Advances in Machine Learning: Valid Inference about High-Dimensional Parameters

Cumulative Dissertation
to obtain the academic degree of a "Doctor rerum oeconomicarum" (Dr. rer. oec.)
according to doctoral degree regulations 2014

at the Faculty of Business Administration (Hamburg Business School)
Moorweidenstr. 18
20148 Hamburg (Germany)
University of Hamburg

submitted by: Jannis Malte Kück
born on February 29, 1992 in Bremervörde

Hamburg, 2020

Thesis Committee:
Chairman: Prof. Dr. Knut Haase
1st Examiner: Prof. Dr. Martin Spindler
2nd Examiner: Prof. Dr. Michael Merz
3rd Examiner: Prof. Dr. Matthew Harding

Date of Disputation: November 12, 2020

Acknowledgments

First and foremost, I would like to thank my advisor Martin Spindler who was a constant source of inspiration for my academic research during the last four years. He was always available and willing to help me when I needed his support for both research and organizational issues. In particular, I am indebted to Martin for his efforts that allowed me to visit the Deep Data Lab at the University of California which had a major impact on both my academic and personal development. Moreover, we are a perfect team at the chair of statistics at Hamburg Business School and I am proud of all the things that we have achieved together. In this context, I would like to thank Philipp Bach and Sven Klaaßen. I highly appreciate to have such great colleagues. I am grateful for their support and all the valuable discussions with them. I also thank Cornelia Hartwig for helping me with the administrative issues.

Furthermore, I thank my coauthors Victor Chernozhukov and Ye Luo, and Michael Merz for agreeing to act as second reviewer. I would like to gratefully mention Matthew Harding for inviting me to visit the Deep Data Lab at the University of California. He also supported me in the last few months, e.g., by providing me reference letters and by agreeing to be on my committee. I am grateful to Natalie Neumeyer for her great teaching and the supervision of my master thesis during my studies of business mathematics. She contributed significantly to my desire of pursuing a Ph.D. in statistics.

This dissertation concludes an important chapter of my life. I have started my studies almost nine years ago and I am grateful to all the people who have accompanied me along this way. I especially thank Moritz Meyer for always supporting me and for being an outstanding friend.

Most of all, I thank Friederike Falk for her unconditional love and support. Friedi, although I am often highly focused on my work, you are always first and foremost in my mind.

Finally, I would like to express my deep gratitude to my family. In particular, I would like to thank my parents. They have supported and encouraged me in every possible way, especially during difficult times. Thank you for always being there for me.

Contents

List of Figures

List of Tables

Chapter 1

General Introduction

1.1 Background

High-dimensional statistical models have become increasingly popular in the last decades. These models allow for settings where the number of variables p is large (or even larger) compared to the sample size n. The rise of digitalization has dramatically lowered the cost of data acquisition and allows to collect data of many variables. Thus, Big Data becomes more and more available which has accelerated the development of new methods. Key methods in high-dimensional linear regression include Lasso (Tibshirani [93]), the Dantzig selector (Candes and Tao [25]) and L_2-Boosting (Friedman [48], Bühlmann and Yu [23]). Popular nonlinear regression methods in high-dimensions are tree-based methods like Random Forests (Breiman [19]) and Neural Nets. All these so-called machine learning (ML) methods are remarkably effective in prediction contexts. Under structural assumptions, such as sparsity, together with various regularization schemes, machine learning methods achieve impressively fast estimation rates. L_1-penalized methods like Lasso and the Dantzig selector are widely discussed in the recent literature and the estimation rates are well understood, see Bickel et al. [13] and Bühlmann and Van De Geer [22], among many others. Estimation rates for L_2-Boosting in sparse linear regression models are presented in Luo and Spindler [73]. Wager and Walther [100] provide concentration results for Regression Trees and Random Forests and Chen and White [29] provide results for Neural Nets.

However, this good performance in prediction does not necessarily translate into good performance for inference about causal parameters that is key for economic applications. In many situations, the interest is on learning about causal relationships and making inference about treatment effects. In a naive approach in regression models with many explanatory variables, one might first select the relevant variables by machine learning methods, like Lasso and Boosting, and then estimate the treatment effect by including only the selected variables and continue with standard inference methods. Although it is frequently used in applied work, this approach leads to invalid results since it relies on perfect variable selection in the first step. This has been excellently highlighted in Leeb and Pötscher [67]. Thus, the challenge is to combine machine learning with causal inference.

The papers that are collected in this dissertation discuss how machine learning methods can be used to conduct valid inference in high-dimensional settings. The methodology that is used in the papers relies on the so-called double machine learning approach. The theoretical framework has been developed by Belloni, Chernozhukov, Hansen, and coauthors, in a series of papers. Chernozhukov et al. [35] first introduced the general double machine learning approach to construct confidence intervals for low-dimensional target parameters in the presence of an unknown, high-dimensional nuisance parameter which can be estimated with machine learning methods. Recent results of Belloni et al. [12] allow for valid inference about high-dimensional target parameters by allowing the number of moment conditions to grow with the sample

size. It is worth to notice that there is a second so-called debiasing approach for inference about low-dimensional parameters in high-dimensions that has been introduced in Van De Geer et al. [96] and Zhang and Zhang [105]. In the following, the conceptual framework of the double machine learning approach will be introduced.

1.2 Conceptual Framework

Let W be a random element with values in a measurable space $(\mathcal{W}, \mathcal{A}_{\mathcal{W}})$ with probability measure $P \in \mathcal{P}_n$. In regression settings, the random variable W often equals the tuple $W = (Y, X)$. Further, it is assumed to observe n independent identically distributed (i.i.d.) observations of W. The (causal) target parameter $\theta_0 \in \mathbb{R}$ is identified by fulfilling the following moment condition

$$\mathbb{E}[\psi(W, \theta_0, \eta_0)] = 0.$$

Here, $\psi(\cdot)$ is a known score function and $\eta_0 \in T$ is an unknown, high-dimensional nuisance parameter, where T is a convex subset of a normed vector space. The double machine learning estimator $\hat{\theta}_0$ solves the empirical version of the moment condition

$$\frac{1}{n} \sum_{i=1}^{n} \psi(W_i, \hat{\theta}_0, \hat{\eta}) = 0,$$

where instead of the unknown nuisance parameter a ML-based estimator $\hat{\eta}$ is plugged in. The following Neyman orthogonality condition is essential for valid inference in high-dimensions. It ensures that the Gateaux derivative with respect to the nuisance parameter vanishes at zero:

$$\partial_r \mathbb{E}\left[\psi(W, \theta_0, \eta_0 + r(\eta - \eta_0))\right]|_{r=0} = 0.$$

Heuristically, the condition implies that the moment condition to identify θ_0 remains valid under local mistakes in the nuisance parameter. This idea can be traced back to Neyman who used a similar condition for robust estimation in low-dimensional settings. The Neyman orthogonality does not need to hold for all $\eta \in T$ but only for the so-called nuisance realization set $\mathcal{T} \subset T$ that includes the ML-based estimator $\hat{\eta}$ with probability converging to one. In the following section, the role of the Neyman orthogonality is discussed in detail and an intuition why this condition is key for valid inference in high-dimensional settings is provided.
There are several extensions of this basic framework, e.g., the near orthogonality condition that only assumes that the Gateaux derivative is closed to zero and vanishes sufficiently fast for increasing sample size. Further, as already mentioned, Belloni et al. [12] provide a framework for valid inference about a high-dimensional target parameter $\theta_0 = (\theta_1, \ldots, \theta_{d_n})$ by allowing the number of moment conditions, $l = 1, \ldots, d_n$, to increase with the sample size. Here, d_n denotes the number of target parameters.

The Role of Neyman Orthogonality

Dealing with high-dimensional parameters requires relying upon regularization that leads to a substantial bias and this bias spreads into the estimation of the target parameters. This is the reason why naive inference approaches tend to fail in high-dimensions. Under weak regularity conditions, the double machine learning estimator $\hat{\theta}$ obeys the expansion

$$J\sqrt{n}\big(\hat{\theta} - \theta_0\big) = A_n + \sqrt{n}DO\big(|\hat{\eta} - \eta_0|\big) + \sqrt{n}O\left(\|\hat{\eta} - \eta_0\|^2\right) + o_p(1),$$

where J is a variance term, A_n is a well-behaved leading term that is approximately zero-mean Gaussian and D is defined as

$$D := \partial_r \mathbb{E}\left[\psi(W, \theta_0, \eta_0 + r(\eta - \eta_0))\right]|_{r=0}.$$

Under Neyman orthogonality, it holds $D = 0$ and therefore

$$\sqrt{n}\mathrm{D}O\big(|\hat{\eta} - \eta_0|\big) = 0.$$

Hence,

$$\sqrt{n}O\left(\|\hat{\eta} - \eta_0\|^2\right) = o_P(1)$$

is sufficient for root-n consistency and asymptotic normality of the double machine learning estimator $\hat{\theta}$ which only requires $\|\hat{\eta} - \eta_0\| = o_P(n^{-1/4})$. As mentioned, machine learning methods achieve impressively fast estimation rates and $o_P(n^{-1/4})$ is often an attainable rate for estimating η_0. The Neyman orthogonality condition ensures that the moment condition is insensitive towards these small estimation errors which leads to valid inference. Considering valid inference about a high-dimensional target parameter $\theta_0 = (\theta_1, \ldots, \theta_{d_n})$ in the presence of high-dimensional nuisance parameters $\eta_{0,1}, \ldots, \eta_{0,d_n}$, this estimation rate needs to hold uniformly over all nuisance parameters, namely

$$\sup_{l=1,\ldots,d_n} \|\hat{\eta}_l - \eta_{0,l}\| = o_P(n^{-1/4}).$$

1.3 Outline

This dissertation consists of four research papers that present a variety of applications of the double machine learning approach with the aim to provide new methodology for valid inference about a potentially high-dimensional target parameter.

The first paper, presented in Chapter 2, analyzes the following high-dimensional linear regression model

$$Y = D\theta_0 + X_1\beta_1 + \ldots + X_p\beta_p + \varepsilon, \quad \mathbb{E}[\varepsilon \mid X, D] = 0$$

with p potentially much larger than the sample size n. Here, D is a treatment variable and $X_1, \ldots, X_p$ are additional covariates. This model is well known in the literature and the double machine learning approach can be used to conduct valid inference. The estimation of the treatment effect θ_0 often relies on Lasso estimation. Contrary to this, results for valid inference when post- or orthogonal L_2-Boosting is applied for variable selection are provided in Chapter 2.

In the second paper, presented in Chapter 3, the following high-dimensional transformation model

$$\Lambda_{\theta_0}(Y) = X^T\beta_{\theta_0} + \varepsilon_{\theta_0}$$

with $\varepsilon_{\theta_0} \sim \mathcal{N}(0, \sigma^2)$ is considered. This model takes up the idea of the high-dimensional linear regression model and combines it with a parametric transformation of the response variable $\Lambda_\theta(\cdot) \in \mathcal{F}_\Lambda$, where $\mathcal{F}_\Lambda = \{\Lambda_\theta(\cdot) : \theta \in \Theta\}$ is a given family of strictly monotone increasing functions. The transformation allows for more flexibility and aims to change the scale preventing incorrect model assumptions, such as by establishing normally distributed errors. In Chapter 3, an estimator for the true transformation parameter θ_0 is proposed and proven to be asymptotically normally distributed.

The third paper, which provides the basis of Chapter 4, considers a generalized additive model. General-

ized additive models are quite popular in statistics, imposing an additive structure of the nonparametric regression function to evade the curse of dimensionality

$$Y = \beta + f_1(X_1) + \ldots + f_p(X_p) + \varepsilon, \quad \mathbb{E}[\varepsilon|X] = 0.$$

Here, β denotes a constant and $f_1(\cdot), \ldots, f_p(\cdot)$ are univariate regression functions. Chapter 4 provides a new methodology for uniform valid confidence bands of the nonparametric target component f_1.
Chapter 5 provides the fourth paper which analyzes high-dimensional Gaussian graphical models of the form

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

Graphical models are key for representing dependencies of a large set of variables. The aim of this paper is to quantify the uncertainty of recovering the support of the precision matrix Σ_X^{-1} by providing a significance test for a set of potential edges in the graphical model.
Finally, Chapter 6 draws general conclusions and provides an outlook on future research. This dissertation has its origin in four joint works with various coauthors and as the dissertation was written in a cumulative way some sections of the chapters are similar, e.g., the sections that introduce the notations. But to avoid confusion and unnecessary cross references, these sections have been retained.

Chapter 2

Estimation and Inference of Treatment Effects with L_2-Boosting in High-Dimensional Settings

2.1 Introduction

Boosting algorithms are very popular in machine learning and have proven to be very useful for prediction and variable selection (Bühlmann and Hothorn [21]). Nevertheless, in many applications the researcher is interested in inference on selected variables. In many cases there are so-called treatment or policy variables which the researcher would like to learn about and make inferences, in particular in a high-dimensional setting. Increasing digitalization in many fields of life makes large data sets available for research. Typical applications are the estimation of a treatment effect after selecting among many control variables and the estimation of instrumental variables when there are potentially many instruments. We provide results for valid inference in these settings when post- or orthogonal L_2-Boosting is applied for the variable selection. Usually, inference after model selection leads to invalid results. This has been highlighted by Leeb and Pötscher in a series of papers, excellently summarized in Leeb and Pötscher [67]. Here, we use orthogonalized moment conditions introduced by Chernozhukov et al. [35] and recent results of Luo and Spindler [73] on the rate of convergence of L_2-Boosting which yields valid post-selection inference.

Boosting algorithms represent one of the major advances in machine learning and statistics in recent years. Freund and Schapire's AdaBoost algorithm for classification (Freund and Schapire [45]) has attracted much attention from the machine learning community as well as in statistics. Many variants of the AdaBoost algorithm have been introduced and proven to be very competitive in terms of prediction accuracy in a variety of applications with a strong resistance to overfitting as shown in Bühlmann and Hothorn [21]. Boosting methods were originally proposed as ensemble methods which rely on the principle of generating multiple predictions and majority voting (averaging) of the individual classifiers. An important step in the analysis of Boosting algorithms was Breiman's interpretation of Boosting as a gradient descent algorithm in a function space inspired by numerical optimization and statistical estimation (Breiman [17], Breiman [18]). Building on this insight, Friedman et al. [46] and Friedman [48] embedded Boosting algorithms into the framework of statistical estimation and additive basis expansion. This also enabled the application of Boosting to regression analysis. Boosting for regression was proposed by Friedman [48], and then Bühlmann and Yu [23] defined and introduced L_2-Boosting. An extensive overview of the development of Boosting and its manifold applications is given in the survey Bühlmann

and Hothorn [21].
In this paper, we present results for valid inference on treatment effects in a high-dimensional setting. Boosting has proven to be particularly valuable for prediction, but we show in this paper that it can also be applied for causal search. In particular, we consider the case of the estimation of a treatment effect with many control variables, and the estimation of instrumental variables (IVs) with many potential instruments. The first case, the estimation of a treatment effect with many control variables, can also be interpreted as inference on a preselected variable in a high-dimensional linear regression model estimated with L_2-Boosting. Our estimation method relies on the so-called orthogonalized moment conditions. This theory was developed by Belloni, Chernozhukov, Hansen, and coauthors, in a series of papers. The case of instrumental variables is analyzed in Belloni et al. [5], the treatment effect case in Belloni et al. [8]. Surveys with extensions of the general idea are Chernozhukov et al. [35] and Chernozhukov et al. [33]. To ground the discussion, we examine a randomized trial of the pulmonary artery catheter (PAC) that was carried out in 65 intensive care units in the UK between 2001 and 2004 (Harvey et al. [53]). This study got a lot of attention from the scientific community under the name the "PAC-man"study. The PAC is a monitoring device commonly inserted into critically ill patients while staying in intensive care units. It provides continuous measurements of cardiac activity. However, the insertion of a PAC is an invasive procedure bringing the risk of complications and imposing significant costs as described in Dalen [39]. An early study based on observational data by Connors et al. [37] found that a PAC had a negative effect on the survival chances of patients and led to increased costs for the health care sector. This finding was the motivation for a randomized trial by Bloniarz et al. [15] to evaluate PAC interventions. In this study, around 1,000 patients (approx. 50% treatment and 50% control groups) participated and a large number of covariates were collected. If, e.g., two-way interactions of the variables are included in the analysis, the number of parameters already exceeds the number of observations. We analyze the PAC-man data and find that the intervention has no significant effect on the outcome variable, namely the number of quality-adjusted years of life.

First, we explain, in Section 2.2, the problems in estimating treatment effects in high-dimensional settings. In Section 2.3, L_2-Boosting and two variants, to which our results apply, are introduced. In Section 2.4, we present the formal results for valid inference on (low-)dimensional treatment effects in a possibly high-dimensional setting. A simulation study and an empirical application are given in Sections 2.5 and 2.6 . Finally, we conclude in Section 2.7.

2.2 Econometric Considerations

The goal is to estimate the treatment effect α of a treatment variable D on an outcome variable Y, namely

$$Y = \gamma + \alpha D + \varepsilon, \tag{2.1}$$

where γ denotes the intercept and ε a statistical error term. There are two reasons for including covariates $X = (X_1, \ldots, X_p)$ in equation (2.1) for the estimation of the treatment effect. First, covariates improve the precision of the estimation of the average treatment effect in randomized control trials (RCTs). This argument has already been made in Cox [38]. Second, in observational studies, additional covariates might establish unconfoundedness, meaning that given the variables in X, the treatment is as randomized and there are no unobserved confounders. For a book length treatment of this argument, we refer to Imbens and Rubin [57]. Formally, this means

$$Y = \gamma + \alpha D + g(X) + \varepsilon, \quad \mathbb{E}(\varepsilon|D, X) = 0,$$

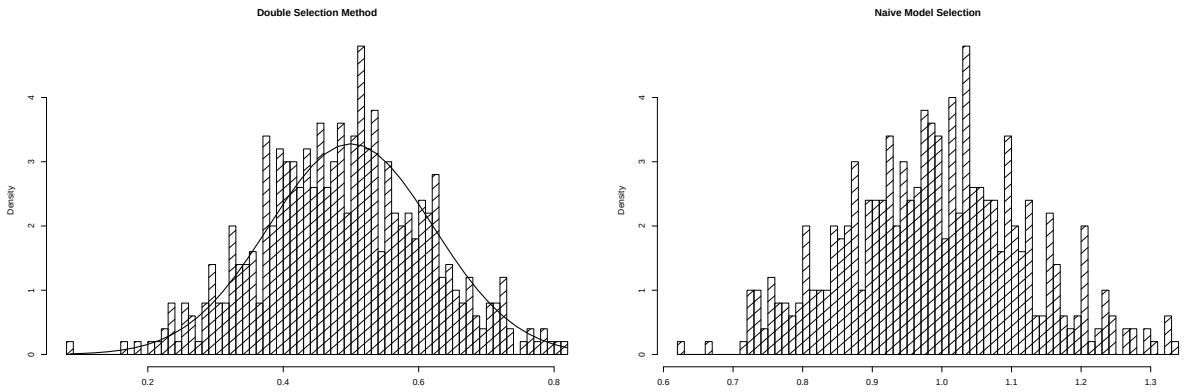
where $g(\cdot)$ is a function of the covariates. The next question is which variables to include in equation (2.1) from a set of potential covariates. In high-dimensional settings, when the number of covariates p is larger than the sample size n, variable selection is inevitable, as, e.g., the least squares estimate is not well defined. Even when p is smaller than n but the ratio p/n is high, ordinary least squares estimates are unreliable and again variable selection is needed. Including too many (noise) covariates might disguise the true treatment effect. For example, the study to evaluate the pulmonary artery catheter (PAC) in Bloniarz et al. [15], which we will also cover, contains 1013 observations and 55 potential covariates. In medical applications, interaction effects might be prevalent leading in all to 500-1000 two-way interactions in this example and to a high-dimensional setting with p very large compared to n, or even $p \gg n$.
In a naive approach, one might first select the relevant covariates by classical t-tests or modern machine learning methods, like Lasso and Boosting, and then estimate the treatment effect by including only the selected variables and continue with standard inference methods. But this procedure, although often used in applied work, might fail to provide a valid post-selection inference. This has been worked out by Leeb and Pötscher [67]. We demonstrate this by a simple simulation study with one treatment variable and one covariate. The data generating process is given by

$$y_i = d_i\alpha + x_i'\beta + \varepsilon_i, \quad d_i = x_i'\gamma + v_i$$

with $\alpha = 0.5$, $\beta = 0.2$ and $\gamma = 0.8$. The noise is normally distributed $\varepsilon_i \sim N(0,1)$ and

$$(d_i, x_i) \sim N\left(0, \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}\right).$$

We apply L_2-Boosting which is explained later in more detail for variable selection in the naive approach. The results for 500 repetitions of the scaled estimate $\hat{\alpha}$ are displayed in Figure 2.1 Panel B. The resulting distribution is highly biased, shows heavy tails and is not in line with a standard normal distribution. To provide valid post-selection inference with Boosting, we apply the double selection approach which is described in Section 2.4 in detail. Figure 2.1 Panel A shows the empirical distribution of the estimates when employing the double selection methods. They are nearly unbiased and can be approximated by a normal distribution. The intuition of the double selection method is that it cures the omitted variables bias which is introduced by imperfect model selection of machine learning methods by running an auxiliary regression/step. As mentioned, details will be provided later in Section 2.4.



Panel A **Panel B**

Figure 2.1: **Histograms of the estimates $\hat{\alpha}$ of the treatment effect with the double selection method and naive approach under a DGP with $\alpha = 0.5$.**

2.3 L_2-Boosting

In this section, we describe the L_2-Boosting algorithm, namely the original Boosting algorithm for regression defined in Bühlmann and Yu [23] and two variants, namely the orthogonal and the post-Boosting algorithm[1]. To define these algorithms for linear models, we consider the following regression setting:

$$y_i = x_i'\beta + \varepsilon_i, \quad i = 1, \dots, n, \tag{2.2}$$

where $x_i' = (x_{i,1}, \dots, x_{i,p_n})$ is a vector that consists of p_n predictor variables. β is a p_n-dimensional coefficient vector and ε_i is a random, zero-mean error term with $\mathbb{E}[\varepsilon_i|x_i] = 0$. We allow the dimension of the predictors p_n to grow with the sample size n. Also, the case $dim(\beta) = p_n \gg n$ is allowed. In this setting, a so-called sparsity condition is unavoidable. This means that there is a large set of potential variables, but the number of variables which have nonzero coefficients, denoted by s, is small compared to the sample size, i.e., $s < n$. This can also be weakened to approximate sparsity. In the following, we will drop the dependence of p_n on the sample size and denote it by p if no confusion will arise. X denotes the $n \times p$ design matrix where the single observations x_i form the rows. X_j denotes the jth column of the design matrix, and $x_{i,j}$ is the jth component of the vector x_i. We assume a fixed design with $\max_{1\leq j\leq p} x_{i,j} \leq C$ for all $i = 1, \dots, n$ and $c \leq \min_{1\leq j\leq p} \mathbb{E}_n[x_{i,j}^2]$ for absolute constants $0 < c < C < \infty$. Without loss of generality, we consider standardized regressors, i.e., $\mathbb{E}_n[x_{i,j}] = 0$ and $\mathbb{E}_n[x_{i,j}^2] = 1$ for $j = 1, \dots, p$. Further assumptions will be imposed in the next sections.

The basic principle of L_2-Boosting works as follows: The criterion function that we would like to minimize is the sum of squared residuals as in the ordinary least squares (OLS) case. We initialize the estimator $\hat{\beta}$ to zero (strictly speaking, a p-dimensional vector consisting of zeros). Then, we calculate the residuals which in this case are equivalent to the observations. Next, we conduct p univariate regressions, namely, we regress the residuals (in the first round, the observations) on each of the p regressors, resulting in p univariate regressions. Then, we select the variable or regression which explains most of the residuals and update this coordinate of our estimated vector in this direction. Now, we repeat this procedure (the calculation of the updated residuals, p univariate regressions, and updating the estimated coefficient vector) until some stopping criterion is reached.

The version above and the orthogonal version, introduced next, are, in deterministic settings, also known as the pure greedy algorithm (PGA) and the orthogonal greedy algorithm (OGA). Boosting is a gradient descent method. In the L_2-case, the (negative) gradient equals the residuals and the residuals are iteratively fitted by a so-called base learner, here componentwise univariate regressions. In the low-dimensional case, the estimator converges to the OLS solution. In the high-dimensional case, overfitting can occur in the absence of early stopping. Hence, early stopping prevents overfitting and is an unusual penalization/regularization scheme.

2.3.1 L_2-Boosting Algorithm

The algorithm for L_2-Boosting with componentwise least squares is given below. The act of stopping is crucial for Boosting algorithms, as stopping too late or never stopping leads to overfitting and therefore some kind of penalization is required. Similar to Lasso, early stopping might induce a bias through shrinkage. A potential way to decrease the bias is by "post-Boosting"which is defined in the next section.

[1] A more detailed exposition of the algorithms can be found in Luo and Spindler [73].

Algorithm 1 L_2-Boosting

(1) Initialization: $\beta^0 = 0$ (p-dimensional vector), $f^0 = 0$, set maximum number of iterations m_{stop} and set iteration index m to 0.

(2) At the $(m+1)^{th}$ step, calculate the residuals $U_i^m = y_i - x_i'\beta^m$.

(3) For each predictor variable $j = 1, \ldots, p$, calculate the correlation with the residuals:

$$\gamma_j^m := \frac{\sum_{i=1}^n U_i^m x_{i,j}}{\sum_{i=1}^n x_{i,j}^2} = \frac{< U^m, X_j >_n}{\mathbb{E}_n[x_{i,j}^2]}.$$

Select the variable j^m that is the most correlated with the residuals, i.e., $\max_{1\leq j\leq p} |corr(U^m, X_j)|$.

(4) Update the estimator: $\beta^{m+1} := \beta^m + \gamma_{j^m}^m e_{j^m}$, where e_{j^m} is the j^mth index vector and $f^{m+1} := f^m + \gamma_{j^m}^m X_{j^m}$.

(5) Increase m by one. If $m < m_{stop}$, continue with (2); otherwise stop.

2.3.2 Post- and Orthogonal L_2-Boosting

Post-L_2-Boosting is a post-model selection estimator that applies ordinary least squares (OLS) to the model selected by the first step, which is L_2-Boosting. To formally define this estimator, we make the following definitions, $T := supp(\beta)$ and $\hat{T} := supp(\beta^m)$, that are the support of the true model and the support of the model estimated by L_2-Boosting as described above with stopping at m, respectively. The superscript C denotes the complement of the set with regard to $\{1, \ldots, p\}$. In the context of Lasso, OLS after model selection was analyzed in Belloni and Chernozhukov [6]. Given the above definitions, the post-model selection estimator or OLS post-L_2-Boosting estimator will take the form

$$\tilde{\beta} = \arg\min_{\beta\in\mathbb{R}^p} Q_n(\beta) : \beta_j = 0 \quad \text{for each} \quad j \in \hat{T}^C. \tag{2.3}$$

$Q_n(\beta)$ denotes the squared sum of residuals defined as $\sum_{i=1}^n (y_i - x_i'\beta)^2$. Another variant of the Boosting algorithm is orthogonal Boosting (oBA), or the orthogonal greedy algorithm in its deterministic version. Only the updating step is changed. An orthogonal projection of the response variable is carried out on all the variables which have been selected up to that point. The advantage of this method is that any variable is selected at most once in this procedure, while in the previous version the same variable might be selected at different steps which makes the analysis far more complicated. More formally, the method can be described as follows by modifying step (4) in Algorithm 1: Define X_o^m as the matrix which consists only of the columns which correspond to the variables selected in the first m steps, i.e., all X_{j_k}, $k = 0, 1, \ldots, m$. Then, we have

$$\beta_o^m = (X_o^{m\prime} X_o^m)^{-1} X_o^{m\prime} y \tag{2.4}$$

$$\hat{y}^{m+1} = f_o^{m+1} = X_o^m \beta_o^m. \tag{2.5}$$

2.3.3 Early Stopping

As already mentioned, early stopping is crucial in Boosting to avoid overfitting. The standard approaches for determining the "optimal" stopping time are cross-validation and a corrected Akaike information criteria (Bühlmann [20]). Both lack a theoretical foundation in a high-dimensional setting, although they are applied by practitioners and often give competitive results. In our analysis, in particular in the simulation study, we rely on theoretical-grounded data driven stopping rules developed in Luo and

Spindler [73]. The idea is to stop the Boosting algorithm when the improvement in fit is below some pre-specified threshold.

2.3.4 Computational Details and Comparison to Lasso

Luo and Spindler [73] showed that post-Boosting and orthogonal Boosting achieve the same rate of convergence as Lasso in a sparse, high-dimensional setting (under slightly stronger assumptions). Also in terms of the empirical performance, Bühlmann and Hothorn [21] did not find an overall superiority of L_2-Boosting over Lasso or vice versa. Friedman et al. [47] pointed out first a strong relationship between L_2-Boosting with componentwise linear least squares and Lasso. Although these methods are not equivalent in general, Efron et al. [42] proofed an approximate equivalence between L_2-Boosting and Lasso and confirmed that L_2-Boosting and Lasso are closely related.
Compared to Lasso, Boosting uses an unusual penalization scheme as discussed above. Hence, L_2-Boosting can be interpreted as an approximate and implicit regularized optimization, whereas Lasso directly solves a complex penalized optimization problem. Further, L_2-Boosting solves univariate regressions that are easily parallelizable. Thus, this form of estimation and variable selection is computationally very efficient. Although there are also efficient algorithms to solve the optimization problem of Lasso, this leads to a computational superiority of L_2-Boosting over Lasso. This has also been observed in Bühlmann and Hothorn [21] who compare the computing time of L_2-Boosting and Lasso in high-dimensional regressions. Hence, Boosting is employed in practice when explicitly solving regularized optimization problems is not practical. This is usually the case in very high-dimensional settings when $p >> n$, see Efron et al. [42].

2.4 Inference for Treatment Effects

In this section, we consider the case where a researcher is interested in estimating the treatment effect α_0 of a treatment variable d. We provide results for valid inference after selecting among very many control variables and in an instrumental variable model with potentially very many instruments when post- or orthogonal L_2-Boosting is used for the variable selection.

2.4.1 Inference after Selection among High-Dimensional Controls

In many situations, the treatment variable is uncorrelated with the error term ε_i only after controlling for sufficient control variables denoted by x_i. It is not clear which set of control variables to include, in particular when many potential control variables are available. In such situations, in particular when the number of variables p is larger than the number of observations n, model selection might be inevitable. Unfortunately, many modern methods like Lasso or Boosting obtain consistent model selection only under very strong, in particular in applications in economics, unrealistic assumptions. Hence, relevant variables might be missed which leads to invalid post-selection inference. To circumvent this problem, we apply the so-called double selection method introduced in Belloni et al. [8], Chernozhukov et al. [35] and Chernozhukov et al. [33]. The key idea of double selection is to introduce and estimate an auxiliary regression which safeguards against model selection errors of moderate size. We consider the model

$$y_i = d_i\alpha_0 + x_i'\beta + \xi_i, \quad \mathbb{E}[\xi_i|d_i, x_i] = 0 \tag{2.6}$$

$$d_i = x_i'\gamma + \nu_i, \quad \mathbb{E}[\nu_i|x_i] = 0. \tag{2.7}$$

The estimation method consists of the following three steps where the first two involve model selection with Boosting:

(1) Run a post- or orthogonal Boosting regression of d_i on x_i. The set of variables which is selected will be denoted by $\hat{I}_1$.

(2) Run a post- or orthogonal Boosting regression of y_i on x_i. The set of variables which is selected will be denoted by $\hat{I}_2$.

(3) Run an OLS regression of y_i on the treatment variable d_i and the set of variables selected in the first two steps. This set might be augmented by additional variables.

The estimated regression coefficient of the treatment variable in step (3) above is the double selection estimator $\check{\alpha}$. To analyze this estimator based on L_2-Boosting, we impose the following assumptions:

A.1. *Let c and C be absolute constants. The following assumptions hold:*

(i) We observe $w_i = (y_i, d_i, x_i)$ i.i.d. on $(\Omega, \mathcal{F}, P)$ obeying (2.6) and (2.7) for $i = 1, \dots, n$.

(ii) The model is sparse: $||\beta||_0 \leq s$ and $||\gamma||_0 \leq s$.

(iii) We have $\mathbb{E}[y^2] \leq C$ and $\mathbb{E}[d^2] \leq C$.

(iv) It holds $c \leq \mathbb{E}[\xi^2|d_i, x_i] \leq C$ a.s. and $c \leq \mathbb{E}[\nu^2|x_i] \leq C$ for all $i = 1, \dots, n$. Further, there exists a absolute constant $4 < q < \infty$ such that $\mathbb{E}\left[|\xi|^q + |\nu|^q\right] \leq C$.

(v) We have

(a) $\frac{s^2 \log^2(p \vee n)}{n} \to 0$,

(b) $\frac{\log^3 p}{n} \to 0$ and

(c) $sn^{-1/2+2/q} \to 0$.

Assumption A.1 imposes standard conditions on the data generating process. Assumption A.1 (ii) imposes sparsity on the two equations. Assumptions A.1 (iii) and (iv) impose technical conditions on the moments of the random variables. Assumption A.1 (v) restricts the growth of the number of parameters.

A.2. *We assume that there exist constants $0 < c < 1$ and C such that $0 < 1 - c \leq \phi_{min}(s', \mathbb{E}_n[x_i' x_i]) \leq \phi_{max}(s', \mathbb{E}_n[x_i' x_i]) \leq C < \infty$ for any $s' \leq M_0$, where M_0 is a sequence such that $M_0 \to \infty$ slowly along with n, and $M_0 \geq s$. $\phi_{min}(s', \mathbb{E}_n[x_i' x_i])$ denotes the minimum eigenvalue of s'-dimensional submatrices of $\mathbb{E}_n[x_i' x_i]$. $\phi_{max}(s', \mathbb{E}_n[x_i' x_i])$ is defined in an analog way for the maximum eigenvalue.*

This condition is standard for the analysis of Lasso and other machine learning methods in a high-dimensional setting. It allows for a more general behavior requiring only that the sparse eigenvalues of the Gram matrix are bounded from above and away from zero. A more restrictive assumption in traditional econometric research is to assume that the (population) Gram matrix has eigenvalues bounded from above and away from zero. The sparse eigenvalues condition is fulfilled for many relevant designs. For examples, we refer to Belloni and Chernozhukov [6]. An extensive overview of different conditions on the matrices and how they are related is given in Van De Geer and Bühlmann [95].

A.3. *With probability greater or equal $1 - \alpha$, we have $\sup_{1 \leq j \leq p} | < x_{ij}, \varepsilon_i >_n | \leq 2\tilde{\sigma} \sqrt{\frac{\log(2p/\alpha)}{n}} =: \lambda_n$ for $\varepsilon_i = \xi_i$ and $\varepsilon_i = \nu_i$. Here, $< x_{ij}, \varepsilon_i >_n$ denotes the empirical inner product and $\tilde{\sigma} := \sqrt{Var(\varepsilon_i)}$.*

Assumption A.3 holds, e.g., if the error terms are i.i.d. normally distributed random variables. This can be weakened to cases of non-normality as discussed in Luo and Spindler [73].

A. 4. $\min_{j\in T}|\beta_j| \geq J$, $\max_{j\in T}|\beta_j| \leq J'$, $|\alpha_0| \leq J'$ *for some constants* $J > 0, J' < \infty$. *The same condition holds for the parameter vector* γ.

This assumption is a so-called beta-min condition for the parameters in both equations. Although it might look quite strong at first glance, it can be weakened so that the sequence of absolute values of the coefficients is decreasing with the sample size. Moreover, we assume that in the Boosting regressions early stopping takes place and the stopping criteria follows the proposals in Luo and Spindler [73] for post- and orthogonal L_2-Boosting, i.e., the procedure is stopped when the improvement in fit is below some pre-specified threshold. With these assumptions, we can now formulate our first main theorem.

Theorem 1. *Let* $\{P_n\}$ *be a sequence of data generating processes for which Assumptions A.1-A.4 hold for* $P = P_n$ *and each* n. *Then, the double-selection estimator based on post-*L_2*-Boosting/orthogonal* L_2*-Boosting* $\check{\alpha}$ *satisfies*

$$\hat{\sigma}_n^{-1}\sqrt{n}(\check{\alpha} - \alpha_0) \to_D N(0,1) \tag{2.8}$$

with

$$\hat{\sigma}_n^2 = [\mathbb{E}_n\hat{\nu}_i^2]^{-1}\mathbb{E}_n[\hat{\nu}_i^2\hat{\xi}_i^2][\mathbb{E}_n\hat{\nu}_i^2]^{-1}$$

for $\hat{\xi}_i := (y_i - d_i\check{\alpha} - x_i'\hat{\beta})(n/(n-\hat{s}-1))^{1/2}$ *and* $\hat{\nu}_i := d_i - x_i'\hat{\gamma}, i = 1,\ldots,n$, *where* $\hat{\beta}$ *denotes the post-double selection estimator and* $\hat{s} = ||\hat{T}||_0$.

