Klaassen, Sven; Kück, Jannis; Spindler, Martin; Chernozhukov, Victor

Working Paper

Uniform inference in high-dimensional gaussian graphical models

cemmap working paper, No. CWP29/19

Provided in Cooperation with:
Institute for Fiscal Studies (IFS), London

Suggested Citation: Klaassen, Sven; Kück, Jannis; Spindler, Martin; Chernozhukov, Victor (2019) : Uniform inference in high-dimensional gaussian graphical models, cemmap working paper, No. CWP29/19, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2019.2919

This Version is available at:
https://hdl.handle.net/10419/211122

cemmap
centre for microdata methods and practice

Uniform Inference in High-Dimensional Gaussian Graphical Models

Sven Klaassen
Jannis Kück
Martin Spindler
Victor Chernozhukov

The Institute for Fiscal Studies
Department of Economics,
UCL

cemmap working paper CWP29/19

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

UNIFORM INFERENCE IN HIGH-DIMENSIONAL GAUSSIAN GRAPHICAL MODELS*

By Sven Klaassen, Jannis Kück,
Martin Spindler and Victor Chernozhukov

Graphical models have become a very popular tool for representing dependencies within a large set of variables and are key for representing causal structures. We provide results for uniform inference on high-dimensional graphical models with the number of target parameters d being possible much larger than sample size. This is in particular important when certain features or structures of a causal model should be recovered. Our results highlight how in high-dimensional settings graphical models can be estimated and recovered with modern machine learning methods in complex data sets. To construct simultaneous confidence regions on many target parameters, sufficiently fast estimation rates of the nuisance functions are crucial. In this context, we establish uniform estimation rates and sparsity guarantees of the square-root estimator in a random design under approximate sparsity conditions that might be of independent interest for related problems in high-dimensions. We also demonstrate in a comprehensive simulation study that our procedure has good small sample properties.

1. Introduction. We provide methodology and theory for uniform inference on high-dimensional graphical models with the number of target parameters being possible much larger than sample size. We demonstrate uniform asymptotic normality of the proposed estimator over d-dimensional rectangles and construct simultaneous confidence bands on all of the d target parameters. The proposed method can be applied to test simultaneously the presence of a large set of edges in the graphical model

$$X = (X_1, \ldots, X_p)^T \sim \mathcal{N}(\mu_X, \Sigma_X).$$

Assuming that the covariance matrix Σ_X is nonsingular, the conditional independence structure of the distribution can be conveniently represented by a graph $G = (V, E)$, where $V = \{1, \ldots, p\}$ is the set of nodes and E the set of edges in $V \times V$. Every pair of variables not contained in the edge set is conditionally independent given all remaining variables. If the vector X

*Version November 2018.

MSC 2010 subject classifications: Primary 60J05, 60J07, 41A25, 49M15

Keywords and phrases: Gaussian Graphical Models, conditional independence, Square-Root Lasso, Post-selection Inference, High-dimensional Setting, Z-estimation

is normally distributed, every edge corresponds to a non-zero entry in the inverse covariance matrix (Lauritzen (1996)) [11].

In the last decade, significant progress has been made on estimation of a large precision matrix in order to analyze the dependence structure of a high-dimensional normal distributed random variable. There are mainly two common approaches to estimate the entries of a precision matrix. The first approach is a penalized likelihood estimation approach with a lasso-type penalty on entries of the precision matrix, typically referred to as the graphical lasso. This approach has been studied in several papers, see e.g Lam and Fan (2009) [10], Rothman et al. (2008) [15], Ravikumar et al. (2011) [13] and Yuan and Lin (2007) [20]. The second approach, first introduced by Meinshausen and Bühlmann (2006) [12], is neighborhood based. It estimates the conditional independence restrictions separately for each node in the graph and is hence equivalent to variable selection for Gaussian linear models. The idea of estimating the precision matrix column by column by running a regression for each variable against the rest of variables was further studied in Yuan (2010) [19], Cai, Liu and Zhou (2011) [5] and Sun and Zhang (2013) [16].

In this paper, we do not aim to estimate the whole precision matrix but we focus on quantifying the uncertainty of recovering its support by providing a significance test for a set of potential edges. In recent years, statistical inference for the precision matrix in high-dimensional settings has been studied, e.g in Janková and van de Geer (2016) [9] and Ren et al. (2015) [14]. Both approaches lead to an estimate that is elementwise asymptotically normal and enables testing for low-dimensional parameters of the precision matrix using standard procedures such as Bonferroni-Holm correction.
In contrast to these existing results, our method explicitly allows for testing a joint hypothesis without correction for multiple testing and conducting inference for a growing number of parameters using high dimensional central limit results. In particular, our results rely on approximate sparsity instead of row sparsity which restricts the number of non-zero entries of each row of the precision matrix to be at most $s \ll n$ that is in many applications a questionable assumption. In order to provide theoretical results, fitting the problem of support discovery in Gaussian graphical models into a general Z-estimation setting with a high-dimensional nuisance function is key. Inference on a (multivariate) target parameter in general Z-estimation problems in high dimensions is covered in Belloni et al. (2014) [3], Belloni et al. (2018) [2] and Chernozhukov et al. (2017) [6]. To conduct inference on a

high-dimensional target parameter, uniform estimation rates and sparsity guarantees of the nuisance function are crucial. In this context, we formally apply recent results from Belloni et al. (2018) [2] to ensure sufficient fast convergence rate of the lasso estimator under approximate sparsity conditions. Moreover, we provide auxiliary results for the square-lasso estimator establishing uniform estimation rates and sparsity guarantees in a random design under approximate sparsity conditions that might be of independent interest for related problems in high-dimensional linear models.

Plan of this Paper. The rest of this paper is organized as follows. In Section 2, we formally define the setting and introduce the notation that will be used fitting the problem of support discovery in Gaussian graphical models into a general Z-estimation problem with a high-dimensional nuisance function. In Section 3, we outline the estimation procedure of the high-dimensional target parameter and the conditions that are needed to achieve our main theorem presented in Section 4. Section 5 provides implementation details and shows how our estimation procedure can be modified by cross-fitting to improve small sample properties. Section 6 provides a simulation study on the proposed method. The supplementary material includes additional technical material. The proof of our main theorem is provided in Appendix A. The uniform nuisance function estimation is discussed in Appendix B. Appendix B.1 formally discusses conditions for the uniform convergence rates of the lasso estimator. Finally, Appendix B.2 provides auxiliary results for the square-lasso estimator.

2. Setting.

Let

$$X = (X_1, \dots, X_p)^T \sim \mathcal{N}(\mu_X, \Sigma_X)$$

be a p-dimensional random variable. For all $(j,k) \in E$ with $j \neq k$, assume that

$$X_j = \sum_{\substack{l=1 \\ l \neq j}}^{p} \beta_l^{(j)} X_l + \varepsilon^{(j)} = \beta^{(j)} X_{-j} + \varepsilon^{(j)}$$

and

$$X_k = \gamma^{(j,k)} X_{-\{j,k\}} + \nu^{(j,k)},$$

where $\mathbb{E}[\varepsilon^{(j)} | X_{-j}] = 0$ and $\mathbb{E}[X_{-\{j,k\}} \nu^{(j,k)}] = 0$. Define the column vector

$$\Gamma^{(j)} = \left(-\beta_1^{(j)}, \dots, -\beta_{j-1}^{(j)}, 1, -\beta_{j+1}^{(j)}, \dots, -\beta_p^{(j)} \right)^T.$$

One may show

$$\Phi_0 = \left(\Phi_0^1, \ldots, \Phi_0^p\right) = \left(\Gamma^{(1)}/Var(\varepsilon^{(1)}), \ldots, \Gamma^{(p)}/Var(\varepsilon^{(p)})\right),$$

where Φ_0^j is the j-th column of the precision matrix $\Phi_0 = \Sigma_X^{-1}$ [9]. Hence

$$\beta_k^{(j)} = 0 \Leftrightarrow \beta_j^{(k)} = 0 \Leftrightarrow X_j \perp X_k | X_{-\{j,k\}} \tag{2.1}$$

for all $j \neq k$. Assume that we are interested in the following set of potential edges

$$\mathcal{M} := \{m_1, \ldots, m_{d_n}\}$$

where the number of edges d_n may increase with sample size n. In the following the dependence on n is omitted to simplify the notation. In order to test whether all variables X_j and X_k are conditionally independent with $m_r = (j_r, k_r)$ for a $r \in \{1, \ldots, d\}$, we have to estimate our target parameter

$$\theta_0 = (\theta_{m_1}, \ldots, \theta_{m_d})^T := (\beta_{k_1}^{(j_1)}, \ldots, \beta_{k_d}^{(j_d)})^T.$$

The setting above fits in the general Z-estimation problem of the form

$$\mathbb{E}\left[\psi_{m_r}\left(X, \theta_{m_r}, \eta_{m_r}\right)\right] = 0$$

for all $r = 1, \ldots, d$ with nuisance parameters

$$\eta_{m_r} = \left(\beta_{-k}^{(j)}, \gamma^{(j,k)}\right)$$

where $\beta_{-k}^{(j)} \equiv \beta^{(m_r)}$ and $\gamma^{(j,k)} \equiv \gamma^{(m_r)}$. The score functions are defined by

$$\psi_{m_r}(X, \theta, \eta) := \left(X_j - \theta X_k - \eta^{(1)} X_{-m_r}\right)\left(X_k - \eta^{(2)} X_{-m_r}\right)$$

for $m_r = (j_r, k_r) \equiv (j, k)$, $\eta = (\eta^{(1)}, \eta^{(2)})$ and $r = 1, \ldots, d$. Without loss of generality we assume $j > k$ for all tuples $m_r \in \mathcal{M}$.

Comment 2.1. The score function ψ is linear, meaning

$$\psi_{m_r}(X, \theta, \eta) = \psi_{m_r}^a(X, \eta^{(2)})\theta + \psi_{m_r}^b(X, \eta)$$

with

$$\psi_{m_r}^a(X, \eta^{(2)}) = -X_k\left(X_k - \eta^{(2)} X_{-m_r}\right)$$

and

$$\psi_{m_r}^b(X, \eta) = \left(X_j - \eta^{(1)} X_{-m_r}\right)\left(X_k - \eta^{(2)} X_{-m_r}\right)$$

for $m_r = (j,k)$ and $r = 1, \ldots, d$.

It is well known that in partially linear regression models θ_0 satisfies the moment condition

$$\mathbb{E}\left[\psi_{m_r}\left(X, \theta_{m_r}, \eta_{m_r}\right)\right] = 0 \tag{2.2}$$

for all $r = 1, \ldots, d$ and also the *Neyman orthogonality* condition

$$\partial_t \left\{\mathbb{E}\left[\psi_{m_r}\left(X, \theta_{m_r}, \eta_{m_r} + t\tilde{\eta}\right)\right]\right\}\Big|_{t=0}$$

for all $\tilde{\eta}$ in an appropriate set where ∂_t denotes the derivate with respect to t. These properties are crucial for valid inference in high dimensional settings. We will show these properties explicitly in the proof of Theorem 1.

3. Estimation. Let $X^{(i)}$, $i = 1, \ldots, n$ be i.i.d. random vectors.
At first we estimate the nuisance parameter $\eta_{m_r} = \left(\eta_{m_r}^{(1)}, \eta_{m_r}^{(2)}\right)$ by running a lasso/ post-lasso/ square-root lasso regression of X_j on X_{-j} to compute $(\tilde{\theta}_{m_r}, \hat{\eta}_{m_r}^{(1)})$ and a lasso/ post-lasso/ square-root lasso regression of X_k on X_{-m_r} to compute $\hat{\eta}_{m_r}^{(2)}$ for each $(j,k) = m_r \in \mathcal{M}$. The estimator $\hat{\theta}_0$ of the target parameter

$$\theta_0 = (\theta_{m_1}, \ldots, \theta_{m_d})^T$$

is defined as the solution of

$$\sup_{r=1,\ldots,d} \left\{ \left| \mathbb{E}_n\left[\psi_{m_r}\left(X, \hat{\theta}_{m_r}, \hat{\eta}_{m_r}\right)\right] \right| - \inf_{\theta \in \Theta_{m_r}} \left| \mathbb{E}_n\left[\psi_{m_r}\left(X, \theta, \hat{\eta}_{m_r}\right)\right] \right| \right\} \leq \epsilon_n, \tag{3.1}$$

where $\epsilon_n = o\left(\delta_n n^{-1/2}\right)$ is the numerical tolerance and $(\delta_n)_{n \geq 1}$ a sequence of positive constants converging to zero.

Assumptions **A1**-**A4**.
Let $a_n := \max(d, p, n, e)$ and C a strictly positive constant independent of n and r. The following assumptions hold uniformly in $n \geq n_0, P \in \mathcal{P}_n$:

A1 *For all $m_r = (j,k) \in \mathcal{M}$ with $j \neq k$ we have the following approximate sparse representations*

(i) It holds

$$\begin{aligned} X_j &= \beta^{(j)} X_{-j} + \varepsilon^{(j)} \\ &= \theta_{m_r} X_k + \left(\beta^{(1,m_r)} + \beta^{(2,m_r)}\right) X_{-m_r} + \varepsilon^{(m_r)} \end{aligned}$$

with

$$\|\beta^{(1,m_r)}\|_0 \leq s, \quad \max_{r=1,\dots,d} \|\beta^{(2,m_r)}\|_1^2 \leq C\sqrt{\frac{s^2\log(a_n)}{n}}$$

and

$$\max_{r=1,\dots,d} \mathbb{E}\left[\left(\beta^{(2,m_r)}X_{-m_r}\right)^2\right] \leq C\frac{s\log(a_n)}{n}.$$

(ii) It holds

$$\begin{aligned} X_k &= \gamma^{(j,k)}X_{-\{j,k\}} + \nu^{(j,k)} \\ &= \left(\gamma^{(1,m_r)} + \gamma^{(1,m_r)}\right)X_{-m_r} + \nu^{(m_r)} \end{aligned}$$

with

$$\|\gamma^{(1,m_r)}\|_0 \leq s, \quad \max_{r=1,\dots,d} \|\gamma^{(2,m_r)}\|_1^2 \leq C\sqrt{\frac{s^2\log(a_n)}{n}}$$

and

$$\max_{r=1,\dots,d} \mathbb{E}\left[\left(\gamma^{(2,m_r)}X_{-m_r}\right)^2\right] \leq C\frac{s\log(a_n)}{n}.$$

A2 *There exist positive numbers $\tilde{q} > 0$ and $\kappa < 1$ such that the following growth conditions are fulfilled:*

$$n^{\frac{1}{\tilde{q}}}\frac{s^2\log^4(a_n)}{n} = o(1), \quad \log(d) = o\left(n^{\frac{1}{9}} \wedge n^{\frac{\kappa}{\tilde{q}}}\right).$$

A3 *For all $m_r = (j,k) \in \mathcal{M}$ it holds*

$$\|\beta^{(m_r)}\|_2 + \|\gamma^{(m_r)}\|_2 \leq C$$

and

$$\sup_{r=1,\dots,d}\sup_{\theta\in\Theta_{m_r}} |\theta| \leq C.$$

Additionally Θ_{m_r} contains a ball of radius $\log(\log(n))n^{-1/2}\log^{1/2}(d)\log(n)$ centered at θ_{m_r}.

A4 *It holds*

$$\inf_{\|\xi\|_2=1} \mathbb{E}\left[(\xi X)^2\right] \geq c \text{ and } \sup_{\|\xi\|_2=1} \mathbb{E}\left[(\xi X)^2\right] \leq C.$$

The condition A1 is a standard approximate sparsity condition that is discussed in more detail in comment 3.1. The number of relevant variables $s_n \equiv s$ captured by the regression coefficient $\beta^{(1,m_r)}$ respectively $\gamma^{(1,m_r)}$ can grow with the sample size. The coefficient $\beta^{(2,m_r)}$ respectively $\gamma^{(2,m_r)}$ is the approximate sparse part of the true regression coefficient. This misspecification of a sparse model is controlled by condition A1. The growth condition A2 ensures that $s^2 \log^4(a_n)/n$ converges towards zero with at least polynomial speed. If this convergence is too slow ($\tilde{q} \geq 9$) the condition on the growth rate of the number of tested edges become more restrictive. In general, both the number of parameters p and the number of relevant variables s can grow with the sample size in a balanced way. If s is fixed, the number of potential parameters p can grow at an exponential rate with the sample size. This means that the set of potential variables can be much larger than the sample size, only the number of relevant variables s has to be smaller than the sample size. This situation is common for Lasso-based estimators. Condition A3 restricts the parameter spaces and ensures that the true coefficients are well behaved. The condition A4 is a standard eigenvalue condition that restricts the correlation between the components of X and bounds the variances of each X_j from below and above. Assumptions A1-A4 combined with the normal distribution of X imply the conditions B1-B4 from theorem 2 which enables us to estimate the nuisance parameter sufficiently fast by lasso and post-lasso. To ensure a sufficiently fast convergence rate and sparsity guarantees of the square-root lasso estimator further model assumptions are needed.

COMMENT 3.1. If we have exact sparsity for each $\beta^{(k)}$ with $(j,k) \in \mathcal{M}_r$ the sparsity of $\gamma^{(m_r)}$ follows directly.
Observe that for $k \in \{1,\dots,p\} \setminus \{j\}$ and $l \in \{1,\dots,p\} \setminus \{j,k\}$ we have

$$\beta_l^{(k)} = 0 \Leftrightarrow X_k \perp X_l | X_{-\{k,l\}} \Leftrightarrow \mathbb{E}[X_k X_l | X_{-\{k,l\}}] = 0$$

which implies

$$\mathbb{E}[X_k X_l | X_{-\{j,k,l\}}] = \mathbb{E}\left[\mathbb{E}[X_k X_l | X_{-\{k,l\}}] | X_{-\{j,k,l\}}\right] = 0$$

and thereby

$$\gamma_l^{(j,k)} = 0 \Leftrightarrow X_k \perp X_l | X_{-\{j,k,l\}} \Leftrightarrow \mathbb{E}[X_k X_l | X_{-\{j,k,l\}}] = 0.$$

Hence, the sparsity conditions for testing on an edge (j,k) are satisfied if each node j and k is only sparsely connected to all other nodes.

4. Main results. We will prove that the assumptions of Corollary 2.2 from Belloni et al. (2018) [2] hold and hence we are able to use their results to construct confidence intervals even for a growing number of hypothesis $d = d_n$. Define

$$J_{m_r} := \partial_\theta \mathbb{E}[\psi_{m_r}(X), \theta, \eta_{m_r}]\big|_{\theta=\theta_{m_r}} = -\mathbb{E}[X_k(X_k - \eta_{m_r}^{(2)} X_{-m_r})]$$
$$\sigma_{m_r}^2 := \mathbb{E}\left[J_{m_r}^{-2}\psi_{m_r}^2(X, \theta_{m_r}, \eta_{m_r})\right]$$

and the corresponding estimators

$$\hat{J}_{m_r} = -\mathbb{E}_n[X_k(X_k - \hat{\eta}_{m_r}^{(2)} X_{-m_r})]$$
$$\hat{\sigma}_{m_r}^2 = \mathbb{E}_n\left[\hat{J}_{m_r}^{-2}\psi_{m_r}^2(X, \hat{\theta}_{m_r}, \hat{\eta}_{m_r})\right]$$

for $r = 1, \ldots, d$. To construct confidence intervals we will employ the Gaussian multiplier bootstrap. Define

$$\hat{\psi}_{m_r}(X) := -\hat{\sigma}_{m_r}^{-1}\hat{J}_{m_r}^{-1}\psi_{m_r}(X, \hat{\theta}_{m_r}, \hat{\eta}_{m_r})$$

and the process

$$\hat{\mathcal{N}} := \left(\hat{\mathcal{N}}_{m_r}\right)_{m_r \in \mathcal{M}} = \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i \hat{\psi}_{m_r}\left(X^{(i)}\right)\right)_{m_r \in \mathcal{M}}$$

where $(\xi_i)_{i=1}^n$ are independent standard normal random variables which are independent from $\left(X^{(i)}\right)_{i=1}^n$. We define c_α as the $(1-\alpha)$-conditional quantile of $\sup_{m_r \in \mathcal{M}} |\hat{\mathcal{N}}_{m_r}|$ given the observations $\left(X^{(i)}\right)_{i=1}^n$. The following theorem is the main result of our paper and establishes simultaneous confidence bands for the target parameter θ_0.

