This paper is concerned with inference about low-dimensional components of a high-dimensional parameter vector $\beta^0$ 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 $\beta^0$ at a rate depending on the mapping properties of $M$ captured by a sparse link condition. Linear combinations of our estimator of $\beta^0$ 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 $\beta^0$. In Monte-Carlo simulations we analyze the finite sample behavior of our estimator.

**Keywords:** Instrumental Variables, sparsity, central limit theorem, lasso, linear model, desparsification, ill-posed estimation problem.
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 \( \beta^0 \) 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 \( \beta^0 \) may be larger than the sample size \( n \).

This paper is concerned with inference on inner products of \( \beta^0 \) 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 \( \beta^0 \) is the parameter of interest and the remaining components of \( \beta^0 \) are considered as nuisance parameters, then inference on \( a^T \beta^0 \) implies inference on this low-dimensional subvector of \( \beta^0 \) for an appropriate choice of the vector \( a \). We also allow the subvector of \( \beta^0 \) of interest to increase slowly with the sample size and provide inference for it. Our main example is when the low-dimensional subvector of \( \beta^0 \) 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 \( \beta^0 \) 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. In this paper, we relate the sparsity of \( M \) to sparsity of the parameter vector \( \beta^0 \) which we capture by a sparsity link condition.

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. Moreover, the relation between the sparsity pattern of \( M \) and the sparsity pattern of the parameter vector \( \beta^0 \) is captured by a coefficient of sparsity ill-posedness which we introduce in this paper and which is used to characterize the sparsity link condition. The coefficient of sparsity ill-posedness generalizes and extends to the high-dimensional setting 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 $\beta^0$. 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 Lasso IV estimator for $\beta^0$ 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 $M$ (opportunistically 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 $\beta^0$ 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 $\beta^0$. As discussed above, inference results on inner products of $\beta^0$ imply inference results on low-dimensional subvectors of $\beta^0$ or even on subvectors of $\beta^0$ slowly increasing with the sample size. In Monte Carlo simulation, 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. In this paper, we will compare the NPIV literature to the high-dimensional IV literature that uses sparsity assumptions.

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]. 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 criteria 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 classical 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. [2016] propose a method to select valid instruments based on hard thresholding in set ups 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 set ups. Kang et al. [2016] use Lasso to instrumental variable selection in the context of invalid instruments, i.e., enter the structural equation. Based on an initial median estimator, Windmeijer et al. [2017] 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 and Tsybakov [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. 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 link 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. [2016] and Gold et al. [2018] that, as we propose in our paper, use two-step estimators using a threshold procedure or 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 $\beta$ 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 $\beta$ 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.

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 $\ell_2$ penalty is used and the asymptotic distribution results can be readily obtained. This approach in high-dimensional Gaussian regression is considered by Bühlmann et al. [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 setup, 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. All proofs can be found in the appendix.
Notation. The \(\ell_p\) 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 \(S\). 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 \(\ell_1\) 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.

2. Model and Methodology

Consider again model (1.1), the high-dimensional instrumental variable model is given by
\[
Y = X^T \beta^0 + U \quad \text{where } \mathbb{E}[UZ] = 0, \tag{1.1}
\]
where \(\beta^0\) 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 \(\beta^0\), we assume throughout the paper that \(q \geq p\).

We also assume throughout the paper that the matrices \(M := \mathbb{E} (ZX^T)\) and \(\Sigma := \mathbb{E} (ZZ^T)\) are of full column rank. Thus, the parameter vector \(\beta^0\) is identified through
\[
\beta^0 = (M^T\Sigma^{-1}M)^{-1}M^T\Sigma^{-1}\mathbb{E}[ZY]. \tag{2.1}
\]
Estimating \(\beta^0\) 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 \(\Sigma\) 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 \( \beta^0 \) 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 \( Y \).

By denoting the orthonormalized excluded regressors as \( \tilde{Z} := \Sigma^{-1/2}Z \), equation (1.1) can be rewritten in its reduced form as

\[
Y = \tilde{Z}^T \beta^* + V,
\]

where \( \beta^* := \Sigma^{-1/2}M\beta^0 \) is the reduced form parameter and \( V = Y - Z^T\Sigma^{-1}\mathbb{E}[YZ] \). Note that \( V \) is mean independent of \( Z \), i.e., \( \mathbb{E}[VZ] = 0 \). A central insight of this paper is to relate the sparsity of the parameter vector \( \beta^0 \) to the sparsity of the transformed parameter vector \( \beta^* \).

Throughout the paper, we assume that there exist two sets \( \tilde{S}_0 \) and \( S_0 \) such that \( \|\beta^0_{\tilde{S}_0}\|_1 \) and \( \|\beta^*_{S_0}\|_1 \) are sufficiently small, as to be specified below. In this sense, we do not impose an exact but only an approximate sparsity condition\(^1\) on the vector \( \beta^0 \) and the reduced form vector \( \beta^* \). The central idea is now to link the sparsity structure of \( \beta^0 \) to the one of \( \beta^* \) in the sense that, roughly speaking, if both sparsity structures are similar we have a well conditioned estimation problem while in the other case it is more ill conditioned.

The sparsity of the structural parameter \( \beta^0 \), captured by \( \tilde{S}_0 \), is then related to the sparsity of the vector \( \beta^* \), captured by \( S_0 \), by the following coefficient of sparsity ill-posedness (COSI)

\[
\omega_1 := \sup_{\beta \in B} \frac{\|\beta_{\tilde{S}_0}\|_2^2}{\|\Sigma^{-1/2}M\beta\|_{S_0}^2},
\]

where \( B := \{\beta : \|\beta_{\tilde{S}_0}\|_1 \leq 3\|\beta_{\tilde{S}_0}\|_1\} \). In the case of exact sparsity, a lower bound for the COSI is given by \( \omega_1 \geq \|\beta^0\|_2^2/\|\beta^*\|_2^2 \). Throughout the paper, we assume that the COSI may increase with the sample size. For our asymptotic results we impose upper bounds on the COSI which depend on the sample size and the underlying sparsity constraints.

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 below). We thus introduce

\[
\omega_2 := 1/\lambda_{\min}\left(M^T\Sigma^{-1}M\right)
\]

which satisfies \( \omega_2 < \infty \) 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, \( \omega_2 \) is strictly positive for all \( n \geq 1 \). On the other hand, in many cases we expect that \( \omega_2 \) 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 \( \Sigma^{-1/2}M \).

---

\(^1\)Approximate sparsity here refers to the fact that most of the elements of \( \beta^* \) are small and not necessarily equal to zero as they are under the classical sparsity condition. Hence, approximate sparsity is a restriction on the \( \ell_1 \) norm of \( \beta^* \) in contrast to the classical sparsity condition which restricts the number of non-zero coefficients.
Similar to Andrews and Cheng [2012], we consider the strongly identified case where \( \omega_2 \) is uniformly bounded above, and the semi-strongly identified case where \( \omega_2 \) is unbounded but satisfies \( n/\omega_2 \to \infty \). We show below that \( \omega_2 \) slows down the rate of convergence of our estimator. In the semi-strong case, the size of the confidence sets increases relative to \( \omega_2 \). There is also a third case which is the weak identified case where \( n/\omega_2 = O(1) \) but the results of our paper do not apply to it. We emphasize that asymptotic behaviors of \( \omega_1 \) and \( \omega_2 \) have very different implications for our inference results.

The asymptotic behavior of \( \omega_1 \) affects only the sparsity requirements of our model but does not affect the convergence rate of the proposed estimator. This is in contrast to \( \omega_2 \), which does not affect the sparsity of the model but might lead to slower rates of convergence. In this paper, we show that under appropriate assumptions the rate of convergence of our estimator for each component of \( \beta^0 \) is \( \sqrt{\omega_2/n} \).

In the case where model (1.1) is seen as an approximation of a more complex nonlinear model, the COSI has the interpretation of an \( \ell_1 \) analog of the sieve measure of ill-posedness introduced in Blundell et al. [2007] and extended in Chen and Pouzo [2012] and Chen and Pouzo [2013]. The sieve measure of ill-posedness relates the strong norm of sieve approximation relative to its weak norm induced by a conditional expectation operator. We also see below that we require a similar link condition as in the sieve nonparametric instrumental variable literature to provide a link between strong and weak norm.

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 \( Y = (Y_1, \ldots, Y_n)^T \), \( X = (X_1, \ldots, X_p) \) with \( X_j = (X_{1,j}, \ldots, X_{n,j})^T \) for \( 1 \leq j \leq p \), and \( Z = (Z_1, \ldots, Z_q) \) with \( Z_j = (Z_{1,j}, \ldots, Z_{n,j})^T \) for \( 1 \leq j \leq q \). Moreover, the \( n \)-vector of unobservables is denoted by \( U = (U_1, \ldots, U_n)^T \). 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 \( \Theta := \Sigma^{-1} M_j \) and \( \Theta^M := (M^T \Theta M)^{-1} \) denoted by \( \hat{\Theta} \), \( \hat{M} \) and \( \hat{\Theta}^M \), respectively, which are introduced in Subsection 2.3 below. We propose the following desparsified IV Lasso estimator of \( \beta^0 \) given by

\[
\hat{\beta} = \hat{\Theta}^M \hat{M}^T \hat{\Theta} Z^T Y/n - (\hat{\Theta}^M \hat{M}^T \hat{\Theta} Z^T X/n - I_p) \tilde{\beta}
\]

(2.3)

where \( \tilde{\beta} \) is a consistent estimator of \( \beta^0 \) that makes use of the underlying sparsity assumption. The first summand on the right hand side of (2.3) corresponds to a regularized empirical analog of \( \beta^0 \) as in (2.1). The second summand of the right hand side of (2.3) 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 \( \hat{\Theta} \), \( \hat{M} \) and \( \hat{\Theta}^M \) are the usual empirical counterparts of \( \Theta \), \( M \) and \( \Theta^M \). In this case it holds \( \hat{\Theta}^M \hat{M}^T \hat{\Theta} Z^T X/n = I_p \) and moreover, \( \hat{\beta} \) coincides with the 2SLS estimator.

The choice of \( \tilde{\beta} \) is motivated by our sparsity assumption given below and by the asymptotic properties for \( \tilde{\beta} \) that we want to obtain. To derive the asymptotic results of
our desparsified IV Lasso estimator we make use of the following key decomposition
\[ \sqrt{n}(\hat{\beta} - \beta^0) = \hat{\Theta}^T \hat{M}^T \hat{\Theta} Z^T U / \sqrt{n} - \Delta, \] (2.4)
for a remainder term \( \Delta \) which is given by
\[ \Delta := \sqrt{n}(\hat{\Theta}^T \hat{M}^T \hat{\Theta} Z^T X / n - I_p)(\hat{\beta} - \beta^0). \]

Then, we have to show that \( \|\Delta\|_\infty \) is asymptotically negligible under regularity assumptions. In particular, to show this we require that \( \|\hat{\beta} - \beta^0\|_1 \) is sufficiently small. This property is satisfied by the Lasso estimator and thus we choose \( \hat{\beta} \) in equation (2.3) to be the Lasso estimator which makes use of the underlying sparsity structure imposed on \( \beta^0 \). Therefore, our estimation procedure is based on the IV Lasso estimator of \( \beta^0 \) given by
\[ \hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ (Z^T Y / n - \hat{M}\beta)^T \hat{\Theta} (Z^T Y / n - \hat{M}\beta) + 2\lambda \|\beta\|_1 \right\} \] (2.5)
for some tuning parameter \( \lambda > 0 \), which we replace in equation (2.3) to obtain \( \hat{\beta} \).

### 2.2. Sparsity Constraints

In this section we introduce some notations and assumptions about sparsity that we tacitly maintain all along the paper. In the following, 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 \( \|\beta^c_{S^0_0}\|_1 \) is sufficiently small as specified by our Assumption 2 below. Moreover, we assume that the set \( \tilde{S}_0 \) is rich enough such that the parameter vector \( \beta^0 \) satisfies
\[ \|\beta^0_{S^0_0}\|_1 = \sum_{j \notin S_0} |\beta^0_j| \leq C \omega_1 s_0 \sqrt{\log(p)/n} \] (2.6)
for some constant \( C > 0 \). We thus restrict the (approximate) sparsity of \( \beta^0 \) through the COSI and the cardinality of \( S_0 \). Below we introduce a sparsity link condition which, together with inequality (2.6), implies an approximate sparsity bound on \( \beta^* \) and characterizes \( S_0 \).

