehat{\Theta} - \Theta\|_{op,\infty} \|M^T \Theta^M\|_1 \|\mathbf{Z}^T \mathbf{U}\|_\infty / \sqrt{\omega n} \Big). \end{aligned}$$

Using Nemirovski's inequality as in proof of Theorem 3.5 we have $\|M^T \Theta \mathbf{Z}^T \mathbf{U}\|_\infty / \sqrt{n} = O_p(\sqrt{\log(q)})$ and $\|\mathbf{Z}^T \mathbf{U}\|_\infty / \sqrt{n} = O_p(\sqrt{\log(q)})$. Further, from $\sqrt{\omega/(a^T\Omega\, a)} \le \underline{\sigma}^{-1} \|a\|_2^{-1}$ we infer

$$II = O_p\Big(\frac{\log(q)}{\sqrt{n}} \frac{\|a\|_1}{\|a\|_2} \big(\omega^{3/2} s_{\max}^M \sqrt{\log(q)} + \max\big(s_M \sqrt{s_{\max}}, s_{\max} \sqrt{s_M} \big) \|\Theta^M\|_1 / \sqrt{\omega} \big) \Big)$$

using $\|\Theta M\|_1 \le \sqrt{s_M s_{\max}}$. The rate requirement imposed on q implies the result.

Second, we establish consistency of covariance matrix estimation. For the covariance matrix estimator $\widehat{\Omega}$ we conclude

$$\begin{aligned} \Big| \frac{a^T \widehat{\Omega}\, a}{a^T \Omega\, a} - 1 \Big| & \le (a^T \Omega\, a)^{-1} \|a\|_1^2 \|\widehat{\Omega} - \Omega\|_\infty \\ & \le \underbrace{\big\| \widehat{\Theta}^M \widehat{M}^T \widehat{\Theta} \big\|_1^2 \|n^{-1} \mathbf{Z}^T \mathrm{diag}(\widehat{\mathbf{U}})^2 \mathbf{Z} - \mathbb{E}[U^2 Z Z^T]\|_\infty}_{=A_1} \\ & + \underbrace{\big\| \widehat{\Theta}^M \widehat{M}^T \widehat{\Theta} - \Theta^M M^T \Theta \big\|_1 \|\Theta^M M^T \Theta\|_1 \|\mathbb{E}[U^2 Z Z^T]\|_\infty}_{=A_2} \\ & + \underbrace{\big\| \widehat{\Theta}^M \widehat{M}^T \widehat{\Theta} - \Theta^M M^T \Theta \big\|_1^2 \|\mathbb{E}[U^2 Z Z^T]\|_\infty}_{=A_3}. \end{aligned}$$

Using again Nemirovski's inequality and $\mathbb{E}[U^2|Z] \le \sigma^2$ we obtain

$$\begin{aligned} & \|n^{-1} \mathbf{Z}^T \mathrm{diag}(\widehat{\mathbf{U}})^2 \mathbf{Z} - \mathbb{E}[U^2 Z Z^T]\|_\infty \\ & \quad = \Big\| n^{-1} \sum_i \big(U_i + X_i^T (\beta^0 - \widetilde{\beta}) \big)^2 Z_i Z_i^T - \mathbb{E}[U^2 Z Z^T] \Big\|_\infty \\ & \quad \le \Big\| n^{-1} \sum_i U_i Z_i Z_i^T - \mathbb{E}[U^2 Z Z^T] \Big\|_\infty + 2 \Big\| (\beta^0 - \widetilde{\beta})^T n^{-1} \sum_i U_i X_i Z_i Z_i^T \Big\|_\infty \\ & \quad + \Big\| n^{-1} \sum_i \big(X_i^T (\beta^0 - \widetilde{\beta}) \big)^2 Z_i Z_i^T \Big\|_\infty \\ & \quad \le O_p\Big(\sqrt{\log(q)/n} \Big) + \|\beta_0 - \widetilde{\beta}\|_1 \times O_p\big(\mathbb{E}\|X\|_\infty^2 \mathbb{E} \max_{1 \le j,l \le q} |Z_j Z_l|^2 \big) \\ & \qquad + \|\beta_0 - \widetilde{\beta}\|_1^2 \times O_p\big(\mathbb{E} \max_{1 \le j,l \le p} |X_j X_l|^2 \mathbb{E} \max_{1 \le j,l \le q} |Z_j Z_l|^2 \big). \end{aligned}$$

Now using $\|\widetilde{\beta} - \beta^0\|_1 = O_p\big(s_0 \sqrt{\log(p)/n} \big)$ we obtain the $A_1 = o_p(1)$. Finally, by using a similar decomposition as for the bound of II, it is easy to see that $A_2 = o_p(1)$ which implies $A_3 = o_p(1)$. □

A.2. Proofs of Bounds on Random Matrices

Proof of Lemma 3.3. The proof of (3.4) is given in van de Geer et al. [2014]. For completeness we provide the following arguments. The KKT condition for $\widehat{\xi}_j$ implies

$\widehat{\tau}_j^2 = \mathbf{Z}_j^T(\mathbf{Z}_j - \mathbf{Z}_{-j}\widehat{\xi}_j)/n$. Consequently, it holds $\mathbf{Z}_j^T\mathbf{Z}\widehat{\Theta}_j/n = \mathbf{Z}_j^T(\mathbf{Z}_j - \mathbf{Z}_{-j}\widehat{\xi}_j)/(n\widehat{\tau}_j^2) = 1$. The KKT conditions also imply $\|\mathbf{Z}_{-j}^T\mathbf{Z}\widehat{\Theta}_j\|_\infty/n \leq \lambda_j^\Theta/\widehat{\tau}_j^2$ or

$$\big\|\widehat{\Sigma}\widehat{\Theta}_j - e_j\big\|_\infty \leq \lambda_j^\Theta/\widehat{\tau}_j^2,$$

where e_j is the j–th unit column vector.

Proof of (3.5). The KKT conditions for the nodewise Lasso (2.11) implies

$$\begin{aligned}
\widetilde{\tau}_j^2 &= \Big((\widehat{\Theta}^{1/2}\widehat{M})_j - (\widehat{\Theta}^{1/2}\widehat{M})_{-j}\widetilde{\gamma}_j\Big)^T\Big((\widehat{\Theta}^{1/2}\widehat{M})_j - (\widehat{\Theta}^{1/2}\widehat{M})_{-j}\widetilde{\gamma}_j\Big) + \lambda_j^M\|\widetilde{\gamma}_j\|_1 \\
&= \Big((\widehat{\Theta}^{1/2}\widehat{M})_j - (\widehat{\Theta}^{1/2}\widehat{M})_{-j}\widetilde{\gamma}_j\Big)^T(\widehat{\Theta}^{1/2}\widehat{M})_j + \lambda_j^M\underbrace{\big(\|\widetilde{\gamma}_j\|_1 - \widetilde{\gamma}_j^T\mathrm{sign}(\widetilde{\gamma}_j)\big)}_{=0} \\
&= \big(\widehat{\Theta}^{1/2}\widehat{M}\,\widetilde{\Gamma}_j\big)^T(\widehat{\Theta}^{1/2}\widehat{M})_j. \qquad \text{(A.8)}
\end{aligned}$$

Consequently, for all $1 \leq j \leq p$:

$$(\widehat{\Theta}^{1/2}\widehat{M})_j^T\,\widehat{\Theta}^{1/2}\widehat{M}\,\widehat{\Theta}_j^M = 1.$$

By the definition of $\widehat{\Theta}_j^M$ we also obtain

$$\begin{aligned}
\big\|(\widehat{\Theta}^{1/2}\widehat{M})_{-j}^T\widehat{\Theta}^{1/2}\widehat{M}\,\widehat{\Theta}_j^M\big\|_\infty &= \big\|(\widehat{\Theta}^{1/2}\widehat{M})_{-j}^T\big((\widehat{\Theta}^{1/2}\widehat{M})_j - (\widehat{\Theta}^{1/2}\widehat{M})_{-j}\widehat{\xi}_j\big)\big\|_\infty/\widetilde{\tau}_j^2 \\
&\leq \lambda_j^M/\widetilde{\tau}_j^2,
\end{aligned}$$

where the last inequality again follows by the KKT conditions for the nodewise Lasso (2.11). □