THEOREM 1.
Under the assumptions A1-A4 with probability $1 - o(1)$ *uniformly in* $P \in \mathcal{P}_n$ *the estimator* $\hat{\theta}$ *in (3.1) obeys*

$$P\left(\hat{\theta}_{m_r} - \frac{c_\alpha \hat{\sigma}_{m_r}}{\sqrt{n}} \leq \theta_{m_r} \leq \hat{\theta}_{m_r} + \frac{c_\alpha \hat{\sigma}_{m_r}}{\sqrt{n}}, r = 1, \ldots, d\right) \to 1 - \alpha. \tag{4.1}$$

Using theorem 1 we are able to construct standard confidence regions which are uniformly valid over a large set of variables and we can check null hypothesis of the form:

$$H_0 : \mathcal{M} \cap E = \emptyset.$$

COMMENT 4.1. Theorem 1 is basically an application of the gaussian approximation and multiplier bootstrap for maxima of sums of high-dimensional random vectors [7]. The central limit theorem and bootstrap in high dimension introduced by Chernozhukov, Chetverikov, Kato et al. (2017) [8] extend these results to more general sets, more precisely sparsely convex sets. Hence our main theorem can be easily generalized to various confidence regions that contain the true target parameter with probability $1-\alpha$. Theorem 1 provides critical regions of the form

$$\sup_{r=1,\dots,d}\left|\sqrt{n}\frac{\hat{\theta}_{m_r}}{\hat{\sigma}_{m_r}}\right| > c_{1-\alpha}. \tag{4.2}$$

Alternatively, we can reject the null hypothesis if

$$\sup_{r=1,\dots,d}\left|\sqrt{n}\frac{\hat{\theta}_{m_r}}{\hat{\sigma}_{m_r}}\right| < c_{\frac{\alpha}{2}} \quad \text{or} \quad \sup_{r=1,\dots,d}\left|\sqrt{n}\frac{\hat{\theta}_{m_r}}{\hat{\sigma}_{m_r}}\right| > c_{1-\frac{\alpha}{2}}. \tag{4.3}$$

Both of these regions are based on the central limit theorem for hyperrectangles in high dimensions. The confidence region (4.3) is motivated by the fact that the standard normal distribution $\mathcal{N}(0, I_d)$ in high dimensions is concentrated in a thin spherical shell around the sphere of radius $\sqrt{d}$ as described by Roman Vershynin (2017) [18] and therefore might have smaller volume. More generally, define

$$\hat{\theta}^*_{m_r}(S, exp) = \sum_{s=1}^{S}\left(\sqrt{n}\frac{\hat{\theta}_{m_{r-s}}}{\hat{\sigma}_{m_{r-s}}}\right)^{exp}$$

for a fix S, $exp \in \{1, 2\}$ and

$$r-s := \begin{cases} r-s & \text{if} \quad r-s>0 \\ d+(r-s) & \text{otherwise} \end{cases}.$$

A test that reject the null hypothesis if

$$\sup_{r=1,\dots,d}\left|\hat{\theta}^*_{m_r}(S, exp)\right| > c^*_{1-\alpha} \tag{4.4}$$

has level α by [8], since the constructed confidence regions correspond to S-sparsely convex sets. Here, $c^*_{1-\alpha}$ is the $(1-\alpha)$-conditional quantile of $\sup_{m_r\in\mathcal{M}}|\hat{\mathcal{N}}^*_{m_r}|$ given the observations $\left(X^{(i)}\right)_{i=1}^n$ with

$$\hat{\mathcal{N}}^*_{m_r} = \sum_{s=1}^{S}\left(\hat{\mathcal{N}}_{m_{r-s}}\right)^{exp}$$

where

$$r - s := \begin{cases} r - s & \text{if} \quad r - s > 0 \\ d + (r - s) & \text{otherwise.} \end{cases}$$

5. Notes on the implementation. We implemented a function that will be added to the *R*-package *hdm* and estimates the target coefficients

$$(\theta_{m_1}, \ldots, \theta_{m_d})^T = (\beta_{k_1}^{(j_1)}, \ldots, \beta_{k_d}^{(j_d)})^T$$

corresponding the considered set of potential edges

$$\mathcal{M} := \{m_1, \ldots, m_{d_n}\}$$

by the proposed method described in section 3. It can be used to perform hypothesis tests with asymptotic level α based on the different confidence regions described in comment 4.1. The nuisance function can be estimated by lasso, post-lasso or square-root lasso.

5.1. *Cross-fitting.* In general Z- estimation problems where a so called debiased or double machine learning (DML) method is used to construct confidence intervals, it is common to use cross-fitting in order to improve small sample properties. A detailed discussion of cross-fitted DML can be found in Chernozhukov et al. (2017) [6]. The following algorithm generalizes our proposed method to a K-fold cross fitted version. We assume that n is divisible by K in order to simplify notation.

ALGORITHM 1. *1) Take a K-fold random partition $(I_k)_{k=1}^K$ of observation indices $[n] = \{1, \ldots, n\}$ such that the size of each fold I_k is N. Also, for each $k \in [K] = \{1, \ldots, K\}$, define $I_k^c := \{1, \ldots, N\} \setminus I_k$. 2) For each $k \in [K]$ and $r = 1, \ldots, d$, construct an estimator*

$$\hat{\eta}_{k,m_r} = \hat{\eta}_{m_r}\left((X_i)_{i \in I_k^c}\right)$$

by lasso/ post-lasso or square-root lasso. 3) For each $k \in [K]$, construct an estimator $\hat{\theta}_k = (\hat{\theta}_{k,m_1}, \ldots, \hat{\theta}_{k,m_d})$ as in 3.1:

$$\sup_{r=1,\ldots,d} \left\{ \left| \mathbb{E}_{N,k}\left[\psi_{m_r}\left(X, \hat{\theta}_{k,m_r}, \hat{\eta}_{k,m_r}\right)\right] \right| - \inf_{\theta \in \Theta_{m_r}} \left| \mathbb{E}_{N,k}\left[\psi_{m_r}\left(X, \theta, \hat{\eta}_{k,m_r}\right)\right] \right| \right\}$$
$$\leq \epsilon_n$$

with $\mathbb{E}_{N,k}[\psi_{m_r}(X_i)] = N^{-1} \sum_{i \in I_k} \psi_{m_r}(X_i)$. 4) Aggregate these estimators:

$$\hat{\theta}^K = \frac{1}{K} \sum_{k=1}^{K} \hat{\theta}_k.$$

5) For $r = 1, \dots, d$ construct the uniform valid confidence interval

$$\left[\hat{\theta}^K_{m_r} - \frac{c_\alpha \hat{\sigma}^K_{m_r}}{n}, \hat{\theta}^K_{m_r} + \frac{c_\alpha \hat{\sigma}^K_{m_r}}{n}\right]$$

with

$$\hat{J}^K_{m_r} = -\frac{1}{K}\sum_{k=1}^{K}(X_k(X_k - \hat{\eta}^{(2)}_{k,m_r} X_{-m_r})),$$
$$\hat{\sigma}^K_{m_r} = \sqrt{(\hat{J}^K_{m_r})^{-2}\frac{1}{K}\sum_{k=1}^{K}\left(\psi^2_{m_r}(X, \hat{\theta}^K_{m_r}, \hat{\eta}_{k,m_r})\right)}.$$

c_α is the $1-\alpha$ bootstrap quantile of $\sup_{r=1,\dots,d} \hat{\mathcal{N}}_{m_r}$ with

$$\hat{\mathcal{N}}_{m_r} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i \hat{\psi}^K_{m_r}(X^{(i)})$$

where $(\xi_i)_{i=1}^n$ are independent standard normal random variables which are independent from $\left(X^{(i)}\right)_{i=1}^n$ and

$$\hat{\psi}^K_{m_r}(X) := -\left(\hat{\sigma}^K_{m_r}\hat{J}^K_{m_r}\right)^{-1}\psi_{m_r}(X, \hat{\theta}^K_{m_r}, \hat{\eta}^K_{m_r}).$$

The confidence region above corresponds to (4.2). Confidence regions corresponding to (4.3) or (4.4) can be constructed in an analogous way.

6. Simulation Study. This section provides a simulation study on the proposed method. In each example the precision matrix of the Gaussian graphical model is generated as in the *R*-package *huge* [21]. Hence, the corresponding adjacency matrix A is generated by setting the nonzero off-diagonal elements to be one and each other element to be zero. To obtain a positive definite pre-version of the precision matrix we set

$$\Phi_{pre} := v \cdot A + (|\Lambda_{\min}(v \cdot A)| + 0.1 + u) \cdot I_{p\times p}.$$

Here $v = 0.3$ and $u = 0.1$ are chosen to control the magnitude of partial correlations. The covariance matrix Σ is generated by inverting Φ_{pre} and scaling the variances to one. The corresponding precision matrix Φ is given by Σ^{-1}. For a given p we generate $n = 200$ independent samples of

$$X = (X_1, \dots, X_p) \sim \mathcal{N}(0, \Sigma)$$

and evaluate whether our test statistic would reject the null hypothesis for a specific set of edges $\mathcal{M}$ which satisfies the null hypothesis. Finally the acceptance rate is calculated over $l = 1000$ independent simulations for a given confidence level $1 - \alpha = 0.95$.

6.1. *Simulation settings.* In our simulation study we estimate the correlation structure of four different designs that are described in the following.

6.1.1. *Example 1: Random Graph.* Each pair of off-diagonal elements of the covariance matrix of the first $p-1$ regressors is randomly set to non-zero with probability $prob = 5/p$. The last regressor is added as an independent random variable. It results in about $(p-1) \cdot (p-2) \cdot prob/2$ edges in the graph. The corresponding precision matrix is of the form

$$\Phi := \begin{pmatrix} & & 0 \\ & B & \vdots \\ & & 0 \\ 0 \cdots 0 & & 1 \end{pmatrix}$$

where B is a sparse matrix. We test the hypothesis, whether the last regressor is independent from all other regressors, corresponding to

$$\mathcal{M} = \{(p, 1), \ldots, (p, p-1)\}.$$

6.1.2. *Example 2: Cluster Graph.* The regressors are evenly partitioned into $g = 4$ disjoint groups. Each pair of off-diagonal elements $\Phi_{(i,j)}$ is set non-zero with probability $prob = 5/p$, if both i and j belong to the same group. It results in about $g \cdot (p/g) \cdot (p/g - 1) \cdot prob/2$ edges in the graph. The precision Matrix is of the form

$$\Phi := \begin{pmatrix} B_1 & & & 0 \\ & B_2 & & \\ & & B_3 & \\ 0 & & & B_4 \end{pmatrix}$$

where each block B_i is a sparse matrix. We test the hypothesis that the first two hubs are conditionally independent. This corresponds to testing the tuples

$$\mathcal{M} = \{(1, p/4+1), \ldots, (1, p/2), (2, p/4+1), \ldots, (p/4, p/2)\}.$$

(a) Random Graph (b) Cluster Graph

Fig 1: Examples

The edges of the graph are colored in black and the edges contained in the hypothesis in red.

6.1.3. *Example 3: Approximately Sparse Random Graph.* In this example we generate a random graph structure as in example 1, but instead of setting the other elements of the adjacency matriy A to zero we generate independent random entries from a uniform distribution on $[-a, a]$ with $a = 1/20$. This results in a precision matrix of the form

$$\Phi := \begin{pmatrix} & & 0 \\ & B & \vdots \\ & & 0 \\ 0 \cdots & 0 & 1 \end{pmatrix}$$

where B is not a sparse matrix anymore. We then again test the hypothesis, whether the last regressor is independent from all other regressors, corresponding to

$$\mathcal{M} = \{(p, 1), \ldots, (p, p-1)\}.$$

6.1.4. *Example 4: Independent Graph.* By setting

$$\Phi := I_{p \times p}$$

we generate samples of p independent normal distributed random variables. We can test the hypothesis whether the regressors are independent by choosing

$$\mathcal{M} = \{(1, 2), \ldots, (1, p), (2, 3), \ldots, (p-1, p)\}.$$

6.2. *Simulation results.* We provide simulated acceptance rates of our proposed estimation procedure with $B = 1000$ bootstrap samples for all of the examples above. Confidence Intervall I corresponds to the standard case in (4.2), whereas Confidence Intervall II is based on the approximation

of the sphere in (4.3). In summary, the results reveal that the empirical acceptance rate is, on average, close to the nominal level of 95% with a mean absolute deviation of 2.581% over all simulations. The Confidence Intervall II has got a mean absolute deviation of 1.875% and performs significantly better than Confidence Intervall I with a mean absolute deviation of 3.287%. More complex S-sparsely convex sets seem to result in better acceptance rates, whereas higher exponents do not improve the rates. The lowest mean absolute deviation (1.138%) is achieved in table 2 for $S = 5$, $exp = 1$ and without cross-fitting.

			Confidence Interval I			Confidence Intervall II		
Model	p	d	lasso	post-lasso	sqrt-lasso	lasso	post-lasso	sqrt-lasso
random	20	19	0.931	0.938	0.936	0.929	0.930	0.935
	50	49	0.915	0.915	0.916	0.926	0.929	0.932
	100	99	0.912	0.912	0.908	0.927	0.930	0.929
cluster	20	25	0.916	0.942	0.918	0.915	0.930	0.921
	40	100	0.916	0.919	0.917	0.934	0.947	0.937
	60	225	0.897	0.893	0.899	0.921	0.922	0.927
approx	20	19	0.931	0.931	0.931	0.947	0.946	0.947
	50	49	0.908	0.908	0.908	0.920	0.920	0.920
	100	99	0.902	0.902	0.902	0.935	0.935	0.935
indepent	5	10	0.931	0.931	0.931	0.933	0.933	0.933
	10	45	0.927	0.927	0.927	0.937	0.937	0.937
	20	190	0.896	0.896	0.896	0.920	0.920	0.920

TABLE 1

Simulation results for S=1,exp=1 and 1-fold

			Confidence Interval I			Confidence Intervall II		
Model	p	d	lasso	post-lasso	sqrt-lasso	lasso	post-lasso	sqrt-lasso
	20	19	0.969	0.925	0.956	0.951	0.932	0.947
random	50	49	0.942	0.944	0.944	0.942	0.954	0.953
	100	99	0.934	0.941	0.940	0.950	0.949	0.952
	20	25	0.972	0.958	0.973	0.914	0.936	0.914
cluster	40	100	0.941	0.937	0.945	0.930	0.936	0.942
	60	225	0.931	0.947	0.942	0.943	0.937	0.950
	20	19	0.958	0.958	0.958	0.965	0.965	0.965
approx	50	49	0.937	0.937	0.937	0.940	0.940	0.940
	100	99	0.920	0.921	0.920	0.936	0.936	0.936
	5	10	0.951	0.951	0.951	0.951	0.951	0.951
indepent	10	45	0.932	0.932	0.932	0.952	0.952	0.952
	20	190	0.926	0.926	0.926	0.947	0.947	0.947

TABLE 2
Simulation results for S=5,exp=1 and 1-fold

			Confidence Interval I			Confidence Intervall II		
Model	p	d	lasso	post-lasso	sqrt-lasso	lasso	post-lasso	sqrt-lasso
	20	19	0.909	0.916	0.921	0.916	0.921	0.930
random	50	49	0.931	0.910	0.926	0.926	0.907	0.927
	100	99	0.907	0.909	0.909	0.917	0.934	0.923
	20	25	0.910	0.905	0.905	0.904	0.898	0.901
cluster	40	100	0.909	0.910	0.910	0.905	0.919	0.921
	60	225	0.885	0.894	0.898	0.912	0.925	0.934
	20	19	0.929	0.928	0.929	0.929	0.928	0.929
approx	50	49	0.888	0.888	0.888	0.911	0.911	0.911
	100	99	0.907	0.907	0.907	0.936	0.936	0.936
	5	10	0.930	0.930	0.930	0.939	0.939	0.939
indepent	10	45	0.921	0.921	0.921	0.933	0.933	0.933
	20	190	0.916	0.916	0.916	0.938	0.938	0.938

TABLE 3
Simulation results for S=5,exp=2 and 1-fold

			Confidence Interval I			Confidence Intervall II		
Model	p	d	lasso	post-lasso	sqrt-lasso	lasso	post-lasso	sqrt-lasso
	20	19	0.917	0.912	0.919	0.919	0.932	0.918
random	50	49	0.927	0.911	0.925	0.938	0.936	0.938
	100	99	0.903	0.894	0.907	0.926	0.933	0.927
	20	25	0.920	0.899	0.918	0.930	0.929	0.929
cluster	40	100	0.920	0.883	0.919	0.927	0.926	0.923
	60	225	0.889	0.885	0.896	0.920	0.930	0.928
	20	19	0.921	0.922	0.921	0.932	0.934	0.932
approx	50	49	0.899	0.899	0.899	0.926	0.926	0.926
	100	99	0.889	0.889	0.889	0.930	0.929	0.930
	5	10	0.922	0.923	0.922	0.935	0.934	0.935
indepent	10	45	0.905	0.905	0.905	0.937	0.937	0.937
	20	190	0.903	0.903	0.903	0.936	0.936	0.936

TABLE 4
Simulation results for S=1,exp=1 and 3-fold

			Confidence Interval I			Confidence Intervall II		
Model	p	d	lasso	post-lasso	sqrt-lasso	lasso	post-lasso	sqrt-lasso
	20	19	0.970	0.919	0.964	0.950	0.932	0.958
random	50	49	0.923	0.911	0.927	0.938	0.951	0.935
	100	99	0.929	0.925	0.930	0.949	0.940	0.948
	20	25	0.971	0.970	0.971	0.915	0.931	0.915
cluster	40	100	0.926	0.915	0.925	0.925	0.917	0.924
	60	225	0.923	0.925	0.926	0.917	0.939	0.930
	20	19	0.959	0.959	0.959	0.958	0.956	0.958
approx	50	49	0.932	0.932	0.932	0.931	0.933	0.931
	100	99	0.929	0.929	0.929	0.949	0.950	0.949
	5	10	0.940	0.940	0.940	0.951	0.951	0.951
indepent	10	45	0.922	0.922	0.922	0.938	0.938	0.938
	20	190	0.930	0.930	0.930	0.938	0.938	0.938

TABLE 5
Simulation results for S=5,exp=1 and 3-fold

			Confidence Interval I			Confidence Intervall II		
Model	p	d	lasso	post-lasso	sqrt-lasso	lasso	post-lasso	sqrt-lasso
random	20	19	0.914	0.897	0.918	0.922	0.921	0.923
	50	49	0.914	0.896	0.911	0.920	0.920	0.921
	100	99	0.891	0.878	0.893	0.918	0.909	0.917
cluster	20	25	0.885	0.882	0.888	0.900	0.896	0.901
	40	100	0.880	0.877	0.879	0.898	0.910	0.907
	60	225	0.886	0.884	0.897	0.915	0.921	0.932
approx	20	19	0.931	0.930	0.931	0.938	0.937	0.938
	50	49	0.914	0.913	0.914	0.932	0.933	0.932
	100	99	0.894	0.894	0.894	0.924	0.924	0.924
indepent	5	10	0.923	0.922	0.923	0.943	0.942	0.943
	10	45	0.917	0.916	0.917	0.934	0.935	0.934
	20	190	0.890	0.890	0.890	0.932	0.932	0.932

TABLE 6

Simulation results for S=5,exp=2 and 3-fold

APPENDIX A: PROOF OF THEOREM 1

PROOF. We want to use corollary 2.2 from Belloni et al. (2018) [2]. Consequently, we will show that their assumptions 2.1-2.4 and the growth conditions of corollary 2.2 hold by modifying the proof of corollary 3.2 in [2]. To make the proof more comparable we try to keep the notation as similar as possible. This implies that we use C for a strictly positive constant, independent of n and r, which may have a different value in each appearance. The notation $a_n \lesssim b_n$ stands for $a_n \leq Cb_n$ for all n for some fixed C. Additionally $a_n = o(1)$ stands for uniform convergence towards zero meaning there exists sequence $(b_n)_{n\geq 1}$ with $|a_n| \leq b_n$, b_n is independent of $P \in \mathcal{P}_n$ for all n and $b_n \to 0$. Finally, the notation $a_n \lesssim_P b_n$ means that for any $\epsilon > 0$, there exists C such that uniformly over all n we have $P_P(a_n > Cb_n) \leq \epsilon$. Let $m_r = (j,k)$ be an arbitrary set in $\mathcal{M}$. We have

$$\max_r \mathbb{E}\left[\left(\nu^{(m_r)}\right)^2\right] \lesssim 1 \text{ and } \max_r \mathbb{E}\left[\left(\varepsilon^{(m_r)}\right)^2\right] \lesssim 1$$

due to the assumptions A3 and A4. Define the convex set

$$T_{m_r} = \{\eta = (\eta^{(1)}, \eta^{(2)}) : \eta^{(1)} \in \mathbb{R}^{p-2}, \eta^{(2)} \in \mathbb{R}^{p-2}\}$$

and endow T_{m_r} with the norm

$$||\eta||_e = ||\eta^{(1)}||_2 \vee ||\eta^{(2)}||_2.$$

Further let $\tau_n := \sqrt{\frac{s\log(a_n)}{n}}$ and define the nuisance realization set

$$\begin{aligned}\mathcal{T}_{m_r} = \Big\{ & \eta \in T_{m_r} : ||\eta^{(1)}||_0 \vee ||\eta^{(2)}||_0 \leq Cs, \\ & ||\eta^{(1)} - \beta^{(m_r)}||_2 \vee ||\eta^{(2)} - \gamma^{(m_r)}||_2 \leq C\tau_n, \\ & ||\eta^{(1)} - \beta^{(m_r)}||_1 \vee ||\eta^{(2)} - \gamma^{(m_r)}||_1 \leq C\sqrt{s}\tau_n \Big\} \cup \left\{\left(\beta^{(m_r)}, \gamma^{(m_r)}\right)\right\}\end{aligned}$$

for a sufficiently large constant $C > 0$. First we verify Assumption 2.1 (i). The moment condition holds since

$$\begin{aligned} & \mathbb{E}[\psi_{m_r}(X, \theta_{m_r}, \eta_{m_r})] \\ = \; & \mathbb{E}[\varepsilon^{(m_r)}\nu^{(m_r)}] \\ = \; & \mathbb{E}[\mathbb{E}[\varepsilon^{(m_r)}\nu^{(m_r)}|X_{-j}]] = \mathbb{E}[\nu^{(m_r)} \underbrace{\mathbb{E}[\varepsilon^{(m_r)}|X_{-j}]}_{=0}] = 0. \end{aligned}$$

In addition, we have

$$\begin{aligned} S_n :&= \mathbb{E}\left[\max_r |\sqrt{n}\mathbb{E}_n[\psi_{m_r}(X,\theta_{m_r},\eta_{m_r})]|\right] \\ &= \mathbb{E}\left[\sup_{f\in\mathcal{F}} \mathbb{G}_n(f)\right] \end{aligned}$$

with $\mathcal{F} = \{\varepsilon^{(m_r)}\nu^{(m_r)} | r = 1,\dots,d\}$ and $\mathbb{G}_n(f) := \sqrt{n}|\mathbb{E}_n[f] - \mathbb{E}[f]|$. By the same arguments as in the beginning of proof of theorem 2 we conclude that the envelope $\sup_{f\in\mathcal{F}} |f|$ of $\mathcal{F}$ fulfills

$$\begin{aligned} ||\max_r |\varepsilon^{(m_r)}\nu^{(m_r)}|||_{P,q} &= \mathbb{E}\left[\max_r \left(|\varepsilon^{(m_r)}\nu^{(m_r)}|\right)^q\right]^{1/q} \\ &\le \mathbb{E}\left[\max_r \left(|\varepsilon^{(m_r)}|\right)^{2q}\right]^{1/2q} \mathbb{E}\left[\max_r \left(|\nu^{(m_r)}|\right)^{2q}\right]^{1/2q} \\ &\le C\log(d), \end{aligned}$$

since the error terms are normal distributed. Using lemma O.2 (Maximal Inequality I) in [2] with $|\mathcal{F}| = d$, we have

$$S_n \le C\log^{1/2}(d) + C\log^{1/2}(d)\left(n^{\frac{2}{q}}\frac{\log^3(d)}{n}\right)^{1/2} \lesssim \log^{1/2}(d)$$

by the assumption A2 for a $q > 2\tilde{q}$. Hence, assumption A3 implies that for all $r = 1,\dots,d$, Θ_{m_r} contains an interval of radius $Cn^{-\frac{1}{2}}S_n\log(n)$ centered at θ_{m_r} for all sufficiently large n for any constant C. Assumption 2.1 (i) follows.