Hereafter, we assume that \( \Theta \) and \( \Theta^M \) exist and assume sparsity with respect to rows of \( \Theta \). To this purpose we define
\[ s_{\text{max}} := \max_{1 \leq i \leq q} |\{k \neq j : \Theta_{jk} \neq 0\}|. \]

The sparsity restriction on \( \Theta \) has the following interpretation: if the \( (jk) \)-th component of \( \Theta \) 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 to impose an \( \ell_1 \)-penalty for the estimation of \( \Sigma^{-1} \), which was proposed by Meinshausen and Bühlmann [2006].

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 $\Theta$ but at the cost of much more technical proofs and notation. For the sparsity of $M$ we introduce the notation

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

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

$$\|\gamma_{j,S_j}\|_1 = \sum_{l \not\in S_j} |\gamma_{jl}| \leq C \log(q) / \sqrt{n} \quad (2.7)$$

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

### 2.3. Regularized Matrix Estimators

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

#### 2.3.1. Construction of $\hat{\Theta}$

Here we construct a regularized estimator of the inverse of $\Theta$ denoted by $\hat{\Theta}$. The basic idea to construct such an estimator is to relate the inversion of a $q \times q$ matrix to $q$ regression problems of $Z_j$ over $Z_{-j}$. This approach was introduced by Meinshausen and Bühlmann [2006]. For every $j = 1, \ldots, q$ we consider the Lasso estimator

$$\hat{\gamma}_j = \arg\min_{\gamma \in \mathbb{R}^{q-1}} \left\{ \| Z_j - Z_{-j} \gamma \|_2^2 / n + 2\lambda_j \|\gamma\|_1 \right\} \quad (2.8)$$

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

$$\hat{\Gamma}_{kj} = \begin{cases} 1 & \text{for } k = j \\ -\hat{\gamma}_{jk} & \text{for } k \neq j \end{cases} \quad (2.9)$$

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

$$\hat{\Theta}_j = \hat{\gamma}_j^2 \hat{\Gamma}_j \quad \text{where} \quad \hat{\gamma}_j^2 = \| Z \hat{\Gamma}_j \|_2^2 / n + \lambda_j \|\hat{\gamma}_j\|_1. \quad (2.10)$$

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

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 \( \hat{M} := Z^T X / n \) that lie below some specified threshold. More precisely, we propose to use the thresholding estimator \( \hat{M} = (\hat{M}_{jk}) \) where

\[
\hat{M}_{jk} := \begin{cases} 
M_{jk} & \text{if } |\hat{M}_{jk}| \geq C_0 \sqrt{\frac{\log q}{n}}, \\
0 & \text{otherwise},
\end{cases} \quad C_0 > 0.
\] (2.11)

For symmetric matrices such a regularization scheme has been considered in Bickel and Levina [2008] and Cai and Zhou [2012] among others.

2.3.3. Construction of \( \hat{\Theta}^M \)

In this section, we construct the estimator \( \hat{\Theta}^M \) which is an approximate inverse of \( \hat{M}^T \hat{\Theta} \hat{M} \). This estimator involves the regularized estimators \( \hat{\Theta} \) and \( \hat{M} \) obtained in Sections 2.3.1 and 2.3.2.

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

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

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

\[
\hat{\gamma}_j = \arg \min_{\gamma \in \mathbb{R}^{p-1}} \left\{ \| (\hat{\Theta}^{1/2} \hat{M})_j - (\hat{\Theta}^{1/2} \hat{M})_{-j} \gamma \|_2^2 + 2\lambda_j^M \| \gamma \|_1 \right\}, \tag{2.12}
\]

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 \( \hat{\Gamma}_j = (\hat{\Gamma}_{kj})_{k=1}^p \) be the \( p \)-column vector determined by

\[
\hat{\Gamma}_{kj} = \begin{cases} 1 & \text{for } k = j, \\
-\hat{\gamma}_{jk} & \text{for } k \neq j
\end{cases}
\]

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

\[
\hat{\Theta}_j^M = \hat{\gamma}_j^{-2} \hat{\Gamma}_j, \quad \hat{\gamma}_j^2 = \| \hat{\Theta}^{1/2} \hat{M} \hat{\Theta}_j^M \|_2^2 + \lambda_j^M \| \hat{\gamma}_j \|_1, \quad 1 \leq j \leq p.
\]

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

In this section, we derive the asymptotic distribution of the desparsified IV Lasso estimator \( \hat{\beta} \) given in (2.3). To obtain asymptotic results on which our inference will be based we have to show that the remainder term \( \Delta \) in the key decomposition (2.4) 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 \( \hat{\beta} \), 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
\[
\mathbb{E}\exp\left(\frac{|v^T W|^2}{C}\right) = O(1)
\]
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 \( \Sigma \) are uniformly bounded away from zero and from infinity. (iv) The smallest eigenvalue \( \lambda_{\text{min}}(M^T \Sigma^{-1} M) \) is bounded from below for each \( n \geq 1 \) and the largest eigenvalue \( \lambda_{\text{max}}(M^T \Sigma^{-1} M) \) 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 \( \Sigma_{jj} \leq \lambda_{\text{max}}(\Sigma) \). Similarly, it also implies that \( \|\Theta_j\|_2 \leq \lambda_{\text{min}}(\Sigma) = O(1) \) uniformly in \( j \) and consequently, \( \|\Theta\|_1 = O(\sqrt{s_{\text{max}}} \) which we use below.

In the following, we aim to relate the sparsity of the structural parameter \( \beta^0 \) to the sparsity of the vector \( \Sigma^{-1/2} M \beta^0 \). Recall the definition of the COSI given by
\[
\omega_1 = \sup_{\beta \in \mathcal{B}} \frac{\|\beta_{\tilde{S}_0}\|_1^2}{\|\left(\Sigma^{-1/2} M \beta\right)_{S_0}\|_1^2},
\]
where \( \mathcal{B} := \{\beta : \|\beta_{\tilde{S}_0}\|_1 \leq 3\|\beta_{\tilde{S}_0}\|_1\} \). We also emphasize that the following analysis regarding the COSI is not required if sparsity constraints are imposed not only on the structural equation but also on linear reduced form equations, as in Chernozhukov et al. [2015], Guo et al. [2016], and Gold et al. [2018]. We assume the following lower bound which involves \( \omega_1 \).

**Assumption 2 (Sparsity Link Condition).** There exists some constant \( \eta > 0 \) such that for all \( \beta \in \mathcal{B} \) we have
\[
\|\beta_{\tilde{S}_0}\|_1^2 \geq \eta \omega_1 \|\left(\Sigma^{-1/2} M \beta\right)_{S_0}\|_1^2.
\]

Assumption 2 is an \( \ell_1 \)–version of the so-called stability condition, which is also referred to as the source condition in inverse problems literature, see for instance Engl et al. [2000]. In the NPIV literature a stability condition in the \( \ell_2 \)–sense was used for instance by Chen and Pouzo [2012, Assumption 5.2]. Assumption 2 links the set \( S_0 \) to the set \( \tilde{S}_0 \) through the parameter \( \omega_1 \). Without this assumption we cannot infer results about the structural parameters \( \beta^0 \) from results on \( \beta^* \). In particular, the approximate
sparsity condition on $\beta^0$ imposed in inequality (2.6) together with Assumption 2 implies for the reduced form parameter $\beta^*$ that
\[
\|\beta^*_S\|_1 \leq C\sqrt{\omega_1/\eta} s_0 \sqrt{\log(p)/n}.
\]
When $\omega_1$ is bounded, the vectors $\beta^*_S$ and $\beta^0_S$ have, in the $\ell_1$-norm, the same upper bound up to a constant. Assumption 2 requires that when $\omega_1$ increases this implies that the ratio $\|\beta^*_S\|/\|\beta^0_S\|$ decreases.

Lemma 3.1. Let Assumptions 1 and 2 be satisfied. If $\log(q)/n = o(1)$ then it holds for all $\beta \in \mathcal{B}$ that
\[
\|\beta^*_S\|^2 \leq \omega_1 s_0 \beta^T M^T \Sigma^{-1} \Sigma^{-1} M \beta/c^2
\]
wp1 for some constant $c > 0$ and where $\Sigma = ZT Z/n$.

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 3. (i) It holds $\lambda \sim \log(q)/\sqrt{n}$, $\lambda_j \sim \sqrt{\log(q)/n}$, and $\lambda^M_j \sim \log(q)/\sqrt{n}$ uniformly in $j$. (ii) It holds $\|M\|_{\infty} = O(1)$, $\mathbb{E} \max(1, |X^T \beta^0|^2) ||M^T \Sigma^{-1} Z||_{\infty}^2 = O(\log(p))$ and $\mathbb{E}[U^2 | Z] \leq \sigma^2$ for a constant $\sigma > 0$. (iii) Assume $s_M \sqrt{s_{\max}} \max(s_M, \|\beta^0\|_1) = O(\sqrt{\log(q)})$ and
\[
\omega_1 s_0 s_M \sqrt{s_{\max}} \max(\sqrt{s_M}, \sqrt{s_{\max}}) \sqrt{\log(p)} \log(q) + \omega_2 s_{\max}^2 = o(\sqrt{n/\log(q)}).
\]

Assumption 3 (i) specifies the rate of the tuning parameters $\lambda$ used for the plug-in Lasso and $\lambda_j, \lambda^M_j$ used for the nodewise Lasso estimators. The rate of the regularization parameters $\lambda$ and $\lambda^M_j$ 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 3 (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 $Z$. Assumption 3 (iii) imposes sparsity restrictions which we require in order to obtain our inference results. In particular, condition (3.2) implies $\log(p)/\sqrt{n} = o(1)$.

For the next assumption, recall that $\gamma_j := \arg\min_{\gamma \in \mathbb{R}^{p-1}} \|(\Theta^{1/2} M)_{\gamma} \|_2$ 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 $\gamma_{kj}$ is the $k$-th entry of $\gamma_j$.

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

Assumption 4 (i) imposes upper bounds on moments associated to $ZX^T$ while Assumption 4 (ii) imposes mild rate conditions on moments of $ZZ^T$. Note that the logarithmic rates in Assumption 4 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 \( \hat{\beta} \). 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 \( \hat{M} \) in the \( \ell_1 \) norm. The first part of the theorem provides a large deviation inequality for the components of \( \tilde{M} \) and it is derived by exploiting sub-Gaussianity of the rows of \( X \) and \( Z \) 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

\[
P(|M_{jk} - \hat{M}_{jk}| \geq v) \leq 4 \exp \left( -cv^2n \right)
\]

for \( 0 \leq v < 1 \). Moreover, let \( \hat{M} \) be the thresholding estimator defined in (2.11) with \( C_0 = \sqrt{8/c} \). If in addition Assumption 3 (i) and (ii) holds, then we have

\[
\|\hat{M} - M\|_1 = O_p(s_M \sqrt{\log(q)/n}).
\]

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

**Lemma 3.3.** We have

\[
\|\hat{\Sigma} \hat{\Theta}_j - e_j\|_\infty \leq \lambda_j / \tilde{\tau}_j^2,
\]

and

\[
\|\hat{M}^T \hat{\Theta} \hat{M} \hat{\Theta}_M^j - e_j\|_\infty \leq \lambda_j^M / \tilde{\tau}_j^2.
\]

We now establish the rate of convergence of the regularized estimators \( \hat{\Theta} \) and \( \hat{\Theta}_M^j \). 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

\[
\|\hat{\Theta} - \Theta\|_{op, \infty} = O_p(s_{\max} \sqrt{\log(q)/n}).
\]

If, in addition, Assumptions 3 and 4 are satisfied then

\[
\|\hat{\Theta}_M^j - \Theta^j\|_{op, \infty} = O_p(\omega_2^2 s_{\max}^M \log(q)/\sqrt{n}).
\]

3.3. Rate of Convergence

In this subsection, we derive the rate of convergence of the desparsified IV Lasso estimator \( \hat{\beta} \). The next theorem provides an asymptotic upper bound of the bias term \( \Delta \), which is key to derive further inference results.
Theorem 3.5. Let Assumptions 1–4 be satisfied. Then, we have
\[ \sqrt{n} (\hat{\beta} - \beta^0) = \omega_2 V + \Delta \]
where
\[ V = \hat{\Theta}^T \hat{M}^T \hat{Z}^T U / (\sqrt{n} \omega_2) \]
and \( \Delta \) satisfies
\[ \| \Delta \|_\infty = O_p(\omega_1 s_0 \max(\omega_2, \| \Theta M \Theta^M \|_1)(\log q)^2 / \sqrt{n}) . \]

From Theorem 3.5 we see that the rate of convergence of the desparsified Lasso estimator \( \hat{\beta} \) is affected by the possibly increasing parameter \( \omega_2 \). In the next result, we show that the bias term \( \Delta \) is indeed asymptotically negligible under additional rate requirements. We also see below that the rate of convergence of our estimator is given by \( \sqrt{\omega_2 / n} \) under a mild assumption.