Proof of Proposition 3.4. The proof of the first result of the proposition is given in van de Geer et al. [2014], and hence the proof is omitted. We now prove the second result. The proof relies on the relation

$$\begin{aligned}
\|\widehat{\Theta}^M - \Theta^M\|_{op,\infty} &= \max_j\|\widehat{\Theta}_j^M - \Theta_j^M\|_1 \\
&= \max_j\|\widetilde{\Gamma}_j/\widetilde{\tau}_j^2 - \Gamma_j/\tau_j^2\|_1 \\
&\leq \max_j\|\widetilde{\gamma}_j - \gamma_j\|_1/\widetilde{\tau}_j^2 + \max_j\|\gamma_j\|_1\max_j\big|1/\widetilde{\tau}_j^2 - 1/\tau_j^2\big| \\
&\leq C\Big(\max_j\|\widetilde{\gamma}_j - \gamma_j\|_1\omega + \omega^2\sqrt{s_{\max}^M}\max_j\big|\tau_j^2 - \widetilde{\tau}_j^2\big|\Big)\max_j\frac{1}{\omega\widetilde{\tau}_j^2},
\end{aligned}$$

for all n sufficiently large. Here, we made use of the lower bound (A.1) and $\|\gamma_j\|_1 \leq C\sqrt{s_j^M}$ for n sufficiently large. We introduce the sets

$$\mathcal{C}_j = \Big\{\|\gamma_{S_j}\|_1^2 \leq Cs_j^M\gamma^T\widehat{M}^T\widehat{\Theta}\widehat{M}\gamma \text{ for all } \|\gamma_{S_j^c}\|_1 \leq 3\|\gamma_{S_j}\|_1\Big\}$$

and

$$\mathcal{T}_j = \Big\{\|\big((\widehat{\Theta}^{1/2}\widehat{M})_j - (\widehat{\Theta}^{1/2}\widehat{M})_{-j}\gamma_j\big)^T(\widehat{\Theta}^{1/2}\widehat{M})_{-j}\|_\infty \leq C\lambda_j^M\Big\}$$

for some sufficiently large constant $C > 0$. Recall $\lambda_j^M \sim \log(q)/\sqrt{n}$. On the set $\mathcal{C}_j \cap \mathcal{T}_j$, it holds

$$\|\widetilde{\gamma}_j - \gamma_j\|_1 \leq C(j)s_j^M\log(q)/\sqrt{n},$$

for some constant $C(j) > 0$, which follows directly from Theorem 2.2 of van de Geer [2016]. Thus, for the proof of the assertion it is sufficient to show

$$\big|\widetilde{\tau}_j^2 - \tau_j^2\big| = O_p\Big(\log(q)\sqrt{s_j^M/n}\Big)$$

which can be seen as follows. Recall from (A.8) that

$$\begin{aligned}
\widetilde{\tau}_j^2 &= \Big((\widehat{\Theta}^{1/2}\widehat{M})_j - (\widehat{\Theta}^{1/2}\widehat{M})_{-j}\widetilde{\gamma}_j\Big)^T(\widehat{\Theta}^{1/2}\widehat{M})_j \\
&= \Big((\widehat{\Theta}^{1/2}\widehat{M})_j - (\widehat{\Theta}^{1/2}\widehat{M})_{-j}\gamma_j\Big)^T(\widehat{\Theta}^{1/2}\widehat{M})_j + \Big((\widehat{\Theta}^{1/2}\widehat{M})_{-j}(\gamma_j - \widetilde{\gamma}_j)\Big)^T(\widehat{\Theta}^{1/2}\widehat{M})_j \\
&= \Big\|(\widehat{\Theta}^{1/2}\widehat{M})_j - (\widehat{\Theta}^{1/2}\widehat{M})_{-j}\gamma_j\Big\|_2^2 + \Big((\widehat{\Theta}^{1/2}\widehat{M})_j - (\widehat{\Theta}^{1/2}\widehat{M})_{-j}\gamma_j\Big)^T(\widehat{\Theta}^{1/2}\widehat{M})_{-j}\gamma_j \\
&\quad + \Big((\widehat{\Theta}^{1/2}\widehat{M})_{-j}(\gamma_j - \widetilde{\gamma}_j)\Big)^T(\widehat{\Theta}^{1/2}\widehat{M})_j \\
&= \Gamma_j^T\,\widehat{M}^T\widehat{\Theta}\widehat{M}\,\Gamma_j + \Big((\widehat{\Theta}^{1/2}\widehat{M})_j - (\widehat{\Theta}^{1/2}\widehat{M})_{-j}\gamma_j\Big)^T(\widehat{\Theta}^{1/2}\widehat{M})_{-j}\gamma_j \\
&\quad + \Big((\widehat{\Theta}^{1/2}\widehat{M})_{-j}(\gamma_j - \widetilde{\gamma}_j)\Big)^T(\widehat{\Theta}^{1/2}\widehat{M})_j
\end{aligned}$$

and recall that $\tau_j^2 = \Gamma_j^T M^T\Theta M\Gamma_j$. We have

$$\begin{aligned}
\big|\widetilde{\tau}_j^2 - \tau_j^2\big| \le{}& \underbrace{\big|\Gamma_j^T\big(\widehat{M}^T\widehat{\Theta}\widehat{M} - M^T\Theta M\big)\Gamma_j\big|}_{I} + \underbrace{\Big|\Big((\widehat{\Theta}^{1/2}\widehat{M})_j - (\widehat{\Theta}^{1/2}\widehat{M})_{-j}\gamma_j\Big)^T(\widehat{\Theta}^{1/2}\widehat{M})_{-j}\gamma_j\Big|}_{II} \\
&+ \underbrace{\big|(\gamma_j - \widetilde{\gamma}_j)^T(\widehat{\Theta}^{1/2}\widehat{M})_{-j}^T(\widehat{\Theta}^{1/2}\widehat{M})_j\big|}_{III}
\end{aligned}$$

where we bound each term on the right hand side as follows. Consider I. We observe

$$I \le \underbrace{\big|\Gamma_j^T(\widehat{M} - M)^T\widehat{\Theta}\widehat{M}\,\Gamma_j\big|}_{T_1} + \underbrace{\big|\Gamma_j^T M^T(\widehat{\Theta} - \Theta)\widehat{M}\,\Gamma_j\big|}_{T_2} + \underbrace{\big|\Gamma_j^T M^T\Theta(\widehat{M} - M)\,\Gamma_j\big|}_{T_3}.$$

In the following, we bound each summand on the right hand side separately. We have

$$T_1 = |(\Theta M\Gamma_j)^T(\widehat{M} - M)\Gamma_j| + o_p\big(\log(q)/\sqrt{n}\big) = O_p\big(\log(q)/\sqrt{n}\big),$$

uniformly in j by using Assumption 3, i.e., $\mathbb{E}\max_j|(\Theta M\Gamma_j)^T ZX^T\Gamma_j|^2 = O(\log(p))$, and following the arguments for the upper bound (A.5). Equivalently, we have $T_3 = O_p\big(\log(q)/\sqrt{n}\big)$. We observe

$$\begin{aligned}
T_2 &\le |\Gamma_j^T M^T\Theta(\widehat{\Sigma}\widehat{\Theta} - I_q)\widehat{M}\,\Gamma_j| + |\Gamma_j^T M^T\Theta(\widehat{\Sigma} - \Sigma)\widehat{\Theta}\widehat{M}\,\Gamma_j| \\
&\le |\Gamma_j^T M^T\Theta(\widehat{\Sigma}\widehat{\Theta} - I_q)M\,\Gamma_j| + |\Gamma_j^T M^T\Theta(\widehat{\Sigma} - \Sigma)\Theta M\,\Gamma_j| + o_p\big(\sqrt{\log(q)/n}\big)
\end{aligned}$$

Due to Assumption 3 (ii), i.e., $\mathbb{E}\max_{1\le j\le p}\|(\Theta M\Gamma_j)^T Z\|_2^4 = O(\log(p)^2)$, it is sufficient to consider the first summand. The KKT condition for the nodewise Lasso estimator $\widehat{\xi}_j$ implies $\mathbf{Z}_{-j}^T\mathbf{Z}\widehat{\Theta}_j/n = \widehat{\tau}_j^{-2}\lambda_j^\Theta\widehat{\kappa}$ and it holds $\mathbf{Z}_j^T\mathbf{Z}\widehat{\Theta}_j/n = e_j$ (see van de Geer et al. [2014]). Consequently, we have

$$\widehat{\Sigma}\widehat{\Theta}_j - e_j = \lambda_j^\Theta\widehat{\kappa}_j/\widehat{\tau}_j^2.$$