For all $m_r \in \mathcal{M}$, the map $(\theta,\eta) \mapsto \psi_{m_r}(X,\theta,\eta)$ is twice continuously Gateaux-differentiable on $\Theta_{m_r}\times\mathcal{T}_{m_r}$, and so is the map $(\theta,\eta) \mapsto \mathbb{E}[\psi_{m_r}(X,\theta,\eta)]$. Further we have

$$\begin{aligned} D_{m_r,0}[\eta,\eta_{m_r}] :&= \partial_t \mathbb{E}[\psi_{m_r}(X,\theta_{m_r},\eta_{m_r} + t(\eta-\eta_{m_r}))]\big|_{t=0} \\ &= \mathbb{E}\Big[\partial_t\Big\{\left(X_j - \theta_{m_r}X_k - \left(\eta^{(1)}_{m_r} + t(\eta^{(1)} - \eta^{(1)}_{m_r})\right)X_{-m_r}\right) \\ &\qquad \left(X_k - \left(\eta^{(2)}_{m_r} + t(\eta^{(2)} - \eta^{(2)}_{m_r})\right)X_{-m_r}\right)\Big\}\Big]\big|_{t=0} \\ &= \mathbb{E}[\varepsilon^{(m_r)}(\eta^{(2)}_{m_r} - \eta^{(2)})X_{-m_r}] + \mathbb{E}[(\eta^{(1)}_{m_r} - \eta^{(1)})X_{-m_r}\nu^{(m_r)}] \\ &= 0. \end{aligned}$$

Therefore, Assumptions 2.1 (ii) and 2.1 (iii) hold. Remark that

$$\begin{aligned}|J_{m_r}| &= |\partial_\theta \mathbb{E}[\psi_{m_r}(X,\theta,\eta_{m_r})]|_{\theta=\theta_{m_r}}| \\ &= |\mathbb{E}[-X_k \nu^{(m_r)}]|| = |\mathbb{E}[(\nu^{(m_r)})^2]| \leq C\end{aligned}$$

and

$$|J_{m_r}| = |\mathbb{E}[(\nu^{(m_r)})^2]| \geq c$$

due to assumption A4. Since the score ψ is linear with respect to θ, we have for all $m_r \in \mathcal{M}$ and $\theta \in \Theta_{m_r}$

$$\mathbb{E}[\psi_{m_r}(X,\theta,\eta_{m_r})] = J_{m_r}(\theta - \theta_{m_r})$$

using the moment condition. This gives us Assumption 2.1 (iv).
For all $t \in [0,1)$, $m_r \in \mathcal{M}$, $\theta \in \Theta_{m_r}$, $\eta \in \mathcal{T}_{m_r}$ we have

$$\begin{aligned}&\mathbb{E}\big[(\psi_{m_r}(X,\theta,\eta) - \psi_{m_r}(X,\theta_{m_r},\eta_{m_r}))^2\big] \\ &= \mathbb{E}\big[(\psi_{m_r}(X,\theta,\eta) - \psi_{m_r}(X,\theta_{m_r},\eta) + \psi_{m_r}(X,\theta_{m_r},\eta) - \psi_{m_r}(X,\theta_{m_r},\eta_{m_r}))^2\big] \\ &\leq C\Bigg(\underbrace{\mathbb{E}\big[(\psi_{m_r}(X,\theta,\eta) - \psi_{m_r}(X,\theta_{m_r},\eta))^2\big]}_{=:I} \\ &\qquad \vee \underbrace{\mathbb{E}\big[(\psi_{m_r}(X,\theta_{m_r},\eta) - \psi_{m_r}(X,\theta_{m_r},\eta_{m_r}))^2\big]}_{=:II}\Bigg)\end{aligned}$$

with

$$\begin{aligned}I &= |\theta - \theta_{m_r}|^2 \mathbb{E}\left[\left(X_k(X_k - \eta^{(2)} X_{-m_r})\right)^2\right] \\ &\leq |\theta - \theta_{m_r}|^2 \left(\mathbb{E}[X_k^2] E[(X_k - \eta^{(2)} X_{-m_r})^2]\right)^{1/2} \\ &\leq C|\theta - \theta_{m_r}|^2\end{aligned}$$

due to assumptions A3, A4 and the definition of $\mathcal{T}_{m_r}$. Additionally we have

$$\begin{aligned} II &= \mathbb{E}\Big[\Big(\big(X_j - \theta_{m_r} X_k - \eta^{(1)} X_{-m_r}\big)\big(X_k - \eta^{(2)} X_{-m_r}\big) \\ &\qquad - \big(X_j - \theta_{m_r} X_k - \eta^{(1)}_{m_r} X_{-m_r}\big)\big(X_k - \eta^{(2)}_{m_r} X_{-m_r}\big)\Big)^2\Big] \\ &= \mathbb{E}\Big[\Big(\big(X_j - \theta_{m_r} X_k - \eta^{(1)} X_{-m_r}\big)\big((\eta^{(2)}_{m_r} - \eta^{(2)}) X_{-m_r}\big) \\ &\qquad + \big(X_k - \eta^{(2)}_{m_r} X_{-m_r}\big)\big((\eta^{(1)}_{m_r} - \eta^{(1)}) X_{-m_r}\big)\Big)^2\Big] \\ &\leq C\Big(\|\eta^{(2)}_{m_r} - \eta^{(2)}\|_2 \vee \|\eta^{(1)}_{m_r} - \eta^{(1)}\|_2\Big)^2 \\ &= C\|\eta_{m_r} - \eta\|_e^2 \end{aligned}$$

with similar arguments as in I above using

$$\sup_{\|\xi\|_2=1} \mathbb{E}\left[(\xi X)^4\right] \leq C$$

due to the normal distributed design. Combining these results gives us Assumption 2.1 (v) (a).
Observe that

$$\begin{aligned} &\Big|\partial_t \mathbb{E}\Big[\psi_{m_r}\big(X, \theta, \eta_{m_r} + t(\eta - \eta_{m_r})\big)\Big]\Big| \\ &= \Big|\mathbb{E}\Big[\Big(X_j - \theta X_k - \big(\eta^{(1)}_{m_r} + t(\eta^{(1)} - \eta^{(1)}_{m_r})\big) X_{-m_r}\Big)\big((\eta^{(2)}_{m_r} - \eta^{(2)}) X_{-m_r}\big) \\ &\qquad + \Big(X_k - \big(\eta^{(2)}_{m_r} + t(\eta^{(2)} - \eta^{(2)}_{m_r})\big) X_{-m_r}\Big)\big((\eta^{(1)}_{m_r} - \eta^{(1)}) X_{-m_r}\big)\Big]\Big| \\ &\leq C\|\eta_{m_r} - \eta\|_e \end{aligned}$$

with the same argument as above, which gives us Assumption 2.1 (v) (b) with $B_{1n} = C$. To complete the Assumption 2.1 (v) (c) with $B_{2n} = C$

observe that

$$
\begin{aligned}
&\Big|\partial_t^2 \mathbb{E}\Big[\psi_{m_r}\big(X, \theta_{m_r} + t(\theta - \theta_{m_r}), \eta_{m_r} + t(\eta - \eta_{m_r})\big)\Big]\Big| \\
=&\Big|\partial_t \mathbb{E}\Big[\Big(X_j - \big(\theta_{m_r} + t(\theta - \theta_{m_r})\big)X_k - \big(\eta_{m_r}^{(1)} + t(\eta^{(1)} - \eta_{m_r}^{(1)})\big)X_{-m_r}\Big) \\
&\qquad \cdot \big((\eta_{m_r}^{(2)} - \eta^{(2)})X_{-m_r}\big) \\
&\qquad + \Big(X_k - \big(\eta_{m_r}^{(2)} + t(\eta^{(2)} - \eta_{m_r}^{(2)})\big)X_{-m_r}\Big) \\
&\qquad \cdot \big((\theta_{m_r} - \theta)X_k + (\eta_{m_r}^{(1)} - \eta^{(1)})X_{-m_r}\big)\Big]\Big| \\
=&\Big|2\mathbb{E}\Big[\big((\eta_{m_r}^{(2)} - \eta^{(2)})X_{-m_r}\big)\big((\theta_{m_r} - \theta)X_k + (\eta_{m_r}^{(1)} - \eta^{(1)})X_{-m_r}\big)\Big]\Big| \\
\leq\ & 2\Bigg(\underbrace{\mathbb{E}\Big[\big((\eta_{m_r}^{(2)} - \eta^{(2)})X_{-m_r}\big)^2\Big]}_{\leq C\|\eta_{m_r}^{(2)} - \eta^{(2)}\|_2^2}\underbrace{\mathbb{E}\Big[\big((\theta_{m_r} - \theta)X_k + (\eta_{m_r}^{(1)} - \eta^{(1)})X_{-m_r}\big)^2\Big]}_{\leq C\left(|\theta_{m_r} - \theta|^2 + \|\eta_{m_r}^{(1)} - \eta^{(1)}\|_2^2\right)}\Bigg)^{1/2} \\
\leq\ & C\big(|\theta_{m_r} - \theta|^2 \vee \|\eta_{m_r} - \eta\|_e^2\big).
\end{aligned}
$$

Therefore Assumption 2.1 holds. Due to the construction of $\mathcal{T}_{m_r}$ Assumptions 2.2 (ii) and (iii) hold. Next, we show that the assumptions of theorem 2 from section B hold which implies Assumption 2.2 (i). Remark that conditions B1 and B4 are satisfied with $\rho = 2$. Condition A1 implies condition B3. Let $\underline{\sigma}^2 > 0$ be a uniform lower bound for the variances of the error terms and the regressors and let $c := \underline{\sigma} z_{\tilde{c}}$, where $z_{\tilde{c}}$ is the $\tilde{c}$-quantile of a standard normal distribution for an arbitrary but fixed $\tilde{c} \in (\frac{1}{2}, \frac{3}{4})$. Uniformly for all $r = 1, \ldots, d$ and $l \in \{1, \ldots, p\} \setminus \{j\}$, it holds

$$
\begin{aligned}
P\left((\varepsilon^{(m_r)})^2 X_l^2 \geq c^4\right) &= 1 - P\left(|\varepsilon^{(m_r)} X_l| \leq c^2\right) \\
&\geq 1 - P\left(|\varepsilon^{(m_r)}| \leq c \vee |X_l| \leq c\right) \\
&\geq 1 - \left(P\left(|\varepsilon^{(m_r)}| \leq c\right) + P\left(|X_l| \leq c\right)\right) \\
&\geq 1 - 2P\left(\underline{\sigma}|Z| \leq c\right) \\
&= 3 - 4\tilde{c} > 0
\end{aligned}
$$

where $Z \sim \mathcal{N}(0, 1)$, which implies that

$$
\min_r \min_l \mathbb{E}[(\varepsilon^{(m_r)})^2 X_l^2] \geq c^4(3 - 4\tilde{c}) > 0.
$$

Analogously

$$
\min_r \min_l \mathbb{E}[(\nu^{(m_r)})^2 X_l^2] > 0.
$$

Combined with condition A4 this implies condition B2. Therefore we are able to estimate the nuisance parameters at a sufficiently fast rate.
Define

$$\mathcal{F}_1 := \Big\{\psi_{m_r}(\cdot,\theta,\eta) : r \in \{1,\dots,d\}, \theta \in \Theta_{m_r}, \eta \in \mathcal{T}_{m_r}\Big\}.$$

To bound the covering entropy of $\mathcal{F}_1$ we at first exclude the true nuisance parameter and define

$$\mathcal{F}_{1,1} := \Big\{\psi_{m_r}(\cdot,\theta,\eta) : r \in \{1,\dots,d\}, \theta \in \Theta_{m_r}, \eta \in \mathcal{T}_{m_r} \setminus \{\eta_{m_r}\}\Big\} \subseteq \mathcal{F}_{1,1}^{(1)}\mathcal{F}_{1,1}^{(2)}$$

with

$$\mathcal{F}_{1,1}^{(1)} = \{X \to (X_j - \theta X_k - \eta^{(1)} X_{-m_r}) : r \in \{1,\dots,d\}, \theta \in \Theta_{m_r}, \eta^{(1)} \in \mathcal{T}_{m_r,1}^*\}$$
$$\mathcal{F}_{1,1}^{(2)} = \{X \to (X_k - \eta^{(2)} X_{-m_r}) : r \in \{1,\dots,d\}, \eta^{(2)} \in \mathcal{T}_{m_r,2}^*\}$$

where $\mathcal{T}_{m_r}^* := \mathcal{T}_{m_r} \setminus \{\eta_{m_r}\}$. Observe that the envelope $F_{1,1}^{(1)}$ of $\mathcal{F}_{1,1}^{(1)}$ fulfills

$$\begin{aligned}
\big\|\big(F_{1,1}^{(1)}\big)^2\big\|_{P,2q} &\le \Big\| \sup_{r\in\{1,\dots,d\}} \sup_{\theta\in\Theta_{m_r}, \|\eta_{m_r}^{(1)} - \eta^{(1)}\|_1 \le C\sqrt{s}\tau_n} \Big(|\varepsilon^{(m_r)}| \\
&\qquad + |(\theta_{m_r} - \theta)X_k| + |(\eta_{m_r}^{(1)} - \eta^{(1)})X_{-m_r}|\Big)^2 \Big\|_{P,2q} \\
&\lesssim \big\| \sup_{r\in\{1,\dots,d\}} \big(\varepsilon^{(m_r)}\big)^2 \big\|_{P,2q} + \big\| \sup_{r\in\{1,\dots,d\}} X_k^2 \big\|_{P,2q} \\
&\qquad + s\tau_n^2 \big\| \sup_{r\in\{1,\dots,d\}} \|X_{-m_r}\|_\infty^2 \big\|_{P,2q} \\
&\lesssim \log(d) + \log(d) + s\tau_n^2 \log(a_n) \\
&\lesssim \log(a_n)
\end{aligned}$$

and with an analogous argument

$$\big\|\big(F_{1,1}^{(2)}\big)^2\big\|_{P,2q} \lesssim \log(a_n).$$

Since we excluded the true nuisance parameter, which does not need to be sparse, we have $\mathcal{F}_{1,1}^{(1)} \subseteq \mathcal{G}_{1,1}$ and $\mathcal{F}_{1,1}^{(2)} \subseteq \mathcal{G}_{1,1}$ with

$$\mathcal{G}_{1,1} := \Big\{X \to \xi X : \xi \in \mathbb{R}^p, \|\xi\|_0 \le Cs, \|\xi\|_2 \le C\Big\}$$

where $\mathcal{G}_{1,1}$ is a union over $\binom{p}{Cs}$ VC-subgraph classes $\mathcal{G}_{1,1,k}$ with VC indices less or equal to $Cs+2$ (Lemma 2.6.15, Van der Vaart and Wellner (1996)[17]).

This implies that $\mathcal{F}_{1,1}^{(1)}$ and $\mathcal{F}_{1,1}^{(2)}$ are unions over $\binom{p}{Cs}$ VC-subgraph classes $\mathcal{F}_{1,1,k}^{(1)}$ and $\mathcal{F}_{1,1,k}^{(2)}$ with VC indices less or equal to $Cs+2$.
Due to theorem 2.6.7 in [17] we obtain

$$\begin{aligned}
&\sup_Q \log N(\varepsilon \|F_{1,1}^{(1)}\|_{Q,2}, \mathcal{F}_{1,1}^{(1)}, \|\cdot\|_{Q,2}) \\
\leq &\sup_Q \log \left(\sum_{k=1}^{\binom{p}{Cs}} N(\varepsilon \|F_{1,1}^{(1)}\|_{Q,2}, \mathcal{F}_{1,1,k}^{(1)}, \|\cdot\|_{Q,2}) \right) \\
\leq &\log \Bigg(\underbrace{\binom{p}{Cs}}_{\leq \left(\frac{e\cdot p}{Cs}\right)^{Cs}} K(Cs+2)(16e)^{Cs+2} \left(\frac{1}{\varepsilon}\right)^{2Cs+2} \Bigg) \\
\leq &\log \left(\left(\frac{e\cdot p}{Cs}\right)^{Cs} K(Cs+2)(16e)^{Cs+2} \left(\frac{1}{\varepsilon}\right)^{2Cs+2} \right) \\
\lesssim &\, s \log\left(\frac{a_n}{\varepsilon}\right)
\end{aligned}$$

where K is an universal constant and with an analogous argument

$$\sup_Q \log N(\varepsilon \|F_{1,1}^{(2)}\|_{Q,2}, \mathcal{F}_{1,1}^{(2)}, \|\cdot\|_{Q,2}) \lesssim s \log\left(\frac{a_n}{\varepsilon}\right).$$

Using basic calculations on covering entropies (see for example Appendix N Lemma N.1 from Belloni et al. (2014) [3]) we can bound the covering entropy of the class $\mathcal{F}_{1,1}$ by

$$\begin{aligned}
&\sup_Q \log N(\varepsilon \|F_{1,1}^{(1)} F_{1,1}^{(2)}\|_{Q,2}, \mathcal{F}_{1,1}, \|\cdot\|_{Q,2}) \\
\leq \quad &\sup_Q \log N\Big(\frac{\varepsilon}{2} \|F_{1,1}^{(1)}\|_{Q,2}, \mathcal{F}_{1,1}^{(1)}, \|\cdot\|_{Q,2}\Big) \\
&\quad + \sup_Q \log N\Big(\frac{\varepsilon}{2} \|F_{1,1}^{(2)}\|_{Q,2}, \mathcal{F}_{1,1}^{(2)}, \|\cdot\|_{Q,2}\Big) \\
\lesssim \quad &s \log\left(\frac{a_n}{\varepsilon}\right)
\end{aligned}$$

where $F_{1,1} := F_{1,1}^{(1)} F_{1,1}^{(2)}$ is an envelope for $\mathcal{F}_{1,1}$ with

$$\|F_{1,1}\|_{P,q} \leq \left(\big\| \big(F_{1,1}^{(1)}\big)^2 \big\|_{P,2q} \big\| \big(F_{1,1}^{(1)}\big)^2 \big\|_{P,2q} \right)^{1/2} \lesssim \log(a_n).$$

Additionally define

$$\mathcal{F}_{1,2} := \Big\{ \psi_{m_r}(\cdot, \theta, \eta_{m_r}) : r \in \{1, \ldots, d\}, \theta \in \Theta_{m_r} \Big\}.$$

With the same argument as above $\mathcal{F}_{1,2}$ is a union over d VC-subgraph classes with VC indices less or equal to 3 implying

$$\sup_Q \log N(\varepsilon \|F_{1,2}\|_{Q,2}, \mathcal{F}_{1,2}, \|\cdot\|_{Q,2}) \leq C \log\left(\frac{d}{\varepsilon}\right) \lesssim \log\left(\frac{a_n}{\varepsilon}\right)$$

where the envelope $F_{1,2}$ of $\mathcal{F}_{1,2}$ obeys

$$\|F_{1,2}\|_{P,q} \lesssim \log(a_n)$$

with an analogous argument as above. Combining these results we obtain

$$\begin{aligned}
&\sup_Q \log N(\varepsilon \|F_1\|_{Q,2}, \mathcal{F}_1, \|\cdot\|_{Q,2}) \\
= \quad &\sup_Q \log N(\varepsilon \|F_{1,1}^{(1)} F_{1,1}^{(2)} \vee F_{1,2}\|_{Q,2}, \mathcal{F}_{1,1} \cup \mathcal{F}_{1,2}, \|\cdot\|_{Q,2}) \\
\leq \quad &\sup_Q \log N(\varepsilon \|F_{1,1}^{(1)} F_{1,1}^{(2)}\|_{Q,2}, \mathcal{F}_{1,1}, \|\cdot\|_{Q,2}) \\
&\quad + \sup_Q \log N(\varepsilon \|F_{1,2}\|_{Q,2}, \mathcal{F}_{1,2}, \|\cdot\|_{Q,2}) \\
\lesssim \quad &s \log\left(\frac{a_n}{\varepsilon}\right)
\end{aligned}$$

where the envelope $F_1 := F_{1,1}^{(1)} F_{1,1}^{(2)} \vee F_{1,2}$ of $\mathcal{F}_1$ satisfies

$$\|F_1\|_{P,q} \lesssim \log(a_n)$$

which gives us Assumption 2.2 (iv). Observe that for all $f \in \mathcal{F}_1$ we have

$$\begin{aligned}
\mathbb{E}[f^2]^{1/2} &\leq \sup_{r,\theta,\eta^{(1)}} \mathbb{E}\left[(X_j - \theta X_k - \eta^{(1)} X_{-m_r})^4\right]^{1/4} \sup_{r,\eta^{(2)}} \mathbb{E}\left[(X_k - \eta^{(2)} X_{-m_r})^4\right]^{1/4} \\
&\lesssim \sup_{\|\xi\|_2=1} \mathbb{E}\left[(\xi X)^4\right]^{1/2} \lesssim C
\end{aligned}$$

and

$$\mathbb{E}[f^2]^{1/2} = \mathbb{E}\Big[\underbrace{(X_j - \theta X_k - \eta^{(1)} X_{-m_r})}_{=:Z_1}{}^2 \underbrace{(X_k - \eta^{(2)} X_{-m_r})}_{=:Z_2}{}^2\Big]^{1/2}.$$

For each Z_i with $i \in \{1, 2\}$ we have

$$E[Z_i^2] \gtrsim \inf_{\|\xi\|_2=1} \mathbb{E}\left[(\xi X)^2\right] \geq c.$$

Therefore Z_1 and Z_2 are both centered normal distributed random variables where the variance is bounded away from zero. This implies

$$E[Z_1^2 Z_2^2]^{1/2} \geq c > 0$$

which gives us Assumption 2.2 (v).
Assumption 2.2 (vi) (a) holds by construction of τ_n and $v_n \lesssim s$. Due to the growth condition A2 we can choose $q = 2\tilde{q}/(1-\kappa)$ such that

$$\begin{aligned} n^{-1/2+1/q} s \log^2(a_n) &= n^{\frac{1-\kappa}{2\tilde{q}}} n^{-1/2} s \log^2(a_n) \\ &= n^{-\frac{\kappa}{2\tilde{q}}} \left(n^{\frac{1}{\tilde{q}}} \frac{s^2 \log^4(a_n)}{n} \right)^{1/2} \lesssim n^{-\frac{\kappa}{2\tilde{q}}}. \end{aligned}$$