Corollary 3.6. Let Assumptions 1–4 be satisfied. In addition, we assume
\[ \omega_1 s_0 (\log q)^2 \max(1, \| \Theta M \Theta^M \|_1 / \omega_2) = o(\sqrt{n}). \] (3.6)
Then, we have
\[ \sqrt{n} (\hat{\beta} - \beta^0) / \omega_2 = V + o_p(1). \]

We will see in the next section that \( V \) converges to a normal distribution with a covariance matrix to be specified below. We also see that only the inverse of \( \omega_2 \) enters the sparsity condition in equation (3.7). We hence conclude that while \( \omega_2 \) does not restrict the sparsity \( s_0 \) nor the dimension \( p \) or \( q \), it affects the rate of convergence. In the strong identified case, the components of \( \hat{\beta} \) are \( \sqrt{n} \) consistent. In the weak identified case, this rate of convergence may slow down depending on the asymptotic behavior of \( \omega_2 \). We also emphasize that \( \omega_1 \) does not affect the rate of convergence but imposes stronger sparsity restriction if \( \omega_1 \) is increasing.

Also the next result is an immediate consequence of Theorem 3.6 and provides a bound of linear functionals of \( \hat{\beta} - \beta^0 \) uniformly over representers \( a \in \mathbb{R}^p \) with \( \ell_1 \) 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 1–4 be satisfied. In addition, we assume
\[ \omega_1 s_0 (\log q)^2 \max(1, \| \Theta M \Theta^M \|_1 / \omega_2) = o(\sqrt{n} / K). \] (3.7)
Then, we have
\[ \sup_{a \in \mathcal{A}_K} \left| \sqrt{n} a^T (\hat{\beta} - \beta^0) / \omega_2 - a^T V \right| = o_p(\sqrt{K}). \]

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 \( \beta \) and vectors \( a \) selecting linear combinations of a finite number of components of \( \beta \), 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 $X_{\text{end}}$. 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. If $X_{\text{end}}$ has bounded support, denoted by $\text{supp}(X_{\text{end}})$, then $\sup_{s \in \text{supp}(X_{\text{end}})} \|\phi^K(s)\|_1 = O(\sqrt{K})$ for CDV wavelets, see Chen and Christensen [2017, Appendix E]. An application of Corollary 3.7 yields that

$$\sup_{s \in \text{supp}(X_{\text{end}})} \left| \sqrt{n}a^K(s)^T(\hat{\beta}_{\text{end}} - \beta^0_{\text{end}})/\omega_2 - \phi^K(s)^TV \right| = o_p(\sqrt{K}).$$

Consequently, for $\phi^K(s)^T(\hat{\beta} - \beta^0)$ we obtain the rate of convergence $\omega_2 \sqrt{K}/n$, provided that $\phi^K(s)^TV = O_p(\sqrt{K})$ 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]. In contrast to the sup-norm convergence results of Chen and Christensen [2017, 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 $\hat{\beta}$. We also see that asymptotic normality of components of $\hat{\beta}$ immediately follows.

To achieve the asymptotic distribution of our estimator $\hat{\beta}$ we consider first a normalization factor to standardize the estimator $\tilde{\beta}$. The covariance matrix for 2SLS estimators is given by

$$\Omega = \Theta^M M^T \Theta E[U^2ZZ^T] \Theta M \Theta^M.$$ 

In the high-dimensional case, we replace the matrices $\Theta^M, M,$ and $\Theta$ by their regularized empirical counterparts to obtain the following estimator of $\Omega$:

$$\hat{\Omega} = \hat{\Theta}^M \hat{M}^T \hat{\Theta} Z^T \text{diag}(\hat{U})^2 \hat{Z} \hat{\Theta} \hat{M} \hat{\Theta}^M,$$

where $\hat{U} = (Y_1 - X_1^T \hat{\beta}, \ldots, Y_n - X_n^T \hat{\beta})$ and $\hat{\beta}$ is the Lasso estimator given in (2.5). We define the set $A = \{ a \in \mathbb{R}^p : a \in \ell_2 \text{ and } \|a\|_1 \leq C \sqrt{\omega_2} \|a\|_2 \}$. We require the following assumption on the covariance matrix $\Omega$.

**Assumption 5.** There exists a constant $\varpi > 0$ such that $\sqrt{a^T \Omega a}/\omega_2 \geq \varpi \|a\|_2$ for all $a \in A$.

Assumption 5 can be easily verified under mild regularity assumptions and given the lower bound $\sqrt{E[U^2|Z]} \geq \varpi$, which is a common condition to derive asymptotic distribution results. This lower bound implies $a^T \Omega a \geq \varpi^2 a^T \Theta^M a$. Assumption 5 holds, for instance, if the eigenvalues of $\Theta^M$ have a polynomial or exponential decay. The next result establishes asymptotic normality of linear combinations of the components of $\hat{\beta}$.

**Theorem 3.8.** Let Assumption 5 and the conditions of Corollary 3.6 be satisfied. Further, assume that $\max(\|X^T\|_2^2, \|ZZ^T\|_\infty^2) = O(1)$. Then, for all $a \in \mathbb{R}^p$ satisfying

$$\omega_2 s_{\text{max}}^M \sqrt{\log(q)} + \sqrt{s_{\text{max}}^M} \max \left( \sqrt{s_{\text{max}}^M}, \sqrt{s_{\text{max}}^2} \right) \|\Theta^M\|_1 / \omega_2 = o\left( \sqrt{n} \|a\|_2 / (\log(q) \|a\|_1) \right),$$

$$\omega_2 s_{\text{max}}^M \sqrt{\log(q)} + \sqrt{s_{\text{max}}^M} \max \left( \sqrt{s_{\text{max}}^M}, \sqrt{s_{\text{max}}^2} \right) \|\Theta^M\|_1 / \omega_2 = o\left( \sqrt{n} \|a\|_2 / (\log(q) \|a\|_1) \right).$$
we have
\[ \sqrt{n/(a^T\hat{\Omega}a)} a^T (\hat{\beta} - \beta^0) \xrightarrow{d} \mathcal{N}(0,1). \]

In the following we discuss several 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 \( \hat{\beta} \). In both cases, it holds \( \|a\|_1/\|a\|_2 \leq \text{const.} \) and the required rate restriction (3.8) simplifies accordingly. But even if the dimension of the low-dimensional subvector of interest increases, the condition \( \|a\|_1/\|a\|_2 \leq \text{const.} \) can be justified as the following example illustrates.

**Example 3.2** (Series Approximation (cont’d)). For most realizations \( x_{\text{end}} \) of \( X_{\text{end}} \) we may assume \( \|\phi^K(x_{\text{end}})\|_2 \sim \sqrt{K} \). Consequently, we may assume that \( \|\phi^K(x_{\text{end}})\|_1/\|\phi^K(x_{\text{end}})\|_2 \) is bounded from above uniformly in \( n \) and the rate restriction in Theorem 3.8 is satisfied under sufficient sparsity restrictions. The corresponding sieve variance \( \phi^K(x_{\text{end}})^T\Omega\phi^K(x_{\text{end}}) \) increases relative to the associated parameter \( \omega_2 \) which is thus related to Chen and Pouzo [2013] or Chen and Christensen [2017].

In the following, we discuss simple implications of the asymptotic normality result. The next theorem establishes asymptotically valid confidence intervals and testing procedures for inner products of \( \beta^0 \). The next results are direct implications of Theorem 3.8 and hence, their proofs are omitted. In the following, \( \Phi \) 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 (3.8) we have that for any \( \alpha \in (0,1) \)
\[
\mathbb{P} \left( a^T\beta^0 \in \left[ a^T\hat{\beta} \pm \Phi^{-1}(1-\alpha/2) (a^T\hat{\Omega}a)^{1/2}/\sqrt{n} \right] \right) = 1 - \alpha + o(1).
\]

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 \( \beta^0_j \) at nominal level \( \alpha \) is given by
\[
\left[ \hat{\beta}_j - \Phi^{-1}(1-\alpha/2) \hat{\Omega}_{jj}^{1/2}/\sqrt{n}, \quad \hat{\beta}_j + \Phi^{-1}(1-\alpha/2) \hat{\Omega}_{jj}^{1/2}/\sqrt{n} \right].
\]
The length of the confidence interval is given by
\[
2\Phi^{-1}(1-\alpha/2) \hat{\Omega}_{jj}^{1/2}/\sqrt{n}.
\]

We thus see that the length of the confidence interval increases relative to the ratio \( \sqrt{\omega_2/n} \). 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 and Tsybakov [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 (3.8)) 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 \( \alpha \in (0,1) \)

\[
\mathbb{P} \left( \frac{\sqrt{n} |a^T(\beta_0 - \beta^H)|}{\sqrt{a^T \Omega a}} \geq \Phi^{-1}(1-\alpha/2) \right) = \alpha + o(1).
\]

### 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 in the parameters of interest. Throughout this section, we consider a sample size of \( n = 100 \) and 1000 Monte Carlo replications.

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

\[
Y = \beta_1 X_1 + \beta_{1,j} X_{-1} + U, \quad X = (X_1, X_{-1})^T, \quad \beta^0 = (\beta_1, \beta_{1,j})^T, \\
X_1 = \alpha_1 Z_1 + \alpha_{1,j} X_{-1} + \sqrt{2 - \alpha_1^2} V, \quad Z = (Z_1, X_{-1})^T, \quad \alpha^0 = (\alpha_1, \alpha_{-1})
\]

with

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

where \( \Sigma = ((0.5)^{|j-k|})_{jk} \) is a \( q \times q \) matrix. The parameter \( \rho \) captures the degree of endogeneity and is varied in the experiments below. We take \( p = 200 \) with one endogenous variable and \( q = 200 \) instruments (included and excluded covariates). The number of observations is varied in the simulations below. The parameters are set in the following way: \( \beta_1 = 2, \beta_{1,j} = 1 + (j-1) * c \) for \( 1 \leq j \leq 40 \), where \( c \) is a constant such that the parameters \( \beta_{1,j} \) are equispaced between 1 and 3, \( \beta_{1,j} = 0 \) for \( 41 \leq j \leq (p - 1) \), and \( \alpha_{1,j} = 4j^{-3} \) for \( 1 \leq j \leq (q - 1) \). The parameter \( \alpha_1 \) 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 \( \sqrt{2 - \alpha_1^2} \), to ensure that the variance of \( X_1 \) does not depend on the value \( \alpha_1 \).