Since $\lambda_j^\Theta \sim \sqrt{\log(q)/n}$ and $\|\Theta M \Gamma_j\|_1 \leq \|\Theta\|_1 \|M\|_1 \|\Gamma_j\|_1 \leq \sqrt{s_{\max} s_M (s_j^M + (\lambda_j^M)^2)}$ we obtain

$$\begin{aligned}
|\Gamma_j^T M^T \Theta(\widehat{\Sigma}\widehat{\Theta} - I_q) M \Gamma_j| &= |\Gamma_j^T M^T \Theta(\lambda_1^\Theta \widehat{\kappa}_1/\widehat{\tau}_1^2, \ldots, \lambda_q^\Theta \widehat{\kappa}_q/\widehat{\tau}_q^2) M \Gamma_j| \\
&\leq \sqrt{\log(q)/n}\, \|\Theta M \Gamma_j\|_1 \|M \Gamma_j\|_1 \max_{1 \leq j \leq q} \widehat{\tau}_j^{-2} \\
&= \sqrt{\log(q)/n}\, \sqrt{s_{\max}} s_M (s_j^M + (\lambda_j^M)^2) \times O_p(1) \\
&= O_p\big(\log(q)/\sqrt{n} \big),
\end{aligned}$$

using that $\widehat{\tau}_j^2$ is a consistent estimator of $1/\Theta_{jj}$ (see the proof of Theorem 2.4 of van de Geer et al. [2014]), Θ_{jj} is bounded uniformly in j, and the first rate condition imposed in Assumption 2 (*iii*). Further, we have on $\mathcal{T}_j$ that

$$\begin{aligned}
II &\leq \|\gamma_j\|_1 \big\| \big((\widehat{\Theta}^{1/2}\widehat{M})_j - (\widehat{\Theta}^{1/2}\widehat{M})_{-j} \gamma_j \big)^T (\widehat{\Theta}^{1/2}\widehat{M})_{-j} \big\|_\infty \\
&= O_p\Big(\log(q) \sqrt{s_j^M/n} \Big).
\end{aligned}$$

Further, the KKT condition for the nodewise Lasso estimator $\widetilde{\gamma}_j$ implies

$$\begin{aligned}
III &= |\Gamma_j^T (\widehat{\Theta}^{1/2}\widehat{M})^T (\widehat{\Theta}^{1/2}\widehat{M})_j| \\
&= |\Gamma_j^T (\widehat{M}\widehat{\Theta}\widehat{M} - M\Theta M) e_j| \\
&= O_p\big(\log(q)/\sqrt{n} \big).
\end{aligned}$$

In the following, we show that $\mathbb{1}_{\mathcal{C}_j \cap \mathcal{T}_j}$ with probability approaching one. To control $\mathcal{C}_j$ we can proceed similarly as in the proof of Theorem 3.5. To control $\mathcal{T}_j$, recall that due to the definition of Γ_j it holds $\Gamma_j^T (\Theta^{1/2} M)^T (\Theta^{1/2} M)_{-j} = 0$. We observe

$$\begin{aligned}
&\| \big((\widehat{\Theta}^{1/2}\widehat{M})_j - (\widehat{\Theta}^{1/2}\widehat{M})_{-j} \gamma_j \big)^T (\widehat{\Theta}^{1/2}\widehat{M})_{-j} \|_\infty = \|\Gamma_j^T \big(\widehat{M}^T \widehat{\Theta} \widehat{M} - M^T \Theta M \big) I_{-j} \|_\infty \\
&\leq \underbrace{\|\Gamma_j^T (\widehat{M} - M)^T \widehat{\Theta} \widehat{M}\|_\infty}_{S_1} + \underbrace{\|\Gamma_j^T M^T (\widehat{\Theta} - \Theta) \widehat{M}\|_\infty}_{S_2} + \underbrace{\|\Gamma_j^T M^T \Theta (\widehat{M} - M)\|_\infty}_{S_3}.
\end{aligned}$$

In the following, we bound each summand on the right hand side separately. We have

$$S_1 = \max_l |(\Theta M)_l^T (\widehat{M} - M) \Gamma_j| + o_p\big(\sqrt{\log(q)/n} \big) = O_p\big(\log(q)/\sqrt{n} \big)$$

by using Assumption 3 (*i*), i.e., $\mathbb{E}\|M^T \Theta Z X^T \Gamma_j\|_\infty^2 = O(\log(p))$ and following the arguments for the upper bound (A.5). We observe

$$\begin{aligned}
S_2 &\leq \|\Gamma_j^T M^T \Theta(\widehat{\Sigma}\widehat{\Theta} - I_q) M\|_\infty + \|\Gamma_j^T M^T \Theta(\widehat{\Sigma} - \Sigma)\widehat{\Theta} M\|_\infty \\
&\leq \|\Gamma_j^T M^T \Theta(\widehat{\Sigma}\widehat{\Theta} - I_q) M\|_\infty + \|\Gamma_j^T M^T \Theta(\widehat{\Sigma} - \Sigma)\Theta M\|_\infty + o_p\big(\sqrt{\log(q)/n} \big)
\end{aligned}$$

where the second summand can be bounded again by using Nemirovski's inequality:

$$\begin{aligned}
\mathbb{E}\|\Gamma_j^T M^T \Theta(\widehat{\Sigma} - \Sigma)\Theta M\|_\infty^2 &= \mathbb{E} \max_{1 \leq l \leq p} \big| n^{-1} \sum_i \Gamma_j^T M^T \Theta Z_i Z_i^T (\Theta M)_l - \Gamma_j^T M^T (\Theta M)_l \big|^2 \\
&\leq 8 \log(2p) n^{-1} \mathbb{E} \max_{1 \leq l \leq p} |\Gamma_j^T M^T \Theta Z Z^T (\Theta M)_l|^2.
\end{aligned}$$

The KKT condition for the nodewise Lasso estimator $\widehat{\xi}_j$ implies $\mathbf{Z}_{-j}^T\mathbf{Z}\widehat{\Theta}_j/n = \widehat{\tau}_j^{-2}\lambda_j^\Theta\widehat{\kappa}_j$ and it holds $\mathbf{Z}_j^T\mathbf{Z}\widehat{\Theta}_j/n = e_j$. Consequently, we have

$$\widehat{\Sigma}\widehat{\Theta}_j - e_j = \lambda_j^\Theta\widehat{\kappa}_j/\widehat{\tau}_j^2.$$

Since $\lambda_j^\Theta \sim \sqrt{\log(q)/n}$ we obtain by employing Theorem 2.4 of van de Geer et al. [2014])

$$\begin{aligned}\|\Gamma_j^T M^T\Theta(\widehat{\Sigma}\widehat{\Theta} - I_q)M\|_\infty &= \|\Gamma_j^T M^T\Theta(\lambda_1^\Theta\widehat{\kappa}_1/\widehat{\tau}_1^2,\ldots,\lambda_q^\Theta\widehat{\kappa}_q/\widehat{\tau}_q^2)M\|_\infty \\ &\leq \sqrt{\log(q)/n}\,\|\Theta M\Gamma_j\|_1\|M\|_1\max_{1\leq j\leq q}\widehat{\tau}_j^{-2} \\ &= \sqrt{\log(q)/n}\,\sqrt{s_{\max}}s_M\sqrt{s_j^M + (\lambda_j^M)^2}\times O_p(1) \\ &= O_p(\log(q)/\sqrt{n}),\end{aligned}$$

by using Assumption 2 (*iii*), i.e., $s_M\sqrt{s_{\max}s_{\max}^M} = O\big(\sqrt{\log(q)}\big)$. Finally, we have

$$S_3 = \|(\Theta M\Gamma_j)^T(\widehat{M} - M)\|_\infty = O_p\big(\log(q)/\sqrt{n}\big),$$

by following again the arguments for the upper bound (A.5), which completes the proof of the result. □

For a random variable W, we introduce the sub-Gaussian norm $\|\cdot\|_{\psi_2}$ as $\|W\|_{\psi_2} := \sup_{q\geq 1} q^{-1/2}(\mathbb{E}|W|^q)^{1/q}$ and the sub-exponential norm $\|\cdot\|_{\psi_1}$ as $\|W\|_{\psi_1} := \sup_{q\geq 1} q^{-1}(\mathbb{E}|W|^q)^{1/q}$, see Vershynin [2012, Definition 5.7 and Lemma 5.5]. If W is sub-Gaussian (see Definition 1) then $\|W\|_{\psi_2}$ is bounded from above. Also note that if W has bounded sub-Gaussian norm then W^2 has bounded sub-exponential norm, see Vershynin [2012, Remark 5.18].