Additionally we have

$$C\tau_n (s \log(a_n))^{1/2} \lesssim \frac{s \log(a_n)}{\sqrt{n}} \lesssim n^{-\frac{1}{2\tilde{q}}},$$

$$\log^{1/2}(d) \frac{\log(n)}{\sqrt{n}} (s \log(a_n))^{1/2} \lesssim \sqrt{\frac{s \log^4(a_n)}{n}} \lesssim n^{-\frac{1}{2\tilde{q}}}$$

and

$$n^{1/2} \tau_n^2 = \frac{s \log(a_n)}{\sqrt{n}} \lesssim n^{-\frac{1}{2\tilde{q}}}$$

which gives us Assumption 2.2 (vi) (b) and (c) with $\delta_n = n^{-\frac{\kappa}{2\tilde{q}}}$. Define the class

$$\mathcal{F}_0 := \{\bar{\psi}_{m_r}(\cdot) : r = 1, \ldots, d\}$$

where $\bar{\psi}_{m_r}(\cdot) := -\sigma_{m_r}^{-1} J_{m_r}^{-1} \psi_{m_r}(\cdot, \theta_{m_r}, \eta_{m_r})$ with $\sigma_{m_r}^2 := J_{m_r}^{-2} \mathbb{E}[\psi_{m_r}^2(X, \theta_{m_r}, \eta_{m_r})]$. Observe that by the Cauchy-Schwarz Inequality for any $q > 0$ the envelope F_0 for $\mathcal{F}_0$ satisfies

$$\begin{aligned} \|F_0\|_{P,q} &= \mathbb{E}\left[\sup_{r=1,\ldots,d} \left(\mathbb{E}[(\varepsilon^{(m_r)} \nu^{(m_r)})^2]^{-1/2} |\varepsilon^{(m_r)} \nu^{(m_r)}| \right)^q \right]^{1/q} \\ &\lesssim \mathbb{E}\left[\sup_{r=1,\ldots,d} \left(|\varepsilon^{(m_r)} \nu^{(m_r)}| \right)^q \right]^{1/q} \\ &\lesssim \log(d). \end{aligned}$$

Since $|\mathcal{F}_0| = d$ we have

$$\sup_Q \log N\big(\varepsilon \|F_0\|_{Q,2}, \mathcal{F}_0, \|\cdot\|_{Q,2}\big) \leq \log\left(\frac{d}{\varepsilon}\right)$$

for all $< \varepsilon \leq 1$. Therefore Assumption 2.3 (i) is satisfied with $\varrho_n = 1$ and $A_n = d \vee n$. Since the errors are centered normal distributed random variables with a uniformly bounded variance we have $E\big[\big(\varepsilon^{(m_r)}\big)^8\big] \lesssim C$ and $E\big[\big(\nu^{(m_r)}\big)^8\big] \lesssim C$. This implies $\mathbb{E}[f^4] \leq C$ for all $f \in \mathcal{F}_0$ which gives us Assumption 2.3 (ii). The growth condititons from corollary 2.1 are satisfied due to Condition A2. Observe that

$$\delta_n^2 \log(n \vee d) \lesssim n^{-\frac{\kappa}{\tilde{q}}} \log(n \vee d) = o(1),$$

$$\log^{2/7}(d) \log(n \vee d) = o(n^{1/7})$$

and we can find a q such that

$$\log^{2/3}(d) \log(n \vee d) = o(n^{1/3-2/(3q)}).$$

Now, we verify Assumption 2.4. Define

$$\tilde{\psi}_{m_r}(X, \eta^{(2)}) := -X_k(X_k - \eta^{(2)} X_{-m_r})$$

and

$$\tilde{m}_{m_r}(\eta^{(2)}) := \mathbb{E}[\tilde{\psi}_{m_r}(X, \eta^{(2)})],$$

where $\hat{J}_{m_r} = -\mathbb{E}_n[\tilde{\psi}_{m_r}(X, \hat{\eta}^{(2)})]$. It holds

$$|\hat{J}_{m_r} - J_{m_r}| \leq |\hat{J}_{m_r} - \tilde{m}_{m_r}(\hat{\eta}^{(2)})| + |\tilde{m}_{m_r}(\hat{\eta}^{(2)}) - \tilde{m}_{m_r}(\eta^{(2)}_{m_r})|$$

with

$$\begin{aligned}
|\tilde{m}_{m_r}(\hat{\eta}^{(2)}) - \tilde{m}_{m_r}(\eta^{(2)}_{m_r})| &= |\mathbb{E}[X_k(\hat{\eta}^{(2)}_{m_r} - \eta^{(2)}_{m_r}) X_{-m_r}]| \\
&= ||\hat{\eta}^{(2)}_{m_r} - \eta^{(2)}_{m_r}||_2 \left| \mathbb{E}\left[X_k \left(\frac{(\hat{\eta}^{(2)}_{m_r} - \eta^{(2)}_{m_r})}{||\hat{\eta}^{(2)}_{m_r} - \eta^{(2)}_{m_r}||_2} X_{-m_r} \right) \right] \right| \\
&\lesssim ||\hat{\eta}^{(2)}_{m_r} - \eta^{(2)}_{m_r}||_2 \lesssim \tau_n.
\end{aligned}$$

Let

$$\tilde{\mathcal{G}}_1 := \{X \mapsto \tilde{\psi}_{m_r}(X, \eta^{(2)}) : r = 1, \ldots, d, \eta^{(2)} \in \mathcal{T}^*_{m_r,2}\}$$

with

$$\sup_r |\hat{J}_{m_r} - J_{m_r}| \lesssim \sup_{g \in \tilde{\mathcal{G}}_1} |\mathbb{E}_n[g(X)] - \mathbb{E}[g(X)]| + \tau_n.$$

The class $\tilde{\mathcal{G}}_1$ has an envelope $\tilde{G}_1$ with

$$\begin{aligned}
\mathbb{E}[\tilde{G}_1^q]^{1/q} &\leq \mathbb{E}\left[\sup_r \sup_{\eta^{(2)}\in\mathcal{T}^*_{m_r,2}} |X_k^q(X_k - \eta^{(2)}X_{m_r})^q|\right]^{1/q} \\
&\leq ||\sup_r X_k||_{P,2q}\mathbb{E}\left[\sup_{r,\eta^{(2)}\in\mathcal{T}^*_{m_r,2}} (X_k - \eta^{(2)}X_{m_r})^{2q}\right]^{1/2q} \\
&\lesssim \log^{\frac{1}{2}}(d)\left(||\sup_r \nu^{(m_r)}||_{P,2q} \vee \mathbb{E}\left[\sup_{r,\eta^{(2)}\in\mathcal{T}^*_{m_r,2}} ((\eta^{(2)}_{m_r} - \eta^{(2)})X_{m_r})^{2q}\right]^{1/2q}\right) \\
&\lesssim \log^{\frac{1}{2}}(d)\left(\log^{\frac{1}{2}}(d) \vee \sqrt{s}\tau_n \sup_r E\left[||X_{m_r}||_\infty^{2q}\right]^{1/2q}\right) \\
&\lesssim \log(a_n).
\end{aligned}$$

for all q. With similar arguments as in the verification of Assumption 2.2. (iv), we obtain

$$\sup_Q \log N\big(\varepsilon\|\tilde{G}_1\|_{Q,2}, \mathcal{G}_1, \|\cdot\|_{Q,2}\big) \lesssim s\log\left(\frac{a_n}{\varepsilon}\right).$$

Therefore, by Lemma O.2, it holds

$$\begin{aligned}
\sup_r |\hat{J}_{m_r} - J_{m_r}| &\lesssim K\left(\sqrt{\frac{s\log(a_n)}{n}} + n^{1/q}\frac{s\log^2(a_n)}{n}\right) + \tau_n \\
&= o\left(\log^{-\frac{3}{2}}(a_n)\right)
\end{aligned}$$

with probability not less then $1 - o(1)$. Next we want to show that

$$\mathbb{E}_n[\psi^2_{m_r}(X, \hat{\theta}_{m_r}, \hat{\eta}_{m_r})] - \mathbb{E}[\psi^2_{m_r}(X, \theta_{m_r}, \eta_{m_r})] = o_P(\log^{-1}(a_n)).$$

By the triangle inequality we have

$$\begin{aligned}
& |\mathbb{E}_n[\psi^2_{m_r}(X, \hat{\theta}_{m_r}, \hat{\eta}_{m_r})] - \mathbb{E}[\psi^2_{m_r}(X, \theta_{m_r}, \eta_{m_r})]| \\
\leq\ & |\mathbb{E}_n[\psi^2_{m_r}(X, \hat{\theta}_{m_r}, \hat{\eta}_{m_r})] - \mathbb{E}[\psi^2_{m_r}(X, \hat{\theta}_{m_r}, \hat{\eta}_{m_r})]| \\
& + |\mathbb{E}[\psi^2_{m_r}(X, \hat{\theta}_{m_r}, \hat{\eta}_{m_r}) - \psi^2_{m_r}(X, \theta_{m_r}, \eta_{m_r})]| \\
\leq\ & |\mathbb{E}_n[\psi^2_{m_r}(X, \hat{\theta}_{m_r}, \hat{\eta}_{m_r})] - \mathbb{E}[\psi^2_{m_r}(X, \hat{\theta}_{m_r}, \hat{\eta}_{m_r})]| \\
& + \mathbb{E}[(\psi_{m_r}(X, \hat{\theta}_{m_r}, \hat{\eta}_{m_r}) + \psi_{m_r}(X, \theta_{m_r}, \eta_{m_r}))^2]^{1/2} \\
& \quad \cdot \mathbb{E}[(\psi_{m_r}(X, \hat{\theta}_{m_r}, \hat{\eta}_{m_r}) - \psi_{m_r}(X, \theta_{m_r}, \eta_{m_r}))^2]^{1/2} \\
\leq\ & |\mathbb{E}_n[\psi^2_{m_r}(X, \hat{\theta}_{m_r}, \hat{\eta}_{m_r})] - \mathbb{E}[\psi^2_{m_r}(X, \hat{\theta}_{m_r}, \hat{\eta}_{m_r})]| \\
& + C(|\theta_{m_r} - \hat{\theta}_{m_r}| \vee \|\eta_{m_r} - \hat{\eta}_{m_r}\|_e)
\end{aligned}$$

due to 2.1(a) and 2.2(v). Observe that with probability $1 - o(1)$

$$\sup_r |\hat{\theta}_{m_r} - \theta_{m_r}| \lesssim \tau_n = o(\log^{-1}(a_n))$$

due to Appendix B from Belloni et al. (2018) [2]. Since the class

$$\tilde{\mathcal{G}}_2 := \left\{ \psi_{m_r}(\cdot, \theta, \eta) : r \in \{1, \dots, d\}, |\theta - \theta_{m_r}| \leq C\tau_n, \eta \in \mathcal{T}^*_{m_r} \right\} \subseteq \mathcal{F}_{1,1}$$

we obtain the same entropy bounds as for $\mathcal{F}_{1,1}$ implying

$$\sup_Q \log N(\varepsilon \|\tilde{G}_2^2\|_{Q,2}, \tilde{\mathcal{G}}_2^2, \|\cdot\|_{Q,2}) \lesssim s \log\left(\frac{a_n}{\varepsilon}\right)$$

where $\tilde{G}_2^2$ is a measurable envelope of $\tilde{\mathcal{G}}_2^2$ with

$$\begin{aligned} \|\tilde{G}_2^2\|_{P,q} &\leq \left\| \left(F_{1,1}\right)^2 \right\|_{P,q} \\ &\leq \left(\left\| \left(F_{1,1}^{(1)}\right)^4 \right\|_{P,q} \left\| \left(F_{1,1}^{(2)}\right)^4 \right\|_{P,q} \right)^{1/2} \\ &\lesssim \log^2(a_n) \end{aligned}$$

due to $\left\| \left(F_{1,1}^{(1)}\right)^4 \right\|_{P,q} \lesssim \log^2(a_n)$ and $\left\| \left(F_{1,1}^{(2)}\right)^4 \right\|_{P,q} \lesssim \log^2(a_n)$. For all $g \in \tilde{\mathcal{G}}_2^2$ we have

$$\begin{aligned} &\sup_{g \in \tilde{\mathcal{G}}_2^2} \mathbb{E}[g(X)^2]^{1/2} \\ &\leq \sup_{r, \theta, \eta^{(1)}} \mathbb{E}\Big[(X_j - \theta X_k - \eta^{(1)} X_{-m_r})^8 \Big]^{1/4} \sup_{r, \eta^{(2)}} \mathbb{E}\Big[(X_k - \eta^{(2)} X_{-m_r})^8 \Big]^{1/4} \\ &\lesssim \sup_{||\xi||_2 = 1} \mathbb{E}\Big[(\xi X)^8 \Big]^{1/2} \leq C. \end{aligned}$$

Therefore we can find a $q > 4$ such that with probability $1 - o(1)$

$$\begin{aligned} \sup_{g \in \tilde{\mathcal{G}}_2^2} |\mathbb{E}_n[g(X)] - \mathbb{E}[g(X)]| &\leq K \left(\sqrt{\frac{s \log(a_n)}{n}} + n^{1/q} \frac{s \log^3(a_n)}{n} \right) \\ &= o(\log^{-1}(a_n)) \end{aligned}$$

which implies

$$\mathbb{E}_n[\psi^2_{m_r}(X, \hat{\theta}_{m_r}, \hat{\eta}_{m_r})] - \mathbb{E}[\psi^2_{m_r}(X, \theta_{m_r}, \eta_{m_r})] = o_P(\log^{-1}(a_n)).$$

Since $1 \lesssim \sigma_{m_r}^2 \lesssim 1$ due to Assumption 2.1 (iv) and 2.2 (v), we have

$$\begin{aligned}
\left|\frac{\hat{\sigma}_{m_r}}{\sigma_{m_r}} - 1\right| &\leq \left|\frac{\hat{\sigma}_{m_r}^2}{\sigma_{m_r}^2} - 1\right| \\
&\lesssim \left|\hat{\sigma}_{m_r}^2 - \sigma_{m_r}^2\right| \\
&\leq \left|\hat{J}_{m_r}^{-2} - J_{m_r}^{-2}\right| \mathbb{E}_n[\psi_{m_r}^2(X, \hat{\theta}_{m_r}, \hat{\eta}_{m_r})] \\
&\quad + J_{m_r}^{-2} |\mathbb{E}_n[\psi_{m_r}^2(X, \hat{\theta}_{m_r}, \hat{\eta}_{m_r})] - \mathbb{E}[\psi_{m_r}^2(X, \theta_{m_r}, \eta_{m_r})]| \\
&\lesssim \left|\hat{J}_{m_r} - J_{m_r}\right| + |\mathbb{E}_n[\psi_{m_r}^2(X, \hat{\theta}_{m_r}, \hat{\eta}_{m_r})] - \mathbb{E}[\psi_{m_r}^2(X, \theta_{m_r}, \eta_{m_r})]| \\
&= o_P(\log^{-1}(a_n))
\end{aligned}$$

uniformly over all $r = 1, \ldots, d$ which gives us Assumption 2.4 with $\Delta_n = o(1)$ and $\varepsilon_n = o(\log^{-1}(a_n))$. Next, we show the Assumption 2.3 (iii). The entropy conditions of the class

$$\hat{\mathcal{F}}_0 = \{\bar{\psi}_{m_r}(\cdot) - \hat{\psi}_{m_r}(\cdot) : r = 1, \ldots, d\}$$

holds by construction with $\bar{A}_n = d \vee n$ and $\bar{\varrho} = 1$. Further it holds for all $f \in \hat{\mathcal{F}}_0$

$$\begin{aligned}
||f||_{P_n,2} &= ||\hat{\sigma}_{m_r}^{-1} \hat{J}_{m_r}^{-1} \psi_{m_r}(X, \hat{\theta}_{m_r}, \hat{\eta}_{m_r}) - \sigma_{m_r}^{-1} J_{m_r}^{-1} \psi_{m_r}(X, \theta_{m_r}, \eta_{m_r})||_{P_n,2} \\
&\leq |\hat{\sigma}_{m_r}^{-1} \hat{J}_{m_r}^{-1} - \sigma_{m_r}^{-1} J_{m_r}^{-1}| \cdot ||\psi_{m_r}(X, \theta_{m_r}, \eta_{m_r})||_{P_n,2} \\
&\quad + \hat{\sigma}_{m_r}^{-1} \hat{J}_{m_r}^{-1} ||\psi_{m_r}(X, \hat{\theta}_{m_r}, \hat{\eta}_{m_r}) - \psi_{m_r}(X, \theta_{m_r}, \eta_{m_r})||_{P_n,2} \\
&:= I + II
\end{aligned}$$

To bound the first term, observe that uniformly over all $r = 1, \ldots, d$

$$|\hat{\sigma}_{m_r}^{-1} \hat{J}_{m_r}^{-1} - \sigma_{m_r}^{-1} J_{m_r}^{-1}| = o_P(\log^{-1}(a_n))$$

since $1 \lesssim J_{m_r} \lesssim 1$ and $1 \lesssim \sigma_{m_r} \lesssim 1$. Define the class

$$\tilde{\mathcal{G}}_3 := \{\psi_{m_r}^2(\cdot, \theta_{m_r}, \eta_{m_r}) : r = 1, \ldots, d\}$$

with cardinality $|\tilde{\mathcal{G}}_3| = d$ and an envelope $\tilde{G}_3$ that fulfills

$$||\tilde{G}_3||_{P,q} \leq \mathbb{E}\left[\sup_r \left(\varepsilon^{(m_r)} \nu^{(m_r)}\right)^{2q}\right]^{1/q} \lesssim \log^2(d).$$

Remark that

$$\sup_r ||\psi_{m_r}(X, \theta_{m_r}, \eta_{m_r})||_{P_n,2} \leq \left(\frac{1}{\sqrt{n}} \sup_{g \in \tilde{G}_3} \mathbb{G}_n(g) + \sup_r \mathbb{E}[\psi_{m_r}^2(X, \theta_{m_r}, \eta_{m_r})]\right)^{\frac{1}{2}}$$

with $\sup_r \mathbb{E}[\psi_{m_r}^2(X, \theta_{m_r}, \eta_{m_r})] \leq C$ and

$$\frac{1}{\sqrt{n}} \sup_{g \in \tilde{G}_3} \mathbb{G}_n(g) \lesssim K \left(\sqrt{\frac{\log(a_n)}{n}} + n^{1/q} \frac{\log^3(a_n)}{n} \right) = o(1)$$

with probability $1 - o(1)$. This implies

$$I = o_P\left(\log^{-1}(a_n)\right)$$

uniformly over all $r = 1, \ldots, d$. To bound the second term, define the class

$$\tilde{\mathcal{G}}_4 := \{\psi_{m_r}(\cdot, \theta, \eta) - \psi_{m_r}(\cdot, \theta_{m_r}, \eta_{m_r}) : r = 1, \ldots, d, \\ |\theta - \theta_{m_r}| \leq C\tau_n, \eta \in \mathcal{T}_{m_r}\}$$

for a sufficiently large constant $C > 0$. Due to Assumption 2.2 (i) we have that

$$\psi_{m_r}(X, \hat{\theta}_{m_r}, \hat{\eta}_{m_r}) - \psi_{m_r}(X, \theta_{m_r}, \eta_{m_r}) \in \tilde{\mathcal{G}}_4$$

with probability $1 - o(1)$. Since $\tilde{\mathcal{G}}_4^2 \subseteq (\mathcal{F}_1 - \mathcal{F}_1)^2$ the covering numbers obey

$$\sup_Q \log N\big(\varepsilon \|\tilde{G}_4^2\|_{Q,2}, \tilde{\mathcal{G}}_4^2, \|\cdot\|_{Q,2}\big) \lesssim s \log\left(\frac{a_n}{\varepsilon}\right)$$

and the envelope

$$\tilde{G}_4^2 = \sup_{r=1,\ldots,d} \sup_{|\theta - \theta_{m_r}| \leq C\tau_n} \sup_{\eta \in \mathcal{T}_{m_r}} \left(\psi_{m_r}(\cdot, \theta, \eta) - \psi_{m_r}(\cdot, \theta_{m_r}, \eta_{m_r})\right)^2$$

satisfies

$$\begin{aligned}
&\|\tilde{G}_4^2\|_{P,q} \\
&\lesssim \| \sup_{r,\theta,\eta^{(2)}} \left((\theta_{m_r} - \theta) X_k (X_k - \eta^{(2)} X_{-m_r})\right)^2 \|_{P,q} \\
&\quad + \| \sup_{r,\eta^{(1)},\eta^{(2)}} \left((X_j - \theta_{m_r} X_k - \eta^{(1)} X_{-m_r})(\eta_{m_r}^{(2)} - \eta^{(2)}) X_{-m_r}\right)^2 \|_{P,q} \\
&\quad + \| \sup_{r,\eta^{(1)}} \left((X_k - \eta_{m_r}^{(2)} X_{-m_r})(\eta_{m_r}^{(1)} - \eta^{(1)}) X_{-m_r}\right)^2 \|_{P,q} \\
&:= T_1 + T_2 + T_3
\end{aligned}$$

with

$$\begin{aligned}
T_1 &\lesssim \tau_n^2 \| \sup_{r,\eta^{(2)}} \left(X_k (X_k - \eta^{(2)} X_{-m_r})\right)^2 \|_{P,q} \\
&\lesssim \tau_n^2 \| \sup_r X_k^2 \|_{P,2q} \| \sup_{r,\eta^{(2)}} (X_k - \eta^{(2)} X_{-m_r})^2 \|_{P,2q} \\
&\lesssim \frac{s \log(a_n)}{n} \log(d)^2 = o(\log^{-1}(a_n)),
\end{aligned}$$

$$
\begin{aligned}
T_2 &\lesssim \| \sup_{r,\eta^{(2)}} ((\eta^{(2)}_{m_r} - \eta^{(2)})X_{-m_r})^2 \|_{P,2q} \| \sup_{r,\eta^{(1)}} (X_j - \theta_{m_r} X_k - \eta^{(1)} X_{-m_r})^2 \|_{P,2q} \\
&\lesssim s\tau_n^2 \| \sup_r \|X_{-m_r}\|_\infty^2 \|_{P,2q} \log(d) \\
&\lesssim \frac{s^2 \log(a_n)}{n} \log(a_n) \log(d) = o(\log^{-1}(a_n))
\end{aligned}
$$

and

$$
\begin{aligned}
T_3 &\lesssim \| \sup_{r,\eta^{(1)}} ((\eta^{(1)}_{m_r} - \eta^{(1)})X_{-m_r})^2 \|_{P,2q} \| \sup_r (\nu_{m_r})^2 \|_{P,2q} \\
&\lesssim s\tau_n^2 \| \sup_r \|X_{-m_r}\|_\infty^2 \|_{P,2q} \log(d) = o(\log^{-1}(a_n)).
\end{aligned}
$$

Since

$$
\sigma := \left(\sup_{g \in \tilde{\mathcal{G}}_4^2} \mathbb{E}[g^2] \right)^{1/2} \lesssim \frac{s^2 \log(a_n)}{n} = o(\log^{-3}(a_n))
$$

it holds

$$
\begin{aligned}
\frac{1}{\sqrt{n}} \sup_{g \in \tilde{G}_4^2} \mathbb{G}_n(g) &\lesssim K \left(\sigma \sqrt{\frac{s \log(a_n)}{n}} + n^{1/q} \|\tilde{G}_4^2\|_{P,q} \frac{s \log(a_n)}{n} \right) \\
&= o(\log^{-4}(a_n))
\end{aligned}
$$

with probability $1 - o(1)$. Hence,

$$
\begin{aligned}
& ||\psi_{m_r}(X, \hat{\theta}_{m_r}, \hat{\eta}_{m_r}) - \psi_{m_r}(X, \theta_{m_r}, \eta_{m_r})||_{P_n,2} \\
\leq \quad & \left(\frac{1}{\sqrt{n}} \sup_{g \in \tilde{G}_4^2} \mathbb{G}_n(g) + \sup_{g \in \tilde{G}_4^2} \mathbb{E}[g(X)] \right)^{\frac{1}{2}} = o(\log^{-3/2}(a_n))
\end{aligned}
$$

with probability $1 - o(1)$ due to Assumption 2.1 (v) (a).