The desparsified IV Lasso estimator \( \hat{\beta} \) is described as in Subsection 2.1. The estimator is based on the initial IV Lasso \( \tilde{\beta} \) given in (2.4) where the tuning parameter \( \lambda \) is chosen via 10-fold cross-validation. The procedure also relies on regularized estimators \( \hat{\Theta}, \hat{M}, \) and \( \hat{\Theta}^M \) which are implemented as described in Subsection 2.3. Regarding the regularized inverses \( \hat{\Theta} \) and \( \hat{\Theta}^M \) we choose the same tuning parameter \( \lambda_j = \lambda_1 \) and \( \lambda_j^M = \lambda_1^M \), respectively, by 10-fold cross-validation among all nodewise regressions. Since \( \mathbb{E}[U^2 | Z] = 1 \) we are in the homoscedastic case where the covariance matrix simplifies to \( \Omega = \mathbb{E}[U^2] \Theta^M \). To estimate the covariance matrix \( \Omega \), we replace the variance of \( U \) by the error variance estimator of the scaled Lasso proposed by Sun and Zhang [2012], which is adjusted to our instrumental variable case, and estimate \( \Theta^M \) using our regularized estimator \( \hat{\Theta}^M \). Given the covariance matrix estimator \( \hat{\Omega} \) we compute confidence intervals for the structural parameter \( \beta_1 \) by following Example 3.3.
In Table 1 we report the absolute values of the mean bias for the desparsified IV estimator $\hat{\beta}_1$ and for the IV Lasso estimator $\tilde{\beta}_1$, for different values of the parameters $\rho$ and $\alpha_1$. The absolute mean is computed over the 1000 Monte Carlo replications. We also report the coverage of our confidence interval for $\beta_1$ at the nominal level 5%.

| Value of $\rho$ | Value of $\alpha_1$ | Absolute mean bias($\hat{\beta}_1$) | Absolute mean bias($\tilde{\beta}_1$) | Coverage for $\beta_1$ |
|----------------|--------------------|-------------------------------|-------------------------------|---------------------|
| 0.7            | 1                  | 0.635                         | 1.574                         | 0.952              |
|                | 0.75               | 0.682                         | 1.636                         | 0.968              |
|                | 0.5                | 0.759                         | 1.652                         | 0.973              |
|                | 0.25               | 0.774                         | 1.709                         | 0.978              |
| 0.5            | 1                  | 0.631                         | 1.574                         | 0.953              |
|                | 0.75               | 0.640                         | 1.632                         | 0.967              |
|                | 0.5                | 0.729                         | 1.656                         | 0.968              |
|                | 0.25               | 0.751                         | 1.714                         | 0.978              |
| 0.3            | 1                  | 0.608                         | 1.585                         | 0.959              |
|                | 0.75               | 0.622                         | 1.626                         | 0.970              |
|                | 0.5                | 0.703                         | 1.667                         | 0.965              |
|                | 0.25               | 0.710                         | 1.736                         | 0.979              |

Table 1: Absolute mean of the bias for the desparsified IV estimator $\hat{\beta}_1$ and the initial IV Lasso estimator $\tilde{\beta}_1$. The right column provides coverages of our confidence interval for $\beta_1$ at the nominal level 5%.

From Table 1 we see that the absolute mean bias of the desparsified IV Lasso estimator $\hat{\beta}_1$ is considerably smaller than the absolute mean bias of the IV Lasso estimator $\tilde{\beta}_1$ for each parameter value of $\rho$ and $\alpha_1$. As $\alpha_1$ decreases, i.e., the strength of instruments declines, we see that the values of the absolute mean bias of $\hat{\beta}_1$ and $\tilde{\beta}_1$ become larger. On the other hand, when $\rho$ increases, i.e., the degree of endogeneity becomes more severe, the absolute mean bias of $\hat{\beta}_1$ becomes somewhat larger while the bias for $\tilde{\beta}_1$ declines. This is due to underestimation of the IV Lasso estimator $\hat{\beta}_1$, which is more severe for small values of $\rho$ while the desparsification term in the estimator $\tilde{\beta}_1$ is not very sensitive to $\rho$. From the last column of Table 1 we see that the coverage, reported at the 5% nominal level, is accurate, in particular, when $\alpha_1$ is close to 1. We also see that the coverage increases as $\alpha_1$ becomes smaller, which is not surprising as our theoretical results indicate that the confidence intervals become larger as the instruments become weaker.

Figure 1 shows the normal Q–Q plots of the desparsified IV Lasso estimator $\tilde{\beta}_1$ when $\rho = 0.5$ for varying values of $\alpha_1$. From this figure we see that the distribution has somewhat heavier tails than the standard normal but otherwise is close to the standard normal even in the case where the instruments have only a weak influence.
Let Assumption 1 hold. By using the Cauchy-Schwartz 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

$$
\|M\|_1 \leq \sqrt{s_M} \max_{1 \leq j \leq p} \|M_j\|_2 \\
= O\left(\sqrt{s_M} \max_{1 \leq j \leq p} \|\Theta^{1/2} M e_j\|_2\right) \\
= O\left(\sqrt{s_M}\right)
$$

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

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

For the next proofs, we require the following notation. For $j = 1, \ldots, p$, recall the definition $\gamma_j = \arg\min_{\gamma \in \mathbb{R}^{p-1}} \|\Theta^{1/2} M_j - (\Theta^{1/2} M_{-j})\gamma\|_2^2$. 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 \(\gamma_{kj}\) is the \(k\)-th entry of \(\gamma_j\). 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 \(\gamma_j\) yields

\[(\Theta^{1/2}M)^T_{-j} \Theta^{1/2}M \Gamma_j = 0,
\]
and thus,

\[M^T \Theta M \Gamma_j = ((\Theta^{1/2}M)^T_{-j} \Theta^{1/2}M)_{-j} e_j = \Gamma_j^T M^T \Theta M \Gamma_j e_j = \tau_j^2 e_j,
\]
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 \(\gamma_j\) to get the second equality. Further, by premultiplying with \(\Theta^M\) 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 \(\omega_2\) we obtain the following lower bound for \(\tau_j\):

\[\tau_j^2 = 1/\Theta_{jj}^M \geq 1/\lambda_{\text{max}}(\Theta^M) = \lambda_{\text{min}}(M^T \Theta M) = \omega_2^{-1},\]  \hspace{1cm} (A.1)

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

\[\|\gamma_j\|_1^2 \leq C\left(s_j^M \|\gamma_j\|_2^2 + (\log(q))^2/n\right) \leq C\left(s_j^M + (\lambda_j^M)^2\right).
\]

where we have used the upper bound \(\|\gamma_j\|_1 \leq \|\gamma_j S_j\|_1 + \|\gamma_j S_j\|_1\), the Cauchy-Schwarz inequality and (2.7) to get the first inequality. Below we also use for matrices \(A\) and \(B\) the inequalities

\[\|AB\|_\infty \leq \|A\|_\infty \|B\|_1 \quad \text{and} \quad \|AB\|_\infty \leq \|B\|_\infty \|A^T\|_1.
\]

### A.1. Proofs of the Main Results

**Proof of Lemma 3.1.** By using the inequality \(\|\Sigma^{-1/2}v\|_2 \leq \|v\|_2/\sqrt{\lambda_{\text{min}}(\Sigma)}\) for all \(v \in \mathbb{R}^q\) and the fact that \(\Sigma\) has eigenvalues uniformly bounded away from zero by Assumption 1 (iii) we have that sub-Gaussianity of \(Z\) implies sub-Gaussianity of \(\tilde{Z} = \Sigma^{-1/2}Z\). Lemma 5.2 (and the proof of Theorem 2.4) in van de Geer et al. [2014] applied to the reduced form model (2.2), together with sub-Gaussianity of \(\Sigma^{-1/2}Z\) and Assumption 1 (iii), implies

\[\|(\Sigma^{-1/2}M\beta)_{S_0}\|_1^2 \leq C s_0^\beta T \Sigma M^T \Sigma M^{-1} \Sigma^{-1} M \beta \]  \hspace{1cm} (A.2)

wpal, for all \(\beta\) satisfying \(\|(\Sigma^{-1/2}M\beta)_{S_0}\|_1 \leq 3\|(\Sigma^{-1/2}M\beta)_{S_0}\|_1\). In addition, from Assumption 2 and the definition of \(\omega_1\) we infer

\[B := \left\{\|\beta_{S_0}\|_1 \leq 3\|\beta_{S_0}\|_1\right\} \subset \left\{\|(\Sigma^{-1/2}M\beta)_{S_0}\|_1 \leq 3\eta^{-1/2}\|(\Sigma^{-1/2}M\beta)_{S_0}\|_1\right\}.
\]

Consequently, (A.2) holds on \(B\) and then by the definition of \(\omega_1\), inequality (3.1) follows for all \(\beta \in B\). \(\square\)
Proof of Theorem 3.5. The proof is based on the decomposition
\[ \Delta = \sqrt{n}(\hat{\Theta}^T \hat{M}^T \hat{\Theta} \hat{M} - I_p)(\bar{\beta} - \beta^0) - \sqrt{n} \hat{\Theta}^T \hat{M}^T \hat{\Theta} (\hat{M} - \bar{M})(\bar{\beta} - \beta^0). \]
We observe
\[ \|\Delta\|_{\infty}/\sqrt{n} \leq \|\hat{\Theta}^T \hat{M}^T \hat{\Theta} \hat{M} - I_p\|_\infty \|\bar{\beta} - \beta^0\|_\infty + \|\hat{\Theta}^T \hat{M}^T \hat{\Theta} (\hat{M} - \bar{M})\|_\infty \|\bar{\beta} - \beta^0\|_\infty \]
\[ \leq \|\hat{\Theta}^T \hat{M}^T \hat{\Theta} - I_p\|_\infty \|\bar{\beta} - \beta^0\|_1 + \|\hat{\Theta}^T \hat{M}^T \hat{\Theta}\|_{\infty,\infty} \|\hat{\Theta}^T (\hat{M} - \bar{M})\|_\infty \|\bar{\beta} - \beta^0\|_1. \]
Further, the upper bound given in (3.5) implies that
\[ \|\Delta\|_{\infty} \leq \sqrt{n} \max_{1 \leq j \leq p} \{ \lambda_j^M / \bar{\tau}_j^2 \} \|\bar{\beta} - \beta^0\|_1 + \sqrt{n} \|\hat{\Theta}^T \hat{M} (\hat{M})^T\|_1 \|\hat{\Theta}^T (\hat{M} - \bar{M})\|_\infty \|\bar{\beta} - \beta^0\|_1. \]
By the definition of the regularized estimator \( \hat{M} \) given in (2.11) it holds for all \( j, k \):
\[ |\hat{M}_{jk} - \hat{M}_{jk}| = |\hat{M}_{jk}| 1 \{ |\hat{M}_{jk}| < C_0 \sqrt{\log(q)/n} \}
\[ < C_0 \sqrt{\log(q)/n}, \]
which implies
\[ \|\hat{M} - \bar{M}\|_\infty < C_0 \sqrt{\log(q)/n}. \] (A.3)
Thus, using that \( \lambda_j^M \sim \log(q)/\sqrt{n} \) uniformly in \( j \), by Assumption 3 (i) we obtain
\[ \|\Delta\|_{\infty} \leq C \log(q) \left( \max_{1 \leq j \leq p} \bar{\tau}_j^{-2} + \|\hat{\Theta}^T \hat{M} (\hat{M})^T\|_1 \|\bar{\beta} - \beta^0\|_1 \right). \]
In the following, we consider the events
\[ \mathcal{C} := \left\{ \|\beta_{\hat{S}_0}\|^2 \leq \omega_1 s_0 \beta^T \hat{\Theta} \hat{M} \beta/c^2 \text{ for all } \|\beta_{\hat{S}_0}\|_1 \leq 3 \|\beta_{\hat{S}_0}\|_1 \right\} \]
and \( \mathcal{T} := \left\{ \|\hat{M}^T \hat{\Theta} Z^T U/n + \hat{\Theta}^T (\hat{M} - \bar{M}) \beta^0\|_{\infty} \leq C \lambda \right\} \) for some sufficiently large constant \( C > 0 \) and recall \( \lambda \sim \log(q)/\sqrt{n} \).
On the event \( \mathcal{C} \cap \mathcal{T} \) we have
\[ \|\hat{\beta} - \beta^0\|_1 \leq C \omega_1 s_0 \log(q)/\sqrt{n}, \]
which follows directly from van de Geer [2016, Theorem 2.2].\(^2\) From the proof of Proposition 3.4 in Appendix A.2 we also have that \( \bar{\tau}_j^2 \) is a consistent estimator of \( \tau_j^2 \). Further, Propositions 3.3 and 3.4 together with the lower bound (A.1) yield
\[ \|\Delta\|_1 \mathbb{1}_{\mathcal{C} \cap \mathcal{T}} = O_p(\omega_1 s_0 \log(q)/\sqrt{n} \max (\omega_2, \|\Theta M \Theta^T\|_1)). \]
It is thus sufficient to show \( \mathbb{1}_{\mathcal{C} \cap \mathcal{T}} = 1 \) w.p.1. We proceed in two steps and control the sets \( \mathcal{T} \) and \( \mathcal{C} \) separately. To handle the set \( \mathcal{T} \) note that
\[ \|\hat{M}^T \hat{\Theta} Z^T U/n + \hat{\Theta}^T (\hat{M} - \bar{M}) \beta^0\|_{\infty} \]
\[ \leq \| U^T Z \Theta M/n \|_{\infty} + \| (\hat{\Theta}^T - \hat{\Theta}^T \hat{\Theta} (\hat{M} - \bar{M}) \beta^0\|_{\infty} \]
\[ + \| \hat{\Theta}^T (\hat{M} - \bar{M}) \beta^0\|_{\infty} . \]
\(^2\)Apply van de Geer [2016, Theorem 2.2] with, in their notation, \( L = 3, \hat{\sigma}^2(3, \hat{\delta}_0) = c^2/\omega_1, X = \hat{\Theta}^{1/2} \hat{M}, Y = \hat{\Theta}^{1/2} Z^T Y, \epsilon = \hat{\Theta}^{1/2} Z^T Y - \hat{\Theta}^{1/2} \hat{M} \beta^0. \)
To bound $I$, we make use of Nemirovski’s inequality (see, for instance, p. 509 in Bühlmann and Van De Geer [2011]) and $E[U^2|Z] \leq \sigma^2$ to get