Proof. The sparsity condition in Assumption A.1 (ii) and Assumptions A.2-A.4 imply, according to Luo and Spindler [73], that condition HLMS(P) in Belloni et al. [8] is satisfied. In the regular fix design setting, Assumption A.1 and Assumption A.4 imply conditions ASTE(P) and SM(P) in Belloni et al. [8]. Condition SE(P) holds due to Assumption A.2. Hence, Theorem 2 in Belloni et al. [8] yields the result. □

This result can be used to conduct valid inference on the regression coefficient α_0. The construction of uniformly valid confidence intervals is given in the following corollary.

Corollary 2.4.1. *Let* $\mathbf{P}_n$ *be the collection of all data generating processes* P *for which the assumptions of Theorem 1 hold for given* n. *Further, let* $\mathbf{P}$ *be the collection of data-generating processes for which the conditions above hold for all* $n \geq n_0$, *and define* $c(1-\xi) := \Phi^{-1}(1-\xi/2)$. *The confidence regions based upon* $\check{\alpha}$ *and* $\hat{\sigma}_n$ *are uniformly valid in* $P \in \mathbf{P}$*:*

$$\lim_{n\to\infty}\sup_{P\in\mathbf{P}}|P(\alpha_0 \in [\check{\alpha} \pm c(1-\xi)\hat{\sigma}_n/\sqrt{n}]) - (1-\xi)| = 0.$$

2.4.2 Inference on Treatment Effects in an Instrumental Variable Model

In this section, we consider the following instrumental variable model with potentially very many instruments

$$\begin{aligned} y_i &= d_i\alpha_0 + \beta' x_i + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i|z_i] = 0 \\ d_i &= \gamma' z_i + \nu_i \end{aligned}$$

with instrument function $D_i = D(z_i) = \mathbb{E}[d_i|z_i] = \gamma' z_i$. For simplicity, in our technical analysis we consider the model above without any controls x_i in the first stage equation and a regular fix design Z with observations z_i:

$$y_i = d_i\alpha_0 + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i|z_i] = 0 \tag{2.9}$$

$$d_i = \gamma' z_i + \nu_i. \tag{2.10}$$

To estimate the coefficient α_0 of the endogenous treatment variable, we employ the following two-stage least squares (tsls) procedure: In the first step, we estimate and predict the instrument $\hat{D}_i = \hat{\gamma}' z_i$ by post- or orthogonal L_2-Boosting. Finally, we estimate $\hat{\alpha}_0$ by a regression of the outcome variable y on the predicted instrument $\hat{D}_i$. To analyze this estimator based on L_2-Boosting, we impose the following assumptions:

B.1. *Let c and C be absolute constants. The following assumptions hold:*

(i) The data (y_i, d_i, z_i) is i.i.d. on $(\Omega, \mathcal{F}, P)$ and obeys the linear IV model in (2.9) and (2.10).

(ii) The optimal instrument function $D_i = \gamma' z_i$ can be approximated by s instruments:

$$||\gamma||_0 \leq s.$$

(iii) We have $\mathbb{E}[d^2] < C$.

(iv) It holds $c \leq \mathbb{E}[\varepsilon^2 | z_i] \leq C$ for all $i = 1, \ldots, n$. Further, there exists a absolute constant $q > 4$ such that $\mathbb{E}[|\varepsilon|^q] + \mathbb{E}[|\nu|^q] \leq C$.

(v) We have

(a) $\frac{s^2 \log^2(p \vee n)}{n} \to 0$,
(b) $\frac{\log^3 p}{n} \to 0$ and
(c) $\frac{s \log(p \vee n)}{n} n^{2/q} \to 0$.

B.2. *We assume that there exist constants $0 < c < 1$ and C such that $0 < 1 - c \leq \phi_{min}(s', \mathbb{E}_n[z_i' z_i]) \leq \phi_{max}(s', \mathbb{E}_n[z_i' z_i]) \leq C < \infty$ for any $s' \leq M_0$, where M_0 is a sequence such that $M_0 \to \infty$ slowly along with n, and $M_0 \geq s$.*

B.3. $\min_{j \in T} |\gamma_j| \geq J$, $\max_{j \in T} |\gamma_j| \leq J'$ *and* $\alpha_0 \leq J'$ *for some constants* $J > 0, J' < \infty$.

B.4. *With probability greater or equal $1 - \alpha$, it holds $\sup_{1 \leq j \leq p} | < z_{ij}, \nu_i >_n | \leq 2\tilde{\sigma} \sqrt{\frac{\log(2p/\alpha)}{n}} =: \lambda_n$ for $\tilde{\sigma} := \sqrt{Var(\nu_i)}$.*

Again, we assume that the stopping criteria in the Boosting regression follows the proposals in Luo and Spindler [73]. The Assumptions B.1-B.4 are essentially the same as the Assumptions A.1-A.4 except some small deviations due to the different underlying setting in Subsection 2.4.1. It is worth to notice that the growth condition B.1(v) is slightly weaker than the growth condition A.1(v) since Assumption A.1(v)(c) implies Assumption B.1(v)(c). With these assumptions, we can show that the IV estimator $\hat{\alpha}$ following the two-stage least squares (tsls) procedure is asymptotically normally distributed. This result is provided by the following theorem.

Theorem 2. *Let $\{P_n\}$ be a sequence of data generating processes for which Assumptions B.1-B.4 hold for $P = P_n$ and each n. Then, the IV estimator $\hat{\alpha}$ based on post-L_2-Boosting or orthogonal L_2-Boosting of the optimal instrument satisfies*

$$(\hat{Q}^{-1} \hat{\Omega} \hat{Q}^{-1})^{-1/2} \sqrt{n} (\hat{\alpha} - \alpha_0) \to_D N(0, 1)$$

for $\hat{\Omega} := \mathbb{E}_n[\hat{\varepsilon}_i^2 \hat{D}(z_i)^2]$ and $\hat{Q} := \mathbb{E}_n[\hat{D}(z_i)^2]$ with $\hat{\varepsilon}_i = y_i - d_i \hat{\alpha}$ and $\hat{D}(z_i) = \hat{\gamma}' z_i$.

This also enables us to construct uniformly valid confidence intervals for the treatment effect as in Corollary 2.4.1.

Proof. Assumptions B.2-B.4 ensure sufficiently fast convergence rates of the fitted optimal instruments $\hat{D}(z_i) = \hat{\gamma}' z_i$ estimated with post- or orthogonal Boosting in first step regression, i.e.,

$$||\hat{D}(z_i) - D(z_i)||_{2,n} \leq C\sqrt{\frac{s\log(p \vee n)}{n}}$$

and

$$\hat{s} \leq Cs$$

with probability $1 - o(1)$ as shown in Luo and Spindler [73]. Since the maximal sparse eigenvalues are uniformly bounded from above due to Assumption B.2, we conclude

$$||\hat{\gamma} - \gamma||_2 \leq C||\hat{D}(z_i) - D(z_i)||_{2,n} \leq C\sqrt{\frac{s\log(p \vee n)}{n}}$$

with probability $1 - o(1)$ which implies

$$||\hat{\gamma} - \gamma||_1 \leq \sqrt{s}||\hat{\gamma} - \gamma||_2 \leq C\sqrt{\frac{s^2\log(p \vee n)}{n}}.$$

This allows us applying Theorem 4 in Belloni et al. [5] since Assumption B.1 implies conditions AS and SM in Belloni et al. [5]. This concludes the proof. □

2.5 Simulation Study

In this section, we present simulation results for both settings.

2.5.1 Setting with High-Dimensional Controls

First, we consider the following data generating process:

$$y_i = d_i\alpha_0 + x_i'\theta_g + \xi_i \tag{2.11}$$

$$d_i = x_i'\theta_m + \nu_i, \tag{2.12}$$

where $(\xi_i, \nu_i)' \sim N(0, I_2)$ with I_2 the 2×2 identity matrix and $x_i \sim N(0, \Sigma)$ with $\Sigma_{kj} = 0.5^{|j-k|}$. The parameter of interest, α_0, is set equal to 0.5. We consider both a sparse setting and an approximate sparse setting where $\theta_g = \theta_m$. In the sparse setting, the first s coefficients are set equal to one and all other parameters $p - s$ are equal to zero. In the approximate sparse setting, the coefficient vectors are of the form $(1, 0.7^2, 0.7^3, \ldots, 0.7^{p-1})'$. We vary the sample size n, the number of covariates p and the sparsity index s. The number of repetitions is $R = 500$ and we set the nominal significance level to 0.05. Tables 2.1 and 2.3 show the results (bias and rejection rates) for the sparse setting with the double selection method. Tables 2.2 and 2.4 show the corresponding results for the approximate sparse setting. Under exact sparsity, the bias of the post-Lasso procedure seems to be slightly smaller than the bias of the Boosting procedures, while the rejection rates seem to be comparable, in particular in the setting with relative small p and s. The pattern in the approximate sparsity setting seems to be similar. The bias of the post-Boosting method is slightly higher than the bias of the orthogonal Boosting method, while the rejection rate of the post-Boosting method is closer to the nominal level in many settings. Finally, we would also like to mention that the classical L_2-Boosting algorithm performs comparable to the other booting algorithms analyzed in the simulation study here, although the results are not included in the tables.

n	p	s	post-Lasso	post-BA	oBA
100	10	5	0.000	-0.005	0.005
100	10	10	0.001	0.000	0.041
100	100	5	-0.008	-0.057	-0.027
200	10	5	-0.001	-0.003	0.001
200	10	10	-0.001	-0.001	0.017
200	100	5	-0.005	-0.033	-0.023
200	200	5	0.002	-0.045	-0.025
400	100	5	-0.004	-0.016	-0.007
400	100	10	-0.003	-0.015	0.002
400	200	5	0.002	-0.021	-0.009
400	200	10	0.002	-0.019	-0.002
400	400	5	0.002	-0.032	-0.017

Table 2.1: **Simulation results: Bias under exact sparsity.**

n	p	post-Lasso	post-BA	oBA
100	10	-0.002	-0.008	-0.004
100	50	-0.004	-0.048	-0.036
100	100	-0.007	-0.069	-0.052
200	10	-0.002	-0.005	-0.002
200	50	0.008	-0.013	-0.006
200	100	-0.004	-0.038	-0.030
200	200	0.004	-0.051	-0.039
400	10	-0.000	-0.002	0.000
400	50	-0.004	-0.013	-0.010
400	100	-0.003	-0.019	-0.016
400	200	0.002	-0.022	-0.018
400	400	0.003	-0.035	-0.030

Table 2.2: **Simulation results: Bias under approximate sparsity.**

n	p	s	post-Lasso	post-BA	oBA
100	10	5	0.046	0.044	0.080
100	10	10	0.044	0.044	0.114
100	100	5	0.044	0.100	0.148
200	10	5	0.052	0.054	0.058
200	10	10	0.056	0.056	0.092
200	100	5	0.040	0.080	0.132
200	200	5	0.054	0.080	0.158
400	100	5	0.056	0.066	0.078
400	100	10	0.052	0.066	0.120
400	200	5	0.034	0.080	0.074
400	200	10	0.034	0.070	0.130
400	400	5	0.062	0.100	0.108

Table 2.3: **Simulation results: Rejection Rate under exact sparsity.**

n	p	post-Lasso	post-BA	oBA
100	10	0.044	0.044	0.050
100	50	0.088	0.118	0.110
100	100	0.046	0.114	0.104
200	10	0.052	0.054	0.048
200	50	0.050	0.064	0.048
200	100	0.044	0.102	0.076
200	200	0.050	0.118	0.094
400	10	0.046	0.042	0.048
400	50	0.052	0.056	0.066
400	100	0.062	0.080	0.088
400	200	0.032	0.074	0.076
400	400	0.060	0.112	0.128

Table 2.4: **Simulation results: Rejection Rate under approximate sparsity.**

2.5.2 IV Estimation with many Instruments

In the setting with many instrumental variables, we consider the following data generating process similar to the simulation experiment in Belloni et al. [5]:

$$y_i = d_i\alpha_0 + \varepsilon_i, \tag{2.13}$$

$$d_i = \gamma' z_i + \nu_i, \tag{2.14}$$

$$(\varepsilon_i, \nu_i) \sim N\left(0, \begin{pmatrix} \sigma_\varepsilon^2 & \sigma_{\varepsilon\nu} \\ \sigma_{\varepsilon\nu} & \sigma_\nu^2 \end{pmatrix}\right) i.i.d., \tag{2.15}$$

where $\alpha_0 = 1$ is the parameter of interest. The regressors $Z_i = (z_{i1}, \ldots, z_{ip})'$ are drawn from a normal distribution $N(0, \Sigma_Z)$ with $\mathbb{E}[z_{ij}^2] = \sigma_z^2$ and $Corr(z_{ij}, z_{ik}) = 0.5^{|j-k|}$. We set $corr(\varepsilon, \nu) = 0.1$ and σ_z^2 and σ_e^2 are set to one. Let $\sigma_v^2 = 1 - \gamma'\Sigma_z\gamma$ such that the unconditional variance of the endogenous variable equals 1. The first stage coefficients are set according to $\gamma = C\tilde{\gamma}$. For $\tilde{\gamma}$ we use a sparse design, i.e., $\tilde{\gamma} = (1, \ldots, 1, 0, \ldots, 0)$ with s coordinates equal to one and all other $p - s$ equal to zero. C is set in such a way that we generate target values for the concentration parameter $\mu^2 = \frac{n\gamma'\Sigma_z\gamma}{\sigma_v^2}$ which determines the behavior of the IV estimators as described in Hansen et al. [51]. We set the concentration parameter equal to 180 and vary the sample size n, the number of covariates p and the sparsity index s. The number of repetitions in the simulations study is again $R = 500$. We estimate the first stage and calculate the first stage predictions with L_2-Boosting and its variants. The simulation results in Tables 2.5 and 2.6 reveal that Boosting performs comparable to post-Lasso in the examined settings concerning both bias and the rejection rates (nominal significance level 0.05). The average bias of the estimated treatment effect is given in Table 2.5, the rejection rates in Table 2.6.

n	p	s	post-Lasso	post-BA	oBA
200	100	5	0.002	0.017	0.017
200	100	10	0.008	0.020	0.020
200	400	5	0.012	0.033	0.036
200	400	10	0.015	0.033	0.033
200	800	5	0.007	0.034	0.035
200	800	10	0.011	0.038	0.040
400	100	5	0.001	0.012	0.013
400	100	10	0.004	0.012	0.013
400	400	5	0.012	0.029	0.032
400	400	10	0.021	0.037	0.039
400	800	5	0.009	0.031	0.033
400	800	10	0.016	0.038	0.039
800	100	5	0.005	0.016	0.016
800	100	10	0.013	0.022	0.023
800	400	5	0.007	0.024	0.025
800	400	10	0.008	0.025	0.026
800	800	5	0.003	0.028	0.030
800	800	10	0.010	0.033	0.034

Table 2.5: **Simulation results: Bias in the IV setting.**

n	p	s	post-Lasso	post-BA	oBA
200	100	5	0.046	0.038	0.046
200	100	10	0.062	0.056	0.062
200	400	5	0.052	0.060	0.062
200	400	10	0.050	0.068	0.066
200	800	5	0.062	0.068	0.072
200	800	10	0.066	0.090	0.094
400	100	5	0.054	0.066	0.060
400	100	10	0.060	0.072	0.076
400	400	5	0.056	0.064	0.068
400	400	10	0.058	0.074	0.092
400	800	5	0.054	0.078	0.084
400	800	10	0.054	0.074	0.096
800	100	5	0.060	0.060	0.060
800	100	10	0.060	0.062	0.064
800	400	5	0.066	0.084	0.009
800	400	10	0.062	0.072	0.084
800	800	5	0.054	0.074	0.074
800	800	10	0.038	0.066	0.056

Table 2.6: **Simulation results: Rejection Rate in the IV setting.**

2.6 Application: Analysis of the PAC-man Study

2.6.1 The PAC-man Study

To illustrate our methodology, we analyze the PAC-man study mentioned in the Section 2.1. There were 1013 patients who took part in this study which was conducted as a randomized control trial. There were 506 patients treated with PAC, and 507 patients formed the control group. The research question was whether the treatment by a PAC increases the patient's number of quality-adjusted life years (QALYs), which is the outcome variable. One QALY represents one year of life in full health whereas an in-hospital death corresponds to a QALY of zero. The data set contains 53 covariates about each individual in the study. There are two reasons, as argued in Section 2.2, to use additional covariates in the analysis of this randomized control trial: First, additional covariates allow a more precise estimation of the treatment effect. Second, despite the randomized design of the study, conditioning on covariates might reinforce unconfoundedness. It might be possible that certain conditions (e.g., acute health conditions) lead to a deviation from the randomized protocol. Using a large set of covariates describing individual specific health conditions, but also hospital specific conditions, might control for such deviations. The PAC-man study was discussed widely in the literature. Bloniarz et al. [15], which is closest to our setting, consider Lasso adjustments of treatment effect estimates in randomized experiments in a high-dimensional setting. We follow their proposal to construct the design matrix X by including all main effects and two-way interactions. Interactions which are highly correlated (with a correlation larger than 0.95) are excluded. Additionally, indicators with very sparse entries (when the number of 1's is less than 20) are also removed. This results in a total of 771 regressors[2]. The covariates contain detailed information on the patient's health conditions, e.g., pre-existing conditions and current health status measured by different biomarkers, and also hospital specific information. For a detailed description, we refer to the documentation of the PAC-man study.

2.6.2 Results

We estimate the following model:

$$y_i = \delta d_i + \beta' x_i + \varepsilon_i, \quad i = 1, \dots, 1013.$$

The number of QALYs are the outcome variable y_i. The treatment variable d_i is a binary variable indicating PAC. ε_i denotes the residuals. We estimate the (constant) treatment effect without any controls (baseline estimator) as it is the standard approach in RCTs, but we also control for covariates. The results are presented in Table 2.7. The baseline estimator gives a negative treatment effect but with a p-value of 0.759. When we control for covariates, the post-Lasso algorithm gives also a negative, but insignificant treatment effect. This is in line with the results presented in Bloniarz et al. [15]. In contrast, the post-Boosting algorithm (post-BA) shows a positive treatment effect, but this effect is also not significant.

	baseline	post-Lasso	post-BA
Est.	−0.062	−0.308	0.224
se	0.201	0.241	0.265
p-value	0.759	0.201	0.397

Table 2.7: **Results of the PAC-man Study.**

[2]Bloniarz et al. [15] have in total 1172 regressors as the data set of the PAC-man study which was provided to them contains six additional variables to which we have no access.

2.7 Conclusion

In this paper, we apply L_2-Boosting, namely the post- and orthogonal version, for estimation of treatment effects in the setting of many controls and many instruments. We derive uniformly valid results for the asymptotic distribution of estimated treatment effects. We use the framework of orthogonalized moment conditions introduces by Belloni, Chernozhukov, Hansen and coauthors in a series of papers to derive the results. The second ingredient are results on the rate of convergence of L_2-Boosting given in Luo and Spindler [73]. In the simulation study, our proposed method performs well and is comparable with Lasso. Finally, we analyze the PAC-man study which stimulated a lot of research in medicine and related fields. We find that the treatment effect is not significantly different from zero.

Chapter 3

Transformation Models in High-Dimensions

3.1 Introduction

Over the last few years, substantial progress has been made in the problem of fitting high-dimensional linear models of the form

$$Y = X^T\beta + \varepsilon, \tag{3.1}$$

where the number of regressors p is much larger than the sample size n. The theoretical properties of penalization approaches, such as Lasso, are now well understood under the assumption that the coefficient vector β is sparse. A detailed summary of the recent results is given in textbook length in Bühlmann and Van De Geer [22].

In this paper, we take up the idea of the high-dimensional linear model in (3.1) and combine it with a parametric transformation of the response variable $\Lambda_\theta(\cdot) \in \mathcal{F}_\Lambda$, where $\mathcal{F}_\Lambda = \{\Lambda_\theta(\cdot) : \theta \in \Theta\}$ is a given family of strictly monotone increasing functions. For every $\theta \in \Theta$, we assume a linear model

$$\Lambda_\theta(Y) = X^T\beta_\theta + \varepsilon_\theta \tag{3.2}$$

with $\mathbb{E}[\varepsilon_\theta] = 0$. Our analysis allows the number of regressors to be much larger than the number of observations, although we require sparsity for every β_θ in (3.2). The goal of data transformation is to change the scale preventing incorrect model assumptions, such as by establishing normally distributed errors. Transformation of the dependent variable is very common in statistics and economics. The Box-Cox power transformations (Box and Cox [16]) or the modification proposed by Yeo and Johnson [102] are very popular transformations. The aim of transformations is typically to achieve symmetry, normality, or independence of the error terms. In labor economics the analysis of wage data is key, and wage data is non-negative and often highly skewed. By default, wage data are transformed by the logarithm and then further processed, for example, as a dependent variable in a Mincer equation. A crucial point for the subsequent analysis is that the applied transformation is correctly specified. Feng et al. [44] list some common scenarios of the misuse and misinterpretation of the log transformation. This underlines the importance of the right transformation to handle the problem of skewed data and non-negative outcomes. In this study, we will present an estimate for the unknown transformation parameter $\theta_0 \in \Theta$ in a high-

dimensional transformation model, which satisfies

$$\Lambda_{\theta_0}(Y) = X^T \beta_{\theta_0} + \varepsilon_{\theta_0} \tag{3.3}$$

with $\varepsilon_{\theta_0} \sim \mathcal{N}(0, \sigma^2)$ and independent from X. This means that, under the true parameter θ_0, the errors are normally distributed with unknown variance. We establish that our estimator is root-n consistent, asymptotically unbiased, and normal. The transformation enables us to establish normality of the error terms and subsequent application of procedures based on normality. Our setting fits into a general Z-estimation problem with a high-dimensional nuisance function which depends on the target parameter θ. Inference on a target parameter in general Z-estimation problems in high dimensions is covered in Belloni et al. [9] and Chernozhukov et al. [35]. In high-dimensional transformation models, the nuisance function depends on the target parameter θ; therefore, in the supplementary material, we establish a theorem regarding inference in a general Z-estimation setting under a different set of entropy conditions where such a dependence is explicitly allowed. This result might be of independent interest for Z-estimation problems with the same underlying structure.

In this paper, we focus on estimation and inference on the transformation parameter because this is the first crucial step and it is important for the interpretation of the model and application of subsequent statistical procedures. A related line of research has focused on inference on the covariates in the model. Given that inference in this case relies on the estimated transformation model, valid post-selection/ estimation inference is crucial, as pointed out by Bickel and Doksum [14] which has led to a vivid discussion on this topic. Bickel and Doksum [14] cover the parametric case, the semiparametric case is covered by Linton et al. [70], and has more recently been examined by Kloodt and Neumeyer [61], amongst others. Inference on the covariates in high-dimensional settings is an interesting problem that we plan to address in future research. The underlying theory is built on Neyman orthogonal moment conditions, as summarized in Chernozhukov et al. [35].

Literature Review

The Box-Cox transformation was introduced in Box and Cox [16], one of the most cited papers in statistics. Since then, transformation models are widely used by empirical researchers and also have stimulated a lot of research on theoretical aspects. Both the transformation parameter and the regression coefficients have been extensively considered in the literature. In this review, we will mostly focus on paper dealing with the transformation parameter. For a thorough review, we refer to Sakia [88] who also mostly focuses on estimation and inference on the transformation parameter. Transformation models are applied in all fields of statistics, including medicine, biostatistics and economics. In economics transformations of the dependent and independent variables are considered. Transformation models are used in labor economics, health economics, macroeconomics and finance, to mention a few, and there the focus has also been on the transformation itself. For example, although Nelson and Granger [77] are interested in forecasting performance, the choice and estimation of the transformation parameter is crucial. For a detailed list of applied papers, we refer to the survey of Sakia [88]. Manning and Mullahy [74] discuss transformation in health economics.

Transformation models also have been challenging and stimulating for theoretical developments. Already Box and Cox [16] propose a test for the transformation parameter based on a likelihood ratio test. Andrews [2] proposes an exact test for the transformation parameter. Amongst many others, Atkinson [4] and Carroll [26] proposed further refinements for inference on the transformation parameter. Transformation models have also been analyzed in a semi-/nonparametric setting (e.g., Linton et al. [70]) and under endogeneity (Vanhems and Van Keilegom [97]). To the best of our knowledge, we are the first to estimate and provide inference results on the transformation parameter in a high-dimensional setting

extending a huge body of literature on the inference on the transformation parameter.
Moreover, we also we contribute to the literature on double machine learning. In this setting, the target of interest is a low-dimensional parameter, while there is also a high-dimensional nuisance parameter involved. The double machine learning framework allows for valid inference on the target parameter in this setting. An excellent overview over this approach is given in Chernozhukov et al. [35] (see also the references therein). The authors mention that in principle the nuisance parameter can depend on the target parameter, but do not pursue this idea further. In this paper, we allow explicitly for this situation which is technically much more involved than the standard case, and give normal conditions for this situation.

Plan of this Paper

The rest of this paper is organized as follows. In Section 3.2, we formally define the setting and propose an estimator for the transformation parameter. In Section 3.3, we prove that a Neyman orthogonality condition holds and we provide theoretical results for the estimation rates of the nuisance functions. We also present the main result for the asymptotic distribution of the estimated transformation parameter. Section 3.4 provides a simulation study and Section 3.5 gives an empirical application. The proofs are provided in Appendix 3.7. The supplementary material includes additional technical material. In Appendix 3.8, conditions for the uniform convergence rates of the Lasso estimator are presented. Finally, Appendix 3.9 provides a theoretical result about inference on a target parameter in general Z-estimation problems with dependent and high-dimensional nuisance functions.

Notation

In what follows, we work with triangular array data $\{(Z_{i,n}, i = 1, ..., n), n = 1, 2, 3, ...\}$ with $Z_{i,n} = (Y_{i,n}, X_{i,n})$ defined on some common probability space $(\Omega, \mathcal{A}, P)$. The law $P_n \in \mathcal{P}_n$ of $\{Z_i, i = 1, ..., n\}$ changes with n. Thus, all parameters that characterize the distribution of $\{Z_i, i = 1, ..., n\}$ are implicitly indexed by the sample size n, but we omit the index n to simplify notation.
The l_2 and l_1 norms are denoted by $||\cdot||_2$ and $||\cdot||_1$. The l_0-norm, $||\cdot||_0$, denotes the number of nonzero components of a vector. We use the notation $a \vee b := \max(a, b)$ and $a \wedge b := \min(a, b)$.
The symbol $\mathbb{E}$ denotes the expectation operator with respect to a generic probability measure. If we need to signify the dependence on a probability measure P, then we use P as a subscript in $\mathbb{E}_P$. For random variables $Z_1, \ldots, Z_n$ and a function $g : \mathcal{Z} \to \mathbb{R}$, we define the empirical expectation

$$\mathbb{E}_n[g(Z)] \equiv \mathbb{E}_{\mathbb{P}_n}[g(Z)] := \frac{1}{n}\sum_{i=1}^{n} g(Z_i)$$

and

$$G_n(g) := \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(g(Z_i) - \mathbb{E}\big[g(Z_i)\big]\Big).$$

For a class of measurable functions $\mathcal{F}$ on a measurable space, let $N(\varepsilon, \mathcal{F}, \|\cdot\|)$ be the minimal number of balls $B_\varepsilon(g) := \{f : \|g - f\| < \varepsilon\}$ of radius ε to cover the set $\mathcal{F}$. Let F be an envelope function of $\mathcal{F}$ with $F(x) \geq |f(x)|$ for all $f \in \mathcal{F}$. The uniform entropy number with respect to the $L_r(Q)$ seminorm $||\cdot||_{Q,r}$ is defined as

$$ent(\mathcal{F}, \varepsilon) := \sup_Q \log N(\varepsilon \|F\|_{Q,r}, \mathcal{F}, L_r(Q)),$$

where the supremum is taken over all probability measures Q with $0 < \mathbb{E}_Q[F^r]^{1/r} < \infty$. For any function $\nu(\theta, u)$ we use the notation $\dot{\nu}_{\theta^*}(u) := \partial\nu(\theta, u)/\partial\theta|_{\theta=\theta^*}$, respectively $\nu'_\theta(u^*) := \partial\nu(\theta, u)/\partial u|_{u=u^*}$.

3.2 Transformation Model

We consider a high-dimensional transformation model where the unknown transformation parameter θ_0 is identified as being the only parameter for which the errors are normally distributed. This assumption is typical for transformation models. Let $\{\Lambda_\theta(\cdot) : \theta \in \Theta\}$ be a given parametric family of strictly monotone increasing and two times differentiable functions and $\Theta \subset \mathbb{R}$ be compact. For every $\theta \in \Theta$, we assume a linear model

$$\Lambda_\theta(Y) = X^T\beta_\theta + \varepsilon_\theta \tag{3.4}$$

with $\mathbb{E}[\varepsilon_\theta|X] = 0$. We write

$$\varepsilon_\theta := \Lambda_\theta(Y) - \underbrace{m_\theta(x)}_{=X^T\beta_\theta}$$

with

$$m_\theta(x) \equiv m(\theta, x) := \mathbb{E}[\Lambda_\theta(Y)|X = x].$$

Additionally, define

$$\sigma_\theta^2 \equiv \sigma^2(\theta) := Var(\varepsilon_\theta).$$

We allow the number of covariates $p = p_n$ to increase with the sample size n, but we require that the index set

$$S_\theta := \{j : \beta_{\theta,j} \neq 0\}$$

is sparse for every $\theta \in \Theta$ with $s := \sup_{\theta\in\Theta} ||\beta_\theta||_0$. The number of relevant variables $s = s_n$ may also increase with the sample size n but it does so at a moderate rate. We assume that β_θ is differentiable in θ. Therefore, we can write

$$\dot{\Lambda}_\theta(Y) = X^T\dot{\beta}_\theta + \dot{\varepsilon}_\theta \tag{3.5}$$

with $\mathbb{E}[\dot{\varepsilon}_\theta|X] = 0$ under regularity conditions (as mentioned later on). The model (3.5) is sparse with $\dot{s} := \sup_{\theta\in\Theta} ||\dot{\beta}_\theta||_0 \leq 2s$.

The assumption that β_θ is sparse and differentiable is common in other applications. For example, in high-dimensional quantile regression, Belloni and Chernozhukov [7] assume that for every quantile $u \in (0, 1)$ the coefficient $\beta(u)$ is sparse and smooth with respect to u.

We estimate θ_0 by a method similar to the "profile likelihood procedure" proposed in Linton et al. [70]. The main idea is to formulate our estimation problem as a Z-estimation problem and then plug-in estimates for all unknown terms.

3.2.1 Transformation Parameter

For the estimation of the transformation parameter, we first determine the likelihood. Since $\Lambda_\theta(\cdot)$ is strictly increasing, we have

$$P(Y \leq y|X) = P(\Lambda_\theta(Y) \leq \Lambda_\theta(y)|X) = P(\varepsilon_\theta \leq \Lambda_\theta(y) - m_\theta(X)|X).$$

For $\theta = \theta_0$, we obtain

$$\begin{aligned} P(Y \leq y|X) &= P(\varepsilon_{\theta_0} \leq \Lambda_{\theta_0}(y) - m_{\theta_0}(X)|X) \\ &= P(\varepsilon_{\theta_0} \leq \Lambda_{\theta_0}(y) - m_{\theta_0}(X)) \\ &= \Phi\left(\frac{\Lambda_{\theta_0}(y) - m_{\theta_0}(X)}{\sigma}\right) \end{aligned}$$

with Φ being the cdf of a standard normal distribution and $\sigma \equiv \sigma_{\theta_0}$.
By transforming the densities, we obtain

$$\begin{aligned} f_{Y|X}(y|x) &= f_{\varepsilon_{\theta_0}}(\Lambda_{\theta_0}(y) - m_{\theta_0}(x))\Lambda'_{\theta_0}(y) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\Lambda_{\theta_0}(y) - m_{\theta_0}(x))^2}{2\sigma^2}\right) \Lambda'_{\theta_0}(y) \end{aligned}$$

and therefore the following log-likelihood function

$$l_{Y|X}(\theta) = -\frac{n}{2}\log(2\pi\sigma_\theta^2) - \frac{1}{2\sigma_\theta^2}\sum_{i=1}^{n}(\Lambda_\theta(Y_i) - m_\theta(X_i))^2 + \sum_{i=1}^{n}\log(\Lambda'_\theta(Y_i)).$$

The maximum likelihood estimator

$$\begin{aligned} \theta^* = \arg\max_{\theta\in\Theta}\Big[&-\frac{1}{2}\log(2\pi\sigma_\theta^2) - \frac{1}{2\sigma_\theta^2 n}\sum_{i=1}^{n}(\Lambda_\theta(Y_i) - m_\theta(X_i))^2 \\ &+ \frac{1}{n}\sum_{i=1}^{n}\log(\Lambda'_\theta(Y_i))\Big] \end{aligned} \tag{3.6}$$

fulfills

$$\begin{aligned} 0 &= \partial\Big(-\frac{1}{2}\log(2\pi\sigma_\theta^2) - \frac{1}{2\sigma_\theta^2 n}\sum_{i=1}^{n}(\Lambda_\theta(Y_i) - m_\theta(X_i))^2 \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\log(\Lambda'_\theta(Y_i))\Big)/\partial\theta\Big|_{\theta=\theta^*} \\ &= \frac{1}{n}\sum_{i=1}^{n}\Big[-\frac{\dot\sigma_{\theta^*}{}^2}{2\sigma_{\theta^*}^2} - \frac{1}{\sigma_{\theta^*}^2}(\Lambda_{\theta^*}(Y_i) - m_{\theta^*}(X_i))(\dot\Lambda_{\theta^*}(Y_i) - \dot m_{\theta^*}(X_i)) \\ &\quad + \frac{\dot\sigma_{\theta^*}{}^2}{2\sigma_{\theta^*}^4}(\Lambda_{\theta^*}(Y_i) - m_{\theta^*}(X_i))^2 + \frac{\dot\Lambda'_{\theta^*}(Y_i)}{\Lambda'_{\theta^*}(Y_i)}\Big] \\ &=: \mathbb{E}_n\Big[\psi\big((Y,X),\theta^*,h_0(\theta^*,X)\big)\Big], \end{aligned}$$

where $h_0 : \Theta \times \mathcal{X} \to \mathbb{R} \times \mathbb{R}^+ \times \mathbb{R} \times \mathbb{R}$ with

$$h_0 \equiv (h_{0,1}, h_{0,2}, h_{0,3}, h_{0,4}) := (m_\theta, \sigma_\theta^2, \dot m_\theta, \dot\sigma_\theta^2)$$

is a nuisance function. We substitute the function h_0 by a Lasso estimator $\hat h_0$, which is defined in Subsection 3.2.2 and analyzed in Subsection 3.3.2.
Finally, we estimate the transformation parameter θ_0 by an estimator $\hat\theta$, which solves

$$\Big|\mathbb{E}_n\Big[\psi\big((Y,X),\hat\theta,\hat h_0(\hat\theta,X)\big)\Big]\Big| = \inf_{\theta\in\Theta}\Big|\mathbb{E}_n\Big[\psi\big((Y,X),\theta,\hat h_0(\theta,X)\big)\Big]\Big| + \epsilon_n, \tag{3.7}$$

where $\epsilon_n = o\left(n^{-1/2}\right)$ is the numerical tolerance.

3.2.2 Nuisance Function

The unknown nuisance function

$$h_0 = (m_\theta, \sigma_\theta^2, \dot{m}_\theta, \dot{\sigma}_\theta^2)$$

can be estimated by

$$\hat{h}_0 = (\hat{m}_\theta, \hat{\sigma}_\theta^2, \hat{\dot{m}}_\theta, \hat{\dot{\sigma}}_\theta^2),$$

where $\hat{m}_\theta(x) = x^T\hat{\beta}_\theta$ with $\hat{\beta}_\theta$ being the Lasso estimate

$$\arg\max_\beta \mathbb{E}_n[(\Lambda_\theta(Y) - x^T\beta)^2] + \frac{\lambda}{n}||\Psi_\theta\beta||_1$$

with penalty term λ and penalty loadings Ψ_θ as in Belloni et al. [11] (p. 260). Analogously, we estimate $\dot{m}_\theta$ by $\hat{\dot{m}}_\theta(x) = x^T\hat{\dot{\beta}}_\theta$ with $\hat{\dot{\beta}}_\theta$ being the Lasso estimate

$$\arg\max_\beta \mathbb{E}_n[(\dot{\Lambda}_\theta(Y) - x^T\beta)^2] + \frac{\tilde{\lambda}}{n}||\tilde{\Psi}_\theta\beta||_1.$$

The unknown variance σ_θ^2 can be estimated by

$$\hat{\sigma}_\theta^2 := \frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_{i,\theta}^2$$

and $\dot{\sigma}_\theta^2$ by

$$\hat{\dot{\sigma}}_\theta^2 := \frac{2}{n}\sum_{i=1}^n \hat{\varepsilon}_{i,\theta}\hat{\dot{\varepsilon}}_{i,\theta}$$

with $\hat{\varepsilon}_{i,\theta} := \Lambda_\theta(Y_i) - \hat{m}_\theta(X_i)$ and $\hat{\dot{\varepsilon}}_{i,\theta} := \dot{\Lambda}_\theta(Y_i) - \hat{\dot{m}}_\theta(X_i)$. Under regularity conditions,

$$\dot{\sigma}_\theta^2 = \partial\mathbb{E}[\varepsilon_\theta^2]/\partial\theta = \mathbb{E}[\partial(\varepsilon_\theta^2)/\partial\theta] = 2\mathbb{E}[\varepsilon_\theta\dot{\varepsilon}_\theta]$$

holds.