This gives us $II = o_p\left(\log^{-1}(a_n)\right)$ with probability $1 - o(1)$ implying Assumption 2.3 (iii) with $\bar{\delta}_n = o(\log^{-1}(a_n)) = o(1)$. It is straightforward to see that the growth conditions of Corollary 2.2 hold. ∎

APPENDIX B: UNIFORM NUISANCE FUNCTION ESTIMATION

Consider the following linear regression model

$$
Y_r = \sum_{j=1}^{p} \beta_{r,j} X_{r,j} + \varepsilon_r = \beta_r X_r + \varepsilon_r
$$

with centered regressors and errors ε_r with $\mathbb{E}[\varepsilon_r] = 0$ for each $r = 1, \ldots, d$. The true parameter obeys

$$\beta_r \in \arg\min_{\beta} \mathbb{E}[(Y_r - \beta X_r)^2]$$

with

$$\beta_r = \beta_r^{(1)} + \beta_r^{(2)}.$$

The parameter $\beta_r^{(2)}$ is the approximate sparse part of the true regression coefficient that captures the misspecification of a sparse model. We show that the lasso, post-lasso and square-root lasso estimators have sufficiently fast estimation rates uniformly for all $r = 1, \ldots, d$. In this setting $d = d_n$ is explicitly allowed to grow with n. In the following analysis, the regressors and errors need to have at least subexponential tails. In this context, we define the Orlicz norm $\|X\|_{\Psi_\rho}$ as

$$\|X\|_{\Psi_\rho} = \inf\{C > 0 : \mathbb{E}[\Psi_\rho(|X|/C)] \leq 1\}$$

with $\Psi_\rho(x) = \exp(x^\rho) - 1$.

B.1. Uniform lasso estimation. Define the weighted lasso estimator

$$\hat{\beta}_r \in \arg\min_{\beta} \left(\frac{1}{2} \mathbb{E}_n \left[(Y_r - \beta X_r)^2 \right] + \frac{\lambda}{n} \|\hat{\Psi}_{r,m} \beta\|_1 \right)$$

with the penalty level

$$\lambda = c_\lambda \sqrt{n} \Phi^{-1} \left(1 - \frac{\gamma}{2pd} \right)$$

for a suitable $c_\lambda > 1$, $\gamma \in [1/n, 1/\log(n)]$ and a fix $m \geq 0$. Define the post-regularized weighted least squares estimator as

$$\tilde{\beta}_r \in \arg\min_{\beta} \left(\frac{1}{2} \mathbb{E}_n \left[(Y_r - \beta X_r)^2 \right] \right) : \quad \operatorname{supp}(\beta) \subseteq \operatorname{supp}(\hat{\beta}_r).$$

The penalty loadings $\hat{\Psi}_{r,m} = \operatorname{diag}(\{\hat{l}_{r,j,m}, j = 1, \ldots, p\})$ are defined by

$$\hat{l}_{r,j,0} = \max_{1 \leq i \leq n} ||X_r^{(i)}||_\infty$$

for $m = 0$ and for all $m \geq 1$ by the following algorithm:

ALGORITHM 2. *Set $\bar{m} = 0$. Compute $\hat{\beta}_r$ based on $\hat{\Psi}_{r,\bar{m}}$. Set $\hat{l}_{r,j,\bar{m}+1} = \mathbb{E}_n \left[\left(\left(Y_r - \hat{\beta}_r X_r \right) X_{r,j} \right)^2 \right]^{1/2}$. If $\bar{m} = m$ stop and report the current value of $\hat{\Psi}_{r,m}$, otherwise set $\bar{m} = \bar{m} + 1$.*

Let $a_n := \max(p, n, d, e)$. In order to establish uniform convergence rates, the following assumptions are required to hold uniformly in $n \geq n_0, P \in \mathcal{P}_n$:

Assumptions **B1**-**B4**:

B1 (Tail conditions)
There exists $1 \leq \rho \leq 2$ *such that*

$$\max_{r=1,\dots,d} \max_{j=1,\dots,p} \|X_{r,j}\|_{\Psi_\rho} \leq C \text{ and } \max_{r=1,\dots,d} \|\varepsilon_r\|_{\Psi_\rho} \leq C.$$

B2 (Uniformly bounded eigenvalues)
For all $r = 1, \dots, d_n$, *it holds*

$$\inf_{\|\xi\|_2=1} \mathbb{E}\left[(\xi X_r)^2\right] \geq c, \ \sup_{\|\xi\|_2=1} \mathbb{E}\left[(\xi X_r)^2\right] \leq C$$

and

$$\min_{r=1,\dots,d} \min_{j=1,\dots,p} \mathbb{E}[\varepsilon_r^2 X_{r,j}^2] \geq c.$$

B3 (Uniform approximate sparsity)
The coefficients obey

$$\max_{r=1,\dots,d} \|\beta_r^{(2)}\|_1^2 \lesssim \sqrt{\frac{s^2 \log(a_n)}{n}}, \quad \max_{r=1,\dots,d} \mathbb{E}\left[(\beta_r^{(2)} X_r)^2\right] \lesssim \frac{s \log(a_n)}{n}$$

and

$$\max_{r=1,\dots,d} \|\beta_r^{(1)}\|_0 \leq s.$$

B4 (Growth conditions)
There exists a positive number $\tilde{q} > 0$ *such that the following growth condition is fulfilled:*

$$n^{\frac{1}{\tilde{q}}} \frac{s \log^{1+\frac{4}{\rho}}(a_n)}{n} = o(1).$$

THEOREM 2. *Under the assumptions B1-B4 the lasso estimator* $\hat{\beta}_r$ *obeys uniformly over all* $P \in \mathcal{P}_n$ *with probability* $1 - o(1)$

$$\max_{r=1,\dots,d} \|\hat{\beta}_r - \beta_r^{(1)}\|_2 \leq C \sqrt{\frac{s \log(a_n)}{n}}, \tag{B.1}$$

$$\max_{r=1,\dots,d} \|\hat{\beta}_r - \beta_r^{(1)}\|_1 \leq C \sqrt{\frac{s^2 \log(a_n)}{n}} \tag{B.2}$$

with

$$\max_{r=1,\dots,d} \|\hat{\beta}_r\|_0 \leq Cs. \tag{B.3}$$

Additionally the post-lasso estimator $\tilde{\beta}_r$ *obeys uniformly over all* $P \in \mathcal{P}_n$ *with probability* $1 - o(1)$

$$\max_{r=1,\dots,d} \|\tilde{\beta}_r - \beta_r^{(1)}\|_2 \leq C\sqrt{\frac{s\log(a_n)}{n}}, \tag{B.4}$$

$$\max_{r=1,\dots,d} \|\tilde{\beta}_r - \beta_r^{(1)}\|_1 \leq C\sqrt{\frac{s^2\log(a_n)}{n}}. \tag{B.5}$$

B.2. Uniform square-root lasso estimation. Now, assume that $X_{r,j}$ are standardized covariates ($\mathbb{E}[X_{r,j}^2] = 1$ for all $j = 1,\dots,p$ and $r = 1,\dots,d$) which are independent from the errors ε_r. Define

$$Q_r(\beta) := \mathbb{E}_n[(Y_r - \beta X_r - \beta_r^{(2)} X_r)^2].$$

The square-root lasso estimator is definded as

$$\hat{\beta}_r \in \arg\min_{\beta} \left(\hat{Q}_r^{1/2}(\beta) + \frac{\lambda}{n}\|\beta\|_1 \right),$$

where $\hat{Q}_r(\beta) := \mathbb{E}_n[(Y_r - \beta X_r)^2]$. $\hat{Q}_r(\beta)$ is a proxy for $Q_r(\beta)$ estimating the approximate sparse part $\beta_r^{(2)}$ by $\hat{\beta}_r^{(2)} = 0$. Let

$$\lambda = c'\sqrt{n}\Phi^{-1}\big(1 - \gamma/(2pd)\big) \tag{B.6}$$

where $1-\gamma$ is a confidence level associated with the probability of the event (B.7), and $c' > c$ is a slack constant. The first part of the analysis is to control the event

$$\frac{\lambda}{n} \geq c \max_{r=1,\dots,d} \|S_r\|_\infty, \tag{B.7}$$

where

$$S_r := \partial_\beta Q^{1/2}(\beta)\big|_{\beta=\beta_r^{(1)}} = -\frac{\mathbb{E}_n[X_r(Y_u - \beta_r^{(1)} X_r - \beta_r^{(2)} X_r)]}{\sqrt{\mathbb{E}_n[(Y_u - \beta_r^{(1)} X_r - \beta_r^{(2)} X_r)^2]}} = -\frac{\mathbb{E}_n[X_r\varepsilon_r]}{\sqrt{\mathbb{E}_n[\varepsilon_r^2]}}$$

is the score of $Q^{1/2}$ at $\beta_r^{(1)}$. Define

$$\hat{S}_r := \partial_\beta \hat{Q}^{1/2}(\beta)\big|_{\beta=\beta_r^{(1)}} = -\frac{\mathbb{E}_n[X_r(\varepsilon_r + \beta_r^{(2)} X_r)]}{\sqrt{\mathbb{E}_n[(\varepsilon_r + \beta_r^{(2)} X_r)^2]}}.$$

The following conditions and lemma 1 are essentially the same as condition WL and lemma L.4. in Belloni et al. (2018) [2]. Let $\underline{C}$ and $\overline{C}$ be some strictly positive constants. Additionally let $(\varphi_n)_{n\geq 1}$, $(\tilde{\varphi}_n)_{n\geq 1}$, $(\bar{\varphi}_n)_{n\geq 1}$ and Δ_n be some sequences of positive constants converging to zero.

Condition WL *The following conditions hold:*

(i) $\max_{r=1,\dots,d}\max_{j=1,\dots,p}\left(\mathbb{E}\left[|X_{r,j}\varepsilon_r|^3\right]\right)^{1/3}\Phi^{-1}(1-\gamma/(2pd)) \leq \varphi_n n^{1/6}$;
(ii) $\underline{C} \leq \mathbb{E}\left[|X_{r,j}\varepsilon_r|^2\right] \leq \overline{C}$, *for all* $r = 1,\dots,d$ *and* $j = 1,\dots,p$;
(iii) with probability at least $1 - \frac{1}{2}\Delta_n$,

$$\max_{r=1,\dots,d}\max_{j=1,\dots,p}|\mathbb{E}_n[X_{r,j}^2\varepsilon_r^2] - \mathbb{E}[X_{r,j}^2\varepsilon_r^2]| \leq \tilde{\varphi}_n$$

and

$$\max_{r=1,\dots,d}|\mathbb{E}_n[\varepsilon_r^2] - \mathbb{E}[\varepsilon_r^2]| \leq \bar{\varphi}_n.$$

The following lemma proves that λ satisfies (B.7) with high probability.

LEMMA 1. *Suppose that condition* ***WL*** *holds. In addition suppose that* λ *satisfies (B.6) for some* $c' > c$ *and* $\gamma = \gamma_n \in [1/n, 1/\log(n)]$. *Then it holds*

$$P\left(\frac{\lambda}{n} \geq c\max_{r=1,\dots,d}\|S_r\|_\infty\right) \geq 1 - \gamma - o(\gamma) - \Delta_n.$$

Under the same uniform sparsity and regularity conditions as in theorem 2 we are able to show that condition **WL** is satisfied and hence we can establish uniform convergence rates of the square-root lasso estimator. In section B.2 we additionally assumed independence between the regressors and the error terms. This eliminates the need to estimate the penalty loadings.

THEOREM 3. *Suppose that the conditions B1-B4 hold. In addition suppose that* λ *satisfies (B.6) for some* $c' > c$ *and* $\gamma = \gamma_n \in [1/n, 1/\log(n)]$. *Then, with probability at least* $1 - o(1)$ *we have*

$$\max_{r=1,\dots,d}\|\hat{\beta}_r - \beta_r^{(1)}\|_2 \leq C\sqrt{\frac{s\log(a_n)}{n}}, \tag{B.8}$$

$$\max_{r=1,\dots,d}\|\hat{\beta}_r - \beta_r^{(1)}\|_1 \leq C\sqrt{\frac{s^2\log(a_n)}{n}} \tag{B.9}$$

with

$$\max_{r=1,\dots,d}\|\hat{\beta}_r\|_0 \leq Cs. \tag{B.10}$$

B.3. Proofs.

Proof of Theorem 2.

Due to condition B1 we can bound the q-th moments of the maxima of the regressors uniformly by

$$\begin{aligned}
\mathbb{E}\left[\max_{r=1,\dots,d}\|X_r\|_\infty^q\right]^{\frac{1}{q}} &= \|\max_{r=1,\dots,d}\max_{j=1,\dots,p}|X_{r,j}|\|_{P,q} \\
&\leq q!\|\max_{r=1,\dots,d}\max_{j=1,\dots,p}|X_{r,j}|\|_{\psi_1} \\
&\leq q!\log^{\frac{1}{\rho}-1}(2)\|\max_{r=1,\dots,d}\max_{j=1,\dots,p}|X_{r,j}|\|_{\psi_\rho} \\
&\leq q!\log^{\frac{1}{\rho}-1}(2)K\log^{\frac{1}{\rho}}(1+dp)\max_{r=1,\dots,d}\max_{j=1,\dots,p}\|X_{r,j}\|_{\psi_\rho} \\
&\leq C\log^{\frac{1}{\rho}}(a_n)
\end{aligned}$$

where C does depend on q and ρ but not on n. For the norm inequalities we refer to van der Vaar and Wellner (1996) [17].

As in the previous proof we use C for a strictly positive constant, independent of n, which may have a different value in each appearance. The notation $a_n \lesssim b_n$ stands for $a_n \leq Cb_n$ for all n for some fixed C. Additionally $a_n = o(1)$ stands for uniform convergence towards zero meaning there exists sequence $(b_n)_{n\geq 1}$ with $|a_n| \leq b_n$, b_n is independent of $P \in \mathcal{P}_n$ for all n and $b_n \to 0$. Finally, the notation $a_n \lesssim_P b_n$ means that for any $\epsilon > 0$, there exists C such that uniformly over all n we have $P_P(a_n > Cb_n) \leq \epsilon$.

We essentially modify the proof from theorem 4.2 from Belloni et al. (2018) [2] to fit our setting and keep the notation as similar as possible.

We set $\mathcal{U} = \{1,\dots,d\}$ and

$$\beta_r^{(1)} \in \arg\min_{\beta\in\mathbb{R}^p}\mathbb{E}\Big[\underbrace{\frac{1}{2}\left(Y_r - \beta X_r - \beta_r^{(2)}X_r\right)^2}_{:=M_r(Y_r,X_r,\beta,a_r)}\Big]$$

with $a_r = \beta_r^{(2)}X_r$ for all $r = 1,\dots,d$. Since the coefficient $\beta^{(2)}$ is approximately sparse by assumption we estimate the nuisance parameter a_r with $\hat{a}_r \equiv 0$. Define

$$M_r(Y_r, X_r, \beta) := M_r(Y_r, X_r, \beta, \hat{a}_r) = \frac{1}{2}\left(Y_r - \beta X_r\right)^2.$$

Then we have

$$\hat{\beta}_r \in \arg\min_{\beta\in\mathbb{R}^p}\left(\mathbb{E}_n\left[M_r(Y_r, X_r, \beta)\right] + \frac{\lambda}{n}\|\hat{\Psi}_r\beta\|_1\right)$$

and

$$\tilde{\beta}_r \in \arg\min_{\beta\in\mathbb{R}^p} \left(\mathbb{E}_n\left[M_r(Y_r, X_r, \beta)\right]\right): \quad \operatorname{supp}(\beta) \subseteq \operatorname{supp}(\hat{\beta}_r).$$

At first we verify the condition WL from Belloni et al. (2018) [2]. Since $N_n = d$ we have $N(\varepsilon, \mathcal{U}, d_{\mathcal{U}}) \leq N_n$ for all $\varepsilon \in (0,1)$ with

$$d_{\mathcal{U}}(i,j) = \begin{cases} 0 & \text{for } i = j \\ 1 & \text{for } i \neq j. \end{cases}$$

To prove WL(i) observe that

$$S_r = \partial_\beta M_r(Y_r, X_r, \beta, a_r)\big|_{\beta=\beta_r^{(1)}} = -\varepsilon_r X_r.$$

Since $\Phi^{-1}(1-t) \lesssim \sqrt{\log(1/t)}$, uniformly over $t \in (0, 1/2)$ we have that

$$\begin{aligned} \|S_{r,j}\|_{P,3}\Phi^{-1}(1-\gamma/2pd) &= \|\varepsilon_r X_{r,j}\|_{P,3}\Phi^{-1}(1-\gamma/2pd) \\ &\leq \left(\|\varepsilon_r\|_{P,6}\|X_{r,j}\|_{P,6}\right)^{1/2}\Phi^{-1}(1-\gamma/2pd) \\ &\leq C\log^{\frac{1}{2}}(a_n) \lesssim \varphi_n n^{\frac{1}{6}} = o(1) \end{aligned}$$

with

$$\varphi_n = O\left(\frac{\log^{\frac{1}{2}}(a_n)}{n^{\frac{1}{6}}}\right)$$

uniformly over all $j = 1, \ldots, p$ and $r = 1, \ldots, d$ by assumption B1 and B4. Further, it holds

$$\begin{aligned} c \leq \mathbb{E}\left[S_{r,j}^2\right] &= \mathbb{E}\left[\varepsilon_r^2 X_{r,j}^2\right] \\ &\leq \left(\mathbb{E}\left[\varepsilon_r^4\right]\mathbb{E}\left[X_{r,j}^4\right]\right)^{1/2} \\ &\leq C \end{aligned}$$

for all $j = 1, \ldots, p$ and $r = 1, \ldots, d$ by assumption B1 and B2 which implies condition $WL(ii)$. Observe that condition $WL(iii)$ reduces to

$$\max_{r=1,\ldots,d}\max_{j=1,\ldots,p} |(\mathbb{E}_n - \mathbb{E})[S_{r,j}^2]| \leq \varphi_n$$

with probability $1-\Delta_n$. We use a maximal inequality, see for example lemma $O.2$ from Belloni et al. (2018) [2]. Let $\mathcal{W} = (\mathcal{Y}, \mathcal{X})$ with $Y = (Y_1, \ldots, Y_d) \in \mathcal{Y}$ and $X = (X_1, \ldots, X_d) \in \mathcal{X}$. Define

$$\mathcal{F} := \{f_{r,j}^2 | r = 1, \ldots, d, j = 1, \ldots, p\}$$

with

$$\begin{aligned} f_{r,j} :&\mathcal{W} = (\mathcal{Y}, \mathcal{X}) \to \mathbb{R} \\ &W = (Y, X) \mapsto (Y_r - \beta_r X_r) X_{r,j} = \varepsilon_r X_{r,j} = S_{r,j}. \end{aligned}$$

Observe that

$$\begin{aligned} \| \sup_{f \in \mathcal{F}} |f| \|_{P,q} &= \| \max_{r=1,\dots,d} \max_{j=1,\dots,p} |f_{r,j}^2| \|_{P,q} \\ &= \mathbb{E}\left[\max_{r=1,\dots,d} \max_{j=1,\dots,p} \varepsilon_r^{2q} X_{r,j}^{2q} \right]^{1/q} \\ &\leq \mathbb{E}\left[\max_{r=1,\dots,d} \varepsilon_r^{2q} \max_{r=1,\dots,d} \max_{j=1,\dots,p} X_{r,j}^{2q} \right]^{1/q} \\ &\leq \left(\mathbb{E}\left[\max_{r=1,\dots,d} \varepsilon_r^{4q} \right]^{1/4q} \mathbb{E}\left[\max_{r=1,\dots,d} \max_{j=1,\dots,p} X_{r,j}^{4q} \right]^{1/4q} \right)^2 \\ &\leq C \log^{\frac{4}{\rho}}(a_n). \end{aligned}$$