$$
\mathbb{E}\left( \max_{1 \leq j \leq p} \| (U^T Z M)_{j} \| / n \right)^{2} \leq 8 \log(2p) \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{E} \max_{1 \leq j \leq p} |U_i(Z^T \Theta M)_{j}|^2
$$

$$
\leq 8 \log(2p) n^{-1} \sigma^2 \mathbb{E} \max_{1 \leq j \leq p} |(Z^T \Theta M)_{j}|^2
$$

and hence, we obtain $I = O_p(\sqrt{E\|Z^T \Theta M\|^2 \log(p)/n}) = O_p(\log(p)/\sqrt{n})$ by using Assumption 3 (ii). Under Assumption 3 (i) we have that $\lambda \sim \log(q)/\sqrt{n}$ and thus, $I = O_p(\lambda)$.

Next, we consider $II$. We have

$$
II = \| \hat{M}^T \Theta - M^T \Theta \|_{op, \infty} \| U^T Z / n + (\hat{M} - \hat{M}) \beta_0 \| \infty
$$

$$
\leq (\| \hat{\Theta} \|_{op, \infty} \| \hat{M} - M \|_1 + \| M \|_1 \| \hat{\Theta} - \Theta \|_{op, \infty}) (\| U^T Z / n \| \infty + \| \hat{M} - \hat{M} \|_\infty \| \beta_0 \|_1).
$$

Again, due to Nemirovski’s inequality, we have $\| U^T Z / n \| \infty = O_p(\sqrt{\log(q)/n})$ under Assumption 1 (iii) and condition $E[U^2|Z] \leq \sigma^2$ imposed in Assumption 3 (ii). Furthermore, $\| \hat{M} - M \|_1 = O(\sqrt{\log(p)/n})$ by inequality (A.3). We also have $\| \hat{M} - \hat{M} \|_1 = O_p(s_M \sqrt{\log(q)/n})$ and $\| \hat{\Theta} - \Theta \|_{op, \infty} = O_p(s_{\max} \sqrt{\log(q)/n})$ from Propositions 3.2 and 3.4. Now using that $\| \beta_0 \|_1 = O(\sqrt{\omega_1 s_0})$ (since $\| \beta_0 \|_1 \leq \omega_1 \| \beta_s_0 \|_1 + C \omega_1 s_0 \sqrt{\log(p)/n} \leq \sqrt{\omega_1 s_0} \| \beta^* \|_1 + o(1)$ by using the fact that $\beta_0 \in B$, (2.6) and (3.2), and $\| \beta^* \|_1 = O(1))$ we obtain

$$
II = O_p((1 + \| \beta_0 \|_1) (s_M \| \Theta \|_1 + s_{\max} \| M \|_1) \log(q)/n)
$$

$$
= O_p(\sqrt{\omega_1 s_0} \max(s_M \sqrt{s_{\max}}, s_{\max} \sqrt{s_M}) \log(q)/n)
$$

$$
= O_p(\sqrt{\log(p)/n})
$$

employing (3.2) in Assumption 3 to get the last equality. Remark that to get the first equality we have used the fact that $\| \hat{\Theta} \|_{op, \infty} \leq \| \hat{\Theta} - \Theta \|_{op, \infty} + \| \Theta \|_1$ because $\Theta$ is symmetric, and by Proposition 3.4 $\| \hat{\Theta} - \Theta \|_{op, \infty} = O_p(s_{\max} \sqrt{\log(q)/n})$ which is negligible with respect to the other terms under (3.2). Consider $III$. We have

$$
III \leq \max_j |(\Theta M)_{j}^T (\hat{M} - M) \beta_0 | + \max_j |(\Theta M)_{j}^T (\hat{M} - M) \beta_0 |,
$$

(A.4)

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

$$
\mathbb{E} \| M^T \Theta (\hat{M} - M) \beta_0 \|^2 = \mathbb{E} \max_{1 \leq j \leq p} |n^{-1} \sum_{i} (\Theta M)_{j}^T Z_i X_i \beta_0^0 - (\Theta M)_{j}^T M \beta_0^0 |^2
$$

$$
\leq 8 \log(2p) n^{-1} \mathbb{E} \max_{1 \leq j \leq p} |(\Theta M)_{j}^T Z X^T \beta_0 |^2
$$

$$
\leq 8 \log(2p) n^{-1} \mathbb{E} [(X^T \beta_0)^2 \| M^T \Theta Z \|^2 \infty]
$$

$$
= O(\log(p)/n),
$$

where we have used Assumption 3 (ii) to get the last line. For the first summand on the right hand side of (A.4) we observe

$$
\max_j |(\Theta M)_{j}^T (\hat{M} - M) \beta_0 | \leq \|\Theta M\|_{\infty} \|\hat{M} - M\|_1 \|\beta_0\|_1
$$

$$
= O_p(\sqrt{s_{\max} s_M \sqrt{\log(q)/n} \sqrt{\omega_1 s_0})
$$

$$
= O_p(\log(q)/\sqrt{n})
$$

(A.5)
due to Assumption 1 (iii) which implies \( \| \Theta \|_1 = O(\sqrt{s_{\max}}) \), Assumption 3 (ii), the second result of Proposition 3.2 and the first rate restriction imposed in Assumption 3 (iii).

It remains to control \( C \). By Lemma 3.1 it holds for all \( \| \beta_{S_0} \|_1 \leq 3 \| \beta_{S_0} \|_1 \) that

\[
\| \beta_{S_0} \|_1^2 \leq \omega_1 s_0 \beta^T M^T \Sigma^{-1} \tilde{\Sigma} \Sigma^{-1} M \beta/\bar{c}^2
\]

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

\[
\omega_1 s_0 \| M^T \Theta (\tilde{\Sigma} - \Sigma) \Theta M \|_\infty \leq c^*/2 \quad \text{wpa1}
\]

and

\[
\omega_1 s_0 \| M^T \Theta - \hat{M}^T \hat{\Theta} \hat{M} \|_\infty \leq c^*/2 \quad \text{wpa1.}
\]

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

\[
\omega_1 s_0 \| \Theta M \|_1^2 \sqrt{\frac{\log(q)}{n}} \leq c^{**}
\]

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 3 (iii).

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

\[
\omega_1 s_0 \| \hat{M} \|_1^2 \| \hat{\Theta} - \Theta \|_\infty \leq 2 \omega_1 s_0 \left( \| M \|_1^2 + \| \hat{M} - M \|_1^2 \right) \| \hat{\Theta} - \Theta \|_\infty
\]

\[
\leq C \omega_1 s_0 s_M^2 \left( 1 + \log(q)/n \right) \sqrt{s_{\max} \log(q)/n},
\]

wpa1, where we have used Assumption 3 (ii) to get \( \| M \|_1 \leq s_M \| M \|_\infty = O(s_M) \), the second result of Proposition 3.2 and the result \( \| \Theta - \hat{\Theta} \|_\infty = O_P(\sqrt{s_{\max} \log(q)/n}) \) (see van de Geer et al. [2014]). Moreover,

\[
\| M^T \Theta M - \hat{M}^T \hat{\Theta} \hat{M} \|_\infty \leq \| M - \hat{M} \|_1 \| \Theta \|_1 \| M \|_\infty + \left( \| M \|_\infty + \| \hat{M} - M \|_1 \right) \| \Theta \|_1 \| \hat{M} - M \|_1 \leq C s_M \sqrt{\log(q)/n} \sqrt{s_{\max}} \left( 1 + s_M \sqrt{\log(q)/n} \right)
\]

wpa1, where we have used Assumptions 1 (iii) and 3 (ii) and the second result of Proposition 3.2. Consequently, by the rate restriction \( \omega_1 s_0 s_M \sqrt{s_{\max}} = o(\sqrt{n/\log(q)}) \) in Assumption 3 (iii), result (A.7) holds wpa1.

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

\[
\sqrt{n/(a^T \Omega a)} a^T (\hat{\beta} - \beta^0) = a^T M^T \Theta M Z^T U / \sqrt{n(a^T \Omega a)} + a^T (\Theta M^T \hat{\Theta} - \theta M^T \Theta) Z^T U / \sqrt{n(a^T \Omega a)} + o_P(\|a\|_1/\sqrt{a^T \Omega a}).
\]
Since $\sqrt{a^T \Omega a} \geq \sigma \sqrt{\sum_{i=1}^n a_i^2}$ it holds $\|a\|_1 / \sqrt{a^T \Omega a} = O(1)$ for all $a \in \mathcal{A}$. We have that $I \Rightarrow \mathcal{N}(0, 1)$ and moreover, $II = o_p(1)$ which can be seen as follows. We observe

$$II \leq \sqrt{\omega_2 / (a^T \Omega a)} \|a\|_1 \left(\|\hat{\Theta}^M - \Theta^M\|_{op, \infty} \|M^T \Theta ZU\|_\infty / (\omega_2 \sqrt{n}) + \|M - M\|_1 \|\Theta^M\|_1 \|\hat{\Theta}\|_{op, \infty} \|Z^T U\|_\infty / (\omega_2 \sqrt{n}) + \|\hat{\Theta} - \Theta\|_{op, \infty} \|M^T \Theta^M\|_1 \|Z^T U\|_\infty / (\omega_2 \sqrt{n})\right).$$

Using Nemirovski’s inequality as in proof of Theorem 3.5 we have $\|M^T \Theta ZU\|_\infty / \sqrt{n} = O_p(\sqrt{\log(q)})$ and $\|Z^T U\|_\infty / \sqrt{n} = O_p(\sqrt{\log(q)})$. Further, from $\sqrt{\omega_2 / (a^T \Omega a)} \leq \sigma^{-1} \|a\|_2^{-1}$ we infer

$$II = O_p\left(\frac{\log(q)}{\sqrt{n}} \|a\|_2 \left(\omega_2 s^M_M \sqrt{\log(q)} + \max\left(s_M \sqrt{s^M_{max}}, s_{max} \sqrt{s^M_M}\right) \|\Theta^M\|_1 / \omega_2\right)$$

using $\|\Theta^M\|_1 \leq \sqrt{s^M_M s_{max}}$. The rate requirement imposed on $q$ implies the result.