3.2.3 Identification of the True Transformation

First, we formulate our estimation problem as a Z-estimation problem (cf. 3.2.1). Let

$$\mathcal{H} = \mathcal{H}_1 \times \mathcal{H}_2 \times \mathcal{H}_3 \times \mathcal{H}_4$$

be a suitable convex space of measurable functions with $\mathcal{H}_1 = \{h_1 : (\theta, x) \mapsto \mathbb{R}\}$, $\mathcal{H}_2 = \{h_2 : \theta \mapsto \mathbb{R}^+\}$, $\mathcal{H}_3 = \{h_3 : (\theta, x) \mapsto \mathbb{R}\}$ and $\mathcal{H}_4 = \{h_4 : \theta \mapsto \mathbb{R}\}$. We obtain the moment function

$$\psi\big((Y, X), \theta, h\big) : (\mathcal{Y} \times \mathcal{X}) \times \Theta \times \mathcal{H} \to \mathbb{R}$$

with

$$\big((Y, X), \theta, h\big) \mapsto \underbrace{-\frac{h_4(\theta)}{2h_2(\theta)}}_{=:I(\theta,h_2,h_4)} \underbrace{-\frac{1}{h_2(\theta)}(\Lambda_\theta(Y) - h_1(\theta, X))(\dot{\Lambda}_\theta(Y) - h_3(\theta, X))}_{=:II(\theta,h_1,h_2,h_3)}$$

$$+ \underbrace{\frac{h_4(\theta)}{2\big(h_2(\theta)\big)^2}(\Lambda_\theta(Y) - h_1(\theta, X))^2}_{=:III(\theta,h_1,h_2,h_4)} + c_\theta,$$

where $c_\theta := \frac{\dot{\Lambda}'_\theta(Y)}{\Lambda'_\theta(Y)}$. The supports of X and Y are given by $\mathcal{X}$ and $\mathcal{Y}$, respectively.

The next lemma ensures the identification of the transformation parameter under weak regularity conditions. The conditions A1, A4, and A5 are only sufficient conditions and they will be required for our main theorem stated in Subsection 3.3.4.

Lemma 1. *Under the conditions A1, A4, and A5, the true parameter θ_0 is identified as a unique solution of the moment condition*

$$\mathbb{E}\Big[\psi\big((Y, X), \theta_0, h_0\big)\Big] = 0.$$

Proof.
We use the same argument as Neumeyer et al. [78]. Define

$$f^{(\theta)}(y|x) := \frac{1}{\sqrt{2\pi\sigma_\theta^2}} \exp\left(-\frac{(\Lambda_\theta(y) - x\beta_\theta)^2}{2\sigma_\theta^2}\right) \Lambda'_\theta(y).$$

The expected Kullback-Leibler-Distance between $f_{Y|X}$ and $f^{(\theta)}$ is greater or equal to zero and equality only holds for the true parameter θ_0. Therefore, the following expression is minimized in θ_0

$$\begin{aligned}
&\int\int \log\left(\frac{f_{Y|X}(y|x)}{f^{(\theta)}(y|x))}\right) f_{Y|X}(y|x)dydF_X(x) \\
&= \underbrace{\int\int \log(f_{Y|X}(y|x) f_{Y|X}(y|x)dydF_X(x)}_{constant} \\
&\quad - \underbrace{\int\int \log(f^{(\theta)}(y|x)) f_{Y|X}(y|x)dydF_X(x)}_{=\mathbb{E}[\log(f^{(\theta)}(Y|X))]}.
\end{aligned}$$

It follows that $\mathbb{E}[\log(f^{(\theta)}(Y|X))]$ is maximized for the true parameter $\theta = \theta_0$. Under the regularity conditions A1, A4 and A5, it holds

$$\begin{aligned}
\mathbb{E}\Big[\psi\big((Y, X), \theta_0, h_0\big)\Big] &= \mathbb{E}\left[\frac{\partial}{\partial\theta} \log(f^{(\theta)}(Y|X))\big|_{\theta=\theta_0}\right] \\
&= \frac{\partial}{\partial\theta}\mathbb{E}[\log(f^{(\theta)}(Y|X))]\big|_{\theta=\theta_0} = 0.
\end{aligned}$$

Here, we used that for all θ

$$0 < c \leq \sigma_\theta^2 \quad \text{and} \quad \sigma_\theta^2 \leq \mathbb{E}\left[\sup_{\theta\in\Theta} \varepsilon_\theta^2\right] \leq C < \infty,$$

which is shown in the proof of Theorem 7. □

3.3 Main Results

This section focuses on the central elements of our Z-estimation problem, which are Neyman orthogonality, uniform estimation of the nuisance function, and a theorem about the asymptotic distribution of the estimated transformation parameter based on an entropy condition. In the following, we consider the model described in Section 3.2.

3.3.1 Neyman Orthogonality Condition

To be able to use plug-in estimators for the nuisance function, the moment condition to identify θ_0 needs to be insensitive towards small changes in the estimated nuisance function. This property is granted by the Neyman orthogonality condition that is defined in Chernozhukov et al. [35]. In this work, the authors describe the condition in great detail and they provide an extensive overview of the settings where the condition holds.
To prove the Neyman orthogonality condition, we define the Gateaux derivative with respect to some $h \in \mathcal{H}$ in h_0

$$D_r[h - h_0] := \partial_r \Big\{ \mathbb{E}\Big[\psi\Big((Y, X), \theta_0, h_0 + r(h - h_0)\Big)\Big]\Big\},$$

where

$$\begin{aligned} &h_0 + r(h - h_0) \\ :=& \Big(m_\theta + r(h_1 - m_\theta), \sigma_\theta^2 + r(h_2 - \sigma_\theta^2), \dot{m}_\theta + r(h_1 - \dot{m}_\theta), \dot{\sigma}_\theta^2 + r(h_2 - \sigma_\theta^2)\Big). \end{aligned}$$

It is important to mention that $\mathcal{H}_1$ to $\mathcal{H}_4$ are assumed to be convex, which ensures that the term $\psi\Big((Y, X), \theta_0, h_0 + r(h - h_0)\Big)$ is well defined and exists for all $r \in [0, 1)$.

Lemma 2. *Let $\mathcal{H}' \subseteq \mathcal{H}$. Under the conditions*

$$\mathbb{E}\left[\sup_{\theta \in \Theta} \varepsilon_\theta^2\right] < \infty \quad \textit{and} \quad \mathbb{E}\left[\sup_{h \in \mathcal{H}'} \left|\psi\Big((Y, X), \theta_0, h\Big)\right|\right] < \infty$$

the Neyman orthogonality condition

$$D_0[h - h_0] = 0$$

is satisfied for all $h \in \mathcal{H}'$.

It is sufficient to restrict the condition onto the nuisance realization set, that is defined in Subsection 3.3.3, which contains the estimated nuisance function with probability $1 - o(1)$.

Our estimation procedure is closely related to the "concentrated out" approach in general likelihood and other M-estimation problems described in Chernozhukov et al. [35] and Newey [79]. In Lemma 2.5, Chernozhukov et al. [35] provide conditions when the score ψ is Neyman orthogonal at (θ_0, h_0). They suppose that the target parameter θ and the nuisance parameter $h_0(\theta)$ solve the optimization problems

$$\max_{\theta \in \Theta, h \in \mathcal{H}} \mathbb{E}[l((Y, X), \theta, h(\theta))] \tag{3.8}$$

and

$$h_0(\theta) = \arg\max_{h \in \mathcal{H}} \mathbb{E}[l((Y, X), \theta, h(\theta))] \tag{3.9}$$

for all $\theta \in \Theta$, where l is a known criterion function. However, our model does not fit in this setting since we set

$$l_{Y|X}(\theta) = -\frac{n}{2} \log(2\pi\sigma_\theta^2) - \frac{1}{2\sigma_\theta^2} \sum_{i=1}^{n} (\Lambda_\theta(Y_i) - m_\theta(X_i))^2 + \sum_{i=1}^{n} \log(\Lambda_\theta'(Y_i)),$$

which is the log-likelihood of our model (3.4) only if $\theta = \theta_0$. Therefore, in general, $h_0(\theta)$ does not satisfy (3.9) and our problem is not covered by this setting.

Next, we give a set of assumptions that are needed for the following theorems and have already been used for Lemma 1 (A1, A4, and A5).

Assumptions **A1**-**A11**.
The following assumptions hold uniformly in $n \geq n_0, P \in \mathcal{P}_n$:

A1

$$\mathbb{E}\left[\sup_{\theta\in\Theta} |\log(\Lambda'_\theta(Y))|\right] < \infty$$

A2 The parameters obey the growth condition

$$s\log(p \vee n) \leq \delta_n n^{1/2}$$

and

$$\log^3(p \vee n) \leq \delta_n n$$

for $\delta_n \searrow 0$ approaching zero from above at a speed at most polynomial in n.

A3 For all $n \in \mathbb{N}$, the regressor $X = (X_1, \ldots, X_p)$ has a bounded support $\mathcal{X}$.

A4 Uniformly in θ, the conditional variance of the error term and its derivation with respect to the transformation parameter are bounded:

$$0 < c \leq \inf_{\theta\in\Theta} \mathbb{E}\big[\varepsilon_\theta^2|X\big] \leq \sup_{\theta\in\Theta} \mathbb{E}\big[\varepsilon_\theta^2|X\big] \leq C < \infty$$
$$0 < c \leq \inf_{\theta\in\Theta} \mathbb{E}\big[\dot{\varepsilon}_\theta^2|X\big] \leq \sup_{\theta\in\Theta} \mathbb{E}\big[\dot{\varepsilon}_\theta^2|X\big] \leq C < \infty.$$

A5 The transformations and its derivations are measurable and the classes of functions

$$\mathcal{F}_\Lambda := \big\{\Lambda_\theta(\cdot)|\theta \in \Theta\big\} \quad \dot{\mathcal{F}}_\Lambda := \big\{\dot{\Lambda}_\theta(\cdot)|\theta \in \Theta\big\}$$

have VC index $C_\Lambda < \infty$ and $\dot{C}_\Lambda < \infty$, respectively. Further, the classes $\mathcal{F}_\Lambda$ and $\dot{\mathcal{F}}_\Lambda$ have envelopes F_Λ and $\dot{F}_\Lambda$, respectively, with

$$\mathbb{E}[F_\Lambda(Y)^{14}] < \infty \quad \text{and} \quad \mathbb{E}[\dot{F}_\Lambda(Y)^8] < \infty.$$

A6 The following condition for the second derivation of the transformation with respect to θ holds:

$$\sup_{\theta\in\Theta} \mathbb{E}\Big[\big(\ddot{\Lambda}_\theta(Y)\big)^2\Big] \leq C.$$

A7 The minimum and maximum sparse eigenvalues of X are bounded away from zero and above, namely

$$0 < \kappa' \leq \inf_{||\delta||_0 \leq s\log(n), ||\delta||=1} ||X^T\delta||_{P,2}$$
$$\leq \sup_{||\delta||_0 \leq s\log(n), ||\delta||=1} ||X^T\delta||_{P,2} \leq \kappa'' < \infty.$$

A8 The class of functions

$$\mathcal{J}_\Lambda := \left\{ c_\theta(\cdot) = \frac{\dot{\Lambda}'_\theta(\cdot)}{\Lambda'_\theta(\cdot)} \Big| \theta \in \Theta \right\}$$

has an envelope J_Λ with

$$\mathbb{E}[J_\Lambda(Y)^6] < \infty.$$

A9 For all $\theta \in \Theta$ and $\tilde{h} \in \tilde{\mathcal{H}}$, it holds that

(i)

$$\mathbb{E}\Big[\Big(\psi\big((Y,X),\theta,h_0(\theta,X)\big) - \psi\big((Y,X),\theta_0,h_0(\theta_0,X)\big)\Big)^2\Big] \leq C|\theta-\theta_0|^2$$

(ii)

$$\begin{aligned}&\mathbb{E}\Big[\Big(\psi\big((Y,X),\theta,\tilde{h}(\theta,X)\big) - \psi\big((Y,X),\theta,h_0(\theta,X)\big)\Big)^2\Big] \\ &\leq C\mathbb{E}\left[\|\tilde{h}(\theta,X) - h_0(\theta,X)\|_2^2\right]\end{aligned}$$

(iii)

$$\begin{aligned}&\sup_{r\in(0,1)}\left|\partial_r^2\left\{\mathbb{E}\Big[\psi\big((Y,X),\theta_0+r(\theta-\theta_0),h_0+r(\tilde{h}-h_0)\big)\Big]\right\}\right| \\ &\leq C\left(|\theta-\theta_0|^2 + \sup_{\theta^*\in\Theta}\mathbb{E}\Big[\|\tilde{h}(\theta^*,X) - h_0(\theta^*,X)\|_2^2\Big]\right)\end{aligned}$$

for a constant C independent from θ and $\tilde{\mathcal{H}}$ defined in Subsection 3.3.3.

A10 For $h \in \tilde{\mathcal{H}}$, the function

$$\theta \mapsto \mathbb{E}\Big[\psi\big((Y,X),\theta,h(\theta,X)\big)\Big]$$

is differentiable in a neighbourhood of θ_0 and, for all $\theta \in \Theta$, the identification relation

$$2|\mathbb{E}[\psi((Y,X)),\theta,h_0(\theta,X)]| \geq |\Gamma(\theta-\theta_0)| \wedge c_0$$

is satisfied with

$$\Gamma := \partial_\theta \mathbb{E}\Big[\psi\big((Y,X),\theta_0,h_0(\theta_0,X)\big)\Big] > c_1.$$

A11 The map $(\theta,h) \mapsto \mathbb{E}[\psi((X,Y),\theta,h)]$ is twice continuously Gateaux-differentiable on $\Theta \times \mathcal{H}$.

Assumptions A1-A11 are a set of sufficient conditions for the main result stated in Theorem 6. Assumption A1 allows us to interchange derivation and integration, which is necessary for the verification of the moment condition. In the sparsity condition A2, 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. Our growth condition is in line with other results in the literature, e.g., with Belloni et al. [9] and many others. Assumptions A2, A6, and A7 are needed for the uniform estimation of the nuisance function. Condition A3 can be relaxed, cf. Assumption 6.1 from Belloni et al. [11]. Assumption A7 is a standard eigenvalue

condition for the Lasso estimation. Condition A4 prevents degenerate distributions in the models (3.4) and (3.5). Assumptions A5, A6 and A8 control the complexity of the class of transformations and bound the moments uniformly over θ. Assumptions A5 and A6 will be discussed in more detail in Comment 3.3.1. Assumption A9 is a set of mild smoothness conditions. Assumption A10 implies sufficient identifiability of the true transformation parameter θ_0. Assumption A11 only requires differentiability of the function $(\theta, h) \mapsto \mathbb{E}[\psi((X,Y),\theta,h)]$ which is a weaker condition than the differentiability of the function $(\theta, h) \mapsto \psi((X,Y),\theta,h)$.

Comment 3.3.1.
Since there exists a true parameter θ_0 such that all moments of $\Lambda_{\theta_0}(Y)$ exist, choosing an appropriate class of transformations $\mathcal{F}_\Lambda$ and restricting the parameter space lead to reasonable assumptions on the moments in A5 and A6. Consider the class of Box-Cox transformations

$$\Lambda_\theta(y) = \begin{cases} \frac{y^\theta - 1}{\theta} & \text{for } \theta \neq 0 \\ \log(y) & \text{for } \theta = 0 \end{cases}$$

and let, without loss of generality, $\Theta = [a,b]$ with $a < 0 < b$. We show that Assumption A5 and A6 are satisfied if $\mathbb{E}[Y^{14a}]$ and $\mathbb{E}[Y^{14b}]$ exist. The envelope fulfills

$$\sup_{\theta\in\Theta} |\Lambda_\theta(y)| = \max_{\theta\in\{a,b\}} |\Lambda_\theta(y)| = \Lambda_b(y)1_{\{y\geq 1\}} - \Lambda_a(y)1_{\{0\leq y<1\}}$$

since $\Lambda_\theta(\cdot)$ is monotonically increasing in θ and positive for all θ if $y \geq 1$. Hence,

$$\begin{aligned} \mathbb{E}\left[\left(\sup_{\theta\in\Theta} |\Lambda_\theta(Y)|\right)^{14}\right] &= \mathbb{E}[\Lambda_b(Y)^{14}1_{\{Y\geq 1\}}] + \mathbb{E}[\Lambda_a(Y)^{14}1_{\{0\leq Y<1\}}] \\ &= \mathbb{E}\left[\left(\frac{Y^b - 1}{b}\right)^{14} 1_{\{Y\geq 1\}}\right] + \mathbb{E}\left[\left(\frac{Y^a - 1}{a}\right)^{14} 1_{\{0\leq Y<1\}}\right] \\ &< \infty. \end{aligned}$$

Analogously, we have

$$\sup_{\theta\in\Theta} |\dot{\Lambda}_\theta(y)| = \max_{\theta\in\{a,b\}} |\dot{\Lambda}_\theta(y)| = \dot{\Lambda}_b(y)1_{\{y\geq 1\}} + \dot{\Lambda}_a(y)1_{\{0\leq y<1\}}$$

since $\dot{\Lambda}_\theta(\cdot)$ is continuously in θ and monotonically increasing for all θ if $y \geq 1$ and monotone decreasing for $y < 1$. Hence,

$$\begin{aligned} \mathbb{E}\left[\left(\sup_{\theta\in\Theta} |\dot{\Lambda}_\theta(Y)|\right)^{8}\right] &= \mathbb{E}[\dot{\Lambda}_b(Y)^8 1_{\{Y\geq 1\}}] + \mathbb{E}[\dot{\Lambda}_a(Y)^8 1_{\{0\leq Y<1\}}] \\ &= \mathbb{E}\left[\left(\frac{1}{b^2}(b\log(Y)Y^b - Y^b + 1)\right)^8 1_{\{Y\geq 1\}}\right] \\ &\quad + \mathbb{E}\left[\left(\frac{1}{a^2}(a\log(Y)Y^a - Y^a + 1)\right)^8 1_{\{0\leq Y<1\}}\right] \\ &< \infty \end{aligned}$$

if $\mathbb{E}[Y^{14a}]$ *and* $\mathbb{E}[Y^{14b}]$ *exist. Further, it holds*

$$\sup_{\theta\in\Theta}\mathbb{E}\left[\left(\ddot{\Lambda}_\theta(Y)\right)^2\right]=\sup_{\theta\in\Theta}\mathbb{E}\left[\left(\frac{1}{\theta^3}\left(\log(Y^\theta)-1\right)^2Y^\theta+Y^\theta-2\right)^2\right].$$

The class of Yeo-Johnson power transformations is an extension of the Box-Cox transformations allowing for negative values in the domain.

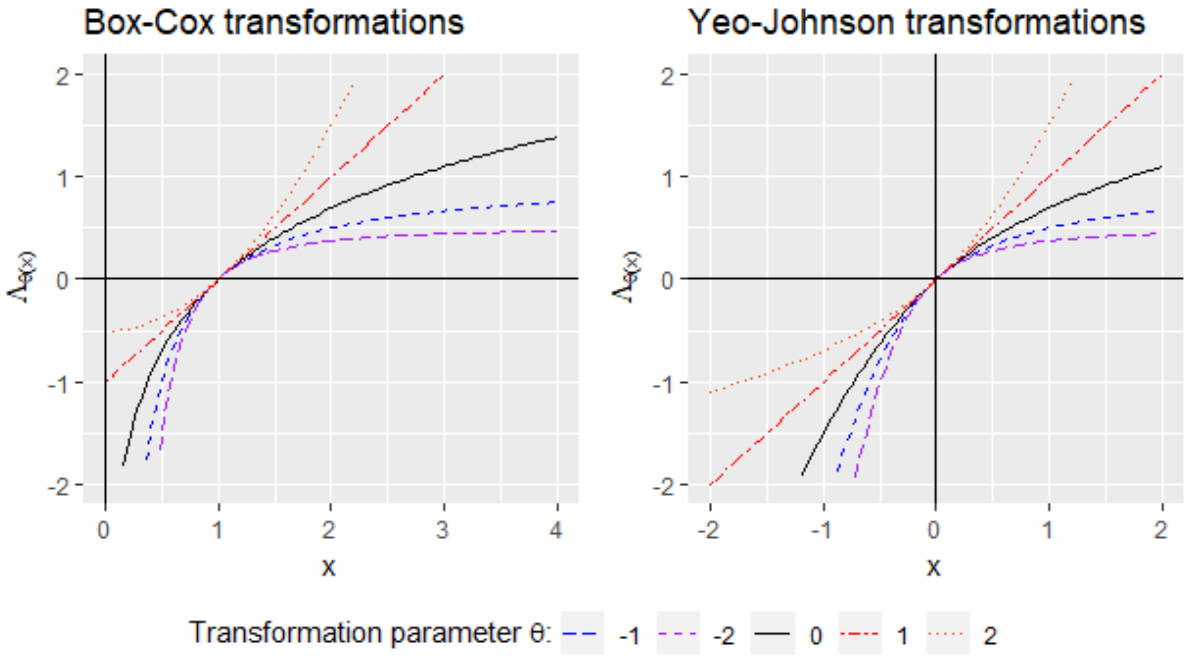


Figure 3.1: **Box-Cox and Yeo-Johnson transformations for different transformation parameters.**

As illustrated in the Figure 3.1, the tail behavior of the Yeo-Johnson power transformations are closely related to the Box-Cox transformations implying similar assumptions on the moments of Y *to ensure the Assumptions A5 and A6. Since the Box-Cox transformations are bounded from below by* $-1/\theta$ *for* $\theta>0$ *and from above by* $-1/\theta$ *for* $\theta<0$*, the transformation* $\Lambda_\theta(Y)$ *cannot be normally distributed except when* $\theta=0$*. This problem has also been discussed in Draper and Cox [41] and in Amemiya and Powell [1]. Hence,* $\theta_0=0$ *is the only possible null hypothesis for this class of transformations. In the class of Yeo-Johnson power transformations the range of valid null hypotheses is given by* $\theta_0\in[0,2]$*.*

The following lemma shows that the first part of condition A5 is satisfied for the popular Box-Cox power transformations and the modification proposed by Yeo and Johnson.

Lemma 3. *The class of Box-Cox transformations* $\mathcal{F}_1=\{\Lambda_\theta(\cdot)|\theta\in\mathbb{R}\}$ *and the class of derivatives* $\mathcal{F}_2=\{\dot{\Lambda}_\theta(\cdot)|\theta\in\mathbb{R}\}$ *with respect to the transformation parameter* θ *are VC classes. The same holds for Yeo-Johnson power transformations.*

The proof of the lemma is given in Appendix 3.7.

3.3.2 Uniform Estimation of the Nuisance Functions

The rates for the estimation of the regression functions m_θ and $\dot{m}_\theta$ can be directly obtained by the uniform prediction rates of the Lasso estimator. The proofs are given in the appendix.

Theorem 3.
Under the Assumptions A1-A7, uniformly for all $P \in \mathcal{P}_n$ with probability $1 - o(1)$, it holds that:

$$\sup_{\theta\in\Theta} ||\hat{\beta}_\theta||_0 = O(s) \tag{3.10}$$

$$\sup_{\theta\in\Theta} ||X^T(\hat{\beta}_\theta - \beta_\theta)||_{\mathbb{P}_{n,2}} \leq \tilde{\delta}_n n^{-\frac{1}{4}} \tag{3.11}$$

$$\sup_{\theta\in\Theta} ||\hat{\beta}_\theta - \beta_\theta||_1 \leq \tilde{\delta}_n \sqrt{s} n^{-\frac{1}{4}}, \tag{3.12}$$

respectively

$$\sup_{\theta\in\Theta} ||\hat{\dot{\beta}}_\theta||_0 = O(s) \tag{3.13}$$

$$\sup_{\theta\in\Theta} ||X^T(\hat{\dot{\beta}}_\theta - \dot{\beta}_\theta)||_{\mathbb{P}_{n,2}} \leq \tilde{\delta}_n n^{-\frac{1}{4}} \tag{3.14}$$

$$\sup_{\theta\in\Theta} ||\hat{\dot{\beta}}_\theta - \dot{\beta}_\theta||_1 \leq \tilde{\delta}_n \sqrt{s} n^{-\frac{1}{4}}, \tag{3.15}$$

where $\tilde{\delta}_n$ is a positive sequence approaching zero from above at a polynomial speed in n.

As a consequence of the uniform rates of the Lasso estimator, we are able to achieve uniform rates for the estimation of the variance σ_θ^2 and its derivation $\dot{\sigma}_\theta^2$.

Theorem 4.
Under the assumptions of Theorem 3, uniformly for all $P \in \mathcal{P}_n$ with probability $1 - o(1)$, it holds that:

$$\sup_{\theta\in\Theta} |\hat{\sigma}_\theta^2 - \sigma_\theta^2| \leq \tilde{\delta}_n n^{-\frac{1}{4}} \tag{3.16}$$

$$\sup_{\theta\in\Theta} |\hat{\dot{\sigma}}_\theta^2 - \dot{\sigma}_\theta^2| \leq \tilde{\delta}_n n^{-\frac{1}{4}}. \tag{3.17}$$

3.3.3 Entropy Condition

At first, we define the following classes of functions

$$\begin{aligned}
\tilde{\mathcal{H}}_1 := \Big\{ & \tilde{h}_1 : \Theta \times \mathcal{X} \to \mathbb{R} |\ \tilde{h}_1(\theta, x) = x^T \tilde{\beta}_\theta, \|\tilde{\beta}_\theta\|_0 \leq Cs, \|\tilde{\beta}_\theta - \beta_\theta\|_1 \leq \tilde{\delta}_n \sqrt{s} n^{-\frac{1}{4}}, \\
& \|X^T(\tilde{\beta}_\theta - \beta_\theta)\|_{P,2} \leq \tilde{\delta}_n n^{-\frac{1}{4}} \Big\}, \\
\tilde{\mathcal{H}}_2 := \Big\{ & \tilde{h}_2 : \Theta \to \mathbb{R}^+ |\ |\tilde{h}_2(\theta) - \sigma_\theta^2| \leq \tilde{\delta}_n n^{-\frac{1}{4}} \Big\}, \\
\tilde{\mathcal{H}}_3 := \Big\{ & \tilde{h}_3 : \Theta \times \mathcal{X} \to \mathbb{R} |\ \tilde{h}_3(\theta, x) = x^T \tilde{\beta}_\theta, \|\tilde{\beta}_\theta\|_0 \leq Cs, \|\tilde{\beta}_\theta - \dot{\beta}_\theta\|_1 \leq \tilde{\delta}_n \sqrt{s} n^{-\frac{1}{4}}, \\
& \|X^T(\tilde{\beta}_\theta - \dot{\beta}_\theta)\|_{P,2} \leq \tilde{\delta}_n n^{-\frac{1}{4}} \Big\}, \\
\tilde{\mathcal{H}}_4 := \Big\{ & \tilde{h}_4 : \Theta \to \mathbb{R} |\ |\tilde{h}_4(\theta) - \dot{\sigma}_\theta^2| \leq \tilde{\delta}_n n^{-\frac{1}{4}} \Big\}
\end{aligned}$$

and

$$\tilde{\mathcal{H}} := \tilde{\mathcal{H}}_1 \times \tilde{\mathcal{H}}_2 \times \tilde{\mathcal{H}}_3 \times \tilde{\mathcal{H}}_4.$$

The set $\tilde{\mathcal{H}}$ is called the nuisance realization set. Theorems 3 and 4 enable us to choose constants C independent from θ and still contain the estimated functions in $\tilde{\mathcal{H}}$ with probability $1-o(1)$. Furthermore, for an arbitrary but fixed $\theta \in \Theta$, we define the following projections

$$\tilde{\mathcal{H}}_1(\theta) := \Big\{\tilde{h}_1 : \mathcal{X} \to \mathbb{R} |\ \tilde{h}_1(x) = \tilde{h}_1(\theta, x) \in \tilde{\mathcal{H}}_1\Big\}$$
$$\tilde{\mathcal{H}}_2(\theta) := \Big\{c \in \mathbb{R}^+ \big|\ |c - \sigma_\theta^2| \leq \tilde{\delta}_n n^{-1/4}\Big\}$$

and $\tilde{\mathcal{H}}_3(\theta)$, $\tilde{\mathcal{H}}_4(\theta)$, respectively, $\tilde{\mathcal{H}}(\theta)$, analogously. We restrict the entropy of $\tilde{\mathcal{H}}(\theta)$ uniformly over θ to use the maximal inequality stated in Theorem 5.1 from Chernozhukov et al. [31]. This enables us to bound the empirical process in the proof of Theorem 8 (step 1).

Theorem 5. *Under the Assumptions A4, A5, and A8, the class of functions*

$$\Psi(\theta) = \Big\{(y, x) \mapsto \psi\big((y, x), \theta, \tilde{h}(\theta, x)\big), \tilde{h} \in \tilde{\mathcal{H}}(\theta)\Big\}$$

has a measurable envelope $\bar{\psi} \geq \sup_{\psi \in \Psi(\theta)} |\psi|$ *independent from* θ *with*

$$\mathbb{E}\Big[(\bar{\psi}(Y, X))^q\Big] \leq C_1$$

for some $q \geq 4$. *The class* $\Psi(\theta)$ *is pointwise measurable and, uniformly for all* $\theta \in \Theta$, *it holds*

$$\sup_Q \log N(\varepsilon ||\bar{\psi}||_{Q,2}, \Psi(\theta), L_2(Q)) \leq C_1 s \log\left(\frac{C_2(p \vee n)}{\varepsilon}\right)$$

with C_1 *and* C_2 *being independent from* θ.

The motivation of the entropy condition stated in Theorem 5 is described in Comment 3.9.1. The entropy condition and the results in Subsection 3.3.1 and Subsection 3.3.2 enable us to establish the asymptotic distribution of our estimated transformation parameter.

3.3.4 Main Theorem

The main theorem provides that our estimator $\hat{\theta}$ converge with rate $1/\sqrt{n}$ and is asymptotic unbiased and normal.

Theorem 6. *Under the Assumptions A1-A11, the estimator* $\hat{\theta}$ *in (3.7) obeys*

$$n^{\frac{1}{2}}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma),$$

where

$$\Sigma := \mathbb{E}\Big[\Gamma^{-2}\psi^2\big((Y, X), \theta_0, h_0(\theta_0, X)\big)\Big]$$

with $\Gamma = \partial_\theta \mathbb{E}\Big[\psi\big((Y, X), \theta_0, h_0(\theta_0, X)\big)\Big]$.

A standard bootstrap can be applied to estimate the unknown variance Σ. Therefore, asymptotic level-α tests for null hypothesis can be constructed based on Theorem 6.

3.4 Simulation

This section provides a simulation study of the proposed estimator. The data generating process is given by

$$\Lambda_{\theta_0}(Y) = X^T \beta_{\theta_0} + \varepsilon.$$

The coefficients are set to

$$\beta_{\theta_0,j} = \begin{cases} 1 & \text{for } j \leq s \\ 0 & \text{for } j > s. \end{cases}$$

Therefore, β_{θ_0} is a sparse vector with $\|\beta_{\theta_0}\|_0 = s$ and $\{\Lambda_\theta : \theta \in \Theta\}$ is a given class of transformations. The design matrix is simulated as

$$X \sim \mathcal{N}\left(0, \Sigma^{(X)}\right)$$

for the following three different correlation structures

$$\Sigma_0^{(X)} = I_p,$$

$$\Sigma_1^{(X)} = (c^{|i-j|})_{i,j\in\{1,\dots,p\}},$$

and

$$\Sigma_2^{(X)} = (1-c^p)I_p + (c^{p-|i-j|})_{i,j\in\{1,\dots,p\}}$$

with $c = 0.35$. The error terms ε are i.i.d. $\mathcal{N}(0,\sigma^2)$-distributed, where σ^2 is chosen according to the correlation matrix $\Sigma^{(X)}$ to keep the signal-to-noise ratio (SNR) at a fixed level. To obtain the simulated values for Y_i, which are used for the estimation of θ_0, we apply the inverse transformation $\Lambda_{\theta_0}^{-1}$ onto the simulated values of $\Lambda_{\theta_0}(Y_i)$. We consider different classes of transformations, correlation structures, and vary the number of the regressors p, the number of observations n, the sparsity index s as well as the SNR. Additional simulations with approximate sparsity and non-normal errors are displayed in Appendix 3.10. The SNR is defined as

$$SNR = \frac{Var(X^T\beta_{\theta_0})}{Var(\varepsilon)}.$$

The number of repetitions is set to $R = 500$. The accuracy of the estimate $\hat{\theta}$ is measured by the mean-absolute-error (MAE)

$$\text{MAE} = \frac{1}{R}\sum_{h=1}^{R} |\hat{\theta}_h - \theta_0|.$$

The accuracy of the predictive performance is measured out-of-sample on an independent sample (test sample) by the relative mean-squared-error

$$\text{MSE} = \frac{1}{R}\sum_{h=1}^{R} \mathbb{E}_{n_t}\left[\left(\Lambda_{\theta_0}(Y) - X^T\hat{\beta}_{\hat{\theta}_h}\right)^2\right] / Var(\varepsilon).$$

The empirical expectation $\mathbb{E}_{n_t}[\cdot]$ is taken over the test sample of size $n_t = 200$. Both measures MAE and MSE are based on the unknown transformation parameter θ_0. Additionally, for a fixed level $\alpha = 0.05$, we validate the significance level (acceptance rate) of a test of the form

$$H_0 : \theta_0 = \theta.$$

We test if a given $\theta \in \Theta$ is the right transformation parameter to guarantee normally distributed errors. Therefore, we estimate the unknown variance Σ via bootstrap by drawing $k = 100$ bootstrap samples and construct a $(1-\alpha)$-confidence interval of the form

$$\left[\theta - \sqrt{\hat{\Sigma}} z_{(1-\alpha/2)}, \theta + \sqrt{\hat{\Sigma}} z_{(1-\alpha/2)}\right],$$

where z_γ is the γ-quantile of the standard normal distribution. The empirical acceptance rate is reported.

3.4.1 Box-Cox Power Transformations

In the first setting, we analyze the class of Box-Cox transformations. The Box-Cox transformations are defined as

$$\Lambda_\theta(y) = \begin{cases} \frac{y^\theta - 1}{\theta} & \text{for } \theta \neq 0 \\ \log(y) & \text{for } \theta = 0. \end{cases}$$

This class and the class of its derivatives with respect to the transformation parameter θ are VC classes by Lemma 3.

In their initial paper "An Analysis of Transformations", Box and Cox [16] proposed to calculate approximate confidence intervals for the transformation parameter based on the quantiles of the chi-squared distribution. The *R* package *MASS* by Venables and Ripley [98] includes the function *boxcox* which computes and optionally plots the profile log-likelihood for the parameter of the Box-Cox power transformation. The plot includes the 95%-confidence intervals for the transformation parameter. Nevertheless, such confidence intervals are only valid in the low-dimensional case. Figure 3.2 displays a short simulation ($n = 100$, $s = 5$, SNR $= 1$, $\theta_0 = 0$ and covariance structure $\Sigma_1^{(X)}$) to emphasize that the coverage of their approach declines with an increasing number of regressors (p/n close to one), whereas our method is able to provide valid confidence intervals.

Figure 3.2: **Coverage for an increasing number of regressors.**

To test the behavior of our proposed estimator in a high-dimensional setting, we set the true transformation parameter $\theta_0 = 0$ and summarize the results for the all settings in the Tables 3.1–3.3. In the three settings, the average of the estimators is close to the true value of 0. The acceptance rate, MAE and relative MSE seem to be comparable for all three settings. In summary, the results reveal that the estimated parameter value is, on average, close to the true one and that the acceptance rate is close to the nominal level of 95%.
Figure 3.3 shows the empirical distribution of $\hat{\theta}$ generated by 10000 independent simulations from the last setting in Table 3.2 with SNR=1. This confirms that our estimator is normally distributed.