Since we have

$$\sup_{f \in \mathcal{F}} \|f\|_{P,2}^2 = \max_{r=1,\dots,d} \max_{j=1,\dots,p} \mathbb{E}\left[S_{r,j}^4\right] \leq \max_{r=1,\dots,d} \max_{j=1,\dots,p} \mathbb{E}\left[\varepsilon_r^8\right]^{1/2} \mathbb{E}\left[X_{r,j}^8\right]^{1/2} \leq C$$

we can choose a constant with

$$\sup_{f \in \mathcal{F}} \|f\|_{P,2}^2 \leq C \leq \| \sup_{f \in \mathcal{F}} |f| \|_{P,2}^2.$$

Additionally $|\mathcal{F}| = dp$ which implies

$$\log \sup_{Q} N(\epsilon \|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}) \leq \log(dp) \lesssim \log(a_n/\epsilon), \quad 0 < \epsilon \leq 1.$$

Using lemma $O.2$ from Belloni et al. (2018) [2] we obtain with probability not less than $1 - o(1)$

$$\begin{aligned} \max_{r=1,\dots,d} \max_{j=1,\dots,p} |(\mathbb{E}_n - \mathbb{E})[S_{r,j}^2]| &= n^{-1/2} \sup_{f \in \mathcal{F}} |\mathbb{G}_n(f)| \\ &\leq n^{-1/2} C \left(\sqrt{\log(a_n)} + n^{-1/2+1/q} \log^{1+\frac{4}{\rho}}(a_n) \right) \\ &= C \left(\sqrt{\frac{\log(a_n)}{n}} + \frac{\log^{1+\frac{4}{\rho}}(a_n)}{n^{1-1/q}} \right) \\ &\leq \varphi_n = o(1) \end{aligned}$$

by the growth condition B4.
We proceed by verifying assumption $L.1$. The function $\beta \mapsto M_r(Y_r, X_r, \beta)$ is convex, which is the first requirement of assumption $L.1$.
We now proceed with a simplified version of proof of $J.1$ from Belloni et al. (2018) [2]. Define

$$\mathcal{G} := \{g_r : X \to (\beta_r^{(2)} X_r)^2 | r = 1, \dots, d\}$$

with envelope

$$G := \max_{r=1,\dots,d} \|X_r\|_\infty^2 \|\beta_r^{(2)}\|_1^2.$$

Note that

$$\begin{aligned}
\|G\|_{P,q} &= \mathbb{E}\left[\max_{r=1,\dots,d} \|X_r\|_\infty^{2q} \|\beta_r^{(2)}\|_1^{2q}\right]^{\frac{1}{q}} \\
&\leq \max_{r=1,\dots,d} \|\beta_r^{(2)}\|_1^2 \mathbb{E}\left[\max_{r=1,\dots,d} \|X_r\|_\infty^{2q}\right]^{\frac{1}{q}} \\
&\lesssim \max_{r=1,\dots,d} \|\beta_r^{(2)}\|_1^2 \log(a_n)^{\frac{2}{\rho}}
\end{aligned}$$

and for all $0 < \varepsilon \leq 1$ we have

$$N(\varepsilon \|G\|_{P,2}, \mathcal{G}, \|\cdot\|_{P,2}) \leq d \leq d/\varepsilon.$$

Since

$$\sup_{g \in \mathcal{G}} \|g\|_{P,2}^2 = \max_{r=1,\dots,d} \mathbb{E}[(\beta_r^{(2)} X_r)^4] \lesssim \max_{r=1,\dots,d} \|\beta_r^{(2)}\|_1^4$$

we can use lemma $O.2$ from Belloni et al. (2018) [2] to obtain with probability not less than $1 - o(1)$

$$\begin{aligned}
&\max_{r=1,\dots,d} |(\mathbb{E}_n - \mathbb{E})[(\beta_r^{(2)} X_r)^2]| \\
&= n^{-1/2} \sup_{g \in \mathcal{G}} |\mathbb{G}_n(g)| \\
&\lesssim C\left(\sqrt{\frac{\log(a_n) \max_{r=1,\dots,d} \|\beta_r^{(2)}\|_1^4}{n}} + n^{-1+1/q} \max_{r=1,\dots,d} \|\beta_r^{(2)}\|_1^2 \log^{1+\frac{2}{\rho}}(a_n)\right) \\
&\lesssim C\left(\sqrt{\frac{\log(a_n)}{n}} \sqrt{\frac{s^2 \log(a_n)}{n}} + \frac{s \log(a_n)}{n} \sqrt{n^{2/q} \frac{\log^{1+\frac{4}{\rho}}(a_n)}{n}}\right) \\
&\lesssim \frac{s \log(a_n)}{n},
\end{aligned}$$

for a suitable choice of q where we used $\max_{r=1,\dots,d} \|\beta_r^{(2)}\|_1^2 \lesssim \sqrt{\frac{s^2 \log(a_n)}{n}}$ from condition B3 and the growth condition B4.
Using the triangle inequality and $\max_{r=1,\dots,d} \mathbb{E}\left[(\beta_r^{(2)} X_r)^2\right] \lesssim \frac{s \log(a_n)}{n}$ from condition B3 we obtain

$$\begin{aligned}\max_{r=1,\dots,d} \mathbb{E}_n[(\beta_r^{(2)} X_r)^2] &\leq \max_{r=1,\dots,d} |(\mathbb{E}_n - \mathbb{E})[(\beta_r^{(2)} X_r)^2]| + \max_{r=1,\dots,d} \mathbb{E}[(\beta_r^{(2)} X_r)^2] \\ &\lesssim_P \frac{s \log(a_n)}{n}. \end{aligned} \tag{B.11}$$

To show assumption $L.1$ (a), note that for all $\delta \in \mathbb{R}^p$

$$\begin{aligned}&\left| \mathbb{E}_n \left[\partial_\beta M_r(Y_r, X_r, \beta_r^{(1)}) - \partial_\beta M_r(Y_r, X_r, \beta_r^{(1)}, a_r) \right]^T \delta \right| \\ &= \left| \mathbb{E}_n \left[X_r (\beta_r^{(2)} X_r) \right]^T \delta \right| \leq ||(\beta_r^{(2)} X_r)||_{\mathbb{P}_n,2} ||X_r^T \delta||_{\mathbb{P}_n,2} \\ &\lesssim_P \sqrt{\frac{s \log(a_n)}{n}} ||X_r^T \delta||_{\mathbb{P}_n,2}\end{aligned}$$

for all $r = 1, \dots, d$. Further we have

$$\begin{aligned}&\mathbb{E}_n \left[\frac{1}{2} \left(Y_r - (\beta_r^{(1)} + \delta^T) X_r \right)^2 \right] - \mathbb{E}_n \left[\frac{1}{2} \left(Y_r - \beta_r^{(1)} X_r \right)^2 \right] \\ &= -\mathbb{E}_n \left[\left(Y_r - \beta_r^{(1)} X_r \right) \delta^T X_r \right] + \frac{1}{2} \mathbb{E}_n \left[(\delta^T X_r)^2 \right],\end{aligned}$$

where

$$-\mathbb{E}_n \left[\left(Y_r - \beta_r^{(1)} X_r \right) \delta^T X_r \right] = \mathbb{E}_n \left[\partial_\beta M_r(Y_r, X_r, \beta_r^{(1)}) \right]^T \delta$$

and

$$\frac{1}{2} \mathbb{E}_n \left[(\delta^T X_r)^2 \right] = ||\sqrt{w_r} \delta^T X_r||_{\mathbb{P}_n,2}^2$$

with $\sqrt{w_r} = 1/4$. This gives us assumption $L.1$ (c) with $\Delta_n = 0$ and $\bar{q}_{A_r} = \infty$. Since condition $WL(ii)$ and $WL(iii)$ hold we have with probability $1 - o(1)$

$$1 \lesssim l_{r,j} = \left(\mathbb{E}_n[S_{r,j}^2] \right)^{1/2} \lesssim 1$$

uniformly over all $r = 1, \dots, d$ and $j = 1, \dots, p$, which directly implies

$$1 \lesssim \|\hat{\Psi}_r^{(0)}\|_\infty := \max_{j=1,\dots,p} |l_{r,j}| \lesssim 1$$

and additionally

$$1 \lesssim \|(\hat{\Psi}_r^{(0)})^{-1}\|_\infty := \max_{j=1,\dots,p} |l_{r,j}^{-1}| \lesssim 1.$$

For now, we suppose that $m = 0$ in algorithm 2. Uniformly over $r = 1, \dots, d$, $j = 1, \dots, p$ we have

$$\hat{l}_{r,j,0} = \left(\mathbb{E}_n[\max_{1 \le i \le n} \|X_r^{(i)}\|_\infty^2] \right)^{1/2} \ge \left(\mathbb{E}_n[\|X_r\|_\infty^2] \right)^{1/2} \gtrsim_P 1$$

where the last inequality holds due to condition B2 and an application of the maximal inequality lemma.
Also uniformly over $r = 1, \dots, d$, $j = 1, \dots, p$ we have for an arbitrary $q > 0$

$$\begin{aligned} \hat{l}_{r,j,0} &= \max_{1 \le i \le n} \|X_r^{(i)}\|_\infty \\ &\le n^{1/q} \left(\frac{1}{n} \sum_{i=1}^n \|X_r^{(i)}\|_\infty^q \right)^{1/q} \\ &= n^{1/q} \left(\mathbb{E}_n[\|X_r\|_\infty^q] \right)^{1/q} \end{aligned}$$

with

$$\mathbb{E}[\|X_r\|_\infty^q]^{1/q} \lesssim \log^{\frac{1}{\rho}}(a_n)$$

By maximal inequality, we have with probability $1 - o(1)$ for a sufficiently large $q' > 0$

$$\begin{aligned} &\max_r |\mathbb{E}_n[\|X_r\|_\infty^q] - \mathbb{E}[\|X_r\|_\infty^q]| \\ &\lesssim C \left(\sqrt{\frac{\log^{\frac{2q}{\rho}+1}(a_n)}{n}} + n^{1/q'-1} \log^{\frac{q}{\rho}+1}(a_n) \right) \\ &\lesssim \log^{\frac{q}{\rho}}(a_n) \end{aligned}$$

since

$$\mathbb{E}[\max_r \|X_r\|_\infty^{qq'}]^{1/q'} \lesssim \log^{\frac{q}{\rho}}(a_n) \text{ and } \max_r \mathbb{E}[\|X_r\|_\infty^{q2}]^{1/2} \lesssim \log^{\frac{q}{\rho}}(a_n).$$

We conclude

$$\begin{aligned} &\hat{l}_{r,j,0} \le n^{1/q} \left(\mathbb{E}_n[\|X_r\|_\infty^q] \right)^{1/q} \\ &\le n^{1/q} \left(|\mathbb{E}_n[\|X_r\|_\infty^q] - \mathbb{E}[\|X_r\|_\infty^q]| + \mathbb{E}[\|X_r\|_\infty^q] \right)^{1/q} \\ &\lesssim_P n^{1/q} \log^{\frac{1}{\rho}}(a_n). \end{aligned}$$

uniformly over r. Therefore assumption $L.1(b)$ holds for some $\Delta_n = o(1)$, $L \lesssim n^{1/q}\log^{\frac{1}{\rho}}(a_n)$ and $l \gtrsim 1$. Hence, we can find a c_l with $l > 1/c_l$. Setting $c_\lambda > c_l$ and $\gamma = \gamma_n \in [1/n, 1/\log(n)]$ in the choice of λ, we have

$$P\left(\frac{\lambda}{n} \geq c_l \max_{r=1,\dots,d} \|(\hat{\Psi}_r^{(0)})^{-1}\mathbb{E}_n[S_r]\|_\infty\right) \geq 1 - \gamma - o(\gamma) - \Delta_n = 1 - o(1)$$

due to lemma $L.4$ from Belloni et al. (2018) [2].
Now we uniformly bound the sparse eigenvalues. Set

$$l_n = \log^{\frac{2}{\rho}}(a_n)n^{2/\bar{q}}$$

for a $\bar{q} > 5\tilde{q}$ with $\tilde{q}$ in B4. We apply Lemma $P.1$ in [2] with $K \lesssim n^{1/\bar{q}}\log^{\frac{1}{\rho}}(a_n)$ and

$$\begin{aligned}\delta_n &\lesssim K\sqrt{sl_n}n^{-1/2}\log(sl_n)\log^{\frac{1}{2}}(a_n)\log^{\frac{1}{2}}(n)\\ &\lesssim \sqrt{n^{\frac{4}{\bar{q}}}\log(n)\log^2(sl_n)\frac{s\log^{1+\frac{4}{\rho}}(a_n)}{n}}\\ &\lesssim \sqrt{n^{\frac{5}{\bar{q}}}\frac{s\log^{1+\frac{4}{\rho}}(a_n)}{n}}\end{aligned}$$

for n large enough. Hence by growth condition B4, it holds

$$\delta_n = o(1)$$

which implies

$$1 \lesssim \min_{\|\delta\|_0 \leq l_n s} \frac{\|\delta X_r\|^2_{\mathbb{P}_n,2}}{\|\delta\|_2^2} \leq \max_{\|\delta\|_0 \leq l_n s} \frac{\|\delta X_r\|^2_{\mathbb{P}_n,2}}{\|\delta\|_2^2} \lesssim 1$$

with probability $1 - o(1)$ uniformly over $r = 1, \dots, d$.
Define $T_r := \text{supp}(\beta_r^{(1)})$ and

$$\tilde{c} := \frac{Lc_l + 1}{lc_l - 1}\max_{r=1,\dots,d}\|\hat{\Psi}_r^{(0)}\|_\infty\|(\hat{\Psi}_r^{(0)})^{-1}\|_\infty \lesssim L.$$

Let the restricted eigenvalues be definied as

$$\bar{\kappa}_{2\tilde{c}} := \min_{r=1,\dots,d}\inf_{\delta \in \Delta_{2\tilde{c},r}} \frac{\|\delta X_r\|_{\mathbb{P}_n,2}}{\|\delta_{T_r}\|_2}$$

where $\Delta_{2\tilde{c},r} := \{\delta : \|\delta^c_{T_r}\|_1 \leq 2\tilde{c}\|\delta_{T_r}\|_1\}$. By the argument given in Bickel et al. (2009) [4] we have

$$\begin{aligned}\bar{\kappa}_{2\tilde{c}} &\geq \left(\min_{\|\delta\|_0 \leq l_n s} \frac{\|\delta X_r\|^2_{\mathbb{P}_n,2}}{\|\delta\|_2^2}\right)^{1/2} - 2\tilde{c}\left(\max_{\|\delta\|_0 \leq l_n s} \frac{\|\delta X_r\|^2_{\mathbb{P}_n,2}}{\|\delta\|_2^2}\right)^{1/2}\left(\frac{s}{sl_n}\right)^{1/2} \\ &\gtrsim \left(\min_{\|\delta\|_0 \leq l_n s} \frac{\|\delta X_r\|^2_{\mathbb{P}_n,2}}{\|\delta\|_2^2}\right)^{1/2} - 2n^{\frac{1}{q}-\frac{1}{\bar{q}}}\left(\max_{\|\delta\|_0 \leq l_n s} \frac{\|\delta X_r\|^2_{\mathbb{P}_n,2}}{\|\delta\|_2^2}\right)^{1/2} \\ &\gtrsim 1\end{aligned}$$

with probability $1 - o(1)$ for a suitable choice of q with $q > \bar{q}$. Since

$$\frac{\lambda}{n} \lesssim n^{-1/2}\Phi^{-1}\left(1 - \gamma/(2dp)\right) \lesssim n^{-1/2}\sqrt{\log(2dp/\gamma)} \lesssim n^{-1/2}\log^{\frac{1}{2}}(a_n)$$

and the penalty loading are uniformly bounded from above and away from zero we have

$$\max_{r=1,\ldots,d} \|(\hat{\beta}_r - \beta_r^{(1)})X_r\|_{\mathbb{P}_n,2} \lesssim_P L\sqrt{\frac{s\log(a_n)}{n}}$$

by lemma $L.1$ from Belloni et al. (2018) [2].

To establish assumption $L.1(b)$ for $m \geq 1$, we proceed by induction. Assume that the assumption holds for $\hat{\Psi}_{r,m-1}$ with some $\Delta_n = o(1)$, $l \gtrsim 1$ and $L \lesssim n^{1/q}\log^{\frac{1}{\rho}}(a_n)$. We have shown that the estimator based on $\hat{\Psi}_{r,m-1}$ obeys

$$\max_{r=1,\ldots,d} \|(\hat{\beta}_r - \beta_r^{(1)})X_r\|_{\mathbb{P}_n,2} \lesssim L\sqrt{\frac{s\log(a_n)}{n}}$$

with probability $1 - o(1)$. Observe that

$$\max_{r=1,\ldots,d} \|\beta_r^{(2)}X_r\|_{\mathbb{P}_n,2} \lesssim_P \sqrt{\frac{s\log(a_n)}{n}}$$

as shown in (B.11). Using the triangle inequality we obtain with probability $1 - o(1)$

$$\begin{aligned}\max_{r=1,\ldots,d} \|(\hat{\beta}_r - \beta_r)X_r\|_{\mathbb{P}_n,2} &\leq \max_{r=1,\ldots,d} \|(\hat{\beta}_r - \beta_r^{(1)})X_r\|_{\mathbb{P}_n,2} + \max_{r=1,\ldots,d} \|\beta_r^{(2)}X_r\|_{\mathbb{P}_n,2} \\ &\lesssim L\sqrt{\frac{s\log(a_n)}{n}}.\end{aligned}$$

This implies

$$
\begin{aligned}
|\hat{l}_{r,j,m} - l_{r,j}| &= \left| \mathbb{E}_n \left[\left(\left(Y_r - \hat{\beta}_r X_r \right) X_{r,j} \right)^2 \right]^{1/2} - \mathbb{E}_n \left[\left(\left(Y_r - \beta_r X_r \right) X_{r,j} \right)^2 \right]^{1/2} \right| \\
&\leq \left| \mathbb{E}_n \left[\left(\left((\hat{\beta}_r - \beta_r) X_r \right) X_{r,j} \right)^2 \right]^{1/2} \right| \\
&\lesssim \|(\hat{\beta}_r - \beta_r) X_r\|_{\mathbb{P}_n,2} \max_{1 \leq i \leq n} \max_{r=1,\ldots,d} \|X_r^{(i)}\|_\infty \\
&\lesssim_P L \sqrt{\frac{s \log(a_n)}{n}} n^{1/q} \log^{\frac{1}{\rho}}(a_n) \\
&\lesssim \sqrt{n^{4/q} \frac{s \log^{1+\frac{4}{\rho}}(a_n)}{n}} = o(1)
\end{aligned}
$$

uniformly over $r = 1, \ldots, d$ and $j = 1, \ldots, p$. Therefore assumption $L.1(b)$ holds for $\hat{\Psi}_{r,m}$ for some $\Delta_n = o(1)$, $l \gtrsim 1$ and $L \lesssim 1$.
Consequently, we have

$$
\max_{r=1,\ldots,d} \|(\hat{\beta}_r - \beta_r^{(1)}) X_r\|_{\mathbb{P}_n,2} \lesssim \sqrt{\frac{s \log(a_n)}{n}}.
$$

and

$$
\max_{r=1,\ldots,d} \|\hat{\beta}_r - \beta_r^{(1)}\|_1 \lesssim \sqrt{\frac{s^2 \log(a_n)}{n}}
$$

with probability $1 - o(1)$ due to lemma $L.1$ from Belloni et al. (2018) [2].
Observe that with probability $1 - o(1)$ uniformly over all $r = 1, \ldots, d$ we have

$$
\begin{aligned}
&\left| \left(\mathbb{E}_n \left[\partial_\beta M_r(Y_r, X_r, \hat{\beta}_r) - \partial_\beta M_r(Y_r, X_r, \beta_r^{(1)}) \right] \right)^T \delta \right| \\
=&\left| \left(\mathbb{E}_n \left[(\hat{\beta}_r - \beta_r^{(1)}) X_r X_r^T \right] \right)^T \delta \right| \\
\leq& \|(\hat{\beta}_r - \beta_r^{(1)}) X_r\|_{\mathbb{P}_n,2} \|\delta X_r\|_{\mathbb{P}_n,2} \leq L_n \|\delta X_r\|_{\mathbb{P}_n,2}
\end{aligned}
$$

where $L_n \lesssim (s \log(a_n)/n)^{1/2}$. Since the maximal sparse eigenvalues

$$
\phi_{max}(l_n s, r) := \max_{\|\delta\|_0 \leq l_n s} \frac{\|\delta X_r\|_{\mathbb{P}_n,2}^2}{\|\delta\|_2^2}
$$

are uniformly bounded from above, lemma $L.2$ from Belloni et al. (2018) [2] directly implies

$$\max_{r=1,\dots,d} \|\hat{\beta}_r\|_0 \lesssim s$$

with probability $1-o(1)$. Combining this result with the uniform restrictions on the sparse eigenvalues from above we directly obtain

$$\max_{r=1,\dots,d} \|\hat{\beta}_r - \beta_r^{(1)}\|_2 \lesssim \max_{r=1,\dots,d} \|(\hat{\beta}_r - \beta_r^{(1)})X_r\|_{\mathbb{P}_n,2} \lesssim \sqrt{\frac{s\log(a_n)}{n}}$$

with probability $1-o(1)$.
We now proceed by using lemma $L.3$ from Belloni et al. (2018) [2]. We obtain uniformly over all $r = 1, \dots, d$

$$\begin{aligned}\mathbb{E}_n[M_r(Y_r, X_r, \tilde{\beta}_r)] - \mathbb{E}_n[M_r(Y_r, X_r, \beta_r)] &\leq \frac{\lambda L}{n}\|\hat{\beta}_r - \beta_r\|_1 \max_{r=1,\dots,d} \|\hat{\Psi}_r^{(0)}\|_\infty \\ &\lesssim \frac{\lambda}{n}\|\hat{\beta}_r - \beta_r\|_1 \\ &\lesssim \frac{s\log(a_n)}{n}\end{aligned}$$

with probability $1-o(1)$, where we used $L \lesssim 1$ and $\max_{r=1,\dots,d} \|\hat{\Psi}_r^{(0)}\|_\infty \lesssim 1$. Since

$$\max_{r=1,\dots,d} \|\mathbb{E}_n[S_r]\|_\infty \leq \max_{r=1,\dots,d} \|\hat{\Psi}_r^{(0)}\|_\infty \|\big(\hat{\Psi}_r^{(0)}\big)^{-1}\mathbb{E}_n[S_r]\|_\infty \lesssim \frac{\lambda}{n} \lesssim n^{-1/2}\log^{\frac{1}{2}}(a_n)$$

with probability $1-o(1)$, we obtain

$$\max_{r=1,\dots,d} \|(\tilde{\beta}_r - \beta_r^{(1)})X_r\|_{\mathbb{P}_n,2} \lesssim \sqrt{\frac{s\log(a_n)}{n}}$$

with probability $1-o(1)$, where we used

$$\max_{r=1,\dots,d} \|\hat{\beta}_r\|_0 \lesssim s,\ C_n \lesssim (s\log(a_n)/n)^{1/2}$$

and that the minimum sparse eigenvalues are uniformly bounded away from zero. With the same argument as above we directly obtain

$$\max_{r=1,\dots,d} \|\tilde{\beta}_r - \beta_r^{(1)}\|_2 \lesssim \max_{r=1,\dots,d} \|(\tilde{\beta}_r - \beta_r^{(1)})X_r\|_{\mathbb{P}_n,2} \lesssim \sqrt{\frac{s\log(a_n)}{n}}$$

This finally completes the proof. ∎

PROOF OF LEMMA 1.