Proof of Proposition 3.2. We start by proving the first part of the theorem. Denote $\varsigma_{zj}^2 := \mathbb{E}[\mathbf{Z}_{1j}^2]$, $\varsigma_{xk}^2 := \mathbb{E}[\mathbf{X}_{1k}^2]$ and $\rho_{jk} := \mathbb{E}[\mathbf{Z}_{1j}\mathbf{X}_{1k}]/(\varsigma_{zj}\varsigma_{xk})$. Let $K_z := \|\mathbf{Z}_{ij}\|_{\psi_2}$ and $K_x := \|\mathbf{X}_{ik}\|_{\psi_2}$, which do not depend on i. Then,

$$\begin{aligned}\mathbb{P}\left(\left|\widetilde{M}_{jk} - M_{jk}\right| \geq v\right) &= \mathbb{P}\left(\left|\sum_{i=1}^n(\mathbf{Z}_{ij}\mathbf{X}_{ik} - M_{jk})\right| \geq nv\right) \\ &= \mathbb{P}\left(\left|\sum_{i=1}^n\left(\frac{\mathbf{Z}_{ij}\mathbf{X}_{ik}}{\varsigma_{zj}\varsigma_{xk}} - \rho_{jk}\right)\right| \geq \frac{nv}{\varsigma_{zj}\varsigma_{xk}}\right).\end{aligned}$$

Moreover,

$$\begin{aligned}\sum_{i=1}^n\left(\frac{\mathbf{Z}_{ij}\mathbf{X}_{ik}}{\varsigma_{zj}\varsigma_{xk}} - \rho_{jk}\right) = \frac{1}{4}\Big[\sum_{i=1}^n[\left(\frac{\mathbf{Z}_{ij}}{\varsigma_{zj}} + \frac{\mathbf{X}_{ik}}{\varsigma_{xk}}\right)^2 - 2(1+\rho_{jk})] \\ - \sum_{i=1}^n[\left(\frac{\mathbf{Z}_{ij}}{\varsigma_{zj}} - \frac{\mathbf{X}_{ik}}{\varsigma_{xk}}\right)^2 - 2(1-\rho_{jk})]\Big].\end{aligned}$$

Because $\mathbf{X}$ and $\mathbf{Z}$ have sub-Gaussian rows then $\mathbf{X}_{ik}$, $\mathbf{Z}_{ij}$, $\left(\frac{\mathbf{Z}_{ij}}{\varsigma_{zj}} + \frac{\mathbf{X}_{ik}}{\varsigma_{xk}}\right)$ and $\left(\frac{\mathbf{Z}_{ij}}{\varsigma_{zj}} - \frac{\mathbf{X}_{ik}}{\varsigma_{xk}}\right)$ are sub-Gaussian (because linear combinations of sub-Gaussian random variables are still sub-Gaussian). The sub-gaussian norms of $\left(\frac{\mathbf{Z}_{ij}}{\varsigma_{zj}} + \frac{\mathbf{X}_{ik}}{\varsigma_{xk}}\right)$ and $\left(\frac{\mathbf{Z}_{ij}}{\varsigma_{zj}} - \frac{\mathbf{X}_{ik}}{\varsigma_{xk}}\right)$ are upper bounded by $\frac{K_z}{\varsigma_{zj}} + \frac{K_x}{\varsigma_{xk}}$.

Therefore, $\left(\frac{\mathbf{Z}_{ij}}{\zeta_{zj}}+\frac{\mathbf{X}_{ik}}{\zeta_{xk}}\right)^2$ and $\left(\frac{\mathbf{Z}_{ij}}{\zeta_{zj}}-\frac{\mathbf{X}_{ik}}{\zeta_{xk}}\right)^2$ are sub-exponential, see *e.g.* Vershynin [2012, Lemma 5.14], whose means are, respectively

$$2(1+\rho_{jk}) \qquad \text{and} \qquad 2(1-\rho_{jk}).$$

Denote $W_{i+} := \left(\frac{\mathbf{Z}_{ij}}{\zeta_{zj}}+\frac{\mathbf{X}_{ik}}{\zeta_{xk}}\right)^2 \frac{1}{2(1+\rho_{jk})} - 1$ and $W_{i-} := \left(\frac{\mathbf{Z}_{ij}}{\zeta_{zj}}-\frac{\mathbf{X}_{ik}}{\zeta_{xk}}\right)^2 \frac{1}{2(1-\rho_{jk})} - 1$ which are also sub-exponential by Vershynin [2012, Remark 5.18] with mean zero. In fact, by using the moment condition characterization of sub-Gaussianity we obtain, for some constant $K>0$ and all $p \geq 1$:

$$\begin{aligned}(\mathbb{E}|W_{i+}|^p)^{1/p} &\leq \left\| \left(\frac{\mathbf{Z}_{ij}}{\zeta_{zj}}+\frac{\mathbf{X}_{ik}}{\zeta_{xk}}\right)^2 \frac{1}{2(1+\rho_{jk})} \right\|_p + \|1\|_p \\ &\leq 2 \left\| \left(\frac{\mathbf{Z}_{ij}}{\zeta_{zj}}+\frac{\mathbf{X}_{ik}}{\zeta_{xk}}\right)^2 \frac{1}{2(1+\rho_{jk})} \right\|_p \leq 2Kp\end{aligned}$$

where we have use the triangle inequality to get the first inequality, the Jensen inequality to get the second inequality and sub-exponentiality of $\left(\frac{\mathbf{Z}_{ij}}{\zeta_{zj}}+\frac{\mathbf{X}_{ik}}{\zeta_{xk}}\right)^2 \frac{1}{2(1+\rho_{jk})}$ to get the last inequality.
The sub-exponential norm of W_{i+} can be upper bounded as follows:

$$\begin{aligned}\|W_{i+}\|_{\psi_1} &\leq \sup_{q\geq 1} q^{-1}\|W_{i+}\|_q \leq \sup_{q\geq 1} q^{-1} \left(\left\| \left(\frac{\mathbf{Z}_{ij}}{\zeta_{zj}}+\frac{\mathbf{X}_{ik}}{\zeta_{xk}}\right)^2 \frac{1}{2(1+\rho_{jk})} \right\|_q + \|1\|_q \right) \\ &\leq \left\| \left(\frac{\mathbf{Z}_{ij}}{\zeta_{zj}}+\frac{\mathbf{X}_{ik}}{\zeta_{xk}}\right)^2 \frac{1}{2(1+\rho_{jk})} \right\|_{\psi_1} + \sup_{q\geq 1} q^{-1} \left\| \left(\frac{\mathbf{Z}_{ij}}{\zeta_{zj}}+\frac{\mathbf{X}_{ik}}{\zeta_{xk}}\right)^2 \frac{1}{2(1+\rho_{jk})} \right\|_q \\ &= 2 \left\| \left(\frac{\mathbf{Z}_{ij}}{\zeta_{zj}}+\frac{\mathbf{X}_{ik}}{\zeta_{xk}}\right)^2 \frac{1}{2(1+\rho_{jk})} \right\|_{\psi_1} \leq 4 \left\| \left(\frac{\mathbf{Z}_{ij}}{\zeta_{zj}}+\frac{\mathbf{X}_{ik}}{\zeta_{xk}}\right) \frac{1}{\sqrt{2(1+\rho_{jk})}} \right\|_{\psi_2}^2 \\ &\leq 4 \left(\frac{K_z}{\zeta_{zj}}+\frac{K_x}{\zeta_{xk}}\right)^2 \frac{1}{2(1+\rho_{jk})} \qquad \text{(A.9)}\end{aligned}$$

where we have first used the triangle inequality, then the Jensen's inequality and, to get the third inequality we have used Vershynin [2012, Lemma 5.14]. In a similar way, we can show that the sub-exponential norm of W_{i-} is upper bounded by

$$\|W_{i-}\|_{\psi_1} \leq 4 \left(\frac{K_z}{\zeta_{zj}}+\frac{K_x}{\zeta_{xk}}\right)^2 \frac{1}{2(1-\rho_{jk})} \qquad \text{(A.10)}$$

and the right hand side does not depend on i. Therefore, for every i, $\|(1+\rho_{jk})W_{i+}\|_{\psi_1} \leq 2\left(\frac{K_z}{\zeta_{zj}}+\frac{K_x}{\zeta_{xk}}\right)^2$ and $\|(1-\rho_{jk})W_{i-}\|_{\psi_1} \leq 2\left(\frac{K_z}{\zeta_{zj}}+\frac{K_x}{\zeta_{xk}}\right)^2$. Let $K := \max_i \|(1-\rho_{jk})W_{i-}\|_{\psi_1}$. For every $t \geq 0$, define the event $\mathcal{A} := \{|\sum_{i=1}^n (1-\rho_{jk})W_{i-}| \geq t\}$ which by using