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

$$\left|\frac{a^T \hat{\Theta} a}{a^T \Omega a} - 1\right| \leq (a^T \Omega a)^{-1} \|a\|_2^2 \|\hat{\Theta} - \Theta\|_\infty$$

$$\leq \left\|\hat{\Theta}^M M^T \hat{\Theta} \|_1^2 \|n^{-1} Z^T \text{diag}(\hat{U})^2 Z - E[U^2 \Sigma ZZ]^T\|_\infty$$

$$= A_1$$

$$+ \left\|\hat{\Theta}^M M^T \hat{\Theta} - \Theta^M M^T \Theta\|_1 \|\Theta^M M^T \Theta\|_1 \|E[U^2 \Sigma ZZ]^T\|_\infty$$

$$= A_2$$

$$+ \left\|\hat{\Theta}^M M^T \hat{\Theta} - \Theta^M M^T \Theta\|_1^2 \|E[U^2 \Sigma ZZ]^T\|_\infty.$$

Using again Nemirovski’s inequality and $E[U^2 | Z] \leq \sigma^2$ we obtain

$$\|n^{-1} Z^T \text{diag}(\hat{U})^2 Z - E[U^2 \Sigma ZZ]^T\|_\infty$$

$$= \left\|n^{-1} \sum_i \left(U_i + X_i^T (\beta^0 - \tilde{\beta})\right)^2 Z_i Z_i^T - E[U^2 \Sigma ZZ]^T\right\|_\infty$$

$$\leq \left\|n^{-1} \sum_i U_i Z_i Z_i^T - E[U^2 \Sigma ZZ]^T\right\|_\infty + 2 \left\|(\beta^0 - \tilde{\beta})^T n^{-1} \sum_i U_i X_i Z_i Z_i^T\right\|_\infty$$

$$+ \left\|n^{-1} \sum_i (X_i^T (\beta^0 - \tilde{\beta})^2 Z_i Z_i^T\right\|_\infty$$

$$\leq O_p\left(\sqrt{\log(q) / n}\right) + \|\beta_0 - \tilde{\beta}\|_1 \times O_p(E\|X\|_\infty^2 \max_{1 \leq j \leq q} |Z_j Z_i|^2)$$

$$+ \|\beta_0 - \tilde{\beta}\|_2^2 \times O_p(E \max_{1 \leq j \leq p} |X_j X_i|^2 \max_{1 \leq j \leq q} |Z_j Z_i|^2).$$

Now using $\|\beta - \beta^0\|_1 = O_p(\omega_1 s_0 \sqrt{\log(p) / n})$ 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 $A_3 = o_p(1)$. \qed
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 \( \hat{r}_j \) implies

\[ \hat{r}_j^2 = Z_j^T (Z_j - Z_{-j} \hat{r}_j)/n. \]

Consequently, it holds \( Z_j^T Z \hat{r}_j/n = Z_j^T (Z_j - Z_{-j} \hat{r}_j)/(n \hat{r}_j^2) = 1. \) The KKT conditions also imply \( \| Z_j^T Z \hat{r}_j \|_\infty/n \leq \lambda_j / \hat{r}_j^2 \) or

\[ \| \Sigma \hat{r}_j - e_j \|_\infty \leq \lambda_j / \hat{r}_j^2, \]

where \( e_j \) is the \( j \)-th unit column vector.

Proof of (3.5). The KKT conditions for the nodewise Lasso (2.12) implies

\[ \hat{r}_j^2 = (\hat{\Theta}^{1/2} \hat{M})_j - (\hat{\Theta}^{1/2} \hat{M})_{-j} \hat{r}_j^2 + \lambda_j^M \| \hat{r}_j \|_1 \]

where \( \hat{\tau}_j \) is the \( n \times 1 \) vector.

\[ (\hat{\Theta}^{1/2} \hat{M})_j - (\hat{\Theta}^{1/2} \hat{M})_{-j} \hat{r}_j^2 = (\hat{\Theta}^{1/2} \hat{M})_j - (\hat{\Theta}^{1/2} \hat{M})_{-j} \hat{r}_j^2 + \lambda_j^M \| \hat{r}_j \|_1 \]

Consequently, for all \( 1 \leq j \leq p \):

\[ (\hat{\Theta}^{1/2} \hat{M})_j - (\hat{\Theta}^{1/2} \hat{M})_{-j} \hat{r}_j^2 = 1. \]

By the definition of \( \hat{\Theta}^M \) we also obtain

\[ \| (\hat{\Theta}^{1/2} \hat{M}^T \hat{M} \hat{M}^T) \|_\infty = \| (\hat{\Theta}^{1/2} \hat{M}^T \hat{M} \hat{M}^T)_{-j} \|_\infty / \hat{r}_j^2 \]

\[ \leq \lambda_j^M / \hat{r}_j^2, \]

where the last inequality again follows by the KKT conditions for the nodewise Lasso (2.12).

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

\[ \| \hat{\Theta}^M - \Theta^M \|_{op, \infty} = \max_j \| \hat{\Theta}^M - \Theta^M \|_1 \]

\[ = \max_j \| \hat{\Theta}^M - \Theta^M \|_1 \]

\[ \leq \max_j \| \hat{\Theta}^M - \Theta^M \|_1 \]

\[ \leq C \left( \max_j \| \hat{\Theta}^M - \Theta^M \|_1 \right) \max_j \frac{1}{\hat{r}_j^2}, \]

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

\[ C_j = \left\{ \| \gamma_j \|_1^2 \leq C s_j^M \| \hat{\Theta}^T \hat{M} \|_\infty \text{ for all } \| \gamma_j \|_1 \leq 3 \| \gamma_j \|_1 \right\} \]

and

\[ T_j = \left\{ \| (\hat{\Theta}^{1/2} \hat{M})_j - (\hat{\Theta}^{1/2} \hat{M})_{-j} \|_\infty \leq C \lambda_j^M \right\} \]
for some sufficiently large constant \(C > 0\). Recall \(\lambda_j^M \sim \log(q)/\sqrt{n}\). On the set \(C_j \cap \mathcal{T}_j\), it holds
\[
\|\tilde{\gamma}_j - \gamma_j\|_1 \leq C(j)s_j^M \log(q)/\sqrt{n},
\]
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
\[
|\tilde{\tau}_j^2 - \tau_j^2| = O_p\left(\log(q)s_j^M/n\right)
\]
which can be seen as follows. Recall from (A.8) that
\[
\tilde{\tau}_j^2 = \left(\left(\hat{\Theta}^{1/2}\hat{M}_j - (\hat{\Theta}^{1/2}\hat{M})_{-j}\tilde{\gamma}_j\right)^T \left(\hat{\Theta}^{1/2}\hat{M}\right)_j
\]
\[
= \left(\left(\hat{\Theta}^{1/2}\hat{M}_j - (\hat{\Theta}^{1/2}\hat{M})_{-j}\gamma_j\right)^T \left(\hat{\Theta}^{1/2}\hat{M}\right)_j + \left((\hat{\Theta}^{1/2}\hat{M})_{-j}(\gamma_j - \tilde{\gamma}_j)\right)^T \left(\hat{\Theta}^{1/2}\hat{M}\right)_j
\]
\[
= \left(\left(\hat{\Theta}^{1/2}\hat{M}_j - (\hat{\Theta}^{1/2}\hat{M})_{-j}\gamma_j\right)^T \left(\hat{\Theta}^{1/2}\hat{M}\right)_j + \left((\hat{\Theta}^{1/2}\hat{M})_{-j}(\gamma_j - \tilde{\gamma}_j)\right)^T \left(\hat{\Theta}^{1/2}\hat{M}\right)_j
\]
\[
= \Gamma_j^T \hat{M}^T \hat{\Theta} \hat{M} \Gamma_j + \left((\hat{\Theta}^{1/2}\hat{M})_{-j}\gamma_j\right)^T \left(\hat{\Theta}^{1/2}\hat{M}\right)_{-j}\gamma_j
\]
\[
+ \left((\hat{\Theta}^{1/2}\hat{M})_{-j}(\gamma_j - \tilde{\gamma}_j)\right)^T \left(\hat{\Theta}^{1/2}\hat{M}\right)_j
\]
and recall that \(\tau_j^2 = \Gamma_j^T M^T \Theta M \Gamma_j\). We have
\[
|\tilde{\tau}_j^2 - \tau_j^2| \leq \underbrace{I}_{T_1} + \underbrace{II}_{T_2} + \underbrace{III}_{T_3}
\]
where we bound each term on the right hand side as follows. Consider \(I\). We observe
\[
I \leq \left\{\Gamma_j^T (\hat{M} - M)^T \hat{\Theta} \hat{M} \Gamma_j + \left|\Gamma_j^T M^T (\hat{\Theta} - \Theta) \hat{M} \Gamma_j\right| + \left|\Gamma_j^T M^T \Theta (\hat{M} - M) \Gamma_j\right|\right\}
\]
In the following, we bound each summand on the right hand side separately. We have
\[
T_1 = |(\Theta \Gamma_j)^T (\hat{M} - M) \Gamma_j| = O_p\left(\log(q)/\sqrt{n}\right),
\]
uniformly in \(j\) by using Assumption 4, i.e., \(\mathbb{E} \max_j |(\Theta \Gamma_j)^T X \Gamma_j|^2 = O(\log(p))\), and following the arguments for the upper bound (A.5). Equivalently, we have \(T_3 = O_p\left(\log(q)/\sqrt{n}\right)\). We observe
\[
T_2 \leq |\Gamma_j^T M^T \Theta (\hat{\Sigma} - \Sigma) \hat{M} \Gamma_j| + |\Gamma_j^T M^T \Theta (\hat{\Sigma} - \Sigma) \hat{M} \Gamma_j|
\]
\[
\leq |\Gamma_j^T M^T \Theta (\hat{\Sigma} - \Sigma) \hat{M} \Gamma_j| + |\Gamma_j^T M^T \Theta (\hat{\Sigma} - \Sigma) \Theta M \Gamma_j| + o_p\left(\log(q)/n\right)
\]
Due to Assumption 4 \((ii)\), i.e., \(\mathbb{E} \max_{1 \leq j \leq p} \| (\Theta \Gamma_j)^T Z \|^2_I = O(\log(p)^2)\), it is sufficient to consider the first summand. The KKT condition for the nodewise Lasso estimator \(\tilde{\gamma}_j\)
Further, the KKT condition for the nodewise Lasso estimator \( \hat{\gamma}_j \) by using Assumption 4 (In the following, we bound each summand on the right hand side separately. We have

\[
\hat{\Sigma} \hat{\gamma}_j - e_j = \lambda_j \hat{\gamma}_j / \hat{\gamma}_j^2.
\]

Since \( \lambda_j \sim \sqrt{\log(q)/n} \) 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

\[
\left| \Gamma_j^T M^T \Theta (\hat{\Sigma}_j - I_q) M \Gamma_j \right| = \left| \Gamma_j^T M^T \Theta (\lambda_1 \hat{\gamma}_1 / \hat{\gamma}_1^2, \ldots, \lambda_q \hat{\gamma}_q / \hat{\gamma}_q^2) M \Gamma_j \right| \\
\leq \sqrt{\log(q)/n} \|\Theta M \Gamma_j\|_1 \|M \Gamma_j\|_1 \max_{1 \leq j \leq q} \hat{\gamma}_j^2 \\
= \sqrt{\log(q)/n} \sqrt{s_{\max} s_M (s_j^M + (\lambda_j^M)^2)} \times O_p(1) \\
= O_p \left( \log(q)/\sqrt{n} \right).
\]

using that \( \hat{\gamma}_j^2 \) is a consistent estimator of \( 1/\Theta_{jj} \) (see the proof of Theorem 2.4 of van de Geer et al. [2014]), \( \Theta_{jj} \) is bounded uniformly in \( j \), and the first rate condition imposed in Assumption 3 (iii). Further, we have on \( T_j \) that

\[
III = \left| \Gamma_j^T \left( \hat{\Theta}_{1/2} M \right) - \left( \hat{\Theta}_{1/2} M \right) - \gamma_j \right| \left( \hat{\Theta}_{1/2} M \right) - \gamma_j \right|_\infty \\
= O_p \left( \log(s_j^M/n) \right).
\]