See the proof of lemma L.4 from Belloni et al. (2018) [2]. Since the regressors are standardized for all $j = 1, \dots, p$ and independent from the error terms for all $r = 1, \dots, d$, observe that

$$\frac{\mathbb{E}[X_{r,j}^2 \varepsilon_r^2]}{\mathbb{E}[\varepsilon_r^2]} = \frac{\mathbb{E}[X_{r,j}^2]\mathbb{E}[\varepsilon_r^2]}{\mathbb{E}[\varepsilon_r^2]} = \mathbb{E}[X_{r,j}^2] = 1.$$

We have due to **WL**(iii)

$$\begin{aligned}
&P\left(\max_{r=1,\dots,d}\max_{j=1,\dots,p}\frac{\mathbb{E}_n[X_{r,j}^2\varepsilon_r^2]}{\mathbb{E}_n[\varepsilon_r^2]} > 1+\varphi_n\right)\\
\leq& P\left(\max_{r=1,\dots,d}\max_{j=1,\dots,p}\frac{\mathbb{E}[X_{r,j}^2\varepsilon_r^2]+\tilde{\varphi}_n}{\mathbb{E}[\varepsilon_r^2]-\bar{\varphi}_n} > 1+\varphi_n\right)+\Delta_n\\
\leq& P\left(\max_{r=1,\dots,d}\left|\frac{\mathbb{E}[\varepsilon_r^2]+\tilde{\varphi}_n}{\mathbb{E}[\varepsilon_r^2]-\bar{\varphi}_n}-1\right| > \varphi_n\right)+\Delta_n\\
=& P\left(\max_{r=1,\dots,d}\left|\frac{\mathbb{E}[\varepsilon_r^2]+\tilde{\varphi}_n}{\mathbb{E}[\varepsilon_r^2]-\bar{\varphi}_n}-\frac{\mathbb{E}[\varepsilon_r^2]}{\mathbb{E}[\varepsilon_r^2]}\right| > \varphi_n\right)+\Delta_n\\
=& P\left(\max_{r=1,\dots,d}\left|\frac{\left(\mathbb{E}[\varepsilon_r^2]+\tilde{\varphi}_n\right)\mathbb{E}[\varepsilon_r^2]-\mathbb{E}[\varepsilon_r^2]\left(\mathbb{E}[\varepsilon_r^2]-\bar{\varphi}_n\right)}{\left(\mathbb{E}[\varepsilon_r^2]-\bar{\varphi}_n\right)\mathbb{E}[\varepsilon_r^2]}\right| > \varphi_n\right)+\Delta_n\\
=& \underbrace{P\left(\left|\frac{((1+\tilde{\varphi}_n')-(1-\bar{\varphi}_n'))}{(1-\bar{\varphi}_n')}\right| > \varphi_n\right)}_{=0}+\Delta_n,
\end{aligned}$$

for an suitable choice of $\varphi_n = o(1)$, where $\bar{\varphi}_n' \geq \underline{C}\bar{\varphi}_n$ and $\tilde{\varphi}_n' \leq \overline{C}\tilde{\varphi}_n$ due to **WL**(ii) .

Next, for each $j = 1, \dots, p$ and $r = 1, \dots, d$, we apply lemma O.1 from Belloni et al. (2018) [2] with $\mu = 1$ and $\ell_n = c''\varphi_n^{-1}$, where c'' is a small constant that can be chosen to depend only on $\underline{C}$ and $\overline{C}$. Then conditions **WL**(i) and **WL**(ii) imply

$$0 \leq \Phi^{-1}\left(1-\frac{\gamma}{2pd}\right) \leq \frac{n^{1/6}M_n(j,r)}{\ell_n} - 1$$

for $M_n(j,r) = \mathbb{E}[X_{r,j}^2\varepsilon_r^2]^{1/2}/\mathbb{E}[|X_{r,j}\varepsilon_r|^3]^{1/3}$ for each $r = 1, \dots, d$ and $j = 1, \dots, p$.

Therefore, we have

$$
\begin{aligned}
&P\left(c \max_{r=1,\ldots,d} \|S_r\|_\infty > c' n^{-1/2} \Phi^{-1}\left(1-\frac{\gamma}{2pd}\right)\right) \\
=&P\left(c \max_{r=1,\ldots,d} \max_{j=1,\ldots,p} \frac{|\mathbb{E}_n[X_{r,j}\varepsilon_r]|}{\sqrt{\mathbb{E}_n[\varepsilon_r^2]}} > c' n^{-1/2} \Phi^{-1}\left(1-\frac{\gamma}{2pd}\right)\right) \\
\leq& \sum_{r=1}^{d}\sum_{j=1}^{p} P\left(c\frac{|n^{1/2}\mathbb{E}_n[X_{r,j}\varepsilon_r]|}{\sqrt{\mathbb{E}_n[\varepsilon_r^2]}} > c' \Phi^{-1}\left(1-\frac{\gamma}{2pd}\right)\right) \\
=& \sum_{r=1}^{d}\sum_{j=1}^{p} P\left(c\frac{|n^{1/2}\mathbb{E}_n[X_{r,j}\varepsilon_r]|}{\sqrt{\mathbb{E}_n[X_{r,j}^2\varepsilon_r^2]}}\sqrt{\frac{\mathbb{E}_n[X_{r,j}^2\varepsilon_r^2]}{\mathbb{E}_n[\varepsilon_r^2]}} > c' \Phi^{-1}\left(1-\frac{\gamma}{2pd}\right)\right) \\
\leq& \sum_{r=1}^{d}\sum_{j=1}^{p} P\left(\frac{|n^{1/2}\mathbb{E}_n[X_{r,j}\varepsilon_r]|}{\sqrt{\mathbb{E}_n[X_{r,j}^2\varepsilon_r^2]}} c\sqrt{1+\varphi_n} > c' \Phi^{-1}\left(1-\frac{\gamma}{2pd}\right)\right) + \Delta_n \\
\leq& 2pd\frac{\gamma}{2pd}\left(1+O(\varphi_n^{1/3})\right) + \Delta_n \\
\leq& \gamma + o(\gamma) + \Delta_n
\end{aligned}
$$

for a sufficiently large n (implying $c\sqrt{1+\varphi_n} \leq c'$). ∎

Proof of Theorem 3.

The proof is derived from the proof of lemma L.1. from Belloni et al. (2018) [2]. At first we show that condition **WL** is fulfilled. Conditions **WL** (i), **WL** (ii) and the first part of condition **WL** (iii) have been verified in the proof of Theorem 2. Hence, we need to show

$$\max_{r=1,\ldots,d} |\mathbb{E}_n[\varepsilon_r^2] - \mathbb{E}[\varepsilon_r^2]| \leq \bar{\varphi}_n$$

with probability converging to one.

Let $\mathcal{W} = (\mathcal{Y}, \mathcal{X})$ with $Y = (Y_1, \ldots, Y_d) \in \mathcal{Y}$ and $X = (X_1, \ldots, X_d) \in \mathcal{X}$. Define $\mathcal{F} := \{f_r | r = 1, \ldots, d\}$ with

$$
\begin{aligned}
f_r :& \mathcal{W} = (\mathcal{Y}, \mathcal{X}) \to \mathbb{R} \\
& W = (Y, X) \mapsto (Y_r - \beta_r X_r)^2 = \varepsilon_r^2.
\end{aligned}
$$

For a constant C that does depend on q but not on n, observe that

$$F := \|\sup_{f\in\mathcal{F}} |f|\|_{P,q} = \|\max_{r=1,\ldots,d} \varepsilon_r^2\|_{P,q} = \left(\mathbb{E}\left[\max_{r=1,\ldots,d} \varepsilon_r^{2q}\right]^{1/2q}\right)^2 \leq C\log(d)^{\frac{2}{\rho}}$$

where we used the same argument as in the beginning of the proof of Theorem 2.
Due to Assumption B1 the second moments of the error terms are uniformly bounded and hence we can choose a constant C such that

$$\max_{r=1,\dots,d} \|\varepsilon_r\|_{P,2}^2 \leq C \leq \| \max_{r=1,\dots,d} \varepsilon_r^2 \|_{P,q}$$

and since $|\mathcal{F}| = d$ we have

$$\log \sup_Q N(\varepsilon \|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}) \leq \log(d).$$

Therefore we are able to use lemma O.2 from Belloni et al. (2018) [2], which implies that with probability $1 - o(1)$

$$\begin{aligned}\max_{r=1,\dots,d} |\mathbb{E}_n[\varepsilon_r^2] - \mathbb{E}[\varepsilon_r^2]| &= n^{-1/2} \sup_{f\in\mathcal{F}} |\mathbb{G}_n(f)| \\ &\lesssim \left(\sqrt{\frac{\log(d)}{n}} + \frac{\log^{1+\frac{2}{\rho}}(d)}{n^{1-1/q}} \right) \leq \bar{\varphi}_n.\end{aligned}$$

Due to the definition of $\hat{\beta}_r$ we have

$$\hat{Q}_r^{1/2}(\hat{\beta}_r) + \frac{\lambda}{n}\|\hat{\beta}_r\|_1 \leq \hat{Q}_r^{1/2}(\beta_r^{(1)}) + \frac{\lambda}{n}\|\beta_r^{(1)}\|_1$$

implying

$$\hat{Q}_r^{1/2}(\hat{\beta}_r) - \hat{Q}_r^{1/2}(\beta_r^{(1)}) \leq \frac{\lambda}{n}\left(\|\delta_{r,T_r}\|_1 - \|\delta_{r,T_r^c}\|_1\right) \tag{B.12}$$

with $\delta_r := \hat{\beta}_r - \beta_r^{(1)}$. Due to the convexity of $\beta \mapsto \hat{Q}_r^{1/2}(\beta)$ we have with probability $1 - o(1)$:

$$\hat{Q}_r^{1/2}(\hat{\beta}_r) - \hat{Q}_r^{1/2}(\beta_r^{(1)}) \geq \delta_r \hat{S}_r.$$

For a sequence $C_n \lesssim \sqrt{\frac{s\log(a_n)}{n}}$ independent from r, it holds

$$\begin{aligned}|\delta_r \hat{S}_r| &\leq |\delta_r S_r| + |\delta_r(\hat{S}_r - S_r)| \\ &\lesssim_P \|\delta_r\|_1 \frac{\lambda}{nc} + |\delta_r(\hat{S}_r - S_r)| \\ &\lesssim_P \|\delta_r\|_1 \frac{\lambda}{nc} + C_n \|\delta_r X_r\|_{\mathbb{P}_n,2}.\end{aligned}$$

To obtain the last inequality observe that

$$\begin{aligned}\mathbb{E}_n[(\varepsilon_r+\beta_r^{(2)}X_r)^2] &= \mathbb{E}_n[\varepsilon_r^2]+2\mathbb{E}_n[\varepsilon_r\beta_r^{(2)}X_r]+\underbrace{\mathbb{E}_n[(\beta_r^{(2)}X_r)^2]}_{\geq 0}\\ &\gtrsim \min_{r=1,\dots,d}\mathbb{E}[\varepsilon_r^2]+o_P(1)\\ &\gtrsim c+o_P(1)\end{aligned}$$

is uniformly bounded away from zero, since with probability $1-o(1)$

$$\begin{aligned}\min_{r=1,\dots,d}\mathbb{E}_n[\varepsilon_r\beta_r^{(2)}X_r] &\geq -\max_{r=1,\dots,d}|\mathbb{E}_n[\varepsilon_r\beta_r^{(2)}X_r]|\\ &\geq -\max_{r=1,\dots,d}\sqrt{\mathbb{E}_n[\varepsilon_r^2]\mathbb{E}_n[(\beta_r^{(2)}X_r)^2]}\\ &\gtrsim -\sqrt{\left(\max_{r=1,\dots,d}\mathbb{E}[\varepsilon_r^2]+\bar{\varphi}_n\right)\left(\max_{r=1,\dots,d}\mathbb{E}[(\beta_r^{(2)}X_r)^2]+\frac{s\log(a_n)}{n}\right)}\\ &\gtrsim -\sqrt{\frac{s\log(a_n)}{n}}\end{aligned}$$

uniformly converges towards zero where we used that

$$\max_{r=1,\dots,d}|\mathbb{E}_n[(\beta_r^{(2)}X_r)^2]-\mathbb{E}[(\beta_r^{(2)}X_r)^2]|\lesssim_P \frac{s\log(a_n)}{n}$$

as shown in proof of Theorem 2.
This implies that

$$\begin{aligned}|\delta_r(\hat{S}_r-S_r)| &= \left|\delta_r\left(\frac{\mathbb{E}_n[X_r(\varepsilon_r+\beta_r^{(2)}X_r)]}{\sqrt{\mathbb{E}_n[(\varepsilon_r+\beta_r^{(2)}X_r)^2]}}-\frac{\mathbb{E}_n[X_r\varepsilon_r]}{\sqrt{\mathbb{E}_n[\varepsilon_r^2]}}\right)\right|\\ &= \left|\delta_r\frac{\mathbb{E}_n[X_r(\varepsilon_r+\beta_r^{(2)}X_r)]\sqrt{\mathbb{E}_n[\varepsilon_r^2]}-\mathbb{E}_n[X_r\varepsilon_r]\sqrt{\mathbb{E}_n[(\varepsilon_r+\beta_r^{(2)}X_r)^2]}}{\sqrt{\mathbb{E}_n[(\varepsilon_r+\beta_r^{(2)}X_r)^2]\mathbb{E}_n[\varepsilon_r^2]}}\right|\\ &\lesssim_P \Bigg|\delta_r\Bigg(\mathbb{E}_n[X_r(\beta_r^{(2)}X_r)]\sqrt{\mathbb{E}_n[\varepsilon_r^2]}\\ &\quad+\mathbb{E}_n[X_r\varepsilon_r]\left(\sqrt{\mathbb{E}_n[\varepsilon_r^2]}-\sqrt{\mathbb{E}_n[(\varepsilon_r+\beta_r^{(2)}X_r)^2]}\right)\Bigg)\Bigg|\\ &\leq \left|\mathbb{E}_n[(\delta_rX_r)(\beta_r^{(2)}X_r)]\sqrt{\mathbb{E}_n[\varepsilon_r^2]}\right|\\ &\quad+|\mathbb{E}_n[(\delta_rX_r)\varepsilon_r]|\underbrace{\left|\left(\sqrt{\mathbb{E}_n[\varepsilon_r^2]}-\sqrt{\mathbb{E}_n[(\varepsilon_r+\beta_r^{(2)}X_r)^2]}\right)\right|}_{\leq\sqrt{\mathbb{E}_n[(\beta_r^{(2)}X_r)^2]}}\end{aligned}$$

$$\lesssim \sqrt{\mathbb{E}_n[(\delta_r X_r)^2]\mathbb{E}_n[(\beta_r^{(2)} X_r)^2]\mathbb{E}_n[\varepsilon_r^2]}$$
$$\lesssim_P C_n \|\delta_r X_r\|_{\mathbb{P}_n,2}$$

with an analogous argument as above. Hence, we have with probability $1 - o(1)$

$$\hat{Q}_r^{1/2}(\hat{\beta}_r) - \hat{Q}_r^{1/2}(\beta_r^{(1)}) \geq \delta_r \hat{S}_r \gtrsim -\|\delta_r\|_1 \frac{\lambda}{nc} - C_n \|\delta_r X_r\|_{\mathbb{P}_n,2}. \tag{B.13}$$

Combining the inequalities (B.12) and (B.13) we obtain

$$-\|\delta_r\|_1 \frac{\lambda}{nc} - C_n \|\delta_r X_r\|_{\mathbb{P}_n,2} \lesssim_P \frac{\lambda}{n} \left(\|\delta_{r,T_r}\|_1 - \|\delta_{r,T_r^c}\|_1 \right)$$
$$\iff \|\delta_{r,T_r^c}\|_1 \lesssim_P \underbrace{\frac{c+1}{c-1}}_{:=\tilde{c}} \|\delta_{r,T_r}\|_1 + \frac{n}{\lambda} \frac{c}{c-1} C_n \|\delta_r X_r\|_{\mathbb{P}_n,2}. \tag{B.14}$$

Further we have

$$\hat{Q}_r(\hat{\beta}_r) - \hat{Q}_r(\beta_r^{(1)}) = \|\delta_r X_r\|_{\mathbb{P}_n,2}^2 - 2\mathbb{E}_n[(Y_r - \beta_r^{(1)} X_r)\delta_r X_r]$$

with

$$\mathbb{E}_n[(Y_r - \beta_r^{(1)} X_r)\delta_r X_r] = \mathbb{E}_n[\varepsilon_r \delta_r X_r] + \mathbb{E}_n[(\beta_r^{(2)} X_r)\delta_r X_r]$$
$$\lesssim_P Q_r^{1/2}(\beta_r^{(1)}) ||S_r||_\infty ||\delta_r||_1 + C_n \|\delta_r X_r\|_{\mathbb{P}_n,2}$$

by Hölder inequality. Due to Lemma $P.1$ in [2] with $K \lesssim n^{1/\bar{q}} \log^{\frac{1}{\rho}}(a_n)$, $k \lesssim s$ for a suitable $\bar{q} > \tilde{q}$ and

$$\delta_n \lesssim K\sqrt{s} n^{-1/2} \log(s) \log^{1/2}(a_n) \log^{1/2}(n)$$
$$\lesssim \sqrt{n^{\frac{1}{\bar{q}}} \frac{s \log^{1+\frac{2}{\rho}}(a_n)}{n}} = o(1)$$

by growth condition B4, it holds

$$c \leq \phi_{min}(k, r) \leq \phi_{max}(k, r) \leq C$$

with probability $1 - o(1)$ uniformly over $r = 1, \ldots, d$. Hence, the restricted eigenvalue

$$\kappa_{2\tilde{c}} = \min_{r=1,\ldots,d} \inf_{\delta \in \Delta_{2\tilde{c},r}} \frac{\|\delta X_r\|_{\mathbb{P}_n,2}}{\|\delta\|_2}$$

is bounded away from zero with probability $1 - o(1)$ where

$$\Delta_{2\tilde{c},r} = \{\delta : ||\delta_{T_r^c}||_1 \leq 2\tilde{c}||\delta_{T_r}||_1\}.$$

If $\delta_r \in \Delta_{2\tilde{c},r}$, then

$$\begin{aligned}
\|\delta_r X_r\|^2_{\mathbb{P}_n,2} &= 2\mathbb{E}_n[(Y_r - \beta_r^{(1)} X_r)\delta_r X_r] + [\hat{Q}_r^{1/2}(\hat{\beta}_r) + \hat{Q}_r^{1/2}(\beta_r^{(1)})][\hat{Q}_r^{1/2}(\hat{\beta}_r) - \hat{Q}_r^{1/2}(\beta_r^{(1)})] \\
&\lesssim_P 2Q_r^{1/2}(\beta_r^{(1)})||S_r||_\infty||\delta_r||_1 + 2C_n\|\delta_r X_r\|_{\mathbb{P}_n,2} \\
&\quad + [\hat{Q}_r^{1/2}(\hat{\beta}_r) + \hat{Q}_r^{1/2}(\beta_r^{(1)})]\frac{\lambda}{n}\left(\frac{\sqrt{s}||\delta_r X_r||_{\mathbb{P}_n,2}}{\kappa_{2\tilde{c}}} - ||\delta_{r,T_r^c}||_1)\right).
\end{aligned}$$

Using

$$\hat{Q}_r^{1/2}(\hat{\beta}_r) \leq \hat{Q}_r^{1/2}(\beta_r^{(1)}) + \frac{\lambda}{n}\frac{\sqrt{s}||\delta_r X_r||_{\mathbb{P}_n,2}}{\kappa_{2\tilde{c}}}$$

we conclude

$$\begin{aligned}
\|\delta_r X_r\|^2_{\mathbb{P}_n,2} &\lesssim_P 2Q_r^{1/2}(\beta_r^{(1)})||S_r||_\infty||\delta_r||_1 \\
&\quad + \left[2\hat{Q}_r^{1/2}(\beta_r^{(1)}) + \frac{\lambda}{n}\frac{\sqrt{s}||\delta_r||_{\mathbb{P}_n,2}}{\kappa_{2\tilde{c}}}\right]\frac{\lambda}{n}\left(\frac{\sqrt{s}||\delta_r||_{\mathbb{P}_n,2}}{\kappa_{2\tilde{c}}} - ||\delta_{r,T_r^c}||_1)\right) \\
&\quad + 2C_n\|\delta_r X_r\|_{\mathbb{P}_n,2} \\
&\lesssim_P 2\frac{\lambda}{n}\left(Q_r^{1/2}(\beta_r^{(1)})||\delta_r||_1 - \hat{Q}_r^{1/2}(\beta_r^{(1)})||\delta_{r,T_r^c}||_1\right) \\
&\quad + 2\hat{Q}_r^{1/2}(\beta_r^{(1)})\frac{\lambda}{n}\frac{\sqrt{s}||\delta_r X_r||_{\mathbb{P}_n,2}}{\kappa_{2\tilde{c}}} + \left(\frac{\lambda}{n}\frac{\sqrt{s}||\delta_r X_r||_{\mathbb{P}_n,2}}{\kappa_{2\tilde{c}}}\right)^2 + 2C_n\|\delta_r X_r\|_{\mathbb{P}_n,2}
\end{aligned}$$

with

$$\begin{aligned}
&\left(Q_r^{1/2}(\beta_r^{(1)})||\delta_r||_1 - \hat{Q}_r^{1/2}(\beta_r^{(1)})||\delta_{r,T_r^c}||_1\right) \\
&= \hat{Q}_r^{1/2}(\beta_r^{(1)})||\delta_{r,T_r}||_1 + \left(Q_r^{1/2}(\beta_r^{(1)}) - \hat{Q}_r^{1/2}(\beta_r^{(1)})\right)||\delta_r||_1 \\
&\leq \hat{Q}_r^{1/2}(\beta_r^{(1)})||\delta_{r,T_r}||_1 + \|\beta_r^{(2)} X_r\|_{\mathbb{P}_n,2}||\delta_r||_1 \\
&\lesssim_P \hat{Q}_r^{1/2}(\beta_r^{(1)})||\delta_{r,T_r}||_1 + C_n 3\tilde{c}||\delta_{r,T_r}||_1.
\end{aligned}$$