Vershynin [2012, Proposition 5.16] has probability upper bounded by

$$\mathbb{P}(\mathcal{A}) \leq 2\exp\left\{-c\min\left\{\frac{t^2}{K^2 n}, \frac{t}{K}\right\}\right\} \leq 2\exp\left\{-c\min\left\{\frac{t^2}{4n\left(\frac{K_z}{\zeta_{zj}}+\frac{K_x}{\zeta_{xk}}\right)^4}, \frac{t}{2\left(\frac{K_z}{\zeta_{zj}}+\frac{K_x}{\zeta_{xk}}\right)^2}\right\}\right\} \quad \text{(A.11)}$$

where $c > 0$ is an absolute constant. The probability that we want to upper bound is the following:

$$\begin{aligned}
&\mathbb{P}\left(\left|\widetilde{M}_{jk} - M_{jk}\right| \geq v\right) = \mathbb{P}\left(\frac{1}{2}\left|\sum_{i=1}^n W_{i+}(1+\rho_{jk}) - \sum_{i=1}^n W_{i-}(1-\rho_{jk})\right| \geq \frac{nv}{\zeta_{zj}\zeta_{xk}}\right) \\
&\leq \mathbb{P}\left(\left|\sum_{i=1}^n W_{i+}(1+\rho_{jk})\right| \geq 2\frac{nv}{\zeta_{zj}\zeta_{xk}} - \left|\sum_{i=1}^n W_{i-}(1-\rho_{jk})\right| \cap \mathcal{A}^c\right) + \mathbb{P}(\mathcal{A}) \\
&\leq \mathbb{P}\left(\left|\sum_{i=1}^n W_{i+}(1+\rho_{jk})\right| \geq \frac{nv}{\zeta_{zj}\zeta_{xk}} \cap \mathcal{A}^c\right) + \mathbb{P}(\mathcal{A}) \\
&\leq \mathbb{P}\left(\left|\sum_{i=1}^n W_{i+}(1+\rho_{jk})\right| \geq \frac{nv}{\zeta_{zj}\zeta_{xk}}\right) + \mathbb{P}(\mathcal{A}).
\end{aligned}$$

Therefore, by using (A.11) with $t = nv/(\zeta_{zj}\zeta_{xk})$ and $0 \leq v \leq 1$ in $\mathcal{A}$, and applying again Vershynin [2012, Proposition 5.16] to upper bound the first probability in the last line of the previous display, we obtain

$$\begin{aligned}
&\mathbb{P}\left(\left|\widetilde{M}_{jk} - M_{jk}\right| \geq v\right) \\
&\leq 2\exp\left\{-c\min\left\{\frac{v^2}{4\zeta_{zj}^2\zeta_{xk}^2\left(\frac{K_z}{\zeta_{zj}}+\frac{K_x}{\zeta_{xk}}\right)^4}, \frac{v}{2\zeta_{zj}\zeta_{xk}\left(\frac{K_z}{\zeta_{zj}}+\frac{K_x}{\zeta_{xk}}\right)^2}\right\} n\right\} \\
&\qquad + 2\exp\left\{-c\min\left\{\frac{v^2}{4\zeta_{zj}^2\zeta_{xk}^2\left(\frac{K_z}{\zeta_{zj}}+\frac{K_x}{\zeta_{xk}}\right)^4}, \frac{v}{2\zeta_{zj}\zeta_{xk}\left(\frac{K_z}{\zeta_{zj}}+\frac{K_x}{\zeta_{xk}}\right)^2}\right\} n\right\} \\
&= 4\exp\left\{-c\min\left\{\frac{v^2}{4\zeta_{zj}^2\zeta_{xk}^2\left(\frac{K_z}{\zeta_{zj}}+\frac{K_x}{\zeta_{xk}}\right)^4}, \frac{v}{2\zeta_{zj}\zeta_{xk}\left(\frac{K_z}{\zeta_{zj}}+\frac{K_x}{\zeta_{xk}}\right)^2}\right\} n\right\} \\
&\leq 4\exp\{-Cv^2 n\}
\end{aligned}$$

where for the last inequality we have used thatmin$(a/b, c/d) \geq \min(a,c)/\max(b,d)$ for any constants a, b, c, d. This proves (3.3). To prove the second part of the theorem notice that, by definition of s_M and under Assumption 2 *(iii)*, M belongs to the class of matrices

$$\mathcal{G}_\chi(\rho, s_M) = \left\{M \in \mathbb{R}^{q\times p} : \max_{1\leq k\leq p} |M_{[j]k}|^\chi \leq s_M/j \text{ for all } j \text{ and } \max_{1\leq j\leq (p\wedge q)} M_{jj} \leq \rho\right\}$$

(A.12)

with $\chi = 0$ and $|M_{[j]k}|$ denoting the j-th largest element in magnitude of the k-th column $(M_{jk})_{1\leq j\leq q}$ of M. This is the extension to rectangular matrices of the class of matrices considered in Cai and Zhou [2012] for $0 \leq \chi < 1$. Hence, the second part of the theorem follows from the proof of Cai and Zhou [2012, Theorem 4] and (3.3). We give some elements of this proof in Appendix B. □

B. Appendix: Technical Results

Recall the notation $\widetilde{M} = \mathbf{Z}^T\mathbf{X}/n$ and the thresholding estimator: $\widehat{M} = (\widehat{M}_{jk})$ with

$$\widehat{M}_{jk} := \widetilde{M}_{jk}\, \mathbb{1}\left\{|\widetilde{M}_{jk}| \geq C_0\sqrt{\frac{\log(q)}{n}}\right\}, \qquad C_0 > 0. \tag{B.1}$$

In the following theorem, we provide the rate for its ℓ_1-norm. The minimax rate for the ℓ_1-norm of the thresholding estimator of quadratic matrix is studied in Cai and Zhou [2012]. Here, we slightly extend their proof to account for the rectangular case and only report the main steps that contain the differences with respect to Cai and Zhou [2012]. We will establish this result for the more general class of matrices $\mathcal{G}_\chi(\rho, s_M)$ defined in (A.12) for $0 \leq \chi < 1$ where $|M_{[j]k}|$ denotes the j-th largest element in magnitude of the k-th column $(M_{jk})_{1\leq j\leq q}$. Every matrix in $\mathcal{G}_\chi(\rho, s_M)$ has columns $(M_{jk})_{1\leq j\leq q}$ that are in a (approximate) sparse weak ℓ_χ ball. The case $\chi = 0$ is the case considered in the paper. Moreover, define the class of distributions $\mathcal{P}(\mathcal{G}_\chi(\rho, s_M))$ as the set of distributions of (Z, X) satisfying (A.12) and such that the rows of $\mathbf{Z}$ and $\mathbf{X}$ are sub-Gaussian.

Theorem B.1. *Let Assumption 1 (ii) hold. Then, the thresholding estimator $\widehat{M}$ satisfies*

$$\sup_{\mathcal{P}(\mathcal{G}_\chi(\rho,s_M))} \mathbb{E}\|\widehat{M} - M\|_1^2 \leq C s_M^2 \left(\frac{\log(p\vee q)}{n}\right)^{1-\chi}$$

for some constant $C > 0$.