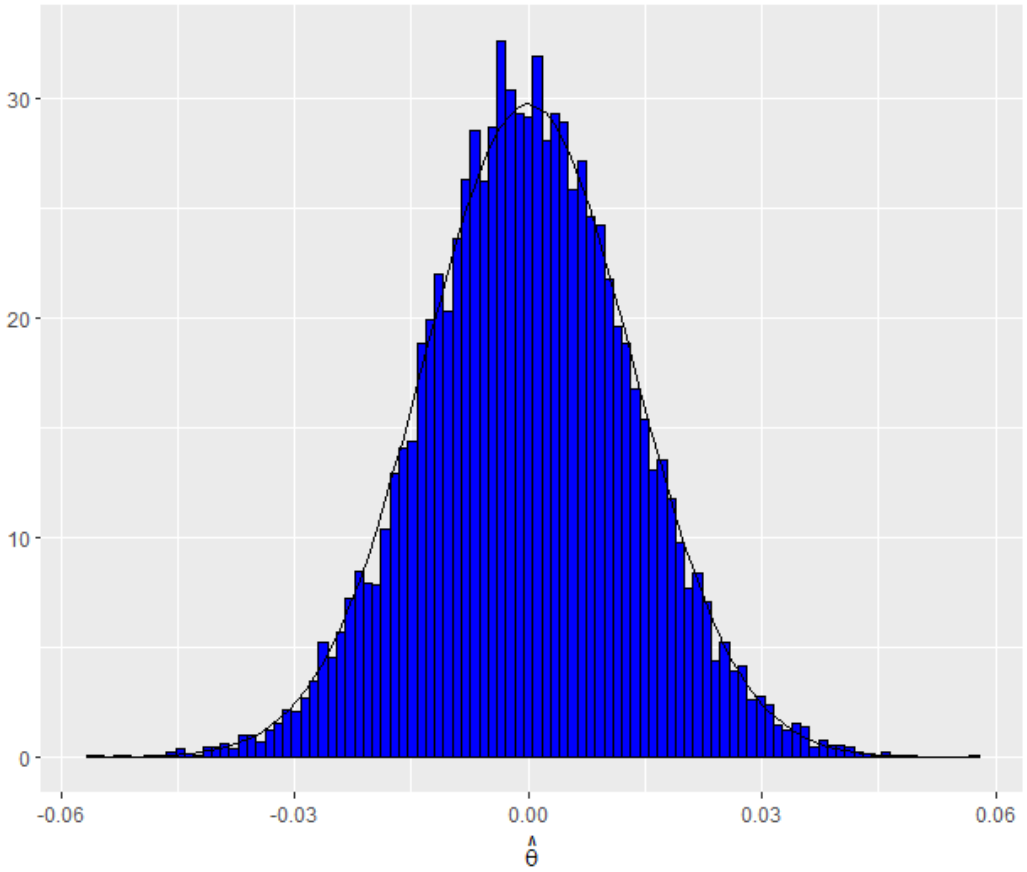


Figure 3.3: **Empirical distribution of the estimator.**

n	p	s	SNR	Estimator	Acceptance rate	MAE	rel. MSE
100	20	5	1.0	-0.00055821	0.928	0.0216	1.8060
100	20	5	3.0	-0.00109542	0.958	0.0209	1.2233
100	20	10	1.0	-0.00029139	0.938	0.0151	1.8626
100	20	10	3.0	0.00025676	0.952	0.0184	3.3238
100	20	20	1.0	-0.00004602	0.946	0.0103	1.7504
100	20	20	3.0	0.00073964	0.950	0.0131	3.5360
100	50	5	1.0	-0.00010804	0.940	0.0217	1.9122
100	50	5	3.0	-0.00140765	0.966	0.0205	1.5162
100	50	10	1.0	0.00011707	0.948	0.0148	1.8327
100	50	10	3.0	-0.00014981	0.970	0.0170	3.5025
100	50	20	1.0	-0.00039691	0.956	0.0101	1.8685
100	50	20	3.0	0.00015181	0.926	0.0131	3.7509
100	100	5	1.0	0.00004696	0.946	0.0208	1.6940
100	100	5	3.0	0.00018713	0.972	0.0209	1.6054
100	100	10	1.0	-0.00050597	0.938	0.0156	2.0206
100	100	10	3.0	0.00091064	0.940	0.0186	3.8155
100	100	20	1.0	0.00063239	0.934	0.0112	1.8887
100	100	20	3.0	-0.00047847	0.952	0.0130	3.6160
100	200	5	1.0	0.00126572	0.944	0.0222	1.8953
100	200	5	3.0	-0.00027602	0.974	0.0220	1.8385
100	200	10	1.0	0.00061581	0.954	0.0155	1.9005
100	200	10	3.0	0.00051767	0.966	0.0184	3.7384
100	200	20	1.0	-0.00161354	0.960	0.0112	1.7789
100	200	20	3.0	-0.00055394	0.944	0.0138	3.5126
200	20	5	1.0	0.00031100	0.938	0.0141	1.1520
200	20	5	3.0	-0.00059165	0.942	0.0133	0.9418
200	20	10	1.0	-0.00039975	0.926	0.0107	1.7806
200	20	10	3.0	-0.00027409	0.962	0.0098	1.2756
200	20	20	1.0	-0.00010337	0.934	0.0069	1.7440
200	20	20	3.0	0.00030134	0.936	0.0089	3.3897
200	50	5	1.0	-0.00072350	0.924	0.0140	1.3226
200	50	5	3.0	0.00023937	0.934	0.0130	0.9610
200	50	10	1.0	0.00037448	0.928	0.0107	1.7955
200	50	10	3.0	0.00001872	0.986	0.0100	1.5800
200	50	20	1.0	-0.00000121	0.936	0.0074	1.8593
200	50	20	3.0	-0.00101499	0.930	0.0095	3.7088
200	100	5	1.0	-0.00098745	0.918	0.0151	1.4290
200	100	5	3.0	-0.00147473	0.942	0.0128	1.0899
200	100	10	1.0	-0.00070476	0.918	0.0107	1.9961
200	100	10	3.0	0.00074576	0.964	0.0106	2.0574
200	100	20	1.0	0.00029025	0.924	0.0076	1.8810
200	100	20	3.0	0.00147962	0.950	0.0085	3.6083
200	200	5	1.0	-0.00099558	0.952	0.0133	1.5255
200	200	5	3.0	0.00039661	0.936	0.0139	0.9671
200	200	10	1.0	-0.00037153	0.920	0.0110	1.8903
200	200	10	3.0	0.00108176	0.958	0.0111	2.2986
200	200	20	1.0	-0.00025439	0.942	0.0072	1.7740
200	200	20	3.0	0.00016864	0.918	0.0095	3.4971
200	500	5	1.0	-0.00012946	0.942	0.0137	1.6904
200	500	5	3.0	0.00037737	0.946	0.0136	1.0846
200	500	10	1.0	-0.00103134	0.942	0.0104	1.9300
200	500	10	3.0	-0.00003818	0.958	0.0121	2.6936
200	500	20	1.0	-0.00004502	0.930	0.0078	2.1176
200	500	20	3.0	-0.00063815	0.946	0.0089	4.1060

Table 3.1: **Box-Cox: Simulation results for $\Sigma^{(X)} = I_p$.**

n	p	s	SNR	Estimator	Acceptance rate	MAE	rel. MSE
100	20	5	1.0	0.00035548	0.932	0.0161	1.2439
100	20	5	3.0	-0.00097029	0.956	0.0147	1.0356
100	20	10	1.0	-0.00046551	0.960	0.0098	1.6340
100	20	10	3.0	0.00092449	0.962	0.0107	1.5360
100	20	20	1.0	0.00001787	0.904	0.0080	1.7442
100	20	20	3.0	0.00025592	0.940	0.0094	2.7994
100	50	5	1.0	-0.00025392	0.942	0.0154	1.3759
100	50	5	3.0	0.00003743	0.960	0.0152	1.0898
100	50	10	1.0	-0.00032109	0.936	0.0111	1.7082
100	50	10	3.0	0.00047115	0.966	0.0112	1.7237
100	50	20	1.0	-0.00002715	0.934	0.0078	1.8586
100	50	20	3.0	-0.00019370	0.952	0.0093	3.2917
100	100	5	1.0	-0.00095235	0.938	0.0147	1.4640
100	100	5	3.0	-0.00019675	0.940	0.0159	1.2573
100	100	10	1.0	0.00029316	0.946	0.0108	1.9001
100	100	10	3.0	-0.00095316	0.964	0.0120	2.0610
100	100	20	1.0	-0.00013236	0.962	0.0075	1.8949
100	100	20	3.0	0.00021921	0.960	0.0093	3.3810
100	200	5	1.0	0.00093168	0.966	0.0150	1.4550
100	200	5	3.0	0.00031272	0.982	0.0142	1.1441
100	200	10	1.0	0.00056643	0.956	0.0111	1.8399
100	200	10	3.0	-0.00082101	0.974	0.0125	2.2858
100	200	20	1.0	-0.00038676	0.968	0.0073	1.7771
100	200	20	3.0	-0.00019953	0.966	0.0088	3.3579
200	20	5	1.0	-0.00030945	0.938	0.0102	0.9889
200	20	5	3.0	0.00070246	0.940	0.0102	0.9428
200	20	10	1.0	-0.00073690	0.940	0.0068	1.1929
200	20	10	3.0	0.00004800	0.946	0.0071	1.0059
200	20	20	1.0	-0.00044286	0.936	0.0052	1.5740
200	20	20	3.0	-0.00007465	0.972	0.0051	1.4465
200	50	5	1.0	0.00006431	0.944	0.0101	1.0327
200	50	5	3.0	0.00007992	0.932	0.0095	0.9615
200	50	10	1.0	0.00045421	0.948	0.0072	1.2865
200	50	10	3.0	0.00075009	0.930	0.0074	1.0532
200	50	20	1.0	-0.00021506	0.934	0.0052	1.7684
200	50	20	3.0	0.00004748	0.966	0.0053	1.7234
200	100	5	1.0	0.00022107	0.938	0.0103	1.1767
200	100	5	3.0	-0.00031888	0.944	0.0100	1.0891
200	100	10	1.0	0.00009239	0.928	0.0075	1.4996
200	100	10	3.0	0.00009438	0.942	0.0070	1.2155
200	100	20	1.0	-0.00029642	0.926	0.0053	1.8474
200	100	20	3.0	0.00051585	0.958	0.0054	2.0451
200	200	5	1.0	-0.00001408	0.946	0.0104	1.0535
200	200	5	3.0	-0.00036084	0.930	0.0102	0.9629
200	200	10	1.0	0.00034062	0.930	0.0074	1.4832
200	200	10	3.0	-0.00009682	0.962	0.0066	1.1283
200	200	20	1.0	-0.00004576	0.936	0.0053	1.7540
200	200	20	3.0	-0.00119200	0.962	0.0059	2.2391
200	500	5	1.0	0.00022032	0.948	0.0100	1.1980
200	500	5	3.0	-0.00096893	0.948	0.0104	1.0818
200	500	10	1.0	0.00042015	0.944	0.0077	1.6414
200	500	10	3.0	0.00000990	0.958	0.0072	1.2561
200	500	20	1.0	0.00056736	0.936	0.0053	2.1036
200	500	20	3.0	-0.00013055	0.952	0.0061	3.0273

Table 3.2: **Box-Cox: Simulation results for $\Sigma^{(X)} = \Sigma_1^{(X)}$.**

n	p	s	SNR	Estimator	Acceptance rate	MAE	rel. MSE
100	20	5	1.0	-0.00029259	0.952	0.0203	1.7828
100	20	5	3.0	-0.00092329	0.964	0.0196	1.1910
100	20	10	1.0	0.00012073	0.950	0.0145	1.8485
100	20	10	3.0	0.00241193	0.934	0.0186	3.2932
100	20	20	1.0	-0.00010180	0.920	0.0104	1.7404
100	20	20	3.0	-0.00013027	0.926	0.0132	3.4969
100	50	5	1.0	0.00058965	0.938	0.0209	1.9268
100	50	5	3.0	-0.00013828	0.968	0.0203	1.4251
100	50	10	1.0	-0.00141342	0.928	0.0157	1.8490
100	50	10	3.0	0.00082478	0.948	0.0189	3.5726
100	50	20	1.0	0.00048334	0.918	0.0111	1.8691
100	50	20	3.0	0.00132362	0.948	0.0126	3.7501
100	100	5	1.0	-0.00160281	0.938	0.0226	1.6711
100	100	5	3.0	0.00193019	0.978	0.0203	1.5980
100	100	10	1.0	0.00111299	0.960	0.0157	1.9941
100	100	10	3.0	0.00109876	0.956	0.0186	3.7599
100	100	20	1.0	-0.00159371	0.938	0.0111	1.8879
100	100	20	3.0	0.00029929	0.948	0.0132	3.6296
100	200	5	1.0	0.00110187	0.960	0.0212	1.9099
100	200	5	3.0	-0.00106521	0.984	0.0216	1.8807
100	200	10	1.0	0.00014970	0.966	0.0145	1.9117
100	200	10	3.0	-0.00030942	0.964	0.0196	3.7486
100	200	20	1.0	0.00014200	0.952	0.0108	1.7889
100	200	20	3.0	-0.00065953	0.972	0.0133	3.5190
200	20	5	1.0	-0.00066470	0.956	0.0122	1.1421
200	20	5	3.0	-0.00007279	0.942	0.0124	0.9439
200	20	10	1.0	0.00023762	0.924	0.0100	1.7593
200	20	10	3.0	0.00056502	0.958	0.0101	1.2129
200	20	20	1.0	0.00062338	0.924	0.0073	1.7263
200	20	20	3.0	0.00019359	0.956	0.0084	3.2262
200	50	5	1.0	0.00131883	0.946	0.0141	1.3133
200	50	5	3.0	0.00028543	0.946	0.0135	0.9588
200	50	10	1.0	-0.00068094	0.938	0.0107	1.8158
200	50	10	3.0	-0.00046929	0.952	0.0104	1.5506
200	50	20	1.0	-0.00047593	0.948	0.0072	1.8612
200	50	20	3.0	-0.00020800	0.932	0.0091	3.7182
200	100	5	1.0	-0.00080455	0.938	0.0138	1.3802
200	100	5	3.0	0.00006808	0.954	0.0129	1.0877
200	100	10	1.0	-0.00077886	0.928	0.0106	1.9660
200	100	10	3.0	0.00072632	0.968	0.0102	2.0464
200	100	20	1.0	-0.00003686	0.944	0.0073	1.8808
200	100	20	3.0	-0.00028688	0.950	0.0085	3.6061
200	200	5	1.0	-0.00010218	0.912	0.0142	1.5353
200	200	5	3.0	0.00138252	0.946	0.0136	0.9609
200	200	10	1.0	0.00000420	0.932	0.0101	1.8956
200	200	10	3.0	-0.00004370	0.962	0.0109	2.2782
200	200	20	1.0	-0.00028175	0.932	0.0075	1.7815
200	200	20	3.0	0.00030455	0.946	0.0085	3.5073
200	500	5	1.0	-0.00004556	0.930	0.0143	1.6798
200	500	5	3.0	0.00023905	0.940	0.0139	1.0843
200	500	10	1.0	0.00016880	0.932	0.0110	1.9023
200	500	10	3.0	-0.00135486	0.960	0.0118	2.7441
200	500	20	1.0	-0.00005315	0.936	0.0078	2.0996
200	500	20	3.0	-0.00009986	0.932	0.0093	4.0653

Table 3.3: **Box-Cox: Simulation results for $\Sigma^{(X)} = \Sigma_2^{(X)}$.**

3.4.2 Yeo-Johnson Power Transformations

Next, we consider the class of Yeo-Johnson power transformations. The Yeo-Johnson power transformations are defined as

$$\Lambda_\theta(y) = \begin{cases} \frac{(y+1)^\theta - 1}{\theta}, & \text{for } y \geq 0, \theta \neq 0 \\ \log(y+1), & \text{for } y \geq 0, \theta = 0 \\ -\frac{(-y+1)^{2-\theta} - 1}{2-\theta}, & \text{for } y < 0, \theta \neq 2 \\ -\log(-y+1), & \text{for } y < 0, \theta = 2. \end{cases}$$

We set the true transformation parameter $\theta_0 = 1$ and summarize the results in the Tables 3.4–3.6. We get similar patterns as under the Box-Cox transformation.
In summary, the empirical acceptance rate is close to the nominal level of 95% and the transformation parameter is estimated accurately.

n	p	s	SNR	Estimator	Acceptance rate	MAE	rel. MSE
100	20	5	1.0	0.99989085	0.956	0.0531	1.7955
100	20	5	3.0	0.99701139	0.956	0.0480	1.2454
100	20	10	1.0	0.99732572	0.946	0.0440	1.8621
100	20	10	3.0	1.00201985	0.954	0.0492	3.2990
100	20	20	1.0	1.00463863	0.932	0.0402	1.7570
100	20	20	3.0	0.99917208	0.940	0.0411	3.5338
100	50	5	1.0	1.00133561	0.942	0.0532	1.9071
100	50	5	3.0	0.99966997	0.982	0.0478	1.4648
100	50	10	1.0	1.00094266	0.940	0.0450	1.8360
100	50	10	3.0	1.00174621	0.960	0.0469	3.4738
100	50	20	1.0	0.99837549	0.964	0.0371	1.8729
100	50	20	3.0	0.99618151	0.938	0.0426	3.7640
100	100	5	1.0	1.00272122	0.956	0.0533	1.6976
100	100	5	3.0	0.99727378	0.954	0.0509	1.6225
100	100	10	1.0	1.00291725	0.944	0.0461	2.0220
100	100	10	3.0	1.00002603	0.942	0.0499	3.8505
100	100	20	1.0	1.00041509	0.952	0.0385	1.8904
100	100	20	3.0	1.00047398	0.958	0.0404	3.6226
100	200	5	1.0	0.99793573	0.968	0.0529	1.8893
100	200	5	3.0	0.99789085	0.974	0.0481	1.8679
100	200	10	1.0	1.00024902	0.962	0.0460	1.9035
100	200	10	3.0	1.00304432	0.952	0.0494	3.7504
100	200	20	1.0	0.99993355	0.948	0.0386	1.7806
100	200	20	3.0	0.99405454	0.948	0.0442	3.5148
200	20	5	1.0	1.00012449	0.938	0.0361	1.1689
200	20	5	3.0	1.00053857	0.936	0.0313	0.9433
200	20	10	1.0	1.00430230	0.944	0.0306	1.7735
200	20	10	3.0	0.99971274	0.960	0.0265	1.2521
200	20	20	1.0	1.00210897	0.936	0.0276	1.7453
200	20	20	3.0	0.99636422	0.936	0.0307	3.4005
200	50	5	1.0	0.99718193	0.954	0.0328	1.3104
200	50	5	3.0	1.00180263	0.942	0.0301	0.9625
200	50	10	1.0	0.99989134	0.930	0.0322	1.8054
200	50	10	3.0	1.00179570	0.962	0.0278	1.5387
200	50	20	1.0	0.99889235	0.948	0.0264	1.8607
200	50	20	3.0	1.00058817	0.942	0.0287	3.6990
200	100	5	1.0	0.99853711	0.960	0.0365	1.3969
200	100	5	3.0	0.99629466	0.952	0.0284	1.0922
200	100	10	1.0	1.00497857	0.946	0.0326	1.9988
200	100	10	3.0	0.99964084	0.960	0.0286	1.9631
200	100	20	1.0	0.99985907	0.938	0.0273	1.8807
200	100	20	3.0	0.99925090	0.952	0.0294	3.6156
200	200	5	1.0	1.00204386	0.952	0.0363	1.5089
200	200	5	3.0	1.00042547	0.946	0.0318	0.9710
200	200	10	1.0	1.00150067	0.956	0.0312	1.8897
200	200	10	3.0	1.00058871	0.972	0.0295	2.2308
200	200	20	1.0	1.00327436	0.948	0.0268	1.7748
200	200	20	3.0	1.00077426	0.954	0.0293	3.4977
200	500	5	1.0	1.00337876	0.934	0.0394	1.7058
200	500	5	3.0	1.00004203	0.946	0.0319	1.0865
200	500	10	1.0	0.99965877	0.948	0.0303	1.9287
200	500	10	3.0	0.99864655	0.962	0.0327	2.8189
200	500	20	1.0	1.00182442	0.950	0.0271	2.1191
200	500	20	3.0	1.00092662	0.948	0.0297	4.1020

Table 3.4: **Yeo-Johnson: Simulation results for $\Sigma^{(X)} = I_p$.**

n	p	s	SNR	Estimator	Acceptance rate	MAE	rel. MSE
100	20	5	1.0	0.99619347	0.924	0.0463	1.2505
100	20	5	3.0	0.99833394	0.960	0.0371	1.0300
100	20	10	1.0	0.99814415	0.946	0.0403	1.6382
100	20	10	3.0	0.99896106	0.980	0.0346	1.5484
100	20	20	1.0	1.00332660	0.946	0.0321	1.7477
100	20	20	3.0	1.00242808	0.952	0.0345	2.8320
100	50	5	1.0	1.00706413	0.950	0.0448	1.3922
100	50	5	3.0	0.99862564	0.950	0.0374	1.0930
100	50	10	1.0	0.99785945	0.952	0.0400	1.6867
100	50	10	3.0	1.00405369	0.970	0.0361	1.7500
100	50	20	1.0	1.00049886	0.942	0.0333	1.8622
100	50	20	3.0	1.00275828	0.954	0.0354	3.3531
100	100	5	1.0	0.99964792	0.958	0.0436	1.4512
100	100	5	3.0	1.00018733	0.970	0.0378	1.2533
100	100	10	1.0	0.99887052	0.970	0.0365	1.9048
100	100	10	3.0	0.99914381	0.976	0.0365	2.1079
100	100	20	1.0	1.00325414	0.940	0.0341	1.8941
100	100	20	3.0	0.99795875	0.968	0.0348	3.4283
100	200	5	1.0	0.99980596	0.952	0.0443	1.5039
100	200	5	3.0	1.00318283	0.966	0.0385	1.1348
100	200	10	1.0	0.99710813	0.966	0.0376	1.8528
100	200	10	3.0	0.99990709	0.966	0.0375	2.2494
100	200	20	1.0	0.99974527	0.960	0.0328	1.7787
100	200	20	3.0	0.99885191	0.972	0.0378	3.3678
200	20	5	1.0	0.99989759	0.934	0.0296	0.9913
200	20	5	3.0	0.99927178	0.938	0.0266	0.9426
200	20	10	1.0	0.99825191	0.946	0.0263	1.1789
200	20	10	3.0	1.00157433	0.958	0.0224	1.0135
200	20	20	1.0	0.99910827	0.960	0.0222	1.5879
200	20	20	3.0	1.00413773	0.972	0.0214	1.4732
200	50	5	1.0	0.99924325	0.954	0.0279	1.0390
200	50	5	3.0	0.99789757	0.938	0.0274	0.9633
200	50	10	1.0	0.99936383	0.934	0.0255	1.2709
200	50	10	3.0	1.00053949	0.946	0.0230	1.0500
200	50	20	1.0	0.99867395	0.940	0.0231	1.7741
200	50	20	3.0	0.99961880	0.960	0.0218	1.7622
200	100	5	1.0	1.00275165	0.948	0.0286	1.1760
200	100	5	3.0	1.00077422	0.938	0.0268	1.0959
200	100	10	1.0	0.99893851	0.946	0.0259	1.4928
200	100	10	3.0	0.99928876	0.954	0.0215	1.2266
200	100	20	1.0	1.00268155	0.950	0.0232	1.8406
200	100	20	3.0	1.00002885	0.936	0.0240	2.0552
200	200	5	1.0	0.99787872	0.946	0.0300	1.0546
200	200	5	3.0	1.00001108	0.948	0.0254	0.9609
200	200	10	1.0	1.00056662	0.952	0.0254	1.4767
200	200	10	3.0	0.99799385	0.954	0.0221	1.1000
200	200	20	1.0	1.00192636	0.950	0.0225	1.7554
200	200	20	3.0	1.00334912	0.944	0.0245	2.2427
200	500	5	1.0	0.99802326	0.944	0.0285	1.1891
200	500	5	3.0	1.00012563	0.950	0.0249	1.0809
200	500	10	1.0	0.99925335	0.942	0.0266	1.6549
200	500	10	3.0	1.00042866	0.968	0.0231	1.2702
200	500	20	1.0	1.00126983	0.956	0.0230	2.0999
200	500	20	3.0	0.99924659	0.966	0.0240	3.0074

Table 3.5: **Yeo-Johnson: Simulation results for $\Sigma^{(X)} = \Sigma_1^{(X)}$.**

n	p	s	SNR	Estimator	Acceptance rate	MAE	rel. MSE
100	20	5	1.0	1.00367279	0.940	0.0537	1.7783
100	20	5	3.0	1.00025425	0.962	0.0434	1.1977
100	20	10	1.0	1.00152401	0.928	0.0453	1.8476
100	20	10	3.0	0.99637531	0.940	0.0500	3.2971
100	20	20	1.0	0.99787166	0.950	0.0373	1.7399
100	20	20	3.0	1.00308084	0.958	0.0402	3.4867
100	50	5	1.0	1.00006547	0.956	0.0517	1.9355
100	50	5	3.0	0.99994812	0.970	0.0440	1.4556
100	50	10	1.0	1.00093756	0.952	0.0438	1.8509
100	50	10	3.0	0.99763201	0.942	0.0477	3.5271
100	50	20	1.0	0.99963396	0.952	0.0381	1.8688
100	50	20	3.0	1.00081743	0.946	0.0409	3.7607
100	100	5	1.0	0.99656704	0.948	0.0540	1.6708
100	100	5	3.0	1.00272331	0.980	0.0490	1.6782
100	100	10	1.0	1.00153317	0.942	0.0463	1.9951
100	100	10	3.0	1.00006603	0.948	0.0501	3.7814
100	100	20	1.0	0.99852324	0.944	0.0358	1.8897
100	100	20	3.0	1.00049696	0.958	0.0427	3.6400
100	200	5	1.0	1.00081946	0.964	0.0536	1.9168
100	200	5	3.0	0.99650061	0.970	0.0533	1.9694
100	200	10	1.0	0.99879653	0.964	0.0438	1.9138
100	200	10	3.0	1.00063372	0.964	0.0477	3.7520
100	200	20	1.0	0.99787393	0.960	0.0385	1.7897
100	200	20	3.0	0.99913458	0.966	0.0414	3.5283
200	20	5	1.0	1.00265436	0.954	0.0326	1.1445
200	20	5	3.0	1.00058134	0.954	0.0301	0.9432
200	20	10	1.0	1.00125043	0.938	0.0321	1.7606
200	20	10	3.0	0.99988890	0.960	0.0257	1.2246
200	20	20	1.0	1.00160783	0.940	0.0264	1.7254
200	20	20	3.0	0.99932267	0.964	0.0274	3.1870
200	50	5	1.0	1.00079590	0.954	0.0341	1.3372
200	50	5	3.0	0.99866985	0.926	0.0324	0.9649
200	50	10	1.0	1.00102807	0.954	0.0310	1.8102
200	50	10	3.0	1.00030919	0.954	0.0283	1.6297
200	50	20	1.0	1.00095764	0.952	0.0270	1.8610
200	50	20	3.0	0.99814638	0.944	0.0300	3.7048
200	100	5	1.0	0.99798485	0.930	0.0373	1.3546
200	100	5	3.0	1.00223459	0.950	0.0294	1.0932
200	100	10	1.0	0.99968150	0.940	0.0318	1.9679
200	100	10	3.0	0.99912236	0.974	0.0286	2.0782
200	100	20	1.0	1.00017655	0.954	0.0258	1.8803
200	100	20	3.0	0.99963970	0.936	0.0296	3.6140
200	200	5	1.0	0.99632996	0.950	0.0345	1.4789
200	200	5	3.0	0.99986194	0.954	0.0303	0.9649
200	200	10	1.0	1.00113110	0.932	0.0320	1.8961
200	200	10	3.0	1.00199368	0.968	0.0294	2.2247
200	200	20	1.0	1.00122702	0.950	0.0275	1.7819
200	200	20	3.0	1.00000603	0.956	0.0287	3.5083
200	500	5	1.0	1.00762721	0.952	0.0384	1.7089
200	500	5	3.0	1.00006396	0.966	0.0298	1.0830
200	500	10	1.0	0.99891396	0.942	0.0318	1.9029
200	500	10	3.0	1.00001533	0.956	0.0327	2.7849
200	500	20	1.0	1.00069240	0.942	0.0267	2.0994
200	500	20	3.0	1.00125883	0.932	0.0300	4.0673

Table 3.6: **Yeo-Johnson: Simulation results for $\Sigma^{(X)} = \Sigma_2^{(X)}$.**

3.5 Application

3.5.1 Econometric Specification of the Wage Equation

In labor economics, the analysis of wage data is key. In addition, labor economics aims to identify the determinants of wages to estimate a so-called Mincer equation and to evaluate the impact of labor market programs on wages. Wages are non-negative and show a high degree of skewness which is not compatible with a normal distribution. Hence, wages are transformed in almost all studies by the logarithm. Figure 3.4 shows the weekly wage distribution (in US dollar) from the US survey data which are described in the next section.

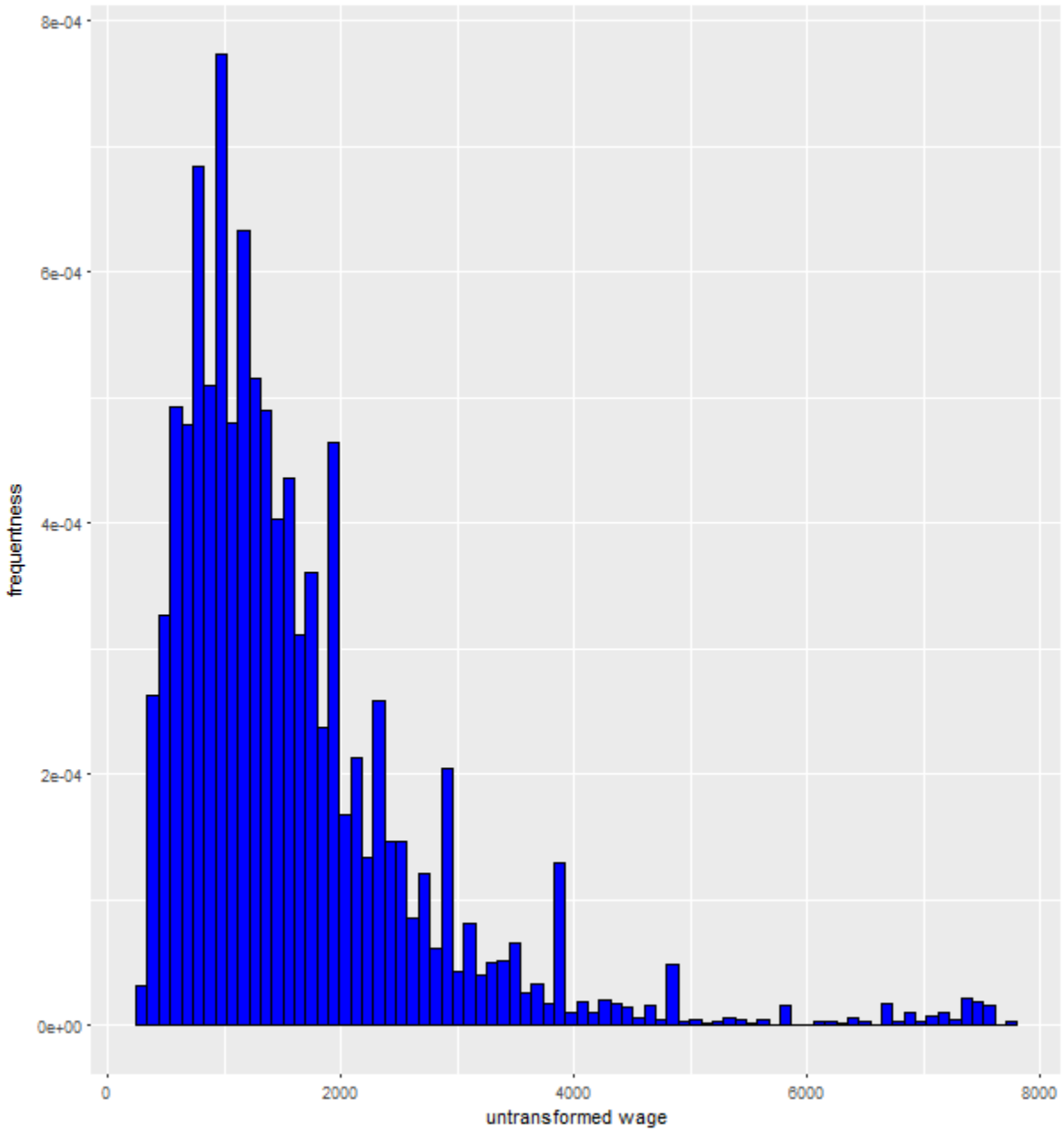


Figure 3.4: **Empirical wage distribution from the US survey data.**

Here, we focus on the estimation of a Mincer type equation. The Mincer equation formulates a relationship between log wages (W) and schooling (S), experience (Exp), and other control variables (X, p-dimensional):

$$\log W = \alpha + \beta S + \gamma Exp + \delta Exp^2 + \mu' X + \varepsilon \tag{3.18}$$

with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. $\alpha, \beta, \gamma, \delta$ are coefficients and μ is a p-dimensional vector of the coefficients of the control variables. We consider a high-dimensional setting where the set of potential control variables is high and we do not take it for granted that the log transformation is appropriate but instead we estimate a transformation model and test for the transformation parameter. The model is given by

$$\Lambda_{\theta_0}(W) = \alpha_{\theta_0} + \beta S_{\theta_0} + \gamma Exp_{\theta_0} + \delta Exp^2_{\theta_0} + \mu'_{\theta_0} X + \varepsilon_{\theta_0} \tag{3.19}$$

with $\varepsilon_{\theta_0} \sim \mathcal{N}(0, \sigma^2)$.

3.5.2 Data Set

Overview

In our empirical study, we use data from the 2015 American Community Survey (ACS) that is provided by Ruggles et al. [87] and extracted from the IPUMS-USA website[1]. The ACS provides a 1%-sample of the US population with mandatory participation. The data offers a large number of socio-economic characteristics at the individual and household level, such as education, industry, occupation, and earnings. We restrict our attention to individuals who graduated from university and are working full time (30+ hours), at least 50 weeks a year. Weekly earnings are computed as annual earnings divided by 52 (weeks). We exclude individuals with experience > 60 and age > 65. Moreover, we discard individuals with a weekly wage of less than \$10 (which is likely to be unreasonable given that we only consider full-time employees). Then, we drop all observations with a weekly wage under the 2.5%-quantile and over the 97.5%-quantile. Our final sample comprises $315,291$ individual observations.
In our analysis, we use 14 initial regressors which are either directly available from the ACS data or have been constructed. We list the variables in Table 3.7. Mostly, we use the categories as provided in the ACS data. This might be particularly informative for the region, occupation, and industry variables for which different definitions exist. Moreover, we construct the variables "years of education" and "labor market experience" from the information available. We construct all of the two-way interactions of the initial regressors, where the categorical variables are transformed to level-wise dummies. Additionally, we include the variable "field of degree" to account for the individual's educational background. Finally, we drop all of the constructed variables that are nearly constant over all observations and we end with a high-dimensional setting with a total of $1,743$ regressors.

Descriptive Statistics

Table 3.8 provides the summary statistics for a selection of the variables that are available in our final sample from the ACS data. Figure 3.4 and the following descriptive statistics illustrate that the mean of weekly wage for university graduates is higher than the median; hence, we have skewed data. Weekly wage is characterized by non-negativity and a high variability.

[1] `https://usa.ipums.org/usa/`

Variable	Type	Baseline Category
Female	binary	
Marital status	six categories	never married, single
Race	four categories	White
English language skills	five categories	speaks only English
Hispanic	binary	
Veteran Status	binary	
Industry	14 categories	wholesale trade
Occupation	26 categories	management, science, arts
Region (US census)	nine categories	New England division
Experience (years)	continuous	
Experience squared	continuous	
Years of Education	continuous	
Family Size	continuous	
Number of own young children	continuous	
Field of degree	37 categories	administration, teaching

Table 3.7: **Application: List of regressors.**

Variable	Mean	SD	Median
Weekly wage	1591.22	1100.20	1307.69
Experience (years)	20.75	11.36	21
Years of Education	16.91	1.24	17
Female	0.48	0.50	-
White	0.84	0.37	-
Black/Negro	0.07	0.25	-
Chinese	0.02	0.15	-
Hispanic	0.05	0.23	-
Veteran Status	0.05	0.21	-
Sample Size	315291		

Table 3.8: **Summary statistics, ACS data.**

3.5.3 Results

The estimated transformation parameter is $\hat{\theta} = -0.1260646$. The confidence interval is based on the asymptotic normality of the estimate $\hat{\theta}$ and the variance is estimated via 300 bootstrap samples. Since the confidence interval $[-0.1307524, -0.1213768]$ does not include 0, we can reject the null hypothesis $\theta = 0$ on a 5% significance level which is equivalent to a log transformation. In Figure 3.5, we compare the Q-Q plot of the untransformed wages with the Q-Q plot of the transformed wages with our estimated parameter (with a normal distribution determined by the sample mean and sample variance) and with the Q-Q plot under the log transformation ($\theta = 0$). The estimated errors without transformation are not normally distributed, whereas after the transformation of the response variable with $\hat{\theta}$ the estimated error terms seem to fit a normal distribution quite well. Considering the transformation function, one can recognize that for a transformation parameter below zero the transformation has a stronger curvature (see figure 3.6). This implies that the wages are more positively skewed towards normal distribution after a log transformation. Although we reject the log transformation, Figure 3.5 reveals that the log

transformation might give a reasonable approximation for applications in labor economics.