Further, the KKT condition for the nodewise Lasso estimator \( \hat{\gamma}_j \) implies

\[
III = \left| \Gamma_j^T \left( \hat{\Theta}_{1/2} M \right) - \left( \hat{\Theta}_{1/2} M \right) - \gamma_j \right| \left( \hat{\Theta}_{1/2} M \right) - \gamma_j \right|_\infty \\
= O_p \left( \log(q)/\sqrt{n} \right).
\]

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

\[
\left\| \left( \hat{\Theta}_{1/2} M \right) - \left( \hat{\Theta}_{1/2} M \right) - \gamma_j \right\|_\infty \\
\leq \left\| \Gamma_j^T \left( M \hat{\Theta}_{1/2} M - M \Theta M \right) \right\|_\infty \\
= O_p \left( \log(s_j^M/n) \right).
\]

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

\[
S_1 = \max_i \|\Theta M\|_i \left( \hat{\Theta}_{1/2} - \hat{\Theta} \right) \Gamma_j + o_p \left( \log(s_j^M/n) \right) = O_p \left( \log(q)/\sqrt{n} \right)
\]

by using Assumption 4 (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

\[
S_2 \leq \left\| \Gamma_j^T M^T \Theta (\hat{\Sigma} - \Sigma) \Gamma_j \right\|_\infty + \left\| \Gamma_j^T M^T \Theta (\hat{\Sigma} - \Sigma) \hat{\Theta} M \right\|_\infty \\
\leq \left\| \Gamma_j^T M^T \Theta (\hat{\Sigma} - \Sigma) \Gamma_j \right\|_\infty + \left\| \Gamma_j^T M^T \Theta (\hat{\Sigma} - \Sigma) \hat{\Theta} M \right\|_\infty + o_p \left( \log(q)/n \right)
\]

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

\[
\mathbb{E} \|\Gamma_j^T M^T \Theta (\hat{\Sigma} - \Sigma) \hat{\Theta} M \|_\infty^2 = \mathbb{E} \max_{1 \leq i \leq p} \left| n^{-1} \sum_{i=1}^n \Gamma_j^T (M^T \Theta Z_i Z_i^T (\hat{\Theta} M) - \Gamma_j^T M^T (\Theta M)) \right|^2 \\
\leq 8 \log(2p)n^{-1} \max_{1 \leq i \leq p} \|\Gamma_j^T M^T \Theta Z Z^T (\Theta M)\|_\infty^2.
\]
The KKT condition for the nodewise Lasso estimator $\hat{\gamma}_j$ implies $Z_j^T Z \hat{\Theta}_j / n = \hat{\gamma}_j^2 \lambda_j \hat{\gamma}_j$ and it holds $Z_j^T Z \hat{\Theta}_j / n = e_j$. Consequently, we have

$$\hat{\Sigma} \hat{\Theta}_j - e_j = \lambda_j \hat{\gamma}_j / \hat{\gamma}_j^2.$$  

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

$$\|T_j^T M^T \Theta (\hat{\Sigma} \hat{\Theta} - I_q) M\|_\infty = \|T_j^T M^T \Theta (\lambda_1 \hat{\gamma}_1 / \hat{\gamma}_1^2, \ldots, \lambda_q \hat{\gamma}_q / \hat{\gamma}_q^2) M\|_\infty$$  

$$\leq \sqrt{\log(q)/n} \|\Theta M\|_1 \max_{1 \leq j \leq q} \hat{\gamma}_j^2$$  

$$= \sqrt{\log(q)/n} \sqrt{s_{\text{max}} s_M} \sqrt{s_M (\lambda^2 M) \times O_p(1)}$$  

$$= O_p(\log(q)/n),$$

by using Assumption 3 (iii), i.e., $s_M \sqrt{s_{\text{max}} s_M} = O(\sqrt{\log(q)})$. Finally, we have

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

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 $\|W\|_{\psi_2} := \sup_{q \geq 1} q^{-1/2} (E|W|^q)^{1/q}$ and the sub-exponential norm $\| \cdot \|_{\psi_1}$ as $\|W\|_{\psi_1} := \sup_{q \geq 1} q^{-1} (E|W|^q)^{1/q}$, 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 $\zeta^2_{ij} := E[Z_{ij}^2], \zeta^2_{ik} := E[X_{ik}^2]$ and $\rho_{jk} := E[Z_{ij} X_{ik}] / (\zeta_{ij} \zeta_{ik})$. Let $K_z := \|Z_{ij}\|_{\psi_2}$ and $K_x := \|X_{ik}\|_{\psi_2}$, which do not depend on $i$. Then,

$$\mathbb{P}\left( \left| \hat{M}_{jk} - M_{jk} \right| \geq v \right) = \mathbb{P}\left( \left\| \sum_{i=1}^n (Z_{ij} X_{ik} - M_{jk}) \right\| \geq nv \right)$$  

$$= \mathbb{P}\left( \left\| \sum_{i=1}^n \left( \frac{Z_{ij} X_{ik}}{\zeta_{ij} \zeta_{ik}} - \rho_{jk} \right) \right\| \geq \frac{nv}{\zeta_{ij} \zeta_{ik}} \right).$$

Moreover,

$$\sum_{i=1}^n \left( \frac{Z_{ij} X_{ik}}{\zeta_{ij} \zeta_{ik}} - \rho_{jk} \right) = \frac{1}{4} \left[ \sum_{i=1}^n \left( \frac{Z_{ij}}{\zeta_{ij}} + \frac{X_{ik}}{\zeta_{ik}} \right)^2 - 2(1 + \rho_{jk}) \right]$$  

$$- \sum_{i=1}^n \left( \frac{Z_{ij}}{\zeta_{ij}} - \frac{X_{ik}}{\zeta_{ik}} \right)^2 - 2(1 - \rho_{jk}).$$

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

28
Therefore, \((\frac{z_{ij}}{\zeta_{jj}} + \frac{x_{ik}}{\zeta_{kk}})^2\) and \((\frac{z_{ij}}{\zeta_{jj}} - \frac{x_{ik}}{\zeta_{kk}})^2\) are sub-exponential, see e.g. Vershynin [2012, Lemma 5.14], whose means are, respectively

\[
2(1 + \rho_{jk}) \quad \text{and} \quad 2(1 - \rho_{jk}).
\]

Denote \(W_{i+} := (\frac{z_{ij}}{\zeta_{jj}} + \frac{x_{ik}}{\zeta_{kk}})^2 \frac{1}{2(1 + \rho_{jk})} - 1\) and \(W_{i-} := (\frac{z_{ij}}{\zeta_{jj}} - \frac{x_{ik}}{\zeta_{kk}})^2 \frac{1}{2(1 - \rho_{jk})} - 1\) 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 \(p \geq 1\):

\[
(\mathbb{E}|W_{i+}|^p)^{1/p} \leq \left\lVert \left(\frac{z_{ij}}{\zeta_{jj}} + \frac{x_{ik}}{\zeta_{kk}}\right)^2 \frac{1}{2(1 + \rho_{jk})} \right\rVert_p + \|1\|_p \\
\leq 2 \left\lVert \left(\frac{z_{ij}}{\zeta_{jj}} + \frac{x_{ik}}{\zeta_{kk}}\right)^2 \frac{1}{2(1 + \rho_{jk})} \right\rVert_p \leq 2Kp
\]

where we have used the triangle inequality to get the first inequality, the Jensen inequality to get the second inequality and sub-exponentiality of \((\frac{z_{ij}}{\zeta_{jj}} + \frac{x_{ik}}{\zeta_{kk}})^2 \frac{1}{2(1 + \rho_{jk})}\) to get the last inequality.

The sub-exponential norm of \(W_{i+}\) can be upper bounded as follows:

\[
\|W_{i+}\|_{\psi_1} \leq \sup_{q \geq 1} q^{-1} \|W_{i+}\|_q \leq \sup_{q \geq 1} q^{-1} \left( \left\lVert \left(\frac{z_{ij}}{\zeta_{jj}} + \frac{x_{ik}}{\zeta_{kk}}\right)^2 \frac{1}{2(1 + \rho_{jk})} \right\rVert_q + \|1\|_q \right) \\
\leq \left\lVert \left(\frac{z_{ij}}{\zeta_{jj}} + \frac{x_{ik}}{\zeta_{kk}}\right)^2 \frac{1}{2(1 + \rho_{jk})} \right\rVert_{\psi_1} + \sup_{q \geq 1} q^{-1} \left( \left\lVert \left(\frac{z_{ij}}{\zeta_{jj}} + \frac{x_{ik}}{\zeta_{kk}}\right)^2 \frac{1}{2(1 + \rho_{jk})} \right\rVert_q \right) \\
= 2 \left\lVert \left(\frac{z_{ij}}{\zeta_{jj}} + \frac{x_{ik}}{\zeta_{kk}}\right)^2 \frac{1}{2(1 + \rho_{jk})} \right\rVert_{\psi_1} \leq 4 \left\lVert \left(\frac{z_{ij}}{\zeta_{jj}} + \frac{x_{ik}}{\zeta_{kk}}\right) \frac{1}{\sqrt{2(1 + \rho_{jk})}} \right\rVert_{\psi_2}^2 \\
\leq 4 \left(\frac{K_z}{\zeta_{jj}} + \frac{K_x}{\zeta_{kk}}\right)^2 \frac{1}{2(1 + \rho_{jk})} \tag{A.9}
\]

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

\[
\|W_{i-}\|_{\psi_1} \leq 4 \left(\frac{K_z}{\zeta_{jj}} + \frac{K_x}{\zeta_{kk}}\right)^2 \frac{1}{2(1 - \rho_{jk})} \tag{A.10}
\]

and the right hand side does not depend on \(i\). Therefore, for every \(i\), \(\|(1 + \rho_{jk})W_{i+}\|_{\psi_1} \leq 2 \left(\frac{K_z}{\zeta_{jj}} + \frac{K_x}{\zeta_{kk}}\right)^2\) and \(\|(1 - \rho_{jk})W_{i-}\|_{\psi_1} \leq 2 \left(\frac{K_z}{\zeta_{jj}} + \frac{K_x}{\zeta_{kk}}\right)^2\). Let \(K := \max_i \|((1 - \rho_{jk})W_{i-})\|_{\psi_1}\). For every \(t \geq 0\), define the event \(A := \{\sum_{i=1}^n (1 - \rho_{jk})W_{i-} \geq t\}\) which by using
Vershynin [2012, Proposition 5.16] has probability upper bounded by

$$
\mathbb{P}(\mathcal{A}) \leq 2 \exp \left\{ -c \min \left\{ \frac{t^2}{K^2 n}, \frac{t}{K} \right\} \right\}
$$

$$
\leq 2 \exp \left\{ -c \min \left\{ \frac{t^2}{4n \left( \frac{K_{ij}}{\zeta_{ij}} + \frac{K_{jk}}{\zeta_{jk}} \right)^2}, \frac{t}{2 \left( \frac{K_{ij}}{\zeta_{ij}} + \frac{K_{jk}}{\zeta_{jk}} \right)^2} \right\} \right\}
$$

(A.11)

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

$$
\mathbb{P} \left( \left| \tilde{M}_{jk} - M_{jk} \right| \geq v \right) = \mathbb{P} \left( \left| \frac{1}{2} \sum_{i=1}^{n} W_{i+} (1 + \rho_{jk}) - \sum_{i=1}^{n} W_{i-} (1 - \rho_{jk}) \right| \geq \frac{nv}{\zeta_{ij} \zeta_{jk}} \right)
$$

$$
\leq \mathbb{P} \left( \left| \frac{1}{2} \sum_{i=1}^{n} W_{i+} (1 + \rho_{jk}) \right| \geq 2 \frac{nv}{\zeta_{ij} \zeta_{jk}} - \left| \sum_{i=1}^{n} W_{i-} (1 - \rho_{jk}) \right| \cap A^c \right) + \mathbb{P}(\mathcal{A})
$$

$$
\leq \mathbb{P} \left( \left| \frac{1}{2} \sum_{i=1}^{n} W_{i+} (1 + \rho_{jk}) \right| \geq \frac{nv}{\zeta_{ij} \zeta_{jk}} \cap A^c \right) + \mathbb{P}(\mathcal{A})
$$

$$
\leq \mathbb{P} \left( \left| \sum_{i=1}^{n} W_{i+} (1 + \rho_{jk}) \right| \geq \frac{nv}{\zeta_{ij} \zeta_{jk}} \right) + \mathbb{P}(\mathcal{A}).
$$