In the following we directly write q instead of $p \vee q$. Therefore, by Theorem B.1 and the Markov's inequality

$$\begin{aligned}\mathbb{P}\left(\left\|\widehat{M} - M\right\|_1 > \varepsilon\right) &\leq \frac{1}{\varepsilon^2}\mathbb{E}\left\|\widehat{M} - M\right\|_1^2 \\ &\leq C\frac{s_M^2}{\varepsilon^2}\left(\frac{\log(q)}{n}\right)^{1-\chi}\end{aligned} \tag{B.2}$$

which implies:

$$\left\|\widehat{M} - M\right\|_1 \leq s_M\left(\frac{\log(q)}{n}\right)^{(1-\chi)/2}$$

with probability approaching one.

Proof. Define the event $A_{jk} := \{|\widehat{M}_{jk} - M_{jk}| \leq 4 \min\left\{|M_{jk}|, C_0\sqrt{\frac{\log(q)}{n}}\right\}\}$ and $D = (d_{jk})$ with $d_{jk} := (\widehat{M}_{jk} - M_{jk})\,\mathbb{1}_{A_{jk}^c}$. Then,

$$\begin{aligned}
\mathbb{E}\|\widehat{M} - M\|_1^2 &= \mathbb{E}\|\widehat{M} - M - D + D\|_1^2 \\
&\leq \mathbb{E}\|\widehat{M} - M - D\|_1^2 + \mathbb{E}\|D\|_1^2 \\
&\leq 2\mathbb{E}\left(\sup_{1\leq k\leq p}\sum_{j=1}^{q}|\widehat{M}_{jk} - M_{jk}|\,\mathbb{1}_{A_{jk}}\right)^2 + 2\mathbb{E}\|D\|_1^2 \\
&\leq 32\left(\sup_{1\leq k\leq p}\sum_{j=1}^{q}\min\left\{|M_{jk}|, C_0\sqrt{\frac{\log(q)}{n}}\right\}\right)^2 + 2\mathbb{E}\|D\|_1^2 \qquad \text{(B.3)}
\end{aligned}$$

where the inequality in the penultimate line is due to $(\widehat{M} - M - D)_{jk} = (\widehat{M}_{jk} - M_{jk})(1 - \mathbb{1}_{A_{jk}^c}) = (\widehat{M}_{jk} - M_{jk})\,\mathbb{1}_{A_{jk}}$.
To control the first term we use exactly the same procedure as in Cai and Zhou [2012] and so we omit it. We find that

$$32\left(\sup_{1\leq k\leq p}\sum_{j=1}^{q}\min\left\{|M_{jk}|, C_0\sqrt{\frac{\log(q)}{n}}\right\}\right)^2 \leq C_1 s_M\left(\frac{\log(q)}{n}\right)^{(1-\chi)/2} \qquad \text{(B.4)}$$

for some positive constant C_1. We now consider the second term in (B.3) and show that it is negligible with respect to the first term. For this we use the following decomposition (also coming from Cai and Zhou [2012]), where we denote by $\|\cdot\|_F$ the Frobenius norm:

$$\begin{aligned}
\mathbb{E}\|D\|_1^2 &= \mathbb{E}\left(\max_{1\leq k\leq p}\sum_{j=1}^{q}|d_{jk}|\right)^2 \leq \mathbb{E}[q\|D\|_F^2] = q\sum_{k=1}^{p}\sum_{j=1}^{q}\mathbb{E}|d_{jk}|^2 \\
&= q\sum_{k=1}^{p}\sum_{j=1}^{q}\mathbb{E}\left(d_{jk}^2\,\mathbb{1}_{\{A_{jk}^c\cap\{\widehat{M}_{jk}=\widetilde{M}_{jk}\}\}} + d_{jk}^2\,\mathbb{1}_{\{A_{jk}^c\cap\{\widehat{M}_{jk}=0\}\}}\right) \\
&= q\sum_{k=1}^{p}\sum_{j=1}^{q}\mathbb{E}\left((\widetilde{M}_{jk} - M_{jk})^2\,\mathbb{1}_{\{A_{jk}^c\}} + M_{jk}^2\,\mathbb{1}_{\{A_{jk}^c\cap\{\widehat{M}_{jk}=0\}\}}\right) =: R_1 + R_2.
\end{aligned}$$

Let us start by term R_1. By the Holder's inequality (with norms L_3 and $L_{3/2}$) we obtain

$$\begin{aligned}
R_1 &\leq p\sum_{k=1}^{p}\sum_{j=1}^{q}\mathbb{E}^{1/3}\left[(\widetilde{M}_{jk} - M_{jk})^6\right]\mathbb{P}^{2/3}(A_{jk}^c) \\
&\leq C_3 p^2 q\frac{1}{n}\mathbb{P}^{2/3}(A_{jk}^c)
\end{aligned}$$

where we have used result (B.5) of Lemma B.2 below that $\mathbb{E}^{1/3}\left[(\widetilde{M}_{jk} - M_{jk})^6\right] = O(n^{-1})$. Finally, by using the result of Lemma B.3 below we get that $\mathbb{P}(A_{jk}^c) \leq 2C_4 q^{-9/2}$ so that

$$R_1 \leq 2\frac{C_2 C_3}{n}q^3 q^{-3} \leq C_5/n.$$

Let us now consider term R_2:

$$\begin{aligned}
R_2 &= p\sum_{k=1}^{p}\sum_{j=1}^{q}\mathbb{E}\left(M_{jk}^2\,\mathbb{1}_{\{|M_{jk}|\geq 4C_0\sqrt{\log(q)/n}\}}\,\mathbb{1}_{\{|\widetilde{M}_{jk}|\leq C_0\sqrt{\log(q)/n}\}}\right)\\
&\leq p\sum_{k=1}^{p}\sum_{j=1}^{q}M_{jk}^2\mathbb{E}\left(\mathbb{1}_{\{|M_{jk}|\geq 4C_0\sqrt{\log(q)/n}\}}\,\mathbb{1}_{\{|M_{jk}|-|\widetilde{M}_{jk}-M_{jk}|\leq C_0\sqrt{\log(q)/n}\}}\right)\\
&= \frac{p}{n}\sum_{k=1}^{p}\sum_{j=1}^{q}nM_{jk}^2\mathbb{P}\left(|\widetilde{M}_{jk}-M_{jk}|\geq -C_0\sqrt{\log(q)/n}+|M_{jk}|\right)\mathbb{1}_{\{|M_{jk}|\geq 4C_0\sqrt{\log(q)/n}\}}\\
&\leq \frac{p}{n}\sum_{k=1}^{p}\sum_{j=1}^{q}nM_{jk}^2\mathbb{P}\left(|\widetilde{M}_{jk}-M_{jk}|\geq -\frac{1}{4}|M_{jk}|+|M_{jk}|\right)\mathbb{1}_{\{|M_{jk}|\geq 4C_0\sqrt{\log(q)/n}\}}
\end{aligned}$$

where to get the inequality in the second line we have used $|\widetilde{M}_{jk}|\geq |M_{jk}|-|\widetilde{M}_{jk}-M_{jk}|$. Therefore, by using result (3.3) in Theorem 3.2 we get:

$$\begin{aligned}
R_2 &\leq \frac{p}{n}\sum_{k=1}^{p}\sum_{j=1}^{q}nM_{jk}^2\mathbb{P}\left(|\widetilde{M}_{jk}-M_{jk}|\geq \frac{3}{4}|M_{jk}|\right)\mathbb{1}_{\{|M_{jk}|\geq 4C_0\sqrt{\log(q)/n}\}}\\
&\leq \frac{pC_0^2}{n}\sum_{k=1}^{p}\sum_{j=1}^{q}\frac{n}{C_0^2}M_{jk}^2 4\exp\{-cn9|M_{jk}|^2/16\}\,\mathbb{1}_{\{|M_{jk}|\geq 4C_0\sqrt{\log(q)/n}\}}\\
&\leq \frac{pC_0^2}{n}\sum_{k=1}^{p}\sum_{j=1}^{q}\exp\{nM_{jk}^2\frac{4}{C_0^2}-cn9|M_{jk}|^2/16\}\,\mathbb{1}_{\{|M_{jk}|\geq 4C_0\sqrt{\log(q)/n}\}}
\end{aligned}$$

where to get the last inequality we have used the inequality $nt/e^{nt}\leq 1$ for all $t>0$. Let $C_0=\sqrt{8/c}$, then

$$\begin{aligned}
R_2 &\leq \frac{pC_0^2}{n}\sum_{k=1}^{p}\sum_{j=1}^{q}\exp\{-nM_{jk}^2c/16\}\,\mathbb{1}_{\{|M_{jk}|\geq 4C_0\sqrt{\log(q)/n}\}}\\
&\leq \frac{pC_0^2}{n}\sum_{k=1}^{p}\sum_{j=1}^{q}\exp\{-n16C_0^2\log(q)c/(16n)\}\\
&\leq \frac{pC_0^2}{n}\sum_{k=1}^{p}\sum_{j=1}^{q}\exp\{-8\log(q)\}=\frac{(q)^3C_0^2}{n}(q)^{-8}\\
&\leq C_6/n.
\end{aligned}$$