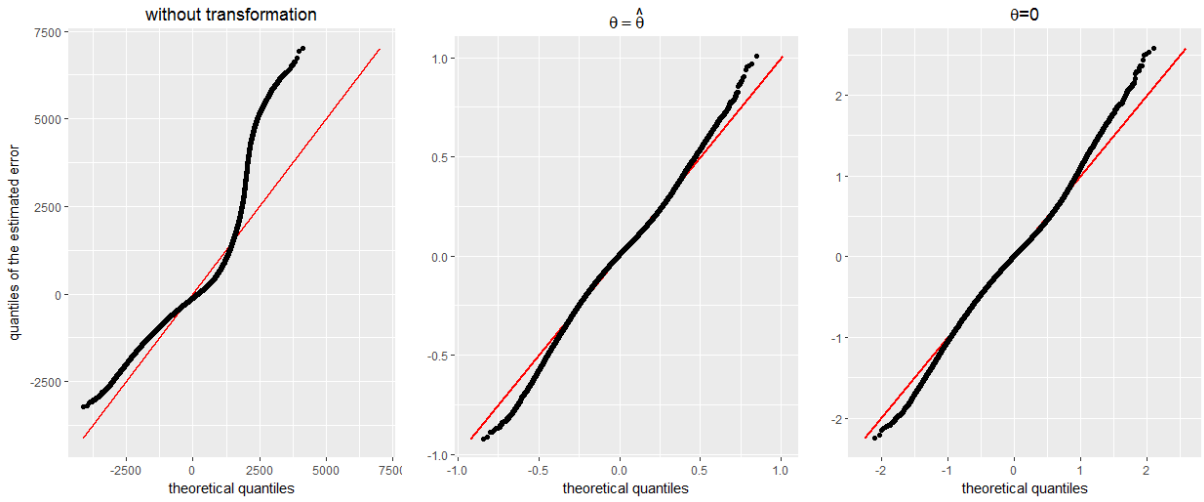


Figure 3.5: **Comparison of Q-Q plots.**

Figure 3.6: **Transformation functions for $\theta = 0$ (black) and $\theta = \hat{\theta}$ (red).**

3.6 Conclusion

In this paper, we propose an estimator for the transformation parameter in a high-dimensional setting. Transformation models, in particular the Box-Cox and Yeo-Johnson transformation, are very popular in applied statistics and econometrics. The rise of digitalization has led to an increased availability of high-dimensional data sets and, hence, make it necessary to extend models for this setting when the number of variables p is large (or even larger) compared to the sample size n. We build on the recent results on the Neyman orthogonality condition to prove the asymptotic normality of our estimator. The nuisance functions are estimated with Lasso.

Our setting fits into a general Z-estimation problem with a high-dimensional nuisance function, which depends on the target parameter θ. We extend the results in Belloni et al. [9] and Chernozhukov et al. [35] to allow for an explicit dependency of the nuisance function on the target parameter θ. This result might be of interest for Z-estimation problems with the same structure.

In labor economics, wage is by default transformed with the logarithm. In our application, by analyzing US survey data, we are able to show that the log transformation is rejected on 5% significance level but the log transformation might give an appropriate approximation.

In future research, we would like to address the problem of estimation and inference on elements of the coefficient vector of the regressors.

Appendix

3.7 Proofs

Proof of Lemma 2.
We refer to Section 3.2.3 for the notation. Let $h = (h_1, h_2, h_3, h_4) \in \mathcal{H}'$ be arbitrary. First, we consider

$$\begin{aligned}&\partial_r\Big(\Lambda_{\theta_0}(Y) - \big(m_{\theta_0}(X) + r\big(h_1(\theta_0, X) - m_{\theta_0}(X)\big)\big)\Big)\Big|_{r=0}\\ &= m_{\theta_0}(X) - h_1(\theta_0, X)\end{aligned}$$

and analogous

$$\begin{aligned}&\partial_r\Big(\dot\Lambda_{\theta_0}(Y) - \big(\dot m_{\theta_0}(X) + r\big(h_3(\theta_0, X) - \dot m_{\theta_0}(X)\big)\big)\Big)\Big|_{r=0}\\ &= \dot m_{\theta_0}(X) - h_3(\theta_0, X).\end{aligned}$$

Additionally, we have

$$\partial_r\left(\left(\sigma^2_{\theta_0} + r(h_2(\theta_0) - \sigma^2_{\theta_0})\right)^{-1}\right)\Big|_{r=0} = -\frac{h_2(\theta_0) - \sigma^2_{\theta_0}}{\left(\sigma^2_{\theta_0}\right)^2}$$

and

$$\partial_r\left(\dot\sigma^2_{\theta_0} + r(h_4(\theta_0) - \dot\sigma^2_{\theta_0})\right)\Big|_{r=0} = h_4(\theta_0) - \dot\sigma^2_{\theta_0}.$$

By the product rule, we obtain

$$\begin{aligned}&\mathbb{E}\left[\partial_r I(\theta_0, \sigma^2_\theta + r(h_2 - \sigma^2_\theta), \dot\sigma^2_\theta + r(h_4 - \dot\sigma^2_\theta))|_{r=0}|X\right]\\ &= \mathbb{E}\left[\frac{h_4(\theta_0) - \dot\sigma^2_{\theta_0}}{2\sigma^2_{\theta_0}} - \dot\sigma^2_{\theta_0}\frac{h_2(\theta_0) - \sigma^2_{\theta_0}}{2\left(\sigma^2_{\theta_0}\right)^2}\Big|X\right]\\ &= \frac{h_4(\theta_0) - \dot\sigma^2_{\theta_0}}{2\sigma^2_{\theta_0}} - \dot\sigma^2_{\theta_0}\frac{h_2(\theta_0) - \sigma^2_{\theta_0}}{2\left(\sigma^2_{\theta_0}\right)^2},\end{aligned}$$

$$\begin{aligned}&\mathbb{E}\left[\partial_r II(\theta_0, m_\theta + r(h_1 - m_\theta), \sigma^2_\theta + r(h_2 - \sigma^2_\theta), \dot m_\theta + r(h_1 - \dot m_\theta))|_{r=0}|X\right]\\ =&\mathbb{E}\left[-\frac{h_2(\theta_0) - \sigma^2_{\theta_0}}{\left(\sigma^2_{\theta_0}\right)^2}\big(\Lambda_{\theta_0}(Y) - m_{\theta_0}(X)\big)\big(\dot\Lambda_{\theta_0}(Y) - \dot m_{\theta_0}(X)\big)\Big|X\right]\\ &+\mathbb{E}\left[\frac{1}{\sigma^2_{\theta_0}}\big(m_{\theta_0}(X) - h_1(\theta_0, X)\big)\big(\dot\Lambda_{\theta_0}(Y) - \dot m_{\theta_0}(X)\big)\Big|X\right]\\ &+\mathbb{E}\left[\frac{1}{\sigma^2_{\theta_0}}\big(\Lambda_{\theta_0}(Y) - m_{\theta_0}(X)\big)\big(\dot m_{\theta_0}(X) - h_3(\theta_0, X)\big)\Big|X\right]\\ =&-\frac{h_2(\theta_0) - \sigma^2_{\theta_0}}{\left(\sigma^2_{\theta_0}\right)^2}\mathbb{E}\left[\varepsilon_{\theta_0}\dot\varepsilon_{\theta_0}\big|X\right] + \frac{m_{\theta_0}(X) - h_1(\theta_0, X)}{\sigma^2_{\theta_0}}\underbrace{\mathbb{E}\left[\dot\varepsilon_{\theta_0}\big|X\right]}_{=0}\\ &+\frac{\dot m_{\theta_0}(X) - h_3(\theta_0, X)}{\sigma^2_{\theta_0}}\underbrace{\mathbb{E}\left[\varepsilon_{\theta_0}\big|X\right]}_{=0}\\ =&-\frac{h_2(\theta_0) - \sigma^2_{\theta_0}}{\left(\sigma^2_{\theta_0}\right)^2}\mathbb{E}\left[\varepsilon_{\theta_0}\dot\varepsilon_{\theta_0}\big|X\right],\end{aligned}$$

and

$$
\begin{aligned}
&\mathbb{E}\left[\partial_r III(\theta_0, m_\theta + r(h_1 - m_\theta), \sigma_\theta^2 + r(h_2 - \sigma_\theta^2), \dot\sigma_\theta^2 + r(h_4 - \dot\sigma_\theta^2))|_{r=0}|X\right] \\
=&\mathbb{E}\left[\frac{h_4(\theta_0) - \dot\sigma_{\theta_0}^2}{2\left(\sigma_{\theta_0}^2\right)^2}\left(\Lambda_{\theta_0}(Y) - m_{\theta_0}(X)\right)^2\Big|X\right] \\
&- \mathbb{E}\left[\dot\sigma_{\theta_0}^2\frac{h_2(\theta_0) - \sigma_{\theta_0}^2}{\left(\sigma_{\theta_0}^2\right)^3}\left(\Lambda_{\theta_0}(Y) - m_{\theta_0}(X)\right)^2\Big|X\right] \\
&+ \mathbb{E}\left[\frac{\dot\sigma_{\theta_0}^2}{\left(\sigma_{\theta_0}^2\right)^2}\left(\Lambda_{\theta_0}(Y) - m_{\theta_0}(X)\right)\left(m_{\theta_0}(X) - h_1(\theta_0, X)\right)\Big|X\right] \\
=&\frac{h_4(\theta_0) - \dot\sigma_{\theta_0}^2}{2\left(\sigma_{\theta_0}^2\right)^2}\underbrace{\mathbb{E}\left[\left(\Lambda_{\theta_0}(Y) - m_{\theta_0}(X)\right)^2|X\right]}_{=\sigma_{\theta_0}^2} \\
&- \dot\sigma_{\theta_0}^2\frac{h_2(\theta_0) - \sigma_{\theta_0}^2}{\left(\sigma_{\theta_0}^2\right)^3}\underbrace{\mathbb{E}\left[\left(\Lambda_{\theta_0}(Y) - m_{\theta_0}(X)\right)^2|X\right]}_{=\sigma_{\theta_0}^2} \\
&+ \frac{\dot\sigma_{\theta_0}^2}{\left(\sigma_{\theta_0}^2\right)^2}\left(m_{\theta_0}(X) - h_1(\theta_0, X)\right)\underbrace{\mathbb{E}\left[\varepsilon_{\theta_0}|X\right]}_{=0} \\
=&\frac{h_4(\theta_0) - \dot\sigma_{\theta_0}^2}{2\sigma_{\theta_0}^2} - \dot\sigma_{\theta_0}^2\frac{h_2(\theta_0) - \sigma_{\theta_0}^2}{\left(\sigma_{\theta_0}^2\right)^2}.
\end{aligned}
$$

The conditions enable us to change derivation and integration, hence we obtain

$$
\begin{aligned}
&D_0[h - h_0] \\
&= \partial_r\Big\{\mathbb{E}\Big[\psi\Big((Y,X), \theta_0, h_0 + r(h - h_0)\Big)\Big]\Big\}\Big|_{r=0} \\
&= \mathbb{E}\Big[\partial_r\psi\Big((Y,X), \theta_0, h_0 + r(h - h_0)\Big)\Big|_{r=0}\Big] \\
&= \mathbb{E}\Big[\mathbb{E}\Big[\partial_r\psi\Big((Y,X), \theta_0, h_0 + r(h - h_0)\Big)\Big|_{r=0}\Big|X\Big]\Big] \\
&= \mathbb{E}\Big[- \mathbb{E}\Big[\partial_r I(\theta_0, \sigma_\theta^2 + r(h_2 - \sigma_\theta^2), \dot\sigma_\theta^2 + r(h_4 - \dot\sigma_\theta^2))|_{r=0}|X\Big] \\
&\quad - \mathbb{E}\Big[\partial_r II(\theta_0, m_\theta + r(h_1 - m_\theta), \sigma_\theta^2 + r(h_2 - \sigma_\theta^2), \dot m_\theta \\
&\quad + r(h_1 - \dot m_\theta))|_{r=0}|X\Big] + \mathbb{E}\Big[\partial_r III(\theta_0, m_\theta + r(h_1 - m_\theta), \sigma_\theta^2 + \\
&\quad r(h_2 - \sigma_\theta^2), \dot\sigma_\theta^2 + r(h_4 - \dot\sigma_\theta^2))|_{r=0}|X\Big] + \underbrace{\partial_r c_{\theta_0}|_{r=0}}_{=0}\Big] \\
&= \mathbb{E}\Big[- \frac{h_4(\theta_0) - \dot\sigma_{\theta_0}^2}{2\sigma_{\theta_0}^2} + \dot\sigma_{\theta_0}^2\frac{h_2(\theta_0) - \sigma_{\theta_0}^2}{2\left(\sigma_{\theta_0}^2\right)^2} + \frac{h_2(\theta_0) - \sigma_{\theta_0}^2}{\left(\sigma_{\theta_0}^2\right)^2}\mathbb{E}\left[\varepsilon_{\theta_0}\dot\varepsilon_{\theta_0}|X\right] \\
&\quad + \frac{h_4(\theta_0) - \dot\sigma_{\theta_0}^2}{2\sigma_{\theta_0}^2} - \dot\sigma_{\theta_0}^2\frac{h_2(\theta_0) - \sigma_{\theta_0}^2}{\left(\sigma_{\theta_0}^2\right)^2}\Big] \\
&= 0,
\end{aligned}
$$

where we used $\dot\sigma_{\theta_0}^2 = 2\mathbb{E}[\varepsilon_{\theta_0}\dot\varepsilon_{\theta_0}]$ in the last step. □

Proof of Theorem 3.
Assumptions A1-A7 directly imply the conditions in Theorem 7 for the model (3.4) and (3.5) except for B7. We need to show that the empirical eigenvalues converge to the restricted sparse eigenvalues defined in A7. By Lemma P.1 in Belloni et al. [12], we have

$$\begin{aligned}&\mathbb{E}\left[\sup_{||\delta||_0 \leq s\log(n), ||\delta||=1} \left| \|X^T\delta\|_{P_n,2}^2 - \|X^T\delta\|_{P,2}^2 \right| \right] \\ &\leq C\left(\frac{s\log^2(n)\log(p)}{n} + \sqrt{\frac{s\log^2(n)\log(p)}{n}}\kappa'' \right) \leq C\sqrt{\frac{s\log^2(n)\log(p)}{n}}\end{aligned}$$

and using Markov's inequality we obtain

$$\sup_{||\delta||_0 \leq s\log(n), ||\delta||=1} \left| \|X^T\delta\|_{P_n,2}^2 - \|X^T\delta\|_{P,2}^2 \right| = o(1)$$

with probability $1 - o(1)$. This implies condition B7 for n large enough since the restricted sparse eigenvalues are bounded away from zero and above. □

Proof of Theorem 4.
As shown in the proof to Theorem 7, we have

$$\sup_{\theta\in\Theta} \left| \frac{1}{n}\sum_{i=1}^n \left(\varepsilon_{\theta,i}^2 - \mathbb{E}[\varepsilon_\theta^2] \right) \right| = O(\log(n)n^{-1/2}),$$

$$\sup_{\theta\in\Theta} \left| \frac{1}{n}\sum_{i=1}^n \left(\dot{\varepsilon}_{\theta,i}^2 - \mathbb{E}[\dot{\varepsilon}_\theta^2] \right) \right| = O(\log(n)n^{-1/2})$$

and with an analogous argument

$$\sup_{\theta\in\Theta} \left| \frac{1}{n}\sum_{i=1}^n \left(\varepsilon_{\theta,i}\dot{\varepsilon}_{\theta,i} - \mathbb{E}[\varepsilon_\theta\dot{\varepsilon}_\theta] \right) \right| = O(\log(n)n^{-1/2})$$

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

$$\begin{aligned}&\sup_{\theta\in\Theta} |\hat{\sigma}_\theta^2 - \sigma_\theta^2| \\ &= \sup_{\theta\in\Theta} \left| \frac{1}{n}\sum_{i=1}^n \Big(\underbrace{\Lambda_\theta(Y_i) - X_i^T\hat{\beta}_\theta}_{=\varepsilon_{\theta,i} - X_i^T\left(\hat{\beta}_\theta - \beta_\theta\right)} \Big)^2 - \mathbb{E}[\varepsilon_\theta^2] \right| \\ &= \sup_{\theta\in\Theta} \left| \frac{1}{n}\sum_{i=1}^n \left(\varepsilon_{\theta,i}^2 - \mathbb{E}[\varepsilon_\theta^2] \right) - \frac{2}{n}\sum_{i=1}^n \varepsilon_{\theta,i} X_i^T\left(\hat{\beta}_\theta - \beta_\theta\right) \right. \\ &\left. + \underbrace{\frac{1}{n}\sum_{i=1}^n \left(X_i^T\left(\hat{\beta}_\theta - \beta_\theta\right) \right)^2}_{=||X^T(\hat{\beta}_\theta - \beta_\theta)||_{\mathbb{P}_n,2}^2} \right| \\ &\leq \underbrace{\sup_{\theta\in\Theta} \left| \frac{1}{n}\sum_{i=1}^n \left(\varepsilon_{\theta,i}^2 - \mathbb{E}[\varepsilon_\theta^2] \right) \right|}_{=O(\log(n)n^{-1/2})} + \underbrace{\sup_{\theta\in\Theta} \left| \frac{2}{n}\sum_{i=1}^n \varepsilon_{\theta,i} X_i^T\left(\hat{\beta}_\theta - \beta_\theta\right) \right|}_{\leq 2\sup_{\theta\in\Theta} \left| ||X^T(\hat{\beta}_\theta - \beta_\theta)||_{\mathbb{P}_n,2} \sqrt{\frac{1}{n}\sum_{i=1}^n \varepsilon_{\theta,i}^2} \right|}\end{aligned}$$

$$
\begin{aligned}
&+ \underbrace{\sup_{\theta\in\Theta}\Big|||X^T(\hat\beta_\theta-\beta_\theta)||^2_{\mathbb{P}_{n,2}}\Big|}_{=O\left(\frac{s\log(p\vee n)}{n}\right)}\\
&\le 2\underbrace{\sup_{\theta\in\Theta}||X^T(\hat\beta_\theta-\beta_\theta)||_{\mathbb{P}_{n,2}}}_{=O\left(\sqrt{\frac{s\log(p\vee n)}{n}}\right)}\underbrace{\sup_{\theta\in\Theta}\sqrt{\frac{1}{n}\sum_{i=1}^n\varepsilon^2_{\theta,i}}}_{=O(1)}+O\Big(\frac{s\log(p\vee n)}{n}\Big)+O(\log(n)n^{-1/2})\\
&= O\left(\max\left(\sqrt{\frac{s\log(p\vee n)}{n}},\frac{\log(n)}{n^{1/2}}\right)\right)\le\tilde\delta_n n^{-\frac14}
\end{aligned}
$$

for a suitable sequence $\tilde\delta_n\searrow 0$, due to the growth condition A2. By the same argument, we obtain with probability $1-o(1)$

$$
\begin{aligned}
&\sup_{\theta\in\Theta}|\hat{\dot\sigma}^2_\theta-\dot\sigma^2_\theta|\\
&=\sup_{\theta\in\Theta}\Big|\frac{2}{n}\sum_{i=1}^n\Big(\underbrace{\Lambda_\theta(Y_i)-\hat m_\theta(X_i)}_{=\varepsilon_{\theta,i}-\left(\hat m_\theta(X_i)-m_\theta(X_i)\right)}\Big)\Big(\underbrace{\Lambda_\theta(Y_i)-\hat{\dot m}_\theta(X_i)}_{=\dot\varepsilon_{\theta,i}-\left(\hat{\dot m}_\theta(X_i)-\dot m_\theta(X_i)\right)}\Big)\\
&\quad-2\mathbb{E}[\varepsilon_\theta\dot\varepsilon_\theta]\Big|\\
&=2\sup_{\theta\in\Theta}\Big|\frac1n\sum_{i=1}^n\big(\varepsilon_{\theta,i}\dot\varepsilon_{\theta,i}-\mathbb{E}[\varepsilon_\theta\dot\varepsilon_\theta]\big)-\frac1n\sum_{i=1}^n\varepsilon_{\theta,i}\big(\hat{\dot m}_\theta(X_i)-\dot m_\theta(X_i)\big)\\
&\quad-\frac1n\sum_{i=1}^n\dot\varepsilon_{\theta,i}\big(\hat m_\theta(X_i)-m_\theta(X_i)\big)\\
&+\quad\underbrace{\frac1n\sum_{i=1}^n\big(\hat m_\theta(X_i)-m_\theta(X_i)\big)\big(\hat{\dot m}_\theta(X_i)-\dot m_\theta(X_i)\big)}\quad\Big|\\
&\quad\le\Big(\tfrac1n\sum_{i=1}^n\big(\hat m_\theta(X_i)-m_\theta(X_i)\big)^2\Big)^{\frac12}\Big(\tfrac1n\sum_{i=1}^n\big(\hat{\dot m}_\theta(X_i)-\dot m_\theta(X_i)\big)^2\Big)^{\frac12}\\
&\le\underbrace{2\sup_{\theta\in\Theta}\Big|\frac1n\sum_{i=1}^n\big(\varepsilon_{\theta,i}\dot\varepsilon_{\theta,i}-\mathbb{E}[\varepsilon_\theta\dot\varepsilon_\theta]\big)\Big|}_{=O(\log(n)n^{-1/2})}+\underbrace{2\sup_{\theta\in\Theta}\Big|\frac1n\sum_{i=1}^n\varepsilon_{\theta,i}\big(\hat{\dot m}_\theta(X_i)-\dot m_\theta(X_i)\big)\Big|}_{\le 2\sup_{\theta\in\Theta}\left|||X^T(\hat{\dot\beta}_\theta-\dot\beta_\theta)||_{\mathbb{P}_{n,2}}\sqrt{\frac1n\sum_{i=1}^n\varepsilon^2_{\theta,i}}\right|}\\
&\quad+\underbrace{2\sup_{\theta\in\Theta}\Big|\frac1n\sum_{i=1}^n\dot\varepsilon_{\theta,i}\big(\hat m_\theta(X_i)-m_\theta(X_i)\big)\Big|}_{\le 2\sup_{\theta\in\Theta}\left|||X^T(\hat\beta_\theta-\beta_\theta)||_{\mathbb{P}_{n,2}}\sqrt{\frac1n\sum_{i=1}^n\dot\varepsilon^2_{\theta,i}}\right|}\\
&\quad+\Big(\underbrace{\sup_{\theta\in\Theta}\Big|||\big(\hat m_\theta(X_i)-m_\theta(X_i)\big)||^2_{\mathbb{P}_{n,2}}\Big|}_{=O\left(\frac{s\log(p\vee n)}{n}\right)}\underbrace{\sup_{\theta\in\Theta}\Big|||\big(\hat{\dot m}_\theta(X_i)-\dot m_\theta(X_i)\big)||^2_{\mathbb{P}_{n,2}}\Big|}_{=O\left(\frac{s\log(p\vee n)}{n}\right)}\Big)^{\frac12}\\
&\le 2\underbrace{\sup_{\theta\in\Theta}||X^T(\hat{\dot\beta}_\theta-\dot\beta_\theta)||_{\mathbb{P}_{n,2}}}_{=O\left(\sqrt{\frac{s\log(p\vee n)}{n}}\right)}\underbrace{\sup_{\theta\in\Theta}\sqrt{\frac1n\sum_{i=1}^n\varepsilon^2_{\theta,i}}}_{=O(1)}+2\underbrace{\sup_{\theta\in\Theta}||X^T(\hat\beta_\theta-\beta_\theta)||_{\mathbb{P}_{n,2}}}_{=O\left(\sqrt{\frac{s\log(p\vee n)}{n}}\right)}\underbrace{\sup_{\theta\in\Theta}\sqrt{\frac1n\sum_{i=1}^n\dot\varepsilon^2_{\theta,i}}}_{=O(1)}\\
&\quad+O\Big(\frac{s\log(p\vee n)}{n}\Big)+O(\log(n)n^{-1/2})
\end{aligned}
$$

and

$$\sup_{\theta\in\Theta}|\hat{\dot\sigma}_\theta^2-\dot\sigma_\theta^2|=O\left(\max\left(\sqrt{\frac{s\log(p\vee n)}{n}},\frac{\log(n)}{n^{1/2}}\right)\right)\leq\tilde\delta_n n^{-\frac14}.$$

□

Proof of Theorem 5.

The strategy of the proof is similar to the proof of Theorem 1 from Belloni et al. [9]. Let C, C_1 and C_2 denote generic positive constants that may differ in each appearance, but do not depend on the sequence $P\in\mathcal{P}_n$.

For every $\theta\in\Theta$, the set $\tilde{\mathcal{H}}_1(\theta)$ consists of unions of p choose Cs sets, where the set of indices $\{i\in\{1,\dots,p\}:\beta_i\neq0\}$ has cardinality not more than Cs, and therefore is a subset of a vector space with dimension Cs. It follows that $\tilde{\mathcal{H}}_1(\theta)$ consists of unions of p choose Cs VC-subgraph classes $\tilde{\mathcal{H}}_{1,k}(\theta)$ with VC indices less or equal to $Cs+2$ (Lemma 2.6.15, Vaart and Wellner [94]).

Using Theorem 2.6.7 in Van der Vaart and Wellner (1996), we obtain

$$\begin{aligned}
&\sup_Q\log N(\varepsilon\|\tilde H_1\|_{Q,2},\tilde{\mathcal{H}}_1(\theta),L_2(Q))\\
&\leq\sup_Q\log\left(\sum_{k=1}^{\binom{p}{Cs}}N(\varepsilon\|\tilde H_1\|_{Q,2},\tilde{\mathcal{H}}_{1,k}(\theta),L_2(Q))\right)\\
&\leq\sup_Q\log\left(\underbrace{\binom{p}{Cs}}_{\leq\left(\frac{e\cdot p}{Cs}\right)^{Cs}}K(Cs+2)(16e)^{Cs+2}\left(\frac1\varepsilon\right)^{2Cs+2}\right)\\
&\leq\log\left(\left(\frac{e\cdot p}{Cs}\right)^{Cs}K(Cs+2)(16e)^{Cs+2}\left(\frac1\varepsilon\right)^{2Cs+2}\right)\\
&\leq Cs\log\left(\frac{p}{\varepsilon}\right)
\end{aligned}$$

with C beeing independent from θ. Since

$$\begin{aligned}
\sup_{h_1(\theta)\in\tilde{\mathcal{H}}_1(\theta)}|h_1(\theta,x)|&\leq\sup_{\tilde\beta:\|\tilde\beta_\theta-\beta_\theta\|_1\leq\tilde\delta_n\sqrt{s}n^{-\frac14}}|x^T\tilde\beta|\\
&\leq\sup_{\tilde\beta:\|\tilde\beta_\theta-\beta_\theta\|_1\leq\tilde\delta_n\sqrt{s}n^{-\frac14}}|x^T\tilde\beta-x^T\beta_\theta|+|x^T\beta_\theta|\\
&\leq KC+\mathbb{E}\left[F_\Lambda|X=x\right]=:\tilde H_1(x),
\end{aligned}$$

the envelope $\tilde H_1$ can be chosen independent from θ. Here and in the following, we omit the dependence from Y in $F_\Lambda\equiv F_\Lambda(Y)$ to simplify notation.

By the same argument, we obtain

$$\sup_Q\log N(\varepsilon\|\tilde H_3\|_{Q,2},\tilde{\mathcal{H}}_3(\theta),L_2(Q))\leq Cs\log\left(\frac{p}{\varepsilon}\right)$$

with envelope $\tilde H_3(x):=KC+\mathbb{E}\left[\dot F_\Lambda|X=x\right]$.

Next, we consider

$$\begin{aligned}
\tilde{\mathcal{H}}_4(\theta):&=\left\{c\in\mathbb{R}|\ |c-\dot\sigma_\theta^2|\leq\tilde\delta_n n^{-1/4}\right\}\subseteq\left[\dot\sigma_\theta^2-Cn^{-1/4},\dot\sigma_\theta^2+Cn^{-1/4}\right]\\
&\subseteq\left[-(c+Cn^{-1/4}),(c+Cn^{-1/4})\right],
\end{aligned}$$

where $c = \sup_{\theta\in\Theta} |\dot{\sigma}^2_\theta| < \infty$. This implies

$$\begin{aligned}&\sup_Q \log N(\varepsilon \|\tilde{H}_4\|_{Q,2}, \tilde{\mathcal{H}}_4(\theta), L_2(Q))\\ &\leq \sup_Q \log N\left(\varepsilon(c+C), \left[-(c+Cn^{-1/4}), c+Cn^{-1/4}\right], |\cdot|\right) \leq \log\left(\frac{C}{\varepsilon}\right)\end{aligned}$$

for all $\theta \in \Theta$ with envelope $\tilde{H}_4 = c + C$ and C independent from θ.

Remark that $0 < c_1 = \inf_{\theta\in\Theta} \sigma^2_\theta$ and $c_2 = \sup_{\theta\in\Theta} \sigma^2_\theta < \infty$ due to Assumptions A4 and A5. For n sufficient large, we find a c_3 with $0 < c_3 \leq c_1 - Cn^{-1/4}$. Therefore, we can define

$$\begin{aligned}\bar{\mathcal{H}}_2(\theta) :&= \left\{ \frac{1}{\tilde{h}_2(\theta)} \middle|\ \tilde{h}_2(\theta) \in \tilde{\mathcal{H}}_2(\theta) \right\}\\ &\subseteq \left\{ 1/c \middle|\ |c - \sigma^2_\theta| \leq Cn^{-1/4} \right\}\\ &= \left\{ 1/c \middle|\ \frac{|c-\sigma^2_\theta|}{|c\sigma^2_\theta|} \leq \frac{1}{|c\sigma^2_\theta|} Cn^{-1/4} \right\}\\ &\subseteq \left\{ 1/c \middle|\ \frac{|c-\sigma^2_\theta|}{|c\sigma^2_\theta|} \leq C^* n^{-1/4} \right\}\\ &= \left\{ \bar{c} \middle|\ |\bar{c} - 1/\sigma^2_\theta| \leq C^* n^{-1/4} \right\}\\ &= \left[1/\sigma^2_\theta - C^* n^{-1/4}, 1/\sigma^2_\theta + C^* n^{-1/4}\right]\\ &\subseteq \left[1/c_2 - C^* n^{-1/4}, 1/c_1 + C^* n^{-1/4}\right]\end{aligned}$$

with $C^* = \frac{C}{c_3 c_1}$. Analogously, we obtain

$$\sup_Q \log N(\varepsilon \|\bar{H}_2\|_{Q,2}, \bar{\mathcal{H}}_2(\theta), L_2(Q)) \leq \log\left(\frac{C}{\varepsilon}\right)$$

for all $\theta \in \Theta$ with envelope $\bar{H}_2 = 1/c_2 + C^*$ and C independent from θ. Define

$$I(\theta, \bar{\mathcal{H}}_2, \tilde{\mathcal{H}}_4) := \left\{ -\frac{1}{2} h_4(\theta) h_2(\theta) \middle|\ h_4(\theta) \in \tilde{\mathcal{H}}_4(\theta), h_2(\theta) \in \bar{\mathcal{H}}_2(\theta) \right\},$$

$$\begin{aligned}&II(\theta, \tilde{\mathcal{H}}_1, \bar{\mathcal{H}}_2, \tilde{\mathcal{H}}_3)\\ &:= \left\{ (y,x) \mapsto -h_2(\theta) \left(\Lambda_\theta(y) - h_1(\theta, x)\right) \left(\dot{\Lambda}_\theta(y) - h_3(\theta, x)\right) \right.\\ &\left. \middle|\ h_1(\theta) \in \tilde{\mathcal{H}}_1(\theta), h_2(\theta) \in \bar{\mathcal{H}}_2(\theta), h_3(\theta) \in \tilde{\mathcal{H}}_3(\theta) \right\}\end{aligned}$$

and

$$\begin{aligned}III(\theta, \tilde{\mathcal{H}}_1, \bar{\mathcal{H}}_2, \tilde{\mathcal{H}}_4) := \Bigg\{ &(y,x) \mapsto \frac{1}{2} h_2^2(\theta) h_4(\theta) \left(\Lambda_\theta(y) - h_1(\theta, x)\right)^2\\ &\Bigg| \ h_1(\theta) \in \tilde{\mathcal{H}}_1(\theta), h_2(\theta) \in \bar{\mathcal{H}}_2(\theta), h_4(\theta) \in \tilde{\mathcal{H}}_4(\theta) \Bigg\}.\end{aligned}$$

By Lemma L.1 in the supplement to Belloni et al. [11], we have

$$\begin{aligned}
&\log N\left(\varepsilon\|1/2\bar{H}_2\tilde{H}_4\|_{Q,2}, I(\theta,\bar{\mathcal{H}}_2,\tilde{\mathcal{H}}_4), L_2(Q)\right)\\
&\leq \log N\left(\frac{\varepsilon}{4}\|\bar{H}_2\|_{Q,2}, \bar{\mathcal{H}}_2(\theta), L_2(Q)\right) + \log N\left(\frac{\varepsilon}{4}\|\tilde{H}_4\|_{Q,2}, \tilde{\mathcal{H}}_4(\theta), L_2(Q)\right)\\
&\leq 2\log\left(\frac{C}{\varepsilon}\right).
\end{aligned}$$

Using A5, we obtain

$$\begin{aligned}
&\log N\left(\varepsilon\|\bar{H}_2(F_\Lambda+\tilde{H}_1)(\dot{F}_\Lambda+\tilde{H}_3)\|_{Q,2}, II(\theta,\tilde{\mathcal{H}}_1,\bar{\mathcal{H}}_2,\tilde{\mathcal{H}}_3), L_2(Q)\right)\\
&\leq \log N\left(\frac{\varepsilon}{2}\|\bar{H}_2(\theta)\|_{Q,2}, \bar{\mathcal{H}}_2(\theta), L_2(Q)\right)\\
&\quad + \log N\left(\frac{\varepsilon}{4}\|(F_\Lambda+\tilde{H}_1)\|_{Q,2}, \mathcal{F}_\Lambda-\tilde{\mathcal{H}}_1(\theta), L_2(Q)\right)\\
&\quad + \log N\left(\frac{\varepsilon}{4}\|(\dot{F}_\Lambda+\tilde{H}_3)\|_{Q,2}, \dot{\mathcal{F}}_\Lambda-\tilde{\mathcal{H}}_3(\theta), L_2(Q)\right)\\
&\leq \log\left(\frac{2C}{\varepsilon}\right) + \log N\left(\frac{\varepsilon}{8}\|F_\Lambda\|_{Q,2}, \mathcal{F}_\Lambda, L_2(Q)\right)\\
&\quad + \log N\left(\frac{\varepsilon}{8}\|\dot{F}_\Lambda\|_{Q,2}, \dot{\mathcal{F}}_\Lambda, L_2(Q)\right)\\
&\quad + \log N\left(\frac{\varepsilon}{8}\|\tilde{H}_1\|_{Q,2}, \tilde{\mathcal{H}}_1(\theta), L_2(Q)\right)\\
&\quad + \log N\left(\frac{\varepsilon}{8}\|\tilde{H}_3\|_{Q,2}, \tilde{\mathcal{H}}_3(\theta), L_2(Q)\right)\\
&\leq \log\left(\frac{2C}{\varepsilon}\right) + C'_\Lambda\log(8C''_\Lambda/\varepsilon) + \dot{C}'_\Lambda\log(8\dot{C}''_\Lambda/\varepsilon) + Cs\log\left(\frac{8p}{\varepsilon}\right)\\
&\quad + Cs\log\left(\frac{8p}{\varepsilon}\right)\\
&\leq C_1 s\log\left(\frac{C_2 p}{\varepsilon}\right)
\end{aligned}$$

and with an analogous argument

$$\begin{aligned}
&\log N\left(\varepsilon\|\frac{1}{2}\bar{H}_2^2\tilde{H}_4(F_\Lambda+\tilde{H}_1)^2\|_{Q,2}, III(\theta,\tilde{\mathcal{H}}_1,\bar{\mathcal{H}}_2,\tilde{\mathcal{H}}_4), L_2(Q)\right)\\
&\leq C_1 s\log\left(\frac{C_2 p}{\varepsilon}\right).
\end{aligned}$$

Since

$$\Psi(\theta) = I(\theta,\bar{\mathcal{H}}_2,\tilde{\mathcal{H}}_4) + II(\theta,\tilde{\mathcal{H}}_1,\bar{\mathcal{H}}_2,\tilde{\mathcal{H}}_3) + III(\theta,\tilde{\mathcal{H}}_1,\bar{\mathcal{H}}_2,\tilde{\mathcal{H}}_4) + c_\theta,$$

we can define the envelope

$$\begin{aligned}
\bar{\psi}(Y,X) := &\frac{1}{2}\bar{H}_2\tilde{H}_4 + \bar{H}_2(F_\Lambda+\tilde{H}_1)(\dot{F}_\Lambda+\tilde{H}_3)\\
&+ \frac{1}{2}\bar{H}_2^2\tilde{H}_4(F_\Lambda+\tilde{H}_1)^2 + J_\Lambda,
\end{aligned}$$

which is independent from θ with

$$\begin{aligned}
&\mathbb{E}\left[\left(\bar{\psi}(Y,X)\right)^4\right]\\
&= \mathbb{E}\left[\left(\frac{1}{2}\bar{H}_2\tilde{H}_4 + \bar{H}_2(F_\Lambda+\tilde{H}_1)(\dot{F}_\Lambda+\tilde{H}_3) + \frac{1}{2}\bar{H}_2^2\tilde{H}_4(F_\Lambda+\tilde{H}_1)^2 + J_\Lambda\right)^4\right]\\
&< \infty,
\end{aligned}$$

where we used A5 and A8. Additionally, by using $N(\varepsilon||J_\Lambda||_{Q,2}, c_\theta, L_2(Q)) = 1$ for all $\theta \in \Theta$ and Lemma L.1 in the supplement to Belloni et al. [11], we obtain

$$\sup_Q \log N(\varepsilon||\bar{\psi}||_{Q,2}, \Psi(\theta), L_2(Q)) \leq C_1 s \log\left(\frac{C_2(p \vee n)}{\varepsilon}\right),$$

where the supremum is taken over all probability measures Q with $\mathbb{E}_Q\left[\left(\bar{\psi}(Y,X)\right)^2\right] < \infty$. □

Proof of Theorem 6.
We demonstrate that the conditions C1-C7 from Theorem 8 are satisfied. Most conditions are already proven in the preceding theorems. The condition C1 is shown in Lemma 1. Due to Theorem 3 and 4 condition C3 is satisfied with $\tilde{\mathcal{H}}$ and $\tilde{\mathcal{H}}(\theta)$ as defined in Subsection 3.3.3. Condition C5 is proved in Theorem 5. Again, choosing $\mathcal{H}' = \tilde{\mathcal{H}}$ as defined in Subsection 3.3.3, the conditions in Lemma 2 hold where we used (3.22) and the envelope in C5 which implies C4. Since conditions C2 and C7 are the same as A11 and A10, we need to verify C6. Due to condition A2, choosing $\rho_n = o(n^{-1/4})$, we have

$$\begin{aligned}
&\sup_{\theta\in\Theta, \tilde{h}\in\tilde{\mathcal{H}}(\theta)} |\mathbb{E}[\psi((Y,X)), \theta, h_0(\theta)] - \mathbb{E}[\psi((Y,X)), \theta, \tilde{h}(\theta)]| \\
&\leq \sup_{\theta\in\Theta, \tilde{h}\in\tilde{\mathcal{H}}(\theta)} \mathbb{E}\Big[\Big(\psi\big((Y,X), \theta, \tilde{h}(\theta,X)\big) - \psi\big((Y,X), \theta, h_0(\theta,X)\big)\Big)^2\Big]^{\frac{1}{2}} \\
&\leq \sup_{\theta\in\Theta, \tilde{h}\in\tilde{\mathcal{H}}(\theta)} C\mathbb{E}\left[\|\tilde{h}(\theta,X) - h_0(\theta,X)\|_2^2\right]^{\frac{1}{2}} \\
&\leq C\rho_n,
\end{aligned}$$

where we used A9 (ii) and

$$\begin{aligned}
\mathbb{E}\Big[\|\tilde{h}(\theta,X) - h_0(\theta,X)\|_2^2\Big] &= \mathbb{E}\big[(\tilde{h}_1(\theta,X) - m_\theta(X))^2\big] + \mathbb{E}\big[(\tilde{h}_2(\theta) - \sigma_\theta^2)^2\big] \\
&\quad + \mathbb{E}\big[(\tilde{h}_3(\theta,X) - \dot{m}_\theta(X))^2\big] + \mathbb{E}\big[(\tilde{h}_4(\theta) - \dot{\sigma}_\theta^2)^2\big] \\
&\leq C\rho_n^2.
\end{aligned}$$

The last inequality follows from the properties of $\tilde{\mathcal{H}}$ and condition A7. We have

$$\begin{aligned}
\mathbb{E}\big[(\tilde{h}_1(\theta,X) - m_\theta(X))^2\big] &= \mathbb{E}\left[\left(X^T(\tilde{\beta}_\theta - \beta_\theta)\right)^2\right] \\
&\leq \sup_{\theta\in\Theta} \|\tilde{\beta}_\theta - \beta_\theta\|_2^2 (\kappa'')^2 \\
&\leq C \sup_{\theta\in\Theta} \left\|X^T\left(\tilde{\beta}_\theta - \beta_\theta\right)\right\|_{\mathbb{P}_n,2}^2 \\
&\leq C\rho_n^2
\end{aligned}$$

and

$$\mathbb{E}\big[(\tilde{h}_2(\theta) - \sigma_\theta^2)^2\big] \leq C\rho_n^2$$

due to the bounded empirical sparse eigenvalue. The same holds for the two remaining terms with an analogous argument. Therefore, C6 (i) holds.