Therefore, by using (A.11) with \( t = nv / (\zeta_{ij} \zeta_{jk}) \) 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

$$
\mathbb{P} \left( \left| \tilde{M}_{jk} - M_{jk} \right| \geq v \right)
$$

$$
\leq 2 \exp \left\{ -c \min \left\{ \frac{v^2}{4 \zeta_{ij}^2 \zeta_{jk}^2 \left( \frac{K_{ij}}{\zeta_{ij}} + \frac{K_{jk}}{\zeta_{jk}} \right)^2}, \frac{v}{2 \zeta_{ij} \zeta_{jk} \left( \frac{K_{ij}}{\zeta_{ij}} + \frac{K_{jk}}{\zeta_{jk}} \right)^2} \right\} n \right\}
$$

$$
+ 2 \exp \left\{ -c \min \left\{ \frac{v^2}{4 \zeta_{ij}^2 \zeta_{jk}^2 \left( \frac{K_{ij}}{\zeta_{ij}} + \frac{K_{jk}}{\zeta_{jk}} \right)^2}, \frac{v}{2 \zeta_{ij} \zeta_{jk} \left( \frac{K_{ij}}{\zeta_{ij}} + \frac{K_{jk}}{\zeta_{jk}} \right)^2} \right\} n \right\}
$$

$$
= 4 \exp \left\{ -c \min \left\{ \frac{v^2}{4 \zeta_{ij}^2 \zeta_{jk}^2 \left( \frac{K_{ij}}{\zeta_{ij}} + \frac{K_{jk}}{\zeta_{jk}} \right)^2}, \frac{v}{2 \zeta_{ij} \zeta_{jk} \left( \frac{K_{ij}}{\zeta_{ij}} + \frac{K_{jk}}{\zeta_{jk}} \right)^2} \right\} n \right\}
$$

$$
\leq 4 \exp \left\{ -Cv^2 n \right\}
$$

where for the last inequality we have used that \( \min(a/b, c/d) \geq \min(a, c) / \max(b, d) \) 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 3 (iii), \( M \) belongs to the class of matrices

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

30
with $\xi = 0$ and $|M_{jk}|$ 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 $0 \leq \xi < 1$. 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. 

\section*{B. Appendix: Technical Results}

Recall the notation $\tilde{M} = Z^T X/n$ and the thresholding estimator: $\hat{M} = (\hat{M}_{jk})$ with
\begin{equation}
\hat{M}_{jk} := \tilde{M}_{jk} \mathbb{1}\left\{ |\tilde{M}_{jk}| \geq C_0 \sqrt{\frac{\log(q)}{n}} \right\}, \quad C_0 > 0. \tag{B.1}
\end{equation}

In the following theorem, we provide the rate for its $\ell_1$-norm. The minimax rate for the $\ell_1$-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 $G_{\xi}(\rho, s_M)$ defined in (A.12) for $0 \leq \xi < 1$ where $|M_{jk}|$ denotes the $j$-th largest element in magnitude of the $k$-th column $(M_{jk})_{1 \leq j \leq q}$. Every matrix in $G_{\xi}(\rho, s_M)$ has columns $(M_{jk})_{1 \leq j \leq q}$ that are in a (approximate) sparse weak $\ell_\xi$ ball. The case $\xi = 0$ is the case considered in the paper. Moreover, define the class of distributions $\mathcal{P}(G_{\xi}(\rho, s_M))$ as the set of distributions of $(Z, X)$ satisfying (A.12) and such that the rows of $Z$ and $X$ are sub-Gaussian.

\textbf{Theorem B.1.} Let Assumption 1 (ii) hold. Then, the thresholding estimator $\hat{M}$ satisfies
\begin{equation}
\sup_{\mathcal{P}(G_{\xi}(\rho, s_M))} \mathbb{E}\left\| \hat{M} - M \right\|_1 \leq C s_M \left( \frac{\log(p \vee q)}{n} \right)^{1-\xi}
\end{equation}
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{equation}
\mathbb{P}\left( \left\| \hat{M} - M \right\|_1 > \varepsilon \right) \leq \frac{1}{\varepsilon^2} \mathbb{E}\left\| \hat{M} - M \right\|_1^2 \leq C s_M^2 \left( \frac{\log(q)}{n} \right)^{1-\xi}
\end{equation}
which implies:
\begin{equation}
\left\| \hat{M} - M \right\|_1 \leq s_M \left( \frac{\log(q)}{n} \right)^{(1-\xi)/2}
\end{equation}
with probability approaching one.
Proof. Define the event $A_{jk} := \{ |\tilde{M}_{jk} - M_{jk}| \leq 4 \min \left\{ |M_{jk}|, C_0 \sqrt{\frac{\log(q)}{n}} \right\} \}$ and $D = (d_{jk})$ with $d_{jk} := (\tilde{M}_{jk} - M_{jk}) \mathbb{1}_{A^c_{jk}}$. Then,

$$\mathbb{E} \left\| \tilde{M} - M \right\|_1^2 = \mathbb{E} \left\| \tilde{M} - M - D + D \right\|_1^2 \leq 2\mathbb{E} \left( \sup_{1 \leq k \leq p} \sum_{j=1}^q |\tilde{M}_{jk} - M_{jk}| \mathbb{1}_{A_{jk}} \right)^2 + 2\mathbb{E} \left\| D \right\|_1^2$$

where the inequality in the penultimate line is due to $(\tilde{M} - M - D)_{jk} = (\tilde{M}_{jk} - M_{jk})(1 - \mathbb{1}_{A_{jk}}) = (\tilde{M}_{jk} - M_{jk}) \mathbb{1}_{A^c_{jk}}$. 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

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

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:

$$\mathbb{E} \left\| D \right\|_1^2 = \mathbb{E} \left( \max_{1 \leq k \leq p} \sum_{j=1}^q |d_{jk}| \right)^2 \leq \mathbb{E} \left[ q \sum_{k=1}^p \sum_{j=1}^q \mathbb{E} |d_{jk}|^2 \right] = q \sum_{k=1}^p \sum_{j=1}^q \mathbb{E} d_{jk}^2$$

$$= \sum_{k=1}^p \sum_{j=1}^q \mathbb{E} \left( d_{jk}^2 \mathbb{1}_{\{A_{jk} \cap \{\tilde{M}_{jk} = M_{jk}\}\}} + d_{jk}^2 \mathbb{1}_{\{A^c_{jk} \cap \{\tilde{M}_{jk} = M_{jk}\}\}} \right)$$

$$= \sum_{k=1}^q \sum_{j=1}^q \mathbb{E} \left( (\tilde{M}_{jk} - M_{jk})^2 + M_{jk}^2 \mathbb{1}_{\{A^c_{jk} \cap \{\tilde{M}_{jk} = M_{jk}\}\}} \right) =: R_1 + R_2.$$

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

$$R_1 \leq \sum_{k=1}^p \sum_{j=1}^q \mathbb{E}^{1/3} \left[ (\tilde{M}_{jk} - M_{jk})^6 \right] \mathbb{P}^{2/3}(A^c_{jk})$$

$$\leq C_3 p^2 q \frac{1}{n} \mathbb{P}^{2/3}(A^c_{jk})$$

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

$$R_1 \leq 2C_2 C_3 \frac{q^3 q^{-3}}{n} \leq C_5 / n.$$
Let us now consider term $R_2$:

\[
R_2 = p \sum_{k=1}^{p} \sum_{j=1}^{q} \mathbb{E} \left( M_{jk}^2 \mathbb{1}_{\{|M_{jk}| \geq 4C_0 \sqrt{\log(q)/n}\}} \right)
\]

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

\[
R_2 \leq p \sum_{k=1}^{p} \sum_{j=1}^{q} nM_{jk}^2 \mathbb{E} \left( \mathbb{1}_{\{|M_{jk}| \geq 4C_0 \sqrt{\log(q)/n}\}} \mathbb{1}_{\{|M_{jk} - \tilde{M}_{jk} - M_{jk}| \leq C_0 \sqrt{\log(q)/n}\}} \right)
\]

\[
= \frac{p}{n} \sum_{k=1}^{p} \sum_{j=1}^{q} nM_{jk}^2 \mathbb{E} \left( \mathbb{1}_{\{|M_{jk}| \geq 4C_0 \sqrt{\log(q)/n}\}} \right)
\]

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

\[
R_2 \leq \frac{pC_0^2}{n} \sum_{k=1}^{p} \sum_{j=1}^{q} \exp\left\{-nM_{jk}^2 c/(16)\right\} \mathbb{1}_{\{|M_{jk}| \geq 4C_0 \sqrt{\log(q)/n}\}}
\]

\[
\leq \frac{pC_0^2}{n} \sum_{k=1}^{p} \sum_{j=1}^{q} \exp\{-n16C_0^2 \log(q)c/(16n)\}
\]

\[
\leq \frac{pC_0^2}{n} \sum_{k=1}^{p} \sum_{j=1}^{q} \exp\{-8 \log(q)\} = \frac{(q)^3C_0^2}{n}(q)^{-8}
\]

\[
\leq C_0/n.
\]

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

\[
\mathbb{E}[|\tilde{M}_{jk} - M_{jk}|^6] \leq \frac{24}{C_0^2n^3}. \tag{B.5}
\]

**Proof.** The 6-th moment can be written as

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

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

\[
\mathbb{E}[|\tilde{M}_{jk} - M_{jk}|^6] \leq 24 \int_0^\infty x^5 \exp\left\{-cnx^2\right\} dx = \frac{24}{c^3n^3}.
\]
Lemma B.3. Define the event $A_{jk}$ as

$A_{jk} := \left\{ |\widehat{M}_{jk} - M_{jk}| \leq 4 \min \{ |M_{jk}, C_0 \sqrt{\frac{\log(q)}{n}}| \} \right\}$

for $C_0 = \sqrt{\frac{n}{C_2}}$ where $C_2$ is as in Lemma B.2. Then,

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

for some constant $C_3 > 0$.

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

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

By the triangular inequality we have:

$$A_1 = \left\{ |\widehat{M}_{jk} - M_{jk} + M_{jk}| \geq C_0 \sqrt{\frac{\log(q)}{n}} \right\} \subset \left\{ |\widehat{M}_{jk} - M_{jk}| \geq C_0 \sqrt{\frac{\log(q)}{n}} - |M_{jk}| \right\}$$

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

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

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.

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.

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. 75(6):1613–1669, 2007.
P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

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

T. T. Cai and H. H. Zhou. Minimax estimation of large covariance matrices under $\ell_1$-norm. *Statistica Sinica*, 22(4):1319–1349, 2012.

X. Chen and T. Christensen. Optimal sup-norm rates, adaptivity and inference in nonparametric instrumental variables estimation. Technical report, Cowles Foundation Discussion Paper no 1923R, 2017.

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 quasi likelihood ratio inference on semi/nonparametric conditional moment models. Technical report, Cowles Foundation for Research in Economics, Yale University, 2013.

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.

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

H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*. Kluwer Academic, Dordrecht, 2000.

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

E. Gautier and A. Tsybakov. 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. T. Cai, and D. S. Small. Confidence intervals for causal effects with invalid instruments using two-stage hard thresholding with voting. *arXiv preprint arXiv:1603.05224*, 2016.

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*, 71:1565–1578, 2003.

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, and N. Davies. On the use of the lasso for instrumental variables estimation with some invalid instruments. 2017.

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.