Lemma B.2. *Let Assumption 1 (ii) hold. Then, there exists a constant $C_2>0$ such that*

$$\mathbb{E}[|\widetilde{M}_{jk}-M_{jk}|^6]\leq\frac{24}{C_2^3n^3}. \tag{B.5}$$

Proof. The 6-th moment can be written as

$$\mathbb{E}[|\widetilde{M}_{jk}-M_{jk}|^6]=6\int_0^\infty x^5\mathbb{P}(|\widetilde{M}_{jk}-M_{jk}|\geq x)dx$$

and by substituting the upper bound in (3.3) and by using integration by parts we get

$$\mathbb{E}[|\widetilde{M}_{jk}-M_{jk}|^6] \leq 24\int_0^\infty x^5\exp\left\{-cnx^2\right\}dx=\frac{24}{c^3n^3}.$$

Lemma B.3. *Define the event A_{jk} as $A_{jk} := \left\{ |\widehat{M}_{jk} - M_{jk}| \leq 4 \min \left\{ |M_{jk}, C_0 \sqrt{\frac{\log(q)}{n}}| \right\} \right\}$ for $C_0 = \sqrt{\frac{8}{C_2}}$ where C_2 is as in Lemma B.2. Then,*

$$\mathbb{P}(A_{jk}) \geq 1 - 2C_3(q)^{-9/2}$$

for some constant $C_3 > 0$.

Proof. Let $A_1 := \left\{ |\widetilde{M}_{jk}| \geq C_0 \sqrt{\frac{\log(p \vee q)}{n}} \right\}$. Then, from the definition of $\widehat{M}_{jk}$ we have

$$|\widehat{M}_{jk} - M_{jk}| = |M_{jk}| \, \mathbb{1}_{A_1^c} + |\widetilde{M}_{jk} - M_{jk}| \, \mathbb{1}_{A_1} .$$

By the triangular inequality we have:

$$\begin{aligned} A_1 &= \left\{ |\widetilde{M}_{jk} - M_{jk} + M_{jk}| \geq C_0 \sqrt{\frac{\log(q)}{n}} \right\} \subset \left\{ |\widetilde{M}_{jk} - M_{jk}| \geq C_0 \sqrt{\frac{\log(q)}{n}} - |M_{jk}| \right\} \\ A_1^c &= \left\{ |\widetilde{M}_{jk} - M_{jk} + M_{jk}| < C_0 \sqrt{\frac{\log(q)}{n}} \right\} \subset \left\{ |\widetilde{M}_{jk} - M_{jk}| > |M_{jk}| - C_0 \sqrt{\frac{\log(q)}{n}} \right\} . \end{aligned}$$

Then, the proof proceed exactly as in Cai and Zhou [2012, Proof of Lemma 8] with $C_0 = \sqrt{\frac{8}{C_2}}$ where C_2 is as in the statement of Lemma B.2.

C. Methodology used for the cross-validation

Implementation of our procedure requires the choice of tuning parameters, namely λ, λ_j^Θ, $j = 1, \ldots, q$, λ_j^M, $j = 1, \ldots, p$, and C_0. These parameters have been chosen by 10-fold cross-validation in our numerical implementation of our procedure. In this section we describe the precise methodology that we have used for the cross-validation.

Consider first the cross-validation procedure to choose λ in the construction of the IV Lasso estimator $\widetilde{\beta}$ in (2.4). The algorithm is the following.

Algorithm 1.

- Randomly divide the set of indices $\{1, \ldots, q\}$ into 10 groups, or folds, of approximately equal size.
- For $i = 1, \ldots, 10$:
  1. construct a submatrix that contains only the rows of $\widehat{\Theta}^{1/2}$ corresponding to the indices in the i-th fold and denote it by $(\widehat{\Theta}^{1/2})^{(i)}$;
  2. construct a submatrix that contains all the rows of $\widehat{\Theta}^{1/2}$ except the ones corresponding to the indices in the i-th fold and denote it by $(\widehat{\Theta}^{1/2})^{(-i)}$;
  3. for a given λ, solve the minimization problem in (2.4) with these submatrices:

$$\widetilde{\beta}^{(-i)}(\lambda) = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \| (\widehat{\Theta}^{1/2})^{(-i)} (\mathbf{Z}^T \mathbf{Y}/n - \widehat{M}\beta) \|_2^2 + 2\lambda \|\beta\|_1 \right\} .$$

  This gives $\widetilde{\beta}^{(-i)}(\lambda)$;

4. compute the mean squared error $MSE^{(-i)}(\lambda)$ associated to the given λ as:

$$MSE^{(-i)}(\lambda) := \|(\widehat{\Theta}^{1/2})^{(i)}(\mathbf{Z}^T\mathbf{Y}/n - \widehat{M}\widetilde{\beta}^{(-i)}(\lambda))\|_2^2.$$

- Compute the 10-fold cross-validation estimate for the test mean squared error as

$$CV_{10}(\lambda) = \frac{1}{10}\sum_{i=1}^{10} MSE^{(-i)}(\lambda). \tag{C.1}$$

- Choose the λ that minimizes $CV_{10}(\lambda)$.

In practice one has to use a grid for λ and select the value in this grid that gives a minimum value for $CV_{10}(\lambda)$. This procedure is automatically produced by the `R` function `cv.glmnet` of the `glmnet` package.

The cross-validation procedure to choose λ_j^Θ, $j = 1, \ldots, q$, is described in the following algorithm.

Algorithm 2.

- Randomly divide the set of observations $Z_1, \ldots, Z_n$ into 10 groups, or folds, of approximately equal size.
- For $i = 1, \ldots, 10$:
  1. construct a subvector and a submatrix of $\mathbf{Z}_j$ and $\mathbf{Z}_{-j}$ that contain only the observations in the held-out i-th fold and denote them by $\mathbf{Z}_j^{(i)}$ and $\mathbf{Z}_{-j}^{(i)}$, respectively;
  2. construct a subvector and a submatrix of $\mathbf{Z}_j$ and $\mathbf{Z}_{-j}$ that contain all the observations except the ones in the i-th fold and denote them by $\mathbf{Z}_j^{(-i)}$ and $\mathbf{Z}_{-j}^{(-i)}$, respectively;
  3. for a given λ_j^Θ, solve the minimization problem in (2.7) by using $\mathbf{Z}_j^{(-i)}$ and $\mathbf{Z}_{-j}^{(-i)}$:
  $$\widehat{\xi}_j^{-i}(\lambda_j^\Theta) = \underset{\xi\in\mathbb{R}^{q-1}}{\operatorname{argmin}} \left\{ \|\mathbf{Z}_j^{(-i)} - \mathbf{Z}_{-j}^{(-i)}\xi\|_2^2/n + 2\lambda_j^\Theta\|\xi\|_1 \right\}.$$
  This gives $\widehat{\xi}_j^{-i}(\lambda_j^\Theta)$;
  4. compute the mean squared error $MSE^{(-i)}(\lambda_j^\Theta)$ associated to the given λ_j^Θ as:
  $$MSE^{(-i)}(\lambda_j^\Theta) := \|\mathbf{Z}_j^{(i)} - \mathbf{Z}_{-j}^{(i)}\widehat{\xi}_j^{-i}(\lambda_j^\Theta)\|_2^2/n.$$
- Compute the 10-fold cross-validation estimate for the test mean squared error as

$$CV_{10}(\lambda_j^\Theta) = \frac{1}{10}\sum_{i=1}^{10} MSE^{(-i)}(\lambda_j^\Theta). \tag{C.2}$$

- Choose the λ_j^Θ that minimizes $CV_{10}(\lambda_j^\Theta)$.