In the following, we take the supremum over all θ with $|\theta - \theta_0| \leq C\rho_n$ and $\tilde{h} \in \tilde{\mathcal{H}}(\theta)$, meaning

$$\sup \equiv \sup_{\theta:|\theta-\theta_0|\leq C\rho_n, \tilde{h}\in\tilde{\mathcal{H}}(\theta)}.$$

By A9 (i) and (ii), we have

$$\begin{aligned}
&\sup \mathbb{E}\Big[\Big(\psi\big((Y,X),\theta,\tilde{h}(\theta,X)\big) - \psi\big((Y,X),\theta_0,h_0(\theta_0,X)\big)\Big)^2\Big]^{1/2} \\
&= \sup \mathbb{E}\Big[\Big(\psi\big((Y,X),\theta,\tilde{h}(\theta,X)\big) - \psi\big((Y,X),\theta,h_0(\theta,X)\big) \\
&\qquad + \psi\big((Y,X),\theta,h_0(\theta,X)\big) - \psi\big((Y,X),\theta_0,h_0(\theta_0,X)\big)\Big)^2\Big]^{1/2} \\
&\leq \sup \mathbb{E}\Big[\Big(\psi\big((Y,X),\theta,\tilde{h}(\theta,X)\big) - \psi\big((Y,X),\theta,h_0(\theta,X)\big)\Big)^2 \\
&\quad + \Big(\psi\big((Y,X),\theta,h_0(\theta,X)\big) - \psi\big((Y,X),\theta_0,h_0(\theta_0,X)\big)\Big)^2 \\
&\quad + 2\Big(\psi\big((Y,X),\theta,\tilde{h}(\theta,X)\big) - \psi\big((Y,X),\theta,h_0(\theta,x)\big)\Big) \\
&\qquad \Big(\psi\big((Y,X),\theta,h_0(\theta,X)\big) - \psi\big((Y,X),\theta_0,h_0(\theta_0,X)\big)\Big)\Big]^{1/2} \\
&\leq \sup C\Bigg(\mathbb{E}\Big[\|\tilde{h}(\theta,X) - h_0(\theta,X)\|_2^2\Big] + |\theta-\theta_0|^2 \\
&\quad + |\theta-\theta_0|\sqrt{\mathbb{E}\Big[\|\tilde{h}(\theta,X) - h_0(\theta,X)\|_2^2\Big]}\Bigg)^{1/2} \\
&\leq C\rho_n \\
&\leq Cn^{-1/4}.
\end{aligned}$$

Due to growth condition A2, we have

$$n^{-1/4}s^{\frac{1}{2}}\log\Big(\frac{(p\vee n)}{n^{-1/4}}\Big)^{\frac{1}{2}} + n^{-\frac{1}{2}+\frac{1}{q}}s\log\Big(\frac{(p\vee n)}{n^{-1/4}}\Big) = o(1)$$

and C6 (ii) follows.

Condition C6 (iii) follows directly from A9 (iii):

$$\begin{aligned}
&\sup_{r\in(0,1)} \sup \Big|\partial_r^2\Big\{\mathbb{E}\Big[\psi\big((Y,X),\theta_0 + r(\theta-\theta_0), h_0 + r(\tilde{h}-h_0)\big)\Big]\Big\}\Big| \\
&\leq \sup_{\theta:|\theta-\theta_0|\leq C\rho_n,\tilde{h}\in\tilde{\mathcal{H}}(\theta)} C\left(|\theta-\theta_0|^2 + \sup_{\theta^*\in\Theta}\mathbb{E}\Big[\|\tilde{h}(\theta^*,X) - h_0(\theta^*,X)\|_2^2\Big]\right) \\
&\leq C\rho_n^2 \\
&= o(n^{-1/2}).
\end{aligned}$$

□

Proof of Lemma 3.

Comment 3.7.1. *The proof for Box-Cox transformations is from Vaart and Wellner [94] who refer to Quiroz et al. [82]. It heavily relies on the properties of the dual density from Assouad [3]. We give a detailed version of the proof of Quiroz et al. [82] and extend the idea to the class of derivatives and Yeo-Johnson power transformations.*

Since adding a single function to a class of functions can increase the VC index at most by one, we exclude the parameter $\theta = 0$ from the proof and restrict the class to

$$\mathcal{F}_1' = \big\{\Lambda_\theta(\cdot)|\theta \in \mathbb{R} \setminus \{0\}\big\}.$$

At first, recall that $\mathcal{F}_1'$ is a VC class if and only if the between graph set

$$\mathcal{C} := \big\{C_\theta|\theta \in \mathbb{R} \setminus \{0\}\big\}$$

with

$$C_\theta := \big\{(x,t) \in \mathbb{R}^+ \times \mathbb{R}|0 \le t \le \Lambda_\theta(x) \text{ or } \Lambda_\theta(x) \le t \le 0\big\}$$

is a VC class (cf. Vaart and Wellner [94], page 152). We now consider the dual class (cf. Assouad [3]) of $\mathcal{C}$ given by

$$\mathcal{D} := \big\{D_{(x,t)}|(x,t) \in \mathbb{R}^+ \times \mathbb{R}\big\}$$

with

$$\begin{aligned} D_{(x,t)} :&= \big\{\theta \in \mathbb{R} \setminus \{0\}|(x,t) \in C_\theta\big\} \\ &= \big\{\theta \in \mathbb{R} \setminus \{0\}|0 \le t \le \Lambda_\theta(x) \text{ or } \Lambda_\theta(x) \le t \le 0\big\}. \end{aligned}$$

For the derivative of $\Lambda_\theta(x)$, we have

$$\begin{aligned} & \dot{\Lambda}_\theta(x) = \frac{1}{\theta^2}\left((\theta\log(x) - 1)\, x^\theta + 1\right) \ge 0 \\ \Leftrightarrow \quad & (\theta\log(x) - 1)\, x^\theta \ge -1 \\ \Leftrightarrow \quad & \log(x^\theta) \ge \frac{x^\theta - 1}{x^\theta}, \end{aligned}$$

which is true for all x and θ. Since $\Lambda_\theta(x)$ is continuous and monotone increasing in θ, the set $D_{(x,t)}$ is the union of at most two intervals in $\mathbb{R} \setminus \{0\}$ and therefore $\mathcal{D}$ is a VC class, which by Proposition 2.12 in Assouad [3] implies that $\mathcal{C}$ is a VC class.
By the same argument as above, we have to prove that

$$\mathcal{D}' = \big\{D'_{(x,t)}|(x,t) \in \mathbb{R}^+ \times \mathbb{R}\big\}$$

is a VC class with

$$D'_{(x,t)} := \big\{\theta \in \mathbb{R} \setminus \{0\}|0 \le t \le \dot{\Lambda}_\theta(x)\big\},$$

since $\dot{\Lambda}_\theta(x) \ge 0$. The second derivative with respect to θ is given by

$$\ddot{\Lambda}_\theta(x) = \frac{1}{\theta^3}\Big(\underbrace{\big(\log(x^\theta) - 1\big)^2 x^\theta + x^\theta - 2}_{=:f(x^\theta)}\Big).$$

The case $x = 1$ directly implies $\ddot{\Lambda}_\theta(x) = 0$. Substitute $z = x^\theta$ in $f(x^\theta)$ and note that

$$f'(z) = (\log(z) - 1)^2 + 2\,(\log(z) - 1) + 1 = (\log(z))^2 \ge 0.$$

This together with $f(1) = 0$ implies $f(z) \geq 0$ for $z \geq 1$ and $f(z) < 0$ for $z < 1$. The four cases

$$\left.\begin{aligned} x > 1,\ \theta > 0 \\ 0 < x < 1,\ \theta < 0 \end{aligned}\right\} \Rightarrow x^\theta > 1$$
$$\left.\begin{aligned} x > 1,\ \theta < 0 \\ 0 < x < 1,\ \theta > 0 \end{aligned}\right\} \Rightarrow 0 < x^\theta < 1$$

and the coefficient $1/\theta^3$ imply

$$\ddot{\Lambda}_\theta(x) = \begin{cases} \geq 0 \text{ for } x \geq 1 \\ < 0 \text{ for } x < 1. \end{cases}$$

We have that $\dot{\Lambda}_\theta(x)$ is continuous in θ, monotone increasing for $x \geq 1$ and monotone decreasing for $x < 1$. This again implies that the set $D_{(x,t)}$ is the union of at most two intervals in $\mathbb{R} \setminus \{0\}$. Now, we consider the class of Yeo-Johnson power transformations

$$\mathcal{F}_2 = \big\{\Psi_\theta(\cdot)|\theta \in \mathbb{R} \setminus \{0, 2\}\big\},$$

where we exclude the parameters $\theta = 0$ and $\theta = 2$. The between graph set is given by

$$\tilde{\mathcal{C}} := \big\{\tilde{C}_\theta|\theta \in \mathbb{R} \setminus \{0, 2\}\big\}$$

with

$$\tilde{C}_\theta := \big\{(x, t) \in \mathbb{R} \times \mathbb{R}|0 \leq t \leq \Psi_\theta(x) \text{ or } \Psi_\theta(x) \leq t \leq 0\big\}.$$

Since $\Psi_\theta(x) \geq 0$ for $x \geq 0$ and $\Psi_\theta(x) < 0$ for $x < 0$, we have

$$\begin{aligned} \tilde{C}_\theta :&= \big\{(x, t) \in \mathbb{R} \times \mathbb{R}|0 \leq t \leq \Psi_\theta(x) \text{ or } \Psi_\theta(x) \leq t \leq 0\big\} \\ &= \big\{(x, t) \in \mathbb{R}_0^+ \times \mathbb{R}|0 \leq t \leq \Psi_\theta(x)\big\} \cup \big\{(x, t) \in \mathbb{R}^- \times \mathbb{R}|\Psi_\theta(x) \leq t \leq 0\big\} \\ &= \underbrace{\big\{(x, t) \in \mathbb{R}_0^+ \times \mathbb{R}|0 \leq t \leq \Lambda_\theta(x + 1)\big\}}_{=:\tilde{C}_{\theta,1}} \\ &\quad \cup \underbrace{\big\{(x, t) \in \mathbb{R}^- \times \mathbb{R}| - \Lambda_{2-\theta}(-x + 1) \leq t \leq 0\big\}}_{=:\tilde{C}_{\theta,2}}. \end{aligned}$$

The sets

$$\tilde{\mathcal{C}}_1 := \big\{\tilde{C}_{\theta,1}|\theta \in \mathbb{R} \setminus \{0, 2\}\big\} \text{ and } \tilde{\mathcal{C}}_2 := \big\{\tilde{C}_{\theta,2}|\theta \in \mathbb{R} \setminus \{0, 2\}\big\}$$

are VC-classes as shown above. Using Lemma 2.6.17 (iii) from Vaart and Wellner [94],

$$\tilde{\mathcal{C}}_1 \sqcup \tilde{\mathcal{C}}_2 = \big\{\tilde{C}_{\theta,1} \cup \tilde{C}_{\theta,2}|\tilde{C}_{\theta,1} \in \tilde{\mathcal{C}}_1, \tilde{C}_{\theta,2} \in \tilde{\mathcal{C}}_2\big\}$$

is a VC-class which contains $\tilde{\mathcal{C}}$. The statement for the class of derivatives can be shown analogously. □

Supplementary Material

3.8 Uniform Convergence Rates for the Lasso

Assumptions **B1**-**B7**.
The following assumptions hold uniformly in $n \geq n_0$ and $P \in \mathcal{P}_n$:

B1 Uniformly in θ, the model is sparse, namely $\sup_{\theta \in \Theta} \|\beta_\theta\|_0 \leq s$.

B2 The parameters obey the growth conditions $n^{-1/3} \log(p \vee n) \leq \delta_n$ and $s \log(p \vee n) \leq \delta_n n$ for $\delta_n \searrow 0$ approaching zero from above at a speed at most polynomial in n.

B3 For all $n \in \mathbb{N}$, the regressor $X = (X_1, \ldots, X_p)$ has a bounded support $\mathcal{X}$.

B4 Uniformly in θ, the conditional variance of the error term is bounded

$$0 < c \leq \inf_{\theta \in \Theta} \mathbb{E}\big[\varepsilon_\theta^2 | X\big] \leq \sup_{\theta \in \Theta} \mathbb{E}\big[\varepsilon_\theta^2 | X\big] \leq C < \infty.$$

B5 The transformations are measurable and the class of transformations

$$\mathcal{F}_\Lambda := \big\{\Lambda_\theta(\cdot) | \theta \in \Theta\big\}$$

has VC index C_Λ and an envelope F_Λ with

$$\mathbb{E}[F_\Lambda(Y)^6] < \infty.$$

B6 The transformations are differentiable with respect to θ and the following condition holds:

$$\sup_{\theta \in \Theta} \mathbb{E}\Big[\big(\dot{\Lambda}_\theta(Y)\big)^2\Big] \leq C < \infty.$$

B7 With probability $1 - o(1)$, the empirical minimum and maximum sparse eigenvalues are bounded from zero and above, namely

$$\begin{aligned} 0 < \kappa' &\leq \inf_{||\delta||_0 \leq s \log(n), ||\delta|| = 1} ||X^T \delta||_{\mathbb{P}_{n,2}} \\ &\leq \sup_{||\delta||_0 \leq s \log(n), ||\delta|| = 1} ||X^T \delta||_{\mathbb{P}_{n,2}} \leq \kappa'' < \infty. \end{aligned}$$

Theorem 7.
Under Assumptions B1-B7 above, uniformly for all $P \in \mathcal{P}_n$ with probability $1 - o(1)$, it holds:

1. $\sup_{\theta\in\Theta} ||\hat{\beta}_\theta||_0 = O(s)$,
2. $\sup_{\theta\in\Theta} ||X^T(\hat{\beta}_\theta - \beta_\theta)||_{\mathbb{P}_{n,2}} = O\left(\sqrt{\frac{s\log(p\vee n)}{n}}\right)$,
3. $\sup_{\theta\in\Theta} ||\hat{\beta}_\theta - \beta_\theta||_1 = O\left(\sqrt{\frac{s^2\log(p\vee n)}{n}}\right)$.

Proof. We verify Assumption 6.1 from Belloni et al. [11]. Due to Assumption B1 and Assumption B2, the condition in 6.1(i) is satisfied. Needless to say, the Assumption 6.1(ii) holds for a compact $\Theta \subset \mathbb{R}$. Remark that Assumption B4 and Assumption B5 imply the conditions in 6.1 (iii). Due to Assumption B3, the conditions in 6.1(iv)(a) are satisfied and we can omit the X in the technical conditions in 6.1(iv)(b). The eigenvalue condition 6.1(iv)(c) is the same as in B7. Therefore, we have to show with probability $1 - o(1)$:

(1) $\sup_{\theta\in\Theta} |(\mathbb{E}_n - \mathbb{E})\varepsilon_\theta^2 \vee (\mathbb{E}_n - \mathbb{E})\Lambda_\theta(Y)^2| = O(\delta_n)$

(2) $n^{1/2} \sup_{|\theta-\theta'|\le 1/n} |\mathbb{E}_n[\varepsilon_\theta - \varepsilon_{\theta'}]| = O(\delta_n)$ and

(3) $\log(p\vee n)^{1/2} \sup_{|\theta-\theta'|\le 1/n} \mathbb{E}_n\left[(\varepsilon_\theta - \varepsilon_{\theta'})^2\right]^{1/2} = O(\delta_n)$.

Since $\mathcal{F}_\Lambda$ is a VC-class of functions with VC index C_Λ, we have by Theorem 2.6.7 in Vaart and Wellner [94]

$$\log N(\varepsilon\|F_\Lambda\|_{Q,2}, \mathcal{F}_\Lambda, L_2(Q)) \le C'_\Lambda \log(C''_\Lambda/\varepsilon) \tag{3.20}$$

for any Q with $\|F_\Lambda\|_{Q,2}^2 = \mathbb{E}_Q[F_\Lambda^2] < \infty$, where the constants C'_Λ and C''_Λ only depend on the VC index. Define

$$\mathcal{F}'_\Lambda := \left\{\mathbb{E}[\Lambda_\theta(\cdot)|X] \,|\theta \in \Theta\right\}$$

with envelope $F'_\Lambda := E[F_\Lambda|X]$ and

$$\mathcal{E}_\Lambda^2 := \left\{ (\Lambda_\theta(\cdot) - \mathbb{E}[\Lambda_\theta(\cdot)|X])^2 \,|\theta \in \Theta\right\}$$

with envelope $(F_\Lambda + F'_\Lambda)^2$. By Lemma L.2 in the supplement to Belloni et al. [11], we have

$$\sup_{Q'} \log N(\varepsilon\|F'_\Lambda\|_{Q',2}, \mathcal{F}'_\Lambda, L_2(Q')) \le \sup_{Q} \log N((\varepsilon/4)^2\|F_\Lambda\|_{Q,2}, \mathcal{F}_\Lambda, L_2(Q)), \tag{3.21}$$

where the supremum on the left-hand side is taken over all probability measures Q' with

$$\|F'_\Lambda\|_{Q',2}^2 := \mathbb{E}_{Q'}\left[\left(\mathbb{E}[F_\Lambda(Y)|X]\right)^2\right] \equiv \mathbb{E}_{Q'}\left[\left(\mathbb{E}[F_\Lambda|X]\right)^2\right] < \infty.$$

Since $\mathcal{E}_\Lambda^2 \subset (\mathcal{F}_\Lambda - \mathcal{F}'_\Lambda)^2$, it follows by Lemma L.1 in the supplement to Belloni et al. [11] for any $\tilde{Q}$ with $\mathbb{E}_{\tilde{Q}}[(F_\Lambda + F'_\Lambda)^4] < \infty$ and $0 < \varepsilon \le 1$

$$\begin{aligned}
&\log N(\varepsilon\|(F_\Lambda + F'_\Lambda)^2\|_{\tilde{Q},2}, \mathcal{E}_\Lambda^2, L_2(\tilde{Q})) \\
&\le 2\log N\left(\frac{\varepsilon}{2}\|F_\Lambda + F'_\Lambda\|_{\tilde{Q},2}, \mathcal{F}_\Lambda - \mathcal{F}'_\Lambda, L_2(\tilde{Q})\right) \\
&\le 2\log N\left(\frac{\varepsilon}{4}\|F_\Lambda\|_{\tilde{Q},2}, \mathcal{F}_\Lambda, L_2(\tilde{Q})\right) + 2\log N\left(\frac{\varepsilon}{4}\|F'_\Lambda\|_{\tilde{Q},2}, \mathcal{F}'_\Lambda, L_2(\tilde{Q})\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq 2 \sup_Q \log N\left(\frac{\varepsilon}{4}\|F_\Lambda\|_{Q,2}, \mathcal{F}_\Lambda, L_2(Q)\right) \\
&\quad + 2 \sup_{Q'} \log N\left(\frac{\varepsilon}{4}\|F'_\Lambda\|_{Q',2}, \mathcal{F}'_\Lambda, L_2(Q')\right) \\
&\leq 4 \sup_Q \log N\left(\frac{\varepsilon^2}{256}\|F_\Lambda\|_{Q,2}, \mathcal{F}_\Lambda, L_2(Q)\right),
\end{aligned}
$$

where we used (3.21) in the last step. We conclude

$$
\begin{aligned}
\log N(\varepsilon\|(F_\Lambda + F'_\Lambda)^2\|_{\tilde{Q},2}, \mathcal{E}^2_\Lambda, L_2(\tilde{Q})) &\leq 4C'_\Lambda \log(256 C''_\Lambda/\varepsilon^2) \\
&= 16 C'_\Lambda \log\left(16\sqrt{C''_\Lambda}/\varepsilon\right)
\end{aligned}
$$

by (3.20). Under Assumption B5, for all $r \in \{1, 2, 3\}$, it holds

$$
\mathbb{E}\left[F'^{2r}_\Lambda\right] = \mathbb{E}\left[\left(\mathbb{E}\left[F_\Lambda|X\right]\right)^{2r}\right] \leq \mathbb{E}\left[\mathbb{E}\left[\left(F_\Lambda\right)^{2r}|X\right]\right] = \mathbb{E}\left[F^{2r}_\Lambda\right] < \infty,
$$

which implies

$$
\begin{aligned}
\mathbb{E}\left[(F_\Lambda + F'_\Lambda)^4\right] &= \mathbb{E}\left[F^4_\Lambda\right] + \underbrace{\mathbb{E}\left[F'^4_\Lambda\right]}_{\leq \mathbb{E}\left[F^4_\Lambda\right]} + 6 \underbrace{\mathbb{E}\left[F^2_\Lambda F'^2_\Lambda\right]}_{\leq \sqrt{\mathbb{E}\left[F^4_\Lambda\right]\mathbb{E}\left[F'^4_\Lambda\right]}} \\
&\quad + 4 \underbrace{\mathbb{E}\left[F^3_\Lambda F'_\Lambda\right]}_{\leq \sqrt{\mathbb{E}\left[F^6_\Lambda\right]\mathbb{E}\left[F'^2_\Lambda\right]}} + 4 \underbrace{\mathbb{E}\left[F_\Lambda F'^3_\Lambda\right]}_{\leq \sqrt{\mathbb{E}\left[F^2_\Lambda\right]\mathbb{E}\left[F'^6_\Lambda\right]}} \\
&\leq C < \infty.
\end{aligned}
$$

Remark that

$$
\mathbb{E}\left[\sup_{\theta\in\Theta} \varepsilon^2_\theta\right] \leq \mathbb{E}\left[(F_\Lambda + F'_\Lambda)^2\right] \leq C < \infty. \tag{3.22}
$$

We have

$$
\sqrt{n} \sup_{\theta\in\Theta} |(\mathbb{E}_n - \mathbb{E})\varepsilon^2_\theta| = \sup_{g\in\mathcal{E}^2_\Lambda} |G_n(g)|.
$$

For every σ^2_C with $\sup_{g\in\mathcal{E}^2_\Lambda} \mathbb{E}[g^2] \leq \sigma^2_C \leq \mathbb{E}\left[(F_\Lambda + F'_\Lambda)^4\right] := G_1 < \infty$ and universal constants K and K_2 with probability not less than $1 - (1/\log(n))$, it holds

$$
\begin{aligned}
&\sup_{g\in\mathcal{E}^2_\Lambda} |G_n(g)| \\
&\leq 2K\left[\left(S\sigma^2_C \log(AG_1^{1/2}/\sigma_C)\right)^{1/2} + SG_1^{1/2}\log(AG_1^{1/2}/\sigma_C)\right] \\
&\quad + K_2(\sigma_C \log(n)^{1/2} + G_1^{1/2}\log(n)) \\
&= O(\log(n))
\end{aligned}
$$

by Lemma 1 in Belloni et al. [9] with $q = 2$, $t = \log(n)$, $A = 16\sqrt{C''_\Lambda}$ and $S = 16C'_\Lambda$. Therefore, it follows with probability $1 - o(1)$

$$
\sup_{\theta\in\Theta} |(\mathbb{E}_n - \mathbb{E})\varepsilon^2_\theta| = O\left(\frac{\log(n)}{\sqrt{n}}\right).
$$

Analogously, it can be shown with probability $1 - o(1)$

$$\sup_{\theta \in \Theta} |(\mathbb{E}_n - \mathbb{E})\Lambda_\theta(Y)^2| = O\left(\frac{\log(n)}{\sqrt{n}}\right).$$

(1) follows by Assumption B2.

Further, we have

$$\sup_{|\theta-\theta'|\leq 1/n} |\mathbb{E}_n\left[\varepsilon_\theta - \varepsilon_{\theta'}\right]| = \sup_{|\theta-\theta'|\leq 1/n} \frac{1}{\sqrt{n}} |G_n(\varepsilon_\theta - \varepsilon'_\theta)|.$$

Define $\mathcal{E}'_\Lambda := \left\{\varepsilon_\theta - \varepsilon_{\theta'} | \theta, \theta' \in \Theta\right\}$ and $\mathcal{E}_\Lambda := \left\{\varepsilon_\theta = (\Lambda_\theta(\cdot) - \mathbb{E}\left[\Lambda_\theta(\cdot)|X\right]) | \theta \in \Theta\right\}$. Using the same argument as above, we obtain

$$\begin{aligned}
&\log N(\varepsilon \|2(F_\Lambda + F'_\Lambda)\|_{\tilde{Q},2}, \mathcal{E}'_\Lambda, L_2(\tilde{Q})) \\
&\leq \log N\left(\frac{\varepsilon}{2}\|2F_\Lambda\|_{\tilde{Q},2}, \mathcal{F}_\Lambda, L_2(\tilde{Q})\right) + \log N\left(\frac{\varepsilon}{2}\|2F'_\Lambda\|_{\tilde{Q},2}, \mathcal{F}'_\Lambda, L_2(\tilde{Q})\right) \\
&\leq \sup_Q \log N\left(\varepsilon\|F_\Lambda\|_{Q,2}, \mathcal{F}_\Lambda, L_2(Q)\right) + \sup_{Q'} \log N\left(\varepsilon\|F'_\Lambda\|_{Q',2}, \mathcal{F}'_\Lambda, L_2(Q')\right) \\
&\leq 2 \sup_Q \log N\left(\left(\frac{\varepsilon}{4}\right)^2 \|F_\Lambda\|_{Q,2}, \mathcal{F}_\Lambda, L_2(Q)\right) \\
&\leq 4C'_\Lambda \log(4\sqrt{C''_\Lambda}/\varepsilon).
\end{aligned}$$

Since

$$\mathcal{E}''_\Lambda := \left\{\varepsilon_\theta - \varepsilon_{\theta'} | \theta, \theta' \in \Theta, |\theta - \theta'| \leq 1/n\right\} \subset \mathcal{E}'_\Lambda,$$

we can use Lemma 1 again, since we obtain the same envelope and bound for the entropy as for $\mathcal{E}'_\Lambda$. We achieve for every σ_n^2 with $\sup_{g \in \mathcal{E}''_\Lambda} \mathbb{E}[g^2] \leq \sigma_n^2 \leq \mathbb{E}[4(F_\Lambda + F'_\Lambda)^2] := G_2$ and universal constants K and K_2 with probability at least $1 - (1/\log(n))$

$$\begin{aligned}
&\sup_{g \in \mathcal{E}''_\Lambda} |G_n(g)| \\
&\leq 2K\left[\left(S\sigma_n^2 \log(AG_2^{1/2}/\sigma_n)\right)^{1/2} + n^{-\frac{1}{4}} S 2\mathbb{E}[(F_\Lambda + F'_\Lambda)^4]^{1/4} \log(AG_2^{1/2}/\sigma_n)\right] \\
&+ K_2(\sigma_n \log(n)^{1/2} + n^{-\frac{1}{4}} 2\mathbb{E}[(F_\Lambda + F'_\Lambda)^4]^{1/4} \log(n))
\end{aligned}$$

by Lemma 1 with $q = 4$, $t = \log(n)$, $A = 4\sqrt{C''_\Lambda}$, $S = 4C'_\Lambda$.
We have

$$\begin{aligned}
&\sup_{|\theta-\theta'|\leq \frac{1}{n}} \mathbb{E}[(\varepsilon_\theta - \varepsilon_{\theta'})^2] \\
&= \sup_{|\theta-\theta'|\leq \frac{1}{n}} \mathbb{E}\left[\left(\Lambda_\theta(Y) - \mathbb{E}[\Lambda_\theta(Y)|X] - \Lambda_{\theta'}(Y) + \mathbb{E}[\Lambda_{\theta'}(Y)|X]\right)^2\right] \\
&= \sup_{|\theta-\theta'|\leq \frac{1}{n}} \mathbb{E}\left[\left(\left(\Lambda_\theta(Y) - \Lambda_{\theta'}(Y)\right) - \left(\mathbb{E}[\Lambda_\theta(Y)|X] - \mathbb{E}[\Lambda_{\theta'}(Y)|X]\right)\right)^2\right] \\
&= \sup_{|\theta-\theta'|\leq \frac{1}{n}} \left(\mathbb{E}\left[\left(\Lambda_\theta(Y) - \Lambda_{\theta'}(Y)\right)^2\right] + \mathbb{E}\Bigg[\underbrace{\mathbb{E}\Big[\left(\Lambda_\theta(Y) - \Lambda_{\theta'}(Y)\right)|X\Big]^2}_{\leq \mathbb{E}\left[\left(\Lambda_\theta(Y) - \Lambda_{\theta'}(Y)\right)^2|X\right]}\Bigg]\right.
\end{aligned}$$

$$
\begin{aligned}
&\quad - 2\,\underbrace{\mathbb{E}\Big[\big(\Lambda_\theta(Y)-\Lambda_{\theta'}(Y)\big)\mathbb{E}\Big[\big(\Lambda_\theta(Y)-\Lambda_{\theta'}(Y)\big)|X\Big]\Big]}_{\geq 0}\Big)\\
&\leq \sup_{|\theta-\theta'|\leq\frac{1}{n}} 2\mathbb{E}\left[\big(\Lambda_\theta(Y)-\Lambda_{\theta'}(Y)\big)^2\right]\\
&= \sup_{|\theta-\theta'|\leq\frac{1}{n}} 2\mathbb{E}\Big[(\theta-\theta')^2\big(\dot\Lambda_{\bar\theta}(Y)\big)^2\Big]\\
&\leq \frac{2}{n^2}\underbrace{\sup_{\theta\in\Theta}\mathbb{E}\Big[\big(\dot\Lambda_\theta(Y)\big)^2\Big]}_{\leq C} = O(n^{-2}).
\end{aligned}
$$

Therefore, we can choose $\sigma_n^2 = O(n^{-2})$ and obtain with probability $1-o(1)$

$$
\begin{aligned}
n^{1/2}\sup_{|\theta-\theta'|\leq 1/n}|\mathbb{E}_n\left[\varepsilon_\theta-\varepsilon_{\theta'}\right]| &= \sup_{|\theta-\theta'|\leq 1/n}|G_n(\varepsilon_\theta-\varepsilon'_\theta)|\\
&= \sup_{g\in\mathcal{E}''_\Lambda}|G_n(g)|\\
&= O\left(\frac{\log(n)}{n^{1/4}}\right) = O\left(\delta_n\right).
\end{aligned}
$$

To verify (3), we can use the same arguments as above and we remark that

$$
\begin{aligned}
\sup_{|\theta-\theta'|\leq 1/n}\mathbb{E}_n\left[(\varepsilon_\theta-\varepsilon_{\theta'})^2\right] &\leq \sup_{|\theta-\theta'|\leq 1/n}\mathbb{E}\left[(\varepsilon_\theta-\varepsilon_{\theta'})^2\right]\\
&\quad + \left|\sup_{|\theta-\theta'|\leq 1/n}\left(\mathbb{E}_n\left[(\varepsilon_\theta-\varepsilon_{\theta'})^2\right]-\mathbb{E}\left[(\varepsilon_\theta-\varepsilon_{\theta'})^2\right]\right)\right|\\
&\leq \sup_{g\in\mathcal{E}_\Lambda^{2\prime}}\frac{1}{\sqrt{n}}G_n(g)+O(n^{-2})
\end{aligned}
$$

with $\mathcal{E}_\Lambda^{2\prime} := \big\{(\varepsilon_\theta-\varepsilon_{\theta'})^2|\theta,\theta'\in\Theta\big\}$. The entropy of this class is bounded by

$$
\begin{aligned}
&\log N(\varepsilon\|4(F_\Lambda+F'_\Lambda)^2\|_{\tilde Q,2},\mathcal{E}_\Lambda^{2\prime},L_2(\tilde Q))\\
&\leq 2\log N\left(\frac{\varepsilon}{2}\|4(F_\Lambda+F'_\Lambda)\|_{\tilde Q,2},\mathcal{E}'_\Lambda,L_2(\tilde Q)\right)\\
&\leq 2\log N\left(\frac{\varepsilon}{4}\|4F_\Lambda\|_{\tilde Q,2},\mathcal{F}_\Lambda,L_2(\tilde Q)\right)+2\log N\left(\frac{\varepsilon}{4}\|4F'_\Lambda\|_{\tilde Q,2},\mathcal{F}'_\Lambda,L_2(\tilde Q)\right)\\
&\leq 2\sup_Q\log N\left(\varepsilon\|F_\Lambda\|_{Q,2},\mathcal{F}_\Lambda,L_2(Q)\right)+2\sup_{Q'}\log N\left(\varepsilon\|F'_\Lambda\|_{Q',2},\mathcal{F}'_\Lambda,L_2(Q')\right)\\
&\leq 4\sup_Q\log N\left(\left(\frac{\varepsilon}{4}\right)^2\|F_\Lambda\|_{Q,2},\mathcal{F}_\Lambda,L_2(Q)\right)\\
&\leq 8C'_\Lambda\log\left(4\sqrt{C''_\Lambda}/\varepsilon\right).
\end{aligned}
$$

For every σ_C^2 with $\sup_{g\in\mathcal{E}_\Lambda^{2\prime}}\mathbb{E}[g^2]\leq\sigma_C^2\leq\mathbb{E}\big[16(F_\Lambda+F'_\Lambda)^4\big] := G_3<\infty$ and universal constants K and K_2 with probability not less than $1-(1/\log(n))$, it holds

$$
\begin{aligned}
&\sup_{g\in\mathcal{E}_\Lambda^{2\prime}}|G_n(g)|\\
&\leq 2K\left[\left(S\sigma_C^2\log(AG_3^{1/2}/\sigma_C)\right)^{1/2}+SG_3^{1/2}\log(AG_3^{1/2}/\sigma_C)\right]\\
&\quad + K_2(\sigma_C\log(n)^{1/2}+G_3^{1/2}\log(n))
\end{aligned}
$$

$$= O(\log(n))$$

by Lemma 1 in Belloni et al. [9] with $q = 2$, $t = \log(n)$, $A = 4\sqrt{C''_\Lambda}$ and $S = 8C'_\Lambda$.
We conclude

$$\sup_{|\theta-\theta'|\leq 1/n} \mathbb{E}_n\left[(\varepsilon_\theta - \varepsilon_{\theta'})^2\right] = O\left(\frac{\log n}{\sqrt{n}}\right)$$

and

$$\log(p \vee n)^{1/2} \sup_{|\theta-\theta'|\leq 1/n} \mathbb{E}_n\left[(\varepsilon_\theta - \varepsilon_{\theta'})^2\right]^{1/2} = O(\delta_n)$$

since $n^{-1/3}\log(p \vee n) \leq \delta_n$. □

3.9 Inference in Z-Estimation Problems

In this section, we consider a general Z-estimation problem, where the target parameter θ_0 obeys the moment condition

$$\mathbb{E}\Big[\psi\big((Y,X),\theta_0,h_0(\theta_0,X)\big)\Big] = 0.$$

We allow the unknown, high-dimensional nuisance function

$$h_0(\theta,X) = (h_{0,1}(\theta,X),\ldots,h_{0,m}(\theta,X)) \in \mathcal{H}$$

to depend on the target parameter θ. The central theorem is a statement about the asymptotic distribution of an estimator, which solves

$$\left|\mathbb{E}_n\Big[\psi\big((Y,X),\hat{\theta},\hat{h}_0(\hat{\theta},X)\big)\Big]\right| = \inf_{\theta\in\Theta}\left|\mathbb{E}_n\Big[\psi\big((Y,X),\theta,\hat{h}_0(\theta,X)\big)\Big]\right| + \epsilon_n, \tag{3.23}$$

where $\epsilon_n = o(n^{-1/2})$ is the numerical tolerance. We need a more general form of the conditions in Section 3.3.