With probability $1 - o(1)$ we have

$$\begin{aligned}
\|\delta_r X_r\|^2_{\mathbb{P}_n,2} &\lesssim 2\frac{\lambda}{n}||\delta_{r,T_r}||_1\left(\hat{Q}_r^{1/2}(\beta_r^{(1)}) + C_n 3\bar{c}\right) \\
&\quad + 2\hat{Q}_r^{1/2}(\beta_r^{(1)})\frac{\lambda}{n}\frac{\sqrt{s}||\delta_r X_r||_{\mathbb{P}_n,2}}{\kappa_{2\tilde{c}}} + \left(\frac{\lambda}{n}\frac{\sqrt{s}||\delta_r X_r||_{\mathbb{P}_n,2}}{\kappa_{2\tilde{c}}}\right)^2 + 2C_n\|\delta_r X_r\|_{\mathbb{P}_n,2} \\
&\lesssim 2\frac{\lambda}{n}\frac{\sqrt{s}||\delta_r X_r||_{\mathbb{P}_n,2}}{\kappa_{2\tilde{c}}}\left(\hat{Q}_r^{1/2}(\beta_r^{(1)}) + C_n 3\bar{c}\right) \\
&\quad + 2\hat{Q}_r^{1/2}(\beta_r^{(1)})\frac{\lambda}{n}\frac{\sqrt{s}||\delta_r X_r||_{\mathbb{P}_n,2}}{\kappa_{2\tilde{c}}} + \left(\frac{\lambda}{n}\frac{\sqrt{s}||\delta_r X_r||_{\mathbb{P}_n,2}}{\kappa_{2\tilde{c}}}\right)^2 + 2C_n\|\delta_r X_r\|_{\mathbb{P}_n,2}
\end{aligned}$$

and therefore obtain

$$\left(1-\left(\frac{\lambda}{n}\frac{\sqrt{s}}{\kappa_{2\tilde{c}}}\right)^2\right)\|\delta_r X_r\|^2_{\mathbb{P}_n,2} \lesssim_P \left(4\hat{Q}_r^{1/2}(\beta_r^{(1)})\frac{\lambda}{n}\frac{\sqrt{s}}{\kappa_{2\tilde{c}}} + C_n\left(6\tilde{c}\frac{\lambda}{n}\frac{\sqrt{s}}{\kappa_{2\tilde{c}}}+2\right)\right)||\delta_r X_r||_{\mathbb{P}_n,2},$$

which implies

$$\|\delta_r X_r\|_{\mathbb{P}_n,2} \lesssim_P \frac{\lambda\sqrt{s}}{n} + C_n \lesssim \sqrt{\frac{s\log(a_n)}{n}}.$$

Here we used that

$$\hat{Q}_r^{1/2}(\beta_r^{(1)}) = \mathbb{E}_n[(\varepsilon_r + \beta_r^{(2)}X_r)^2]^{1/2} \leq \|\varepsilon_r\|_{\mathbb{P}_n,2} + \|\beta_r^{(2)}X_r\|_{\mathbb{P}_n,2} \lesssim_P C + \bar{\varphi}_n + C_n.$$

If $\delta_r \notin \Delta_{2\tilde{c},r}$ (implying $||\delta_{r,T_r^c}||_1 > 2\tilde{c}||\delta_{r,T_r}||_1$), (B.14) directly implies

$$2\tilde{c}||\delta_{r,T_r}||_1 \lesssim_P \tilde{c}\|\delta_{r,T_r}\|_1 + \frac{n}{\lambda}\frac{c}{c-1}C_n\|\delta_r X_r\|_{\mathbb{P}_n,2}$$

and therefore

$$||\delta_{r,T_r}||_1 \lesssim_P \frac{n}{\lambda}\frac{c}{c-1}C_n\|\delta_r X_r\|_{\mathbb{P}_n,2}$$

due to $\tilde{c} \geq 1$. Additionally (B.14) implies

$$\|\delta_{r,T_r^c}\|_1 \lesssim_P \frac{1}{2}\|\delta_{r,T_r^c}\|_1 + \frac{n}{\lambda}\frac{c}{c-1}C_n\|\delta_r X_r\|_{\mathbb{P}_n,2}$$

and therefore

$$\|\delta_{r,T_r^c}\|_1 \lesssim_P \frac{2n}{\lambda}\frac{c}{c-1}C_n\|\delta_r X_r\|_{\mathbb{P}_n,2},$$

which, combined with the inequality above, implies

$$\|\delta_r\|_1 \lesssim_P \frac{3n}{\lambda}\frac{c}{c-1}C_n\|\delta_r X_r\|_{\mathbb{P}_n,2}.$$

Using

$$\hat{Q}_r^{1/2}(\hat{\beta}_r) - \hat{Q}_r^{1/2}(\beta_r^{(1)}) \leq \frac{\lambda}{n}\left(\|\delta_{r,T_r}\|_1 - \|\delta_{r,T_r^c}\|_1\right) \leq \frac{\lambda}{n}\|\delta_r\|_1$$

and following the same argument as above we obtain with probability $1 - o(1)$:

$$\begin{aligned}
\|\delta_r X_r\|^2_{\mathbb{P}_n,2} &= 2\mathbb{E}_n[(Y_r - \beta_r^{(1)\prime} X_r)\delta_r X_r] + [\hat{Q}_r^{1/2}(\hat{\beta}_r) + \hat{Q}_r^{1/2}(\beta_r^{(1)})][\hat{Q}_r^{1/2}(\hat{\beta}_r) - \hat{Q}_r^{1/2}(\beta_r^{(1)})] \\
&\lesssim 2Q_r^{1/2}(\beta_r^{(1)})||S_r||_\infty||\delta_r||_1 + 2C_n\|\delta_r X_r\|_{\mathbb{P}_n,2} \\
&\quad + \left(2\hat{Q}_r^{1/2}(\beta_r^{(1)}) + \frac{\lambda}{n}\|\delta_r\|_1\right)\frac{\lambda}{n}\|\delta_r\|_1 \\
&\lesssim \left(2\frac{1}{c}\underbrace{\left(Q_r^{1/2}(\beta_r^{(1)}) - \hat{Q}_r^{1/2}(\beta_r^{(1)})\right)}_{\lesssim C_n} + 2\left(\frac{1}{c}+1\right)\hat{Q}_r^{1/2}(\beta_r^{(1)}) + \frac{\lambda}{n}\|\delta_r\|_1\right)\frac{\lambda}{n}\|\delta_r\|_1 \\
&\quad + 2C_n\|\delta_r X_r\|_{\mathbb{P}_n,2} \\
&\leq 6\left(\frac{C_n}{c} + \left(\frac{1}{c}+1\right)\hat{Q}_r^{1/2}(\beta_r^{(1)})\right)\frac{c}{c-1}C_n\|\delta_r X_r\|_{\mathbb{P}_n,2} \\
&\quad + \left(3\frac{c}{c-1}C_n\|\delta_r X_r\|_{\mathbb{P}_n,2}\right)^2 + 2C_n\|\delta_r X_r\|_{\mathbb{P}_n,2}.
\end{aligned}$$

Hence,

$$\begin{aligned}
\left(1 - \left(3\frac{c}{c-1}C_n\right)^2\right)\|\delta_r X_r\|^2_{\mathbb{P}_n,2} \lesssim_P 6\left(\frac{C_n}{c} + \left(\frac{1}{c}+1\right)\hat{Q}_r^{1/2}(\beta_r^{(1)})\right)\frac{c}{c-1}C_n\|\delta_r X_r\|_{\mathbb{P}_n,2} \\
+ 2C_n\|\delta_r X_r\|_{\mathbb{P}_n,2}
\end{aligned}$$

which implies

$$\|\delta_r X_r\|_{\mathbb{P}_n,2} \lesssim_P C_n \lesssim \sqrt{\frac{s\log(a_n)}{n}}.$$

To prove the second claim observe that

$$\begin{aligned}
\|\delta_r\|_1 &= 1_{\{\delta_r \in \Delta_{2\tilde{c},r}\}}\|\delta_r\|_1 + 1_{\{\delta_r \notin \Delta_{2\tilde{c},r}\}}\|\delta_r\|_1 \\
&\leq 1_{\{\delta_r \in \Delta_{2\tilde{c},r}\}}(1+2\tilde{c})\|\delta_{r,T_r}\|_1 + 1_{\{\delta_r \notin \Delta_{2\tilde{c},r}\}}\|\delta_r\|_1 \\
&\lesssim_P \left((1+2\tilde{c})\frac{\sqrt{s}}{\kappa_{2\tilde{c}}} + \frac{3n}{\lambda}\frac{c}{c-1}C_n\right)\|\delta_r X_r\|_{\mathbb{P}_n,2} \\
&\lesssim_P \sqrt{\frac{s^2\log(a_n)}{n}}
\end{aligned}$$

uniformly over all $r = 1, \dots, d$. Now, we proof that

$$\max_{r=1,\dots,d} \|\hat{\beta}_r\|_0 \lesssim s.$$

This proof is derived from the proof of lemma L.2. from Belloni et al. (2018) [2]. At first observe that

$$0 < c \lesssim_P \min_{r=1,\dots,d} \|\varepsilon_r + \beta_r^{(2)} X_r\|_{\mathbb{P}_n,2}^2 \leq \max_{r=1,\dots,d} \|\varepsilon_r + \beta_r^{(2)} X_r\|_{\mathbb{P}_n,2}^2 \lesssim_P C < \infty$$

where the first inequality is shown above and the second follows with an analogous argument. Additionally we obtain

$$\max_{r=1,\dots,d} \left| \|Y_r - \hat{\beta}_r X_r\|_{\mathbb{P}_n,2}^2 - \|\varepsilon_r + \beta_r^{(2)} X_r\|_{\mathbb{P}_n,2}^2 \right| \lesssim_P C_n + C_n^2 = o(1)$$

due to

$$\|Y_r - \hat{\beta}_r X_r\|_{\mathbb{P}_n,2}^2 = \|\varepsilon_r + \beta_r^{(2)} X_r\|_{\mathbb{P}_n,2}^2 - 2\mathbb{E}_n[(\varepsilon_r + \beta_r^{(2)} X_r)\delta_r X_r] + \underbrace{\|\delta_r X_r\|_{\mathbb{P}_n,2}^2}_{\lesssim_P C_n^2}$$

with

$$\begin{aligned} |\mathbb{E}_n[(\varepsilon_r + \beta_r^{(2)} X_r)\delta_r X_r]| &\leq \sqrt{\mathbb{E}_n[(\varepsilon_r + \beta_r^{(2)} X_r)^2]\mathbb{E}_n[(\delta_r X_r)^2]} \\ &\lesssim (C + o_P(1)) \, \|\delta_r X_r\|_{\mathbb{P}_n,2} \\ &\lesssim_P C_n \end{aligned}$$

uniformly over all $r = 1, \dots, d$. This implies

$$\begin{aligned} &|\delta(\partial_\beta \hat{Q}_r^{1/2}(\beta)|_{\beta=\hat{\beta}_r} - \hat{S}_r)| \\ =&\left| \delta \left(\frac{\mathbb{E}_n[X_r(Y_r - \beta_r^{(1)} X_r)]}{\sqrt{\mathbb{E}_n[(Y_r - \beta_r^{(1)} X_r)^2]}} - \frac{\mathbb{E}_n[X_r(Y_r - \hat{\beta}_r X_r)]}{\sqrt{\mathbb{E}_n[(Y_r - \hat{\beta}_r X_r)^2]}} \right) \right| \\ =&\left| \delta \left(\frac{\mathbb{E}_n[X_r(Y_r - \beta_r^{(1)} X_r)]\|Y_r - \hat{\beta}_r X_r\|_{\mathbb{P}_n,2} - \|\varepsilon_r + \beta_r^{(2)} X_r\|_{\mathbb{P}_n,2}\mathbb{E}_n[X_r(Y_r - \hat{\beta}_r X_r)]}{\|\varepsilon_r + \beta_r^{(2)} X_r\|_{\mathbb{P}_n,2}\|Y_r - \hat{\beta}_r X_r\|_{\mathbb{P}_n,2}} \right) \right| \\ \lesssim_P &\left| \delta \left(\mathbb{E}_n[X_r(Y_r - \beta_r^{(1)} X_r)] - \mathbb{E}_n[X_r(Y_r - \hat{\beta}_r X_r)] \right) \right| \\ \leq &\|\delta_r X_r\|_{\mathbb{P}_n,2} \|\delta X_r\|_{\mathbb{P}_n,2} \lesssim_P C_n \|\delta X_r\|_{\mathbb{P}_n,2}. \end{aligned}$$

By the definition of $\hat{\beta}_r$, there exists a subgradient $\partial_\beta \hat{Q}_r^{1/2}(\beta)|_{\beta=\hat{\beta}_r}$ of $\hat{Q}_r^{1/2}(\hat{\beta}_r)$ such that for every j with $|\hat{\beta}_{r,j}| > 0$

$$|(\partial_\beta \hat{Q}_r^{1/2}(\beta)|_{\beta=\hat{\beta}_r})_j| = \frac{\lambda}{n}.$$

Let $\hat{T}_r := \mathrm{supp}(\hat{\beta}_r)$ and $|\hat{T}_r| := \hat{s}_r$. We obtain

$$\begin{aligned}
\frac{\lambda}{n}\sqrt{\hat{s}_r} &= \|(\partial_\beta \hat{Q}_r^{1/2}(\beta)|_{\beta=\hat{\beta}_r})_{\hat{T}_r}\|_2 \\
&\le \|S_{r\hat{T}_r}\|_2 + \|(\hat{S}_r - S_r)_{\hat{T}_r}\|_2 + \|(\partial_\beta \hat{Q}_r^{1/2}(\beta)|_{\beta=\hat{\beta}_r} - \hat{S}_r)_{\hat{T}_r}\|_2 \\
&\lesssim_P \sqrt{\hat{s}_r}\|S_r\|_\infty \\
&\quad + C_n \sup_{\|\delta\|_2=1, \|\delta\|_0 \le \hat{s}_r} \|\delta X_r\|_{\mathbb{P}_n,2} \\
&\quad + \sup_{\|\delta\|_2=1, \|\delta\|_0 \le \hat{s}_r} |\delta(\partial_\beta \hat{Q}_r^{1/2}(\beta)|_{\beta=\hat{\beta}_r} - \hat{S}_r)| \\
&\lesssim_P \sqrt{\hat{s}_r}\frac{\lambda}{nc} + 2C_n \sup_{\|\delta\|_2=1, \|\delta\|_0 \le \hat{s}_r} \|\delta X_r\|_{\mathbb{P}_n,2},
\end{aligned}$$

where we used

$$\|(\hat{S}_r - S_r)_{\hat{T}_r}\|_2 \le \sup_{\|\delta\|_2=1, \|\delta\|_0 \le \hat{s}_r} |\delta(\hat{S}_r - S_r)| \lesssim_P C_n \sup_{\|\delta\|_2=1, \|\delta\|_0 \le \hat{s}_r} \|\delta X_r\|_{\mathbb{P}_n,2}.$$

Hence with probability $1 - o(1)$,

$$\begin{aligned}
\hat{s}_r &\le \left(\frac{2CnC_n}{\lambda(1-1/c)}\right)^2 \sup_{\|\delta\|_2=1, \|\delta\|_0 \le \hat{s}_r} \|\delta X_r\|^2_{\mathbb{P}_n,2} \\
&\le \Bigg(\underbrace{\frac{2CnC_n}{\lambda(1-1/c)}}_{:=L}\Bigg)^2 \phi_{max}(\hat{s}_r, r) \lesssim s\phi_{max}(\hat{s}_r, r)
\end{aligned} \tag{B.15}$$

where

$$\phi_{max}(\hat{s}_r, r) := \max_{\|\delta\|_0 \le \hat{s}_r} \frac{\|\delta X_r\|^2_{\mathbb{P}_n,2}}{\|\delta\|_2^2}.$$

We can find a suitable C such that $M = Cs \in \mathcal{M}_r$ with

$$\mathcal{M}_r := \{m \in \mathbb{N} : m > 2\phi_{max}(m, r)L^2\}.$$

Suppose that $\hat{s}_r > M$. By the sublinearity of the maximum sparse eigenvalue (Lemma 3 in [1]), for any integer $k \ge 0$ and constant $l \ge 0$, we have

$$\phi_{max}(lk, r) \le \lceil l \rceil \phi_{max}(k, r)$$

where $\lceil l \rceil$ denotes the ceiling of l. Since $\lceil k \rceil \le 2k$ for any $k \ge 1$,

$$\begin{aligned}
\hat{s}_r &\le L^2\phi_{max}(\hat{s}_r, r) = L^2\phi_{max}(M\hat{s}_r/M, r) \\
&\le \left\lceil \frac{\hat{s}_r}{M} \right\rceil L^2\phi_{max}(M, r) \le \frac{2\hat{s}_r}{M} L^2\phi_{max}(M, r)
\end{aligned}$$

that violates the condition that $M \in \mathcal{M}_r$. Therefore, we have $\hat{s}_r \leq M$. Applying B.15, we obtain

$$\max_{r=1,\ldots,d} \hat{s}_r \leq \max_{r=1,\ldots,d} \phi_{max}(M, r)s \lesssim s$$

with probability $1 - o(1)$ and the stated claim follows:

$$\max_{r=1,\ldots,d} \|\hat{\beta}_r\|_0 \lesssim s.$$

Since the maximal sparse eigenvalues are uniformly bounded from above, we conclude

$$\max_{r=1,\ldots,d} \|\hat{\beta}_r - \beta_r^{(1)}\|_2 \lesssim \max_{r=1,\ldots,d} \|(\hat{\beta}_r - \beta_r^{(1)})X_r\|_{\mathbb{P}_n,2} \lesssim C_n$$

with probability at least $1 - o(1)$. ∎

REFERENCES

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

[2] Belloni, A., Chernozhukov, V., Chetverikov, D., Wei, Y. et al. (2018). Uniformly valid post-regularization confidence regions for many functional parameters in z-estimation framework. *The Annals of Statistics* **46** 3643–3675.

[3] Belloni, A., Chernozhukov, V. and Kato, K. (2014). Uniform post selection inference for LAD regression and other Z-estimation problems Technical Report, cemmap working paper, Centre for Microdata Methods and Practice.

[4] Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics* **37** 1705–1732.

[5] Cai, T., Liu, W. and Luo, X. (2011). A constrained l1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* **106** 594–607.

[6] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2017). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*.

[7] Chernozhukov, V., Chetverikov, D., Kato, K. et al. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics* **41** 2786–2819.

[8] Chernozhukov, V., Chetverikov, D., Kato, K. et al. (2017). Central limit theorems and bootstrap in high dimensions. *The Annals of Probability* **45** 2309–2352.

[9] Janková, J. and van de Geer, S. (2017). Honest confidence regions and optimality in high-dimensional precision matrix estimation. *Test* **26** 143–162.

[10] Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of statistics* **37** 4254.

[11] Lauritzen, S. L. (1996). *Graphical models* **17**. Clarendon Press.

[12] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The annals of statistics* 1436–1462.

[13] Ravikumar, P., Wainwright, M. J., Raskutti, G., Yu, B. et al. (2011). High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence. *Electronic Journal of Statistics* **5** 935–980.

[14] Ren, Z., Sun, T., Zhang, C.-H., Zhou, H. H. et al. (2015). Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *The Annals of Statistics* **43** 991–1026.

[15] Rothman, A. J., Bickel, P. J., Levina, E., Zhu, J. et al. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* **2** 494–515.

[16] Sun, T. and Zhang, C.-H. (2013). Sparse matrix inversion with scaled lasso. *The Journal of Machine Learning Research* **14** 3385–3418.

[17] Van der Vaart, A. and Wellner, J. (1996). Weak convergence and empirical processes.

[18] Vershynin, R. (2017). *High-Dimensional Probability*. Cambridge University Press (to appear).

[19] Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research* **11** 2261–2286.

[20] Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35.

[21] Zhao, T., Liu, H., Roeder, K., Lafferty, J. and Wasserman, L. (2012). The huge package for high-dimensional undirected graph estimation in R. *Journal of Machine*

Learning Research **13** 1059–1062.

Sven Klaassen
University of Hamburg
Hamburg Business School
Moorweidenstr. 18
20148 Hamburg
Germany
E-mail: sven.klaassen@uni-hamburg.de

Jannis Kück
University of Hamburg
Hamburg Business School
Moorweidenstr. 18
20148 Hamburg
Germany
E-mail: jannis.kueck@uni-hamburg.de

Martin Spindler
University of Hamburg
Hamburg Business School
Moorweidenstr. 18
20148 Hamburg
Germany
E-mail: martin.spindler@uni-hamburg.de

Victor Chernozhukov
Massachusetts Institute of Technology
Department of Economics and
Operations Research Center
50 Memorial Drive
Cambridge, MA 02142
USA
E-mail: vchern@mit.edu