The cross-validation procedure to choose λ_j^M, $j = 1, \ldots, p$ is the same as the one described in Algorithm 1 with the following modification of steps 3-4 in the for loop:

3. for a given λ_j^M, solve the minimization problem in (2.11):

$$\widetilde{\gamma}_j^{(-i)}(\lambda_j^M) = \underset{\gamma\in\mathbb{R}^{p-1}}{\operatorname{argmin}} \left\{ \|((\widehat{\Theta}^{1/2})^{(-i)}\widehat{M})_j - ((\widehat{\Theta}^{1/2})^{(-i)}\widehat{M})_{-j}\gamma\|_2^2 + 2\lambda_j^M\|\gamma\|_1 \right\}.$$

This gives $\widetilde{\gamma}_j^{(-i)}(\lambda_j^M)$;

4. compute the mean squared error $MSE^{(-i)}(\lambda_j^M)$ associated to the given λ_j^M as:

$$MSE^{(-i)}(\lambda_j^M) := \|((\widehat{\Theta}^{1/2})^{(i)}\widehat{M})_j - ((\widehat{\Theta}^{1/2})^{(i)}\widehat{M})_{-j}\widetilde{\gamma}_j^{(-i)}(\lambda_j^M)\|_2^2.$$

Finally, the cross-validation procedure to select the constant $c_n := C_0\sqrt{\log(q)/n}$ for the construction of $\widehat{M}$ is given in the following algorithm.

Algorithm 3.

- For $i = 1, \ldots, 10$:
  1. randomly select a set of observations in $\{(X_1, Z_1), \ldots, (X_n, Z_n)\}$ of size $n_{tr} = \lceil n(1 - 1/log(n))\rceil$. This is the training dataset and is denoted with a tr_i index, the remaining observations will be the validation data denoted with a v_i index;
  2. construct the submatrices of $\mathbf{X}$ and $\mathbf{Z}$ that contain only the observations in the validation fold and denote them by $\mathbf{X}^{(v_i)}$ and $\mathbf{Z}^{(v_i)}$, respectively;
  3. construct the submatrices of $\mathbf{X}$ and $\mathbf{Z}$ that contain only the observations in the training dataset and denote them by $\mathbf{X}^{(tr_i)}$ and $\mathbf{Z}^{(tr_i)}$, respectively;
  4. construct $\widetilde{M}^{(tr_i)} := (\mathbf{Z}^{(tr_i)})^T\mathbf{X}^{(tr_i)}/n_{tr}$;
  5. for a given c_n, compute the thresholding estimator by using (2.10), $\mathbf{X}^{(tr_i)}$ and $\mathbf{Z}^{(tr_i)}$:
  $$\widehat{M}_{jk}^{(i)}(c_n) := \widetilde{M}_{jk}^{(tr_i)}\, \mathbb{1}\left\{|\widetilde{M}_{jk}^{(tr_i)}| \geq c_n\right\}.$$
  This gives $\widehat{M}_{jk}^{(i)}(c_n)$;
  6. compute the Frobenius norm $\|\cdot\|_F$ of the difference between $\widehat{M}_{jk}^{(i)}(c_n)$ and $\widetilde{M}^{(v_i)} := (\mathbf{Z}^{(v_i)})^T\mathbf{X}^{(v_i)}/(n - n_{tr})$:
  $$Loss^{(i)}(c_n) := \|\widehat{M}_{jk}^{(i)}(c_n) - \widetilde{M}^{(v_i)}\|_F.$$
- Compute the mean of the losses as

$$Loss_{10}(c_n) = \frac{1}{10}\sum_{tr=1}^{10} Loss^{(i)}(c_n). \tag{C.3}$$

- Choose the c_n in a grid that minimizes $Loss_{10}(c_n)$.

References

C. Ai and X. Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843, 2003.

D. W. Andrews and X. Cheng. Estimation and inference with weak, semi-strong, and strong identification. *Econometrica*, 80(5):2153–2211, 2012.

J. D. Angrist and A. B. Krueger. Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106(4):979–1014, 1991.

A. Belloni, V. Chernozhukov, and H. Hansen. LASSO methods for gaussian instrumental variables models. Technical report, arXiv:1012.1297, 2011.

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

A. Belloni, V. Chernozhukov, I. Fernández-Val, and C. Hansen. Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298, 2017a.

A. Belloni, V. Chernozhukov, C. Hansen, and W. Newey. Simultaneous confidence intervals for high-dimensional linear models with many endogenous variables. *arXiv preprint arXiv:1712.08102*, 2017b.

S. Berry, J. Levinsohn, and A. Pakes. Automobile prices in market equilibrium. *Econometrica*, 63(4):841–890, 1995.

P. J. Bickel and E. Levina. Covariance regularization by thresholding. *Annals of Statistics*, 36(6):2577–2604, 2008.

R. Blundell, X. Chen, and D. Kristensen. Semi-nonparametric iv estimation of shape-invariant engel curves. *Econometrica*, 75(6):1613–1669, 2007.

C. Breunig and J. Johannes. Adaptive estimation of functionals in nonparametric instrumental regression. *Econometric Theory*, 32(3):612–654, 2016.

P. Bühlmann. Statistical significance in high-dimensional linear models. *Bernoulli*, 19 (4):1212–1242, 2013.

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

T. T. Cai and H. H. Zhou. Minimiax estimation of large covariance matrices under ℓ_1-norm. *Statistica Sinica*, 22(4):1319–1349, 2012.

M. Carrasco and M. Doukali. Efficient estimation using regularized jackknife iv estimator. *Annals of Economics and Statistics*, (128):109–149, 2017.

J. C. Chao and N. R. Swanson. Consistent estimation with a large number of weak instruments. *Econometrica*, 73(5):1673–1692, 2005.

X. Chen and T. M. Christensen. Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric iv regression. *Quantitative Economics*, 9(1): 39–84, 2018.

X. Chen and D. Pouzo. Estimation of nonparametric conditional moment models with possibly nonsmooth moments. *Econometrica*, 80(1):277–322, 2012.

X. Chen and D. Pouzo. Sieve wald and qlr inferences on semi/nonparametric conditional moment models. *Econometrica*, 83(3):1013–1079, 2015.

V. Chernozhukov, C. Hansen, and M. Spindler. Post-selection and post-regularization inference in linear models with many controls and instruments. *American Economic Review*, 105(5):486–90, May 2015.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

S. Darolles, Y. Fan, J.-P. Florens, and E. Renault. Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565, 2011.

J. Fan and Y. Liao. Endogeneity in high dimensions. *Annals of Statistics*, 42(3):872, 2014.

E. Gautier, A. Tsybakov, and C. Rose. High-dimensional instrumental variables regression and confidence sets. *arXiv preprint arXiv:1105.2454*, 2011.

D. Gold, J. Lederer, and J. Tao. Inference for high-dimensional instrumental variables regression. *arXiv:1708.05499v2*, 2018.

Z. Guo, H. Kang, T. Tony Cai, and D. S. Small. Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):793–815, 2018.

C. Hansen and D. Kozbur. Instrumental variables estimation with many weak instruments using regularized jive. *Journal of Econometrics*, 182(2):290–308, 2014.

C. Hansen, J. Hausman, and W. Newey. Estimation with many instrumental variables. *Journal of Business & Economic Statistics*, 26(4):398–422, 2008.

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014a.

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

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

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

W. K. Newey and J. L. Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

M. Neykov, Y. Ning, J. S. Liu, and H. Liu. A unified theory of confidence regions and testing for high-dimensional estimating equations. *Statistical Science*, 33(3):427–443, 2018.

S. Ng and J. Bai. Selecting instrumental variables in a data rich environment. *Journal of Time Series Econometrics*, 1(1), 2009.

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

S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42(3): 1166–1202, 06 2014.

S. A. van de Geer. *Estimation and testing under sparsity*. Springer, 2016.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok, editors, *Compressed Sensing: Theory and Applications*, volume 26, pages 210–268. Cambridge University Press, 2012.

F. Windmeijer, H. Farbmacher, N. Davies, and G. D. Smith. On the use of the lasso for instrumental variables estimation with some invalid instruments. *Journal of the American Statistical Association*, 114(527):1339–1350, 2019.

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