Assumptions **C1-C7**.
The following assumptions hold uniformly in $n \geq n_0$ and $P \in \mathcal{P}_n$:

C1 The true parameter θ_0 obeys the moment condition

$$\mathbb{E}\Big[\psi\big((Y,X),\theta_0,h_0\big)\Big] = 0.$$

C2 The map $(\theta,h) \mapsto \mathbb{E}[\psi((X,Y),\theta,h)]$ is twice continuously Gateaux-differentiable on $\Theta \times \mathcal{H}$.

C3 Let $\tilde{\mathcal{H}} = \{\tilde{h} : \Theta \times \mathcal{X} \mapsto \mathbb{R}^m\} \subseteq \mathcal{H}$ be a suitable set of functions. For every $\theta \in \Theta$, we have a nuisance function estimator $\hat{h}(\theta)$ and a set of functions $\tilde{\mathcal{H}}(\theta) = \{\tilde{h} : \mathcal{X} \mapsto \mathbb{R}^m : \tilde{h}(x) = \tilde{h}(\theta,x) \in \tilde{\mathcal{H}}\}$ with $P(\hat{h}(\theta) \in \tilde{\mathcal{H}}(\theta)) = 1 - o(1)$, where $\tilde{\mathcal{H}}(\theta)$ contains $h_0(\theta,\cdot)$ and is constrained by conditions given below.

C4 For all $\tilde{h} \in \tilde{\mathcal{H}}$, the score ψ obeys the Neyman orthogonality property

$$D_0[\tilde{h} - h_0] = 0.$$

C5 For all $\theta \in \Theta$, the class of functions

$$\Psi(\theta) = \Big\{(y,x) \mapsto \psi\big((y,x),\theta,\tilde{h}(\theta,x)\big), \tilde{h} \in \tilde{\mathcal{H}}(\theta)\Big\}$$

has a measurable envelope $\bar{\psi} \geq \sup_{\psi\in\Psi(\theta)} |\psi|$ independent from θ, such that for some $q \geq 4$

$$\mathbb{E}\Big[(\bar{\psi}(Y,X))^q\Big] \leq C.$$

The class $\Psi(\theta)$ is pointwise measurable and, uniformly for all $\theta \in \Theta$, it holds

$$\sup_Q \log N(\varepsilon||\bar{\psi}||_{Q,2}, \Psi(\theta), L_2(Q)) \leq C_1 s \log\left(\frac{C_2(p\vee n)}{\varepsilon}\right)$$

with C_1 and C_2 being independent from θ.

C6 (i) For a positive sequence $\rho_n \searrow 0$ with

$$n^{-1/2}\left(s^{\frac{1}{2}}\log(p\vee n)^{\frac{1}{2}} + n^{-\frac{1}{2}+\frac{1}{q}} s\log(p\vee n)\right) = O(\rho_n),$$

we have

$$\sup_{\theta\in\Theta,\tilde{h}\in\tilde{\mathcal{H}}(\theta)} |\mathbb{E}[\psi((Y,X)),\theta,h_0(\theta,X)] - \mathbb{E}[\psi((Y,X)),\theta,\tilde{h}(\theta,X)]| \leq C\rho_n.$$

(ii) We define

$$\sup \mathbb{E}\Big[\Big(\psi\big((Y,X),\theta,\tilde{h}(\theta,X)\big) - \psi\big((Y,X),\theta_0,h_0(\theta_0,X)\big)\Big)^2\Big]^{1/2} =: r_n,$$

where the supremum is taken over all θ with $|\theta-\theta_0| \leq C\rho_n$ and $\tilde{h} \in \tilde{\mathcal{H}}$, meaning

$$\sup \equiv \sup_{\theta:|\theta-\theta_0|\leq C\rho_n, \tilde{h}\in\tilde{\mathcal{H}}(\theta)},$$

and it holds $r_n s^{\frac{1}{2}} \log\left(\frac{(p\vee n)}{r_n}\right)^{\frac{1}{2}} + n^{-\frac{1}{2}+\frac{1}{q}} s\log\left(\frac{(p\vee n)}{r_n}\right) = o(1)$ with q from Assumption C5.

(iii) It holds

$$\sup\left|\partial_r^2\left\{\mathbb{E}\Big[\psi\big((Y,X),\theta_0 + r(\theta-\theta_0), h_0 + r(\tilde{h}-h_0)\big)\Big]\right\}\right| = o(n^{-1/2}),$$

where

$$\sup \equiv \sup_{r\in(0,1),\theta:|\theta-\theta_0|\leq C\rho_n,\tilde{h}\in\tilde{\mathcal{H}}(\theta)}.$$

C7 For $h \in \tilde{\mathcal{H}}$, the function

$$\theta \mapsto \mathbb{E}\Big[\psi\big((Y,X),\theta,h(\theta,X)\big)\Big]$$

is differentiable in a neighborhood of θ_0 and, for all $\theta \in \Theta$, the identification relation

$$2|\mathbb{E}[\psi((Y,X)),\theta,h_0(\theta,X)]| \geq |\Gamma(\theta-\theta_0)| \wedge c_0$$

is satisfied with

$$\Gamma := \partial_\theta \mathbb{E}\Big[\psi\big((Y,X),\theta_0,h_0(\theta_0,X)\big)\Big] > c_1.$$

Since the nuisance functions depend on the target parameter, the conditions ensure that they can be estimated uniformly over all θ with a sufficiently fast rate.

Theorem 8. *Under the Assumptions C1-C7, an estimator $\hat{\theta}$ of the form (3.23) obeys*

$$n^{\frac{1}{2}}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma),$$

where

$$\Sigma := \mathbb{E}\Big[\Gamma^{-2}\psi^2\big((Y,X), \theta_0, h_0(\theta_0, X)\big)\Big]$$

with $\Gamma = \partial_\theta \mathbb{E}\Big[\psi\big((Y,X), \theta_0, h_0(\theta_0, X)\big)\Big]$.

Comment 3.9.1.
This setting and the theorem is almost identical to Assumption 3.4 and Theorem 3.3 in Chernozhukov et al. [35]. Their theorem holds for dependent nuisance functions, but the entropy condition may be hard to verify in some settings:
Suppose the unknown nuisance function h_0 is a linear function of X, where the coefficients $\beta_0(\theta)$ ($\|\beta_0(\theta)\|_0 \leq s$ for all θ) are dependent on the target parameter. If $h_0(\theta, X) = X\beta_0(\theta)$ is estimated by the Lasso estimator, the uniform covering entropy of

$$\mathcal{F}_h := \Big\{\psi\big(\cdot, \theta, h(\theta, \cdot)\big), \theta \in \Theta\Big\}$$

may not fulfill the desired condition. This is because the uniform covering entropy of the class

$$\mathcal{H} := \Big\{h(\theta, \cdot) : \mathcal{X} \to \mathbb{R} | h(\theta, X) = \beta(\theta)X, \|\beta(\theta)\|_0 \leq s, \theta \in \Theta\Big\}$$

can not be bounded by representing the class as the union over sets with a bounded VC-index (see, e.g., in Belloni et al. [9]) since the indices which differ from zero may vary for each θ.
In their example, the estimation of the average treatment effect, this problem does not occur, since the estimated nuisance functions do not depend on the target parameter. To bypass this, we rely on a slightly different set of entropy conditions, which enables us to restrict the entropy of the classes uniformly over all $\theta \in \Theta$.

Proof. We are using similar arguments as in proof of Theorem 2 from Belloni et al. [9]. We prove our theorem under an arbitrary sequence $P = P_n \in \mathcal{P}_n$. Therefore, the dependence of P on n can be suppressed. Let ρ_n be a positive sequence converging to zero.

Step 1.
Let $\tilde{\theta}$ be an arbitrary estimator fulfilling $|\tilde{\theta} - \theta_0| \leq C\rho_n$ with probability $1 - o(1)$. We aim to prove that with probability $1 - o(1)$

$$\begin{aligned}
&\mathbb{E}_n\Big[\psi\big((Y,X), \tilde{\theta}, \hat{h}(\tilde{\theta}, X)\big)\Big] \\
&= \mathbb{E}_n\Big[\psi\big((Y,X), \theta_0, h_0(\theta_0, X)\big)\Big] \\
&\quad + \underbrace{\partial_\theta \mathbb{E}\Big[\psi\big((Y,X), \theta_0, h_0(\theta_0, X)\big)\Big]}_{:=\Gamma}(\tilde{\theta} - \theta_0) + o(n^{-\frac{1}{2}}).
\end{aligned}$$

By Assumption C1, we can expand the term

$$\mathbb{E}_n\Big[\psi\big((Y,X), \tilde{\theta}, \hat{h}(\tilde{\theta}, X)\big)\Big]$$

$$
\begin{aligned}
&= \mathbb{E}_n\Big[\psi\big((Y,X),\tilde{\theta},\hat{h}(\tilde{\theta},X)\big)\Big] + \underbrace{\mathbb{E}\Big[\psi\big((Y,X),\theta_0,h_0(\theta_0,X)\big)\Big]}_{=0} \\
&\quad + \mathbb{E}_n\Big[\psi\big((Y,X),\theta_0,h_0(\theta_0,X)\big)\Big] - \mathbb{E}_n\Big[\psi\big((Y,X),\theta_0,h_0(\theta_0,X)\big)\Big] \\
&\quad + \mathbb{E}\Big[\psi\big((Y,X),\tilde{\theta},\hat{h}(\tilde{\theta},X)\big)\Big] - \mathbb{E}\Big[\psi\big((Y,X),\tilde{\theta},\hat{h}(\tilde{\theta},X)\big)\Big] \\
&= \underbrace{\mathbb{E}_n\Big[\psi\big((Y,X),\theta_0,h_0(\theta_0,X)\big)\Big]}_{:=I} + \underbrace{\mathbb{E}\Big[\psi\big((Y,X),\tilde{\theta},\hat{h}(\tilde{\theta},X)\big)\Big]}_{:=II} \\
&\quad + \underbrace{\mathbb{E}_n\Big[\psi\big((Y,X),\tilde{\theta},\hat{h}(\tilde{\theta},X)\big)\Big] - \mathbb{E}\Big[\psi\big((Y,X),\tilde{\theta},\hat{h}(\tilde{\theta},X)\big)\Big]}_{:=III} \\
&\quad - \Bigg(\underbrace{\mathbb{E}_n\Big[\psi\big((Y,X),\theta_0,h_0(\theta_0,X)\big)\Big] - \mathbb{E}\Big[\psi\big((Y,X),\theta_0,h_0(\theta_0,X)\big)\Big]}_{:=IV}\Bigg) \\
&= I + II + III - IV.
\end{aligned}
$$

Considering the last two terms, we have with probability $1 - o(1)$

$$
\begin{aligned}
&n^{\frac{1}{2}}\Big(III - IV\Big) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Bigg(\psi\big((Y,X),\tilde{\theta},\hat{h}(\tilde{\theta},X)\big) - \psi\big((Y,X),\theta_0,h_0(\theta_0,X)\big) \\
&\quad - \Big(\mathbb{E}\Big[\psi\big((Y,X),\tilde{\theta},\hat{h}(\tilde{\theta},X)\big)\Big] - \mathbb{E}\Big[\psi\big((Y,X),\theta_0,h_0(\theta_0,X)\big)\Big]\Big)\Bigg) \\
&\leq \sup_{\theta:|\theta-\theta_0|\leq C\rho_n}\Bigg|\Bigg[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Bigg(\psi\big((Y,X),\theta,\hat{h}(\theta,X)\big) - \psi\big((Y,X),\theta_0,h_0(\theta_0,X)\big) \\
&\quad - \Big(\mathbb{E}\Big[\psi\big((Y,X),\theta,\hat{h}(\theta,X)\big)\Big] - \mathbb{E}\Big[\psi\big((Y,X),\theta_0,h_0(\theta_0,X)\big)\Big]\Big)\Bigg)\Bigg]\Bigg| \\
&\leq \sup_{\theta:|\theta-\theta_0|\leq C\rho_n}\left(\sup_{f\in\Psi'(\theta)}|G_n(f)|\right)
\end{aligned}
$$

with

$$
\Psi'(\theta) = \Big\{(y,x)\mapsto \psi\big((y,x),\theta,\tilde{h}(\theta,x)\big) - \psi\big((y,x),\theta_0,h_0(\theta_0,x)\big), \tilde{h}\in\tilde{\mathcal{H}}(\theta)\Big\}
$$

and envelope $2\bar{\psi}$. Here, we used Assumption C5 and that we have $\hat{h}(\theta,X), h_0(\theta,X)\in\tilde{\mathcal{H}}(\theta)$ for all $\theta\in\Theta$ by Assumption C3 with probability $1-o(1)$. We note that

$$
\sup_Q \log N(\varepsilon||2\bar{\psi}||_{Q,2}, \Psi'(\theta), L_2(Q)) \leq C_1 s\log\left(\frac{C_2(p\vee n)}{\varepsilon}\right)
$$

for constants C_1 and C_2 beeing independent from θ. We want to apply Lemma 1 from Belloni et al. [9]. By Assumption C6, we have

$$
\begin{aligned}
&\sup_{\theta:|\theta-\theta_0|\leq C\rho_n, f\in\Psi'(\theta)}\mathbb{E}\Big[f^2\big((Y,X)\big)\Big] \\
&= \sup_{\theta:|\theta-\theta_0|\leq C\rho_n,\tilde{h}\in\tilde{\mathcal{H}}(\theta)}\mathbb{E}\Big[\Big(\psi\big((Y,X),\theta,\tilde{h}(\theta,X)\big) - \psi\big((Y,X),\theta_0,h_0(\theta_0,X)\big)\Big)^2\Big] \\
&=: r_n^2
\end{aligned}
$$

with $r_n s^{\frac{1}{2}}\log\left(\frac{p\vee n}{r_n}\right)^{\frac{1}{2}} + n^{-\frac{1}{2}+\frac{1}{q}}s\log\left(\frac{p\vee n}{r_n}\right) = o(1)$.

Choosing $\sigma_n^2 = r_n^2$ and $\max_{q\in\{2,4\}} \mathbb{E}[(\bar{\psi}(Y,X))^q] \leq C$, the first inequality of Lemma 1 in Belloni et al. [9] implies

$$\begin{aligned}
&\mathbb{E}\left[\sup_{f\in\Psi'(\theta)} |G_n(f)|\right] \\
&\leq K\left[\left(C_1 s\sigma_n^2 \log\left(\frac{C_2(p\vee n)C^{\frac{1}{2}}}{\sigma_n}\right)\right)^{\frac{1}{2}} + n^{-\frac{1}{2}+\frac{1}{q}} C_1 s C^{\frac{1}{q}} \log\left(\frac{C_2(p\vee n)C^{\frac{1}{2}}}{\sigma_n}\right)\right] \\
&\leq K'\left(\sigma_n\left(s\log\left(\frac{p\vee n}{\sigma_n}\right)\right)^{\frac{1}{2}} + n^{-\frac{1}{2}+\frac{1}{q}} s\log\left(\frac{p\vee n}{\sigma_n}\right)\right).
\end{aligned}$$

Applying the second part of Lemma 1 with $t = \log(n)$, we obtain

$$\begin{aligned}
n^{\frac{1}{2}}|III - IV| &\leq \sup_{\theta:|\theta-\theta_0|\leq C\rho_n}\Big(\sup_{f\in\Psi'(\theta)} |G_n(f)|\Big) \\
&\leq \sup_{\theta:|\theta-\theta_0|\leq C\rho_n}\Bigg(2\mathbb{E}\Big[\sup_{f\in\Psi'(\theta)} |G_n(f)|\Big] \\
&\quad + K_q\Big(\sigma_n \log(n)^{\frac{1}{2}} + n^{-\frac{1}{2}+\frac{1}{q}} C^{\frac{1}{q}} \log(n)\Big)\Bigg) \\
&\leq K_q'\Bigg(\sigma_n\Bigg(s\log\Big(\frac{p\vee n}{\sigma_n}\Big)\Bigg)^{\frac{1}{2}} + n^{-\frac{1}{2}+\frac{1}{q}} s\log\Big(\frac{p\vee n}{\sigma_n}\Big)\Bigg) \\
&= o(1).
\end{aligned}$$

Now, we expand the term II. Let $\tilde{h} \in \tilde{\mathcal{H}}$ and $\tilde{\theta} \in \Theta$.
By Taylor expansion of the function $r \mapsto \mathbb{E}\Big[\psi\big((Y,X), \theta_0 + r(\tilde{\theta}-\theta_0), h_0 + r(\tilde{h}-h_0)\big)\Big]$ and Assumption C2, we have

$$\begin{aligned}
&\mathbb{E}\Big[\psi\big((Y,X),\tilde{\theta},\tilde{h}\big)\Big] \\
&= \mathbb{E}\Big[\psi\big((Y,X),\theta_0,h_0\big)\Big] \\
&\quad + \partial_r\Big\{\mathbb{E}\Big[\psi\big((Y,X),\theta_0 + r(\tilde{\theta}-\theta_0), h_0 + r(\tilde{h}-h_0)\big)\Big]\Big\}\Big|_{r=0} \\
&\quad + \frac{1}{2}\partial_r^2\Big\{\mathbb{E}\Big[\psi\big((Y,X),\theta_0 + r(\tilde{\theta}-\theta_0), h_0 + r(\tilde{h}-h_0)\big)\Big]\Big\}\Big|_{r=\bar{r}}
\end{aligned}$$

for some $\bar{r} \in (0,1)$. Due to the orthogonality condition in C4, it holds

$$\begin{aligned}
&\partial_r\Big\{\mathbb{E}\Big[\psi\big((Y,X),\theta_0 + r(\tilde{\theta}-\theta_0), h_0 + r(\tilde{h}-h_0)\big)\Big]\Big\}\Big|_{r=0} \\
=&\partial_r\Big\{\mathbb{E}\Big[\psi\big((Y,X),\theta_0 + r(\tilde{\theta}-\theta_0), h_0 + r(\tilde{h}-h_0)\big)\Big]\Big\}\Big|_{r=0} - D_0[\tilde{h}-h_0] \\
=&\partial_r\Big\{\mathbb{E}\Big[\psi\big((Y,X),\theta_0 + r(\tilde{\theta}-\theta_0), h_0 + r(\tilde{h}-h_0)\big)\Big] \\
&- \mathbb{E}\Big[\psi\big((Y,X),\theta_0, h_0 + r(\tilde{h}-h_0)\big)\Big]\Big\}\Big|_{r=0} \\
=&\partial_r\Big\{r(\tilde{\theta}-\theta_0)\partial_\theta\mathbb{E}\Big[\psi\big((Y,X),\theta, h_0 + r(\tilde{h}-h_0)\big)\Big]\Big|_{\theta\in[\theta_0,\theta_0+r(\tilde{\theta}-\theta_0)]}\Big\}\Big|_{r=0} \\
=&(\tilde{\theta}-\theta_0)\partial_\theta\mathbb{E}\Big[\psi\big((Y,X),\theta_0,h_0\big)\Big].
\end{aligned}$$

By Assumption C6, we have

$$\left|\partial_r^2\left\{\mathbb{E}\Big[\psi\big((Y,X),\theta_0+r(\tilde{\theta}-\theta_0),h_0+r(\tilde{h}-h_0)\big)\Big]\right\}\Big|_{r=\bar{r}}\right| = o(n^{-1/2})$$

and therefore

$$\mathbb{E}\Big[\psi\big((Y,X),\tilde{\theta},\tilde{h}\big)\Big] = \Gamma(\tilde{\theta}-\theta_0) + o(n^{-1/2}).$$

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

$$\begin{aligned}\mathbb{E}_n\Big[\psi\big((Y,X),\tilde{\theta},\hat{h}(\tilde{\theta},X)\big)\Big] &= \mathbb{E}_n\Big[\psi\big((Y,X),\theta_0,h_0(\theta_0,X)\big)\Big] \\ &\quad + \Gamma(\tilde{\theta}-\theta_0) + o(n^{-\frac{1}{2}}).\end{aligned}$$

Step 2.
We want to prove that with probability $1-o(1)$

$$\inf_{\theta\in\Theta}\left|\mathbb{E}_n\Big[\psi\big((Y,X),\theta,\hat{h}(\theta,X)\big)\Big]\right| = o(n^{-\frac{1}{2}}).$$

Define

$$\theta^* := \theta_0 - \Gamma^{-1}\mathbb{E}_n\Big[\psi\big((Y,X),\theta_0,h_0(\theta_0,X)\big)\Big].$$

By the central limit theorem, it follows directly

$$|\theta^*-\theta_0| = |\Gamma^{-1}|\Big|\underbrace{\mathbb{E}_n\Big[\psi\big((Y,X),\theta_0,h_0(\theta_0,X)\big)\Big]}_{=O\left(n^{-\frac{1}{2}}\right)}\Big| \leq C\rho_n.$$

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

$$\inf_{\theta\in\Theta}\left|\mathbb{E}_n\Big[\psi\big((Y,X),\theta,\hat{h}(\theta,X)\big)\Big]\right| \leq \left|\mathbb{E}_n\Big[\psi\big((Y,X),\theta^*,\hat{h}(\theta^*,X)\big)\Big]\right| = o(n^{-\frac{1}{2}})$$

by inserting the definition of θ^*.

Step 3.
We aim to show that the estimated $\hat{\theta}$ converges towards θ_0, meaning with probability $1-o(1)$

$$|\hat{\theta}-\theta_0| \leq C\rho_n.$$

By definition of $\hat{\theta}$ and Step 2, we have

$$\left|\mathbb{E}_n\Big[\psi\big((Y,X),\hat{\theta},\hat{h}(\hat{\theta},X)\big)\Big]\right| = o(n^{-\frac{1}{2}}).$$

Since $\hat{h}(\theta)\in\tilde{\mathcal{H}}(\theta)$ with probability $1-o(1)$ for all $\theta\in\Theta$, we have

$$\begin{aligned}&\sup_{\theta\in\Theta}\left|\mathbb{E}_n\Big[\psi\big((Y,X),\theta,\hat{h}(\theta,X)\big)\Big] - \mathbb{E}\Big[\psi\big((Y,X),\theta,\hat{h}(\theta,X)\big)\Big]\right| \\ &\leq \sup_{\theta\in\Theta}\left(n^{-\frac{1}{2}}\sup_{g\in\Psi(\theta)}|G_n(g)|\right) \\ &= O\left(n^{-1/2}\left(s^{\frac{1}{2}}\log(p\vee n)^{\frac{1}{2}} + n^{-\frac{1}{2}+\frac{1}{q}}s\log(p\vee n)\right)\right),\end{aligned}$$

where we used Lemma 1 in Belloni et al. [9] and $\mathbb{E}\Big[(\bar{\psi}(Y,X))^2\Big] \leq C$ as in Step 1. Combining this with the triangle inequality, we obtain

$$\begin{aligned}
&\big|\mathbb{E}\Big[\psi\big((Y,X),\hat{\theta},h_0(\hat{\theta},X)\big)\Big]\big| \\
&\leq \sup_{\theta\in\Theta,\tilde{h}\in\tilde{\mathcal{H}}(\theta)} |\mathbb{E}[\psi((Y,X)),\theta,h_0(\theta,X)] - \mathbb{E}[\psi((Y,X)),\theta,\tilde{h}(\theta,X)]| \\
&+ \sup_{\theta\in\Theta,\tilde{h}(\theta)\in\tilde{\mathcal{H}}(\theta)} \Big|\mathbb{E}_n\Big[\psi\big((Y,X),\theta,\tilde{h}(\theta,X)\big)\Big] - \mathbb{E}\Big[\psi\big((Y,X),\theta,\tilde{h}(\theta,X)\big)\Big]\Big| \\
&+ \Big|\mathbb{E}_n\Big[\psi\big((Y,X),\hat{\theta},\hat{h}(\hat{\theta},X)\big)\Big]\Big| \leq C\rho_n
\end{aligned}$$

by Assumption C6. Hence, it follows by Assumption C7 with probability $1-o(1)$

$$|\Gamma(\hat{\theta}-\theta_0)| \wedge c_0 \leq 2\big|\mathbb{E}[\psi((Y,X)),\hat{\theta},h_0(\hat{\theta},X)]\big| \leq C\rho_n$$

and dividing by $\Gamma > c_1$ gives the claim of this step.

Step 4.
Due to Step 3, we are able to use Step 1 for the estimated parameter and obtain with probability $1-o(1)$

$$\begin{aligned}
\mathbb{E}_n\Big[\psi\big((Y,X),\hat{\theta},\hat{h}(\hat{\theta},X)\big)\Big] &= \mathbb{E}_n\Big[\psi\big((Y,X),\theta_0,h_0(\theta_0,X)\big)\Big] \\
&\quad + \Gamma(\hat{\theta}-\theta) + o\left(n^{-\frac{1}{2}}\right).
\end{aligned}$$

By Step 2, we have

$$\begin{aligned}
&\Gamma(\hat{\theta}-\theta) \\
&= \underbrace{\mathbb{E}_n\Big[\psi\big((Y,X),\hat{\theta},\hat{h}(\hat{\theta},X)\big)\Big]}_{=o\left(n^{-\frac{1}{2}}\right)} - \mathbb{E}_n\Big[\psi\big((Y,X),\theta_0,h_0(\theta_0,X)\big)\Big] + o\left(n^{-\frac{1}{2}}\right) \\
&= -\Big(\mathbb{E}_n\Big[\psi\big((Y,X),\theta_0,h_0(\theta_0,X)\big)\Big] - \underbrace{\mathbb{E}\Big[\psi\big((Y,X),\theta_0,h_0(\theta_0,X)\big)\Big]}_{=0}\Big) \\
&\quad + o\left(n^{-\frac{1}{2}}\right).
\end{aligned}$$

Using the central limit theorem, we get with probability $1-o(1)$

$$\begin{aligned}
&n^{\frac{1}{2}}(\hat{\theta}-\theta) \\
&= \underbrace{-\Gamma^{-1}n^{\frac{1}{2}}\Big(\mathbb{E}_n\Big[\psi\big((Y,X),\theta_0,h_0(\theta_0,X)\big)\Big] - \mathbb{E}\Big[\psi\big((Y,X),\theta_0,h_0(\theta_0,X)\big)\Big]\Big)}_{\xrightarrow{\mathcal{D}}\mathcal{N}(0,\Sigma)} \\
&\quad + o(1)
\end{aligned}$$

with

$$\Sigma := \mathrm{Var}\Big(\Gamma^{-1}\psi\big((Y,X),\theta_0,h_0(\theta_0,X)\big)\Big) = \mathbb{E}\Big[\Gamma^{-2}\psi^2\big((Y,X),\theta_0,h_0(\theta_0,X)\big)\Big].$$

□

3.10 Additional Simulations

This section provides additional simulation studies.

3.10.1 Approximately Sparse Setting

In the approximately sparse setting, the coefficients are set to

$$\beta_{\theta_0,j} = \begin{cases} 1 & \text{for } j \leq s \\ \frac{1}{(j-s+1)^2} & \text{for } j > s. \end{cases}$$

The other parameters are chosen as in the simulations in the main text (Section 3.4), but to restrict the calculation time we focus on the correlation structure $\Sigma_1^{(X)}$. The results for Box-Cox transformations ($\theta_0 = 0$) are presented in Table 3.9 and the results for Yeo-Johnson power transformations ($\theta_0 = 1$) in Table 3.10. We note that the case $p = 20$ and $s = 20$ is not contained in both tables since these settings coincide with the exactly sparse setting. The results are similar to the exactly sparse setting and the acceptance rate is close to the nominal level.

n	p	s	SNR	Estimator	Acceptance rate	MAE	rel. MSE
100	20	5	1.0	-0.00072435	0.946	0.0147	1.2232
100	20	5	3.0	0.00065075	0.944	0.0139	1.0325
100	20	10	1.0	-0.00035129	0.950	0.0108	1.6230
100	20	10	3.0	-0.00068912	0.940	0.0114	1.4992
100	50	5	1.0	-0.00052439	0.952	0.0141	1.3855
100	50	5	3.0	-0.00039227	0.970	0.0136	1.1084
100	50	10	1.0	0.00037004	0.952	0.0103	1.7093
100	50	10	3.0	0.00040427	0.952	0.0111	1.7451
100	50	20	1.0	0.00024774	0.948	0.0071	1.8708
100	50	20	3.0	0.00032668	0.946	0.0092	3.3351
100	100	5	1.0	0.00115031	0.958	0.0143	1.4935
100	100	5	3.0	-0.00002014	0.976	0.0147	1.3083
100	100	10	1.0	0.00066524	0.952	0.0105	1.8751
100	100	10	3.0	-0.00072896	0.966	0.0119	2.0950
100	100	20	1.0	-0.00033613	0.936	0.0079	1.8906
100	100	20	3.0	-0.00003507	0.962	0.0091	3.4120
100	200	5	1.0	-0.00067739	0.976	0.0151	1.4985
100	200	5	3.0	-0.00000964	0.966	0.0151	1.1812
100	200	10	1.0	-0.00071120	0.952	0.0105	1.8505
100	200	10	3.0	-0.00149334	0.980	0.0130	2.2642
100	200	20	1.0	-0.00103713	0.946	0.0080	1.7758
100	200	20	3.0	-0.00008740	0.962	0.0094	3.3879
200	20	5	1.0	-0.00104238	0.924	0.0095	0.9621
200	20	5	3.0	0.00119542	0.928	0.0098	0.9285
200	20	10	1.0	0.00034137	0.942	0.0067	1.1898
200	20	10	3.0	-0.00036637	0.932	0.0068	1.0075
200	50	5	1.0	0.00029033	0.938	0.0100	1.0325
200	50	5	3.0	0.00128785	0.946	0.0095	0.9765
200	50	10	1.0	0.00027922	0.948	0.0069	1.2999
200	50	10	3.0	0.00015796	0.950	0.0067	1.0655
200	50	20	1.0	0.00014660	0.932	0.0053	1.7986
200	50	20	3.0	0.00027307	0.948	0.0054	1.8437
200	100	5	1.0	0.00033937	0.944	0.0087	1.1958
200	100	5	3.0	-0.00000796	0.946	0.0090	1.1458
200	100	10	1.0	0.00027762	0.946	0.0069	1.4877
200	100	10	3.0	0.00127725	0.946	0.0070	1.2023
200	100	20	1.0	-0.00028952	0.944	0.0051	1.8415
200	100	20	3.0	0.00049090	0.954	0.0060	2.0796
200	200	5	1.0	0.00045282	0.926	0.0101	1.0675
200	200	5	3.0	0.00005082	0.952	0.0096	0.9836
200	200	10	1.0	0.00062688	0.944	0.0076	1.4918
200	200	10	3.0	0.00052976	0.954	0.0070	1.1149
200	200	20	1.0	0.00011013	0.940	0.0051	1.7536
200	200	20	3.0	-0.00068245	0.966	0.0057	2.2181
200	500	5	1.0	0.00058465	0.906	0.0104	1.2234
200	500	5	3.0	-0.00013389	0.926	0.0097	1.1230
200	500	10	1.0	0.00012013	0.942	0.0075	1.6718
200	500	10	3.0	0.00021128	0.944	0.0070	1.2930
200	500	20	1.0	0.00019671	0.944	0.0052	2.0953
200	500	20	3.0	-0.00008077	0.974	0.0060	3.0412

Table 3.9: **Additional simulations for Box-Cox transformations.**

n	p	s	SNR	Estimator	Acceptance rate	MAE	rel. MSE
100	20	5	1.0	1.00056735	0.962	0.0418	1.2250
100	20	5	3.0	0.99698158	0.932	0.0410	1.0245
100	20	10	1.0	0.99986806	0.942	0.0382	1.6179
100	20	10	3.0	0.99983138	0.960	0.0373	1.5512
100	50	5	1.0	1.00426870	0.950	0.0421	1.3767
100	50	5	3.0	1.00276705	0.962	0.0362	1.1167
100	50	10	1.0	0.99942878	0.950	0.0391	1.7085
100	50	10	3.0	1.00116016	0.966	0.0357	1.7955
100	50	20	1.0	0.99842764	0.946	0.0321	1.8739
100	50	20	3.0	0.99724029	0.966	0.0355	3.4082
100	100	5	1.0	1.00384643	0.942	0.0462	1.4958
100	100	5	3.0	0.99651381	0.962	0.0382	1.2919
100	100	10	1.0	0.99891940	0.950	0.0385	1.8919
100	100	10	3.0	0.99564080	0.968	0.0361	2.1218
100	100	20	1.0	0.99995916	0.950	0.0326	1.8880
100	100	20	3.0	1.00184681	0.950	0.0374	3.4484
100	200	5	1.0	1.00045974	0.978	0.0424	1.5090
100	200	5	3.0	0.99536783	0.968	0.0383	1.1875
100	200	10	1.0	0.99877481	0.972	0.0377	1.8432
100	200	10	3.0	1.00074576	0.962	0.0398	2.3258
100	200	20	1.0	0.99797088	0.952	0.0346	1.7767
100	200	20	3.0	1.00079835	0.976	0.0352	3.3919
200	20	5	1.0	0.99985307	0.928	0.0292	0.9671
200	20	5	3.0	0.99996887	0.962	0.0255	0.9295
200	20	10	1.0	1.00053756	0.924	0.0259	1.1881
200	20	10	3.0	0.99887849	0.942	0.0223	1.0106
200	50	5	1.0	1.00122936	0.916	0.0312	1.0292
200	50	5	3.0	0.99870047	0.944	0.0258	0.9795
200	50	10	1.0	1.00219511	0.946	0.0251	1.2824
200	50	10	3.0	0.99885312	0.938	0.0237	1.0708
200	50	20	1.0	0.99941435	0.936	0.0224	1.7849
200	50	20	3.0	0.99886322	0.960	0.0224	1.8152
200	100	5	1.0	1.00073382	0.932	0.0301	1.1934
200	100	5	3.0	0.99614218	0.926	0.0268	1.1418
200	100	10	1.0	1.00113064	0.936	0.0271	1.5229
200	100	10	3.0	1.00177678	0.956	0.0214	1.1959
200	100	20	1.0	1.00030201	0.948	0.0240	1.8491
200	100	20	3.0	0.99925082	0.948	0.0230	2.1054
200	200	5	1.0	1.00129852	0.938	0.0288	1.0590
200	200	5	3.0	0.99850564	0.940	0.0275	0.9827
200	200	10	1.0	0.99588019	0.938	0.0262	1.5113
200	200	10	3.0	1.00066971	0.954	0.0233	1.1361
200	200	20	1.0	0.99697192	0.940	0.0239	1.7520
200	200	20	3.0	1.00029713	0.962	0.0216	2.2585
200	500	5	1.0	1.00007598	0.962	0.0279	1.2183
200	500	5	3.0	0.99890970	0.936	0.0276	1.1211
200	500	10	1.0	1.00145277	0.936	0.0272	1.6619
200	500	10	3.0	0.99971989	0.962	0.0226	1.2654
200	500	20	1.0	0.99981870	0.940	0.0242	2.1018
200	500	20	3.0	0.99695433	0.964	0.0231	2.9583

Table 3.10: **Additional simulations for Yeo-Johnson transformations.**

3.10.2 Non-Normal Errors

In this section, we evaluate the performance of our proposed method under non-normal errors. The same simulation is run as in Section 3.4 with $n = 100$ observations, but we simulate errors according to a t-distribution with df degrees of freedom

$$\varepsilon \sim t(df).$$

We focus on the correlation structure $\Sigma_1^{(X)}$ and the Box-Cox transformations ($\theta_0 = 0$). We set $s = 20$ and vary the degrees of freedom.

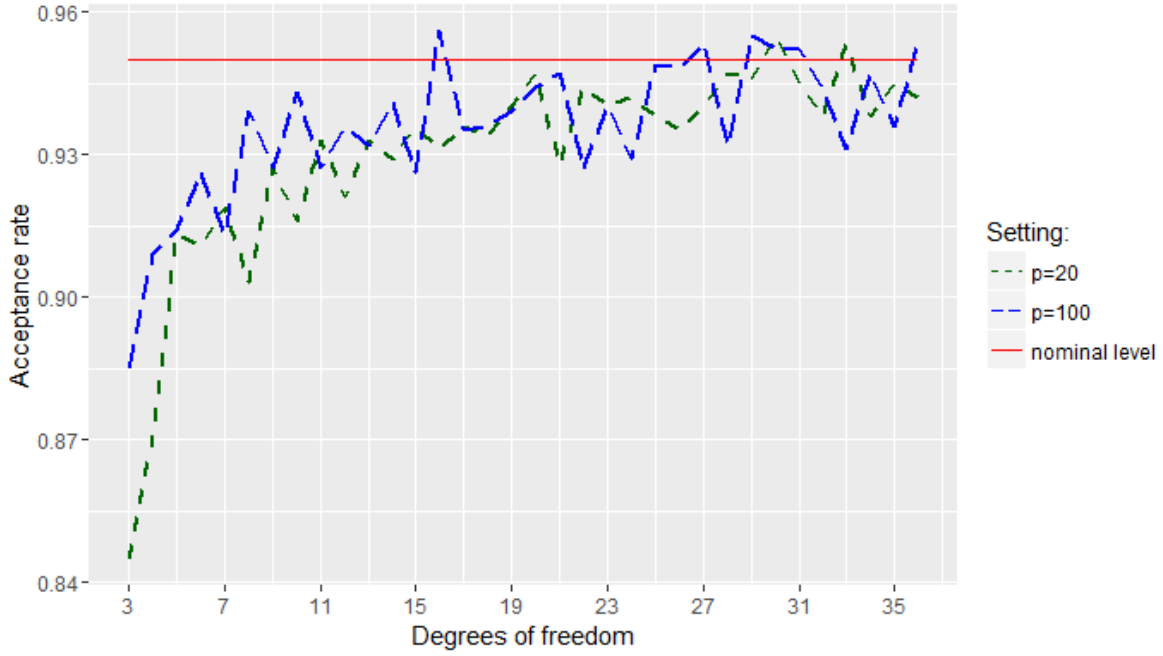


Figure 3.7: **Coverage for an increasing number of degrees of freedom.**

Figure 3.7 displays the effect of non-normal errors on the coverage. If the deviation from the normal distribution is high (low number of degrees of freedom), the coverage differs largely from the nominal level of 95%. With an increasing number of regressors the coverage gradually approaches the nominal level.

Chapter 4

Uniform Inference in High-Dimensional Generalized Additive Models

4.1 Introduction

Nonparametric regression allows the estimation of the relationship f between a target variable Y and input variables $X = (X_1, \ldots, X_p)^T$ without imposing (strong) functional assumptions:

$$Y = f(X_1, \ldots, X_p) + \varepsilon,$$

where ε denotes the random error term satisfying $\mathbb{E}[\varepsilon|X] = 0$. When p is large, estimation of the regression function $f(X_1, \ldots, X_p)$ is practically infeasible due to the curse of dimensionality. One approach to overcome this challenge that has been very popular in statistics and econometrics is to impose additional additive structure leading to generalized additive models (GAMs):

$$Y = \alpha + f_1(X_1) + \ldots + f_p(X_p) + \varepsilon, \tag{4.1}$$

where α is a constant and $f_j(\cdot), j = 1, \ldots, p$, are smooth univariate functions. The idea of GAMs can be traced back to Friedman and Stuetzle [49], Stone [90] and Hastie and Tibshirani [54]. Estimation and inference in the low-dimensional setting with fixed p has been analyzed widely in the literature. For an introduction to GAMs, we refer to the textbook treatments by Hastie and Tibshirani [54] and Wood [101]. In recent years, considerable progress has been made in understanding and analyzing GAMs in high-dimensional settings (i.e., when the number of components can grow with the sample size) under the additional assumption that only a small subset of the components of size s are nonzero. In high-dimensional settings, the focus has been on theoretical results on the estimation rate of sparse additive models. This has been analyzed in Sardy and Tseng [89], Lin and Zhang [69] and many others [83, 75, 56, 62, 59, 81, 71]. How to perform statistical inference for the model has shown to be a much more challenging problem. Confidence bands that measure the uncertainty of the estimation in a setting with fixed dimension have been widely studied by Härdle [55], Sun and Loader [91], Fan and Zhang [43], Claeskens and Keilegom [36] and Zhang and Peng [106]. A standard assumption in high-dimensions is sparsity meaning that only a small subset of s components is different from zero. Results regarding inference for GAMs in a high-dimensional setting have been derived only recently. We discuss these results in the next

paragraphs and emphasize our contribution to the existing literature.

Kozbur [64] proposes an estimation and inference method for a single target component called Post-Nonparametric Double Selection which is an application of the double machine learning approach developed in Belloni et al. [8]. Our work contributes to this expanding literature on high-dimensional inference, especially to the debiased/double machine learning literature. Results for valid confidence intervals for low-dimensional parameters in high-dimensional linear models were also derived in Van De Geer et al. [96] and Zhang and Zhang [105]. For a survey on post-selection inference in high-dimensional settings and generalization, we refer to Chernozhukov et al. [33]. We consider the same setting as Kozbur [64], i.e., a more general additively separable model

$$Y = f_1(X_1) + f_{-1}(X_2, \ldots, X_p) + \varepsilon,$$

that includes the general additive model (GAM)

$$Y = \alpha + f_1(X_1) + \ldots + f_p(X_p) + \varepsilon.$$

Kozbur [64] focuses on inference on functionals of the form $\theta = a(f_1)$ and obtains pointwise confidence intervals based on a penalized series estimator. In contrast, we are able to construct uniformly valid confidence bands for the whole function f_1. Our paper builds on recent results, allowing for inference on high-dimensional target parameters, provided by Belloni et al. [12] and Belloni et al. [10]. Further, Kozbur [64] relies on two high level assumptions on Lasso estimation and variable selection (see Assumptions 9 and 10 in Kozbur [64]) that are hard to verify. We clarify technical requirements and provide results on uniform Lasso estimation that are needed to perform valid inference.

Gregory et al. [50] use the so-called Debiasing approach introduced in Zhang and Zhang [105] to estimate the first component f_1 in a high-dimensional GAM where the number p of additive components may increase with sample size. The estimator is constructed in two steps. The first step is an under-smoothed estimator based on near-orthogonal projections with a group Lasso bias correction. Then, a debiased version of the first step estimator is used to construct pseudo responses $\hat{Y}$. In the second step, a smoothing method is applied to a nonparametric regression problem with $\hat{Y}$ and covariates X_1. Under sparsity assumptions on the number of nonzero additive components, they show the so-called oracle property meaning asymptotic equivalence of their estimator and the oracle estimator where the functions $f_2, \ldots, f_p$ are known. The asymptotics of the oracle estimator are well understood and carry over to the proposed debiasing estimate including methodology to construct uniformly valid confidence intervals for f_1. Nevertheless, Gregory et al. [50] do not explicitly focus on inference and they need much stronger assumptions to let the oracle property hold. For example, they assume normally distributed errors that need to be independent to X. Further, they assume a bounded support of X. As in our paper, they choose a large set of basis functions (e.g., polynomials or splines) to approximate the components f_1 and f_{-1}. However, we allow the degree of approximating functions to grow to infinity with increasing sample size.

Lu et al. [72] provide an explicit procedure for constructing uniformly valid confidence bands for components in high-dimensional additive models. They argue that this is a challenging problem, as a direct generalization of the ideas for the finite-dimensional case is difficult. Confidence bands in the low-dimensional case are mostly built upon kernel methods, while estimators for sparse additive models are sieve estimators based on dictionaries. To derive their results, Lu et al. [72] have to combine both kernel and sieve methods to utilize the advantages of each method resulting in a kernel-sieves hybrid estimator. This also leads to a two-step estimator with many tuning parameters as the bandwidth and penalization

levels that need to be chosen by cross-validation. The advantage of our estimator is that we can stay in the sieves framework and nevertheless derive valid confidence bands. This is possible as we consider the problem as a high-dimensional Z-estimation problem utilizing recent results from Belloni et al. [12]. We also provide a theory driven choice of the penalization level. As in Gregory et al. [50], Lu et al. [72] assume normally distributed errors that are independent to X. This is much more restrictive than in our paper since we only need to assume sub-exponential tails and we allow for heteroscedastic error terms. Further, they assume that the number of nonzero components $s = O(1)$ is bounded. In our setting, s may grow to infinity with increasing sample size. However, their approach differs from ours in that they consider an additive local approximation model with sparsity (ATLAS), in which they only need to impose a local sparsity structure.

The finite sample properties of our estimator are evaluated in a simulation study that is based on the data generating processes in Gregory et al. [50]. The results show that the suggested method is able to perform valid simultaneous inference even in small samples and high-dimensional settings. Finally, we include an empirical application to the Boston housing data and provide evidence on nonlinear effects of certain socio-economic factors on house prices.

4.1.1 Organization of the Paper

The paper is organized as follows. In Section 4.2, the setting is outlined. Section 4.3 introduces the estimation method. In Section 4.4, the main result is provided. A simulation study, highlighting the small sample properties and implementation of our proposed method, is presented in Section 4.5. Section 4.6 illustrates the use of the method in an empirical application to the Boston housing data. The proof of the main theorem is provided in Section 4.8. The Appendix includes additional technical material. In Appendix 4.9, a general result for uniform inference about a high-dimensional linear functional is presented. Appendix 4.10 provides results regarding uniform Lasso estimation rates in high-dimensions. Finally, computational details are presented in Appendix 4.11.

4.1.2 Notation

Throughout the paper, we consider a random element W from some common probability space $(\Omega, \mathcal{A}, P)$. We denote by $P \in \mathcal{P}_n$ a probability measure out of large class of probability measures, which may vary with the sample size (since the model is allowed to change with n), and by $\mathbb{P}_n$ the empirical probability measure. Additionally, let $\mathbb{E}$ respectively $\mathbb{E}_n$ be the expectation with respect to P, respectively $\mathbb{P}_n$, and $\mathbb{G}_n(\cdot)$ denotes the empirical process

$$\mathbb{G}_n(f) := \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} f(W_i) - \mathbb{E}[f(W_i)]\right)$$

for a class of suitably measurable functions $\mathcal{F} : \mathcal{W} \to \mathbb{R}$. $\|\cdot\|_{P,q}$ denotes the $L^q(P)$-norm. In the following, we write $\|\cdot\|_{\Psi_\rho}$ for the Orlicz-norm that is defined as

$$\|W\|_{\Psi_\rho} := \inf\left\{C > 0 : \mathbb{E}\left[\exp((|W|/C)^\rho) - 1\right] \leq 1\right\}$$

for $\rho > 1$. Further, $\|v\|_1 = \sum_{l=1}^{p} |v_l|$ denotes the ℓ_1-norm, $\|v\|_2 = \sqrt{v^T v}$ the ℓ_2-norm and $\|v\|_0$ equals the number of nonzero components of a vector $v \in \mathbb{R}^p$. We define $v_{-l} := (v_1, \ldots, v_{l-1}, v_{l+1} \ldots, v_p)^T \in \mathbb{R}^{p-1}$ for any $1 \leq l \leq p$. $\|v\|_\infty = \sup_{l=1,\ldots,p} |v_l|$ denotes the sup-norm. Let c and C denote positive constants independent of n with values that may change at each appearance. The notation $a_n \lesssim b_n$ means $a_n \leq C b_n$ for all n and some C. Furthermore, $a_n = o(1)$ denotes that there exists a sequence $(b_n)_{\geq 1}$ of positive

numbers such that $|a_n| \leq b_n$ for all n where b_n is independent of $P \in \mathcal{P}_n$ for all n and b_n converges to zero. Finally, $a_n = O_P(b_n)$ means that, for any $\epsilon > 0$, there exists a C such that $P(a_n > Cb_n) \leq \epsilon$ for all n.

4.2 Setting

Consider the following nonparametric additively separable model

$$Y = f(X) + \varepsilon = f_1(X_1) + f_{-1}(X_{-1}) + \varepsilon$$

with $\mathbb{E}[\varepsilon|X] = 0$ and $\text{Var}(\varepsilon|X) \geq c$. Let the scalar response Y and features $X = (X_1, \ldots, X_p)$ take values in $\mathcal{Y}$ and $\mathcal{X} = (\mathcal{X}_1, \ldots, \mathcal{X}_p)$, respectively. We assume to observe n i.i.d. copies $(W^{(i)})_{i=1}^n = (Y^{(i)}, X^{(i)})_{i=1}^n$ of $W = (Y, X)$, where the number of covariates p is allowed to grow with sample size n. For identifiability, we assume $\mathbb{E}[f_{-1}(X_{-1})] = 0$. We aim to construct uniformly valid confidence regions for the first nonparametric component of the regression function, namely we want to find functions $\hat{l}(x)$ and $\hat{u}(x)$ converging to $f_1(x)$ with

$$P\left(\hat{l}(x) \leq f_1(x) \leq \hat{u}(x), \forall x \in I\right) \to 1 - \alpha.$$

Here, $I \subseteq \mathcal{X}_1$ is a bounded interval of interest where we want to conduct inference. We approximate f_1 and f_{-1} by a linear combination of approximating functions $g_1, \ldots, g_{d_1}$ and $h_1, \ldots, h_{d_2}$, respectively. Define

$$g(x) := (g_1(x), \ldots, g_{d_1}(x))^T$$

for $x \in \mathbb{R}$ and

$$h(x) := (h_1(x), \ldots, h_{d_2}(x))^T$$

for $x \in \mathbb{R}^{p-1}$. It is important to note that we allow the number of approximating functions d_1 and d_2 to increase with the sample size. Assume that the approximations are given by

$$f_1(X_1) = \theta_0^T g(X_1) + b_1(X_1), \tag{4.2}$$

where $\theta_{0,l} \in \Theta_l$ and analogously

$$f_{-1}(X_{-1}) := \beta_0^T h(X_{-1}) + b_2(X_{-1}), \tag{4.3}$$

where b_1 and b_2 denote the error terms. Additionally, it is convenient to define the combination

$$z(x) := (g_1(x), \ldots, g_{d_1}(x), h_1(x), \ldots, h_{d_2}(x))^T$$

for $x \in \mathbb{R}^p$, where we abbreviate

$$Z := z(X) = (g_1(X_1), \ldots, g_{d_1}(X_1), h_1(X_{-1}), \ldots, h_{d_2}(X_{-1}))^T.$$

For each element g_l of g, we consider

$$g_l(X_1) = (\gamma_0^{(l)})^T Z_{-l} + b_3^{(l)}(Z_{-l}) + \nu^{(l)} \tag{4.4}$$

and $\mathbb{E}[\nu^{(l)}|Z_{-l}] = 0$ and $\mathrm{Var}(\nu^{(l)}|Z_{-l}) \geq c$. This corresponds to

$$\mathbb{E}[g_l(X_1)|Z_{-l}] = (\gamma_0^{(l)})^T Z_{-l} + b_3^{(l)}(Z_{-l})$$

with approximation error $b_3^{(l)}(Z_{-l})$. The second stage equation (4.4) is used to construct an orthogonal score function for valid inference in a high-dimensional setting as in Chernozhukov et al. [35]. Estimating

$$f_1(\cdot) \approx \theta_0^T g(\cdot)$$

can be recast into a general Z-estimation problem of the form

$$\mathbb{E}\left[\psi_l(W, \theta_{0,l}, \eta_{0,l})\right] = 0, \quad l \in 1, \ldots, d_1$$

with target parameter θ_0, where the score functions are defined by

$$\begin{aligned}\psi_l(W, \theta, \eta) = &\Big(Y - \theta g_l(X_1) - (\eta^{(1)})^T Z_{-l} - \eta^{(3)}(X)\Big) \\ &\cdot \Big(g_l(X_1) - (\eta^{(2)})^T Z_{-l} - \eta^{(4)}(Z_{-l})\Big).\end{aligned}$$

Here,

$$\eta = (\eta^{(1)}, \eta^{(2)}, \eta^{(3)}, \eta^{(4)})^T$$

with $\eta^{(1)} \in \mathbb{R}^{d_1+d_2-1}, \eta^{(2)} \in \mathbb{R}^{d_1+d_2-1}$, $\eta^{(3)} \in \ell^\infty(\mathbb{R}^p)$ and $\eta^{(4)} \in \ell^\infty(\mathbb{R}^{d_1+d_2-1})$. The true nuisance parameter $\eta_{0,l}$ is given by

$$\begin{aligned}\eta_{0,l}^{(1)} &:= \beta_0^{(l)} \\ \eta_{0,l}^{(2)} &:= \gamma_0^{(l)} \\ \eta_{0,l}^{(3)}(X) &:= b_1(X_1) + b_2(X_{-1}) \\ \eta_{0,l}^{(4)}(Z_{-l}) &:= b_3^{(l)}(Z_{-l}),\end{aligned}$$

where $\beta_0^{(l)}$ is defined as

$$\beta_0^{(l)} := (\theta_{0,1}, \ldots, \theta_{0,l-1}, \theta_{0,l+1}, \ldots \theta_{0,d_1}, \beta_{0,1}, \ldots, \beta_{0,d_2})^T.$$

Essentially, the index l determines which coefficient is not contained in $\beta_0^{(l)}$. The third part of the nuisance functions captures the error made by the approximation of f_1 and f_{-1} which is independent from l. Therefore, we sometimes omit l.

Comment 4.2.1. *The score ψ is linear, meaning*

$$\psi_l(W, \theta, \eta) = \psi_l^a(X, \eta^{(2)}, \eta^{(4)})\theta + \psi_l^b(X, \eta)$$

with

$$\psi_l^a(X, \eta^{(2)}, \eta^{(4)}) = -g_l(X_1)(g_l(X_1) - (\eta^{(2)})^T Z_{-l} - \eta^{(4)}(Z_{-l}))$$

and

$$\psi_l^b(X, \eta) = (Y - (\eta^{(1)})^T Z_{-l} - \eta^{(3)}(X))(g_l(X_1) - (\eta^{(2)})^T Z_{-l} - \eta^{(4)}(Z_{-l}))$$

for all $l = 1, \ldots, d_1$.

Comment 4.2.2. *The score function ψ satisfies the moment condition, namely*

$$\mathbb{E}\left[\psi_l(W, \theta_{0,l}, \eta_{0,l})\right] = 0$$

for all $l = 1, \ldots, d_1$, and, given further conditions mentioned in Section 4.4, the near Neyman orthogonality condition

$$D_{l,0}[\eta, \eta_{0,l}] := \partial_t \big\{\mathbb{E}[\psi_l(W, \theta_{0,l}, \eta_{0,l} + t(\eta - \eta_{0,l}))]\big\}\big|_{t=0} \lesssim \delta_n n^{-1/2},$$

where ∂_t denotes the derivative with respect to t and $(\delta_n)_{n\geq 1}$ a sequence of positive constants converging to zero.

4.3 Estimation

In this section, we describe our estimation method and how the uniform valid confidence bands are constructed. The nuisance functions are estimated by Lasso regressions. Finally, they are plugged into the moment conditions that are solved for the target parameters, which yield an estimate $\hat{f}_1$ for the first component. The lower and upper curve of the confidence bands are finally based on the estimated covariance matrix and a critical value which is determined by a multiplier bootstrap procedure. As mentioned, the details are given in this section.

Let

$$g(x) = (g_1(x), \ldots, g_{d_1}(x))^T \in \mathbb{R}^{d_1 \times 1}$$

and

$$\psi(W, \theta, \eta) = (\psi_1(W, \theta_1, \eta_1), \ldots, \psi_{d_1}(W, \theta_{d_1}, \eta_{d_1}))^T \in \mathbb{R}^{d_1 \times 1}$$

for some vector

$$\theta = (\theta_1, \ldots, \theta_{d_1})^T$$

and

$$\eta = (\eta_1, \ldots, \eta_{d_1})^T.$$

For each $l = 1, \ldots, d_1$, let $\hat{\eta}_l = \left(\hat{\eta}_l^{(1)}, \hat{\eta}_l^{(2)}, \hat{\eta}_l^{(3)}, \hat{\eta}_l^{(4)}\right)$ be an estimator of the nuisance function. The estimator $\hat{\theta}_0$ of the target parameter

$$\theta_0 = (\theta_{0,1}, \ldots, \theta_{0,d_1})^T$$

is defined as the solution of

$$\sup_{l=1,\ldots,d_1} \left\{ \left| \mathbb{E}_n\left[\psi_l\big(W, \hat{\theta}_l, \hat{\eta}_l\big)\right] \right| - \inf_{\theta \in \Theta_l} \left| \mathbb{E}_n\left[\psi_l\big(W, \theta, \hat{\eta}_l\big)\right] \right| \right\} \leq \epsilon_n, \tag{4.5}$$

where $\epsilon_n = o\left(\delta_n n^{-1/2}\right)$ is the numerical tolerance. Finally, the target function $f_1(\cdot)$ can be estimated by

$$\hat{f}_1(\cdot) := \hat{\theta}_0^T g(\cdot). \tag{4.6}$$

Define the Jacobian matrix

$$J_0 := \frac{\partial}{\partial\theta}\mathbb{E}[\psi(W,\theta,\eta_0)]\Big|_{\theta=\theta_0} = \operatorname{diag}\left(J_{0,1},\dots,J_{0,d_1}\right) \in \mathbb{R}^{d_1\times d_1}$$

with

$$\begin{aligned}
J_{0,l} &= E[\psi_l^a(W,\eta_{0,l}^{(2)},\eta_{0,l}^{(4)})] \\
&= -\mathbb{E}[((\gamma_0^{(l)})^T Z_{-l} + b_3^{(l)}(Z_{-l}) + \nu^{(l)})\nu^{(l)}] \\
&= -\mathbb{E}\Big[\big((\gamma_0^{(l)})^T Z_{-l} + b_3^{(l)}(Z_{-l})\big)\underbrace{\mathbb{E}[\nu^{(l)}|Z_{-l}]}_{=0}\Big] - \mathbb{E}[(\nu^{(l)})^2] \\
&= -\mathbb{E}[(\nu^{(l)})^2]
\end{aligned}$$

for all $l = 1,\dots,d_1$. Observe that

$$\mathbb{E}\big[\psi(W,\theta_0,\eta_0)\psi(W,\theta_0,\eta_0)^T\big] =: \Sigma_{\varepsilon\nu}$$

is the covariance matrix of $\varepsilon\nu := (\varepsilon\nu^{(1)},\dots,\varepsilon\nu^{(d_1)})$. Define the approximate covariance matrix

$$\begin{aligned}
\Sigma_n :&= J_0^{-1}\mathbb{E}\big[\psi(W,\theta_0,\eta_0)\psi(W,\theta_0,\eta_0)^T\big](J_0^{-1})^T \\
&= J_0^{-1}\Sigma_{\varepsilon\nu}(J_0^{-1})^T \in \mathbb{R}^{d_1\times d_1}
\end{aligned}$$

with

$$\Sigma_n := \begin{pmatrix}
\frac{\mathbb{E}[(\varepsilon\nu^{(1)})^2]}{\mathbb{E}[(\nu^{(1)})^2]^2} & \frac{\mathbb{E}\left[\varepsilon\nu^{(1)}\varepsilon\nu^{(2)}\right]}{\mathbb{E}[(\nu^{(1)})^2]\mathbb{E}[(\nu^{(2)})^2]} & \cdots & \frac{\mathbb{E}\left[\varepsilon\nu^{(1)}\varepsilon\nu^{(d_1)}\right]}{\mathbb{E}[(\nu^{(1)})^2]\mathbb{E}[(\nu^{(d_1)})^2]} \\
\frac{\mathbb{E}\left[\varepsilon\nu^{(2)}\varepsilon\nu^{(1)}\right]}{\mathbb{E}[(\nu^{(2)})^2]\mathbb{E}[(\nu^{(1)})^2]} & \frac{\mathbb{E}[(\varepsilon\nu^{(2)})^2]}{\mathbb{E}[(\nu^{(2)})^2]^2} & \cdots & \frac{\mathbb{E}\left[\varepsilon\nu^{(2)}\varepsilon\nu^{(d_1)}\right]}{\mathbb{E}[(\nu^{(2)})^2]\mathbb{E}[(\nu^{(d_1)})^2]} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\mathbb{E}\left[\varepsilon\nu^{(d_1)}\varepsilon\nu^{(1)}\right]}{\mathbb{E}[(\nu^{(d_1)})^2]\mathbb{E}[(\nu^{(1)})^2]} & \frac{\mathbb{E}\left[\varepsilon\nu^{(d_1)}\varepsilon\nu^{(2)}\right]}{\mathbb{E}[(\nu^{(d_1)})^2]\mathbb{E}[(\nu^{(1)})^2]} & \cdots & \frac{\mathbb{E}[(\varepsilon\nu^{(d_1)})^2]}{\mathbb{E}[(\nu^{(d_1)})^2]^2}
\end{pmatrix}.$$

The approximate covariance matrix can be estimated by replacing every expectation by the empirical analog and plugging in the estimated parameters

$$\begin{aligned}
\hat\Sigma_n :&= \hat J^{-1}\mathbb{E}_n\big[\psi(W,\hat\theta,\hat\eta)\psi(W,\hat\theta,\hat\eta)^T\big](\hat J^{-1})^T \\
&= \hat J^{-1}\hat\Sigma_{\varepsilon\nu}(\hat J^{-1})^T \\
&= \begin{pmatrix}
\frac{\mathbb{E}_n[(\hat\varepsilon\hat\nu^{(1)})^2]}{\mathbb{E}_n[(\hat\nu^{(1)})^2]^2} & \frac{\mathbb{E}_n\left[\hat\varepsilon\hat\nu^{(1)}\hat\varepsilon\hat\nu^{(2)}\right]}{\mathbb{E}_n[(\hat\nu^{(1)})^2]\mathbb{E}_n[(\hat\nu^{(2)})^2]} & \cdots & \frac{\mathbb{E}_n\left[\hat\varepsilon\hat\nu^{(1)}\hat\varepsilon\hat\nu^{(d_1)}\right]}{\mathbb{E}_n[(\hat\nu^{(1)})^2]\mathbb{E}_n[(\hat\nu^{(d_1)})^2]} \\
\frac{\mathbb{E}_n\left[\hat\varepsilon\hat\nu^{(2)}\hat\varepsilon\hat\nu^{(1)}\right]}{\mathbb{E}_n[(\hat\nu^{(2)})^2]\mathbb{E}_n[(\hat\nu^{(1)})^2]} & \frac{\mathbb{E}_n[(\hat\varepsilon\hat\nu^{(2)})^2]}{\mathbb{E}_n[(\hat\nu^{(2)})^2]^2} & \cdots & \frac{\mathbb{E}_n\left[\hat\varepsilon\hat\nu^{(2)}\hat\varepsilon\hat\nu^{(d_1)}\right]}{\mathbb{E}_n[(\hat\nu^{(2)})^2]\mathbb{E}_n[(\hat\nu^{(d_1)})^2]} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\mathbb{E}_n\left[\hat\varepsilon\hat\nu^{(d_1)}\hat\varepsilon\hat\nu^{(1)}\right]}{\mathbb{E}_n[(\hat\nu^{(d_1)})^2]\mathbb{E}_n[(\hat\nu^{(1)})^2]} & \frac{\mathbb{E}_n\left[\hat\varepsilon\hat\nu^{(d_1)}\hat\varepsilon\hat\nu^{(2)}\right]}{\mathbb{E}_n[(\hat\nu^{(d_1)})^2]\mathbb{E}_n[(\hat\nu^{(1)})^2]} & \cdots & \frac{\mathbb{E}_n[(\hat\varepsilon\hat\nu^{(d_1)})^2]}{\mathbb{E}_n[(\hat\nu^{(d_1)})^2]^2}
\end{pmatrix}.
\end{aligned}$$

This estimated covariance matrix can be used to construct the confidence bands

$$\begin{aligned}
\hat u(x) &:= \hat f_1(x) + \frac{(g(x)^T\hat\Sigma_n g(x))^{1/2}c_\alpha}{\sqrt{n}} \\
\hat l(x) &:= \hat f_1(x) - \frac{(g(x)^T\hat\Sigma_n g(x))^{1/2}c_\alpha}{\sqrt{n}},
\end{aligned}$$

where c_α is a critical value determined by the following standard multiplier bootstrap method introduced in Chernozhukov et al. [30]. Define

$$\hat{\psi}_x(\cdot) := (g(x)^T \hat{\Sigma}_n g(x))^{-1/2} g(x)^T \hat{J}_0^{-1} \psi(\cdot, \hat{\theta}_0, \hat{\eta}_0)$$

and let

$$\hat{\mathcal{G}} = \left(\hat{\mathcal{G}}_x\right)_{x \in I} = \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i \hat{\psi}_x(W_i) \right)_{x \in I},$$

where $(\xi_i)_{i=1}^n$ are independent standard normal random variables (especially independent from $(W_i)_{i=1}^n$). The multiplier bootstrap critical value c_α is given by the $(1-\alpha)$-quantile of the conditional distribution of $\sup_{x \in I} |\hat{\mathcal{G}}_x|$ given $(W_i)_{i=1}^n$.

4.4 Main Results

Now, we specify the conditions that are required to construct the uniformly valid confidence bands. Since we would like to represent f_1 and f_{-1} by their approximations in (4.2) and (4.3), we need to choose an appropriate set of approximating functions. Let $\bar{d}_n := \max(d_1, d_2, n, e)$ and C be a strictly positive constant independent of n and l. Additionally, we set $t_1 := \sup_{x \in I} \|g(x)\|_0 \leq d_1$. The following assumptions hold uniformly in $n \geq n_0$ and $P \in \mathcal{P}_n$:

Assumption A. 1.

(i) It holds

$$\inf_{x \in I} \|g(x)\|_2^2 \geq c > 0, \quad \sup_{x \in I} \sup_{l=1,\ldots,d_1} |g_l(x)| \leq C < \infty$$

and, for all $\varepsilon > 0$,

$$\log N(\varepsilon, g(I), \|\cdot\|_2) \leq C t_1 \log\left(\frac{A_n}{\varepsilon}\right).$$

(ii) There exists $1 \leq \rho \leq 2$ such that

$$\max_{l=1,\ldots,d_1} \|b_3^{(l)}(Z_{-l})\|_{\Psi_\rho} \leq C, \quad \|b_1(X_1) + b_2(X_{-1})\|_{\Psi_\rho} \leq C.$$

Additionally, the approximation errors obey

$$\mathbb{E}\Big[\big(b_1(X_1) + b_2(X_{-1})\big)^2\Big] \leq Cs \log(\bar{d}_n)/n,$$
$$\max_{l=1,\ldots,d_1} \mathbb{E}\Big[\big(b_3^{(l)}(Z_{-l})\big)^2\Big] \leq Cs \log(\bar{d}_n)/n$$

and

$$\mathbb{E}_n\Big[\big(b_1(X_1) + b_2(X_{-1})\big)^2\Big] - \mathbb{E}\Big[\big(b_1(X_1) + b_2(X_{-1})\big)^2\Big] \leq Cs \log(\bar{d}_n)/n,$$
$$\max_{l=1,\ldots,d_1} \left(\mathbb{E}_n\Big[\big(b_3^{(l)}(Z_{-l})\big)^2\Big] - \mathbb{E}\Big[\big(b_3^{(l)}(Z_{-l})\big)^2\Big]\right) \leq Cs \log(\bar{d}_n)/n.$$

(iii) We have

$$\sup_{\|\xi\|_2=1} \mathbb{E}\left[(\xi^T Z)^2\big(b_1(X_1)+b_2(X_{-1})\big)^2\right] \leq C\mathbb{E}\left[\big(b_1(X_1)+b_2(X_{-1})\big)^2\right]$$

and

$$\sup_{\|\xi\|_2=1} \mathbb{E}\left[(\xi^T Z)^2\big(b_3^{(l)}(Z_{-l})\big)^2\right] \leq C\mathbb{E}\left[\big(b_3^{(l)}(Z_{-l})\big)^2\right]$$

for $l = 1, \ldots, d_1$.

(iv) It holds

$$\mathbb{E}\Big[\nu^{(l)}\big(b_1(X_1)+b_2(X_{-1})\big)\Big] \leq C\delta_n n^{-1/2}$$

with $\delta_n = o\big(t_1^{-\frac{3}{2}} \log^{-\frac{1}{2}}(A_n)\big)$.

Assumption A.1(i) contains regularity conditions on g. We assume that the infimum of the ℓ_2-norm of $g(x)$ is bounded away from zero, but the supremum is allowed to increase with sample size (affecting the growth conditions in A.2(v)). The lower bound on the infimum is not necessary and can be replaced by a decaying sequence at the cost of stricter growth rates. The Assumptions A.1(ii) and (iii) are tail and moment conditions on the approximation error. These assumptions are mild since the number of approximating functions may increase with sample size. Finally, Assumption A.1(iv) ensures that the violation of the exact Neyman Orthogonality due to the approximation errors is negligible. It is worth to notice that if $b_1(X_1)$ and $b_2(X_{-1})$ are measurable with respect to Z_{-l} (for example in the linear approximate sparse setting for the conditional expectation), the exact Neyman Orthogonality holds. Now, we go more into detail regarding the condition on the covering number of the image of g. Especially, if $t_1 < d_1$, the complexity of the approximating functions is reduced significantly. One obtains

$$g(I) \subseteq \bigcup_{j=1}^{\binom{d_1}{t_1}} g^{(j)}(I),$$

where each $g^{(j)}(I)$ is only dependent on t_1 nonzero components. It is straightforward to see that for each $g^{(j)}(I)$ the covering numbers satisfy

$$N(\varepsilon, g^{(j)}(I), \|\cdot\|_2) \leq \left(\frac{6\sup_{x\in I}\|g(x)\|_2}{\varepsilon}\right)^{t_1}$$

(cf. Vaart and Wellner [94]) implying

$$\begin{aligned}
\log N(\varepsilon, g(I), \|\cdot\|_2) &\leq \log\left(\sum_{j=1}^{\binom{d_1}{t_1}} N(\varepsilon, g^{(j)}(I), \|\cdot\|_2)\right) \\
&\leq \log\left(\left(\frac{e\cdot d_1}{t_1}\right)^{t_1}\left(\frac{6\sup_{x\in I}\|g(x)\|_2}{\varepsilon}\right)^{t_1}\right) \\
&\leq t_1 \log\left(\left(\frac{6ed_1\sup_{x\in I}\|g(x)\|_2}{t_1}\right)\frac{1}{\varepsilon}\right) \\
&\leq Ct_1 \log\left(\frac{d_1}{\varepsilon}\right).
\end{aligned}$$

For specific classes of approximating functions the complexity can be further reduced.

Assumption A. 2.

(i) For all $l = 1, \dots, d_1$, Θ_l *contains a ball of radius*

$$\log(\log(n)) n^{-1/2} \log^{1/2}(d_1 \vee e) \log(n)$$

centered at $\theta_{0,l}$ *with*

$$\sup_{l=1,\dots,d_1} \sup_{\theta_l \in \Theta_l} |\theta_l| \leq C.$$

(ii) It holds

$$\|\beta_0^{(l)}\|_0 \leq s, \quad \|\beta_0^{(l)}\|_2 \leq C$$

for all $l = 1, \dots, d_1$ *and*

$$\max_{l=1,\dots,d_1} \|\gamma_0^{(l)}\|_0 \leq s, \quad \max_{l=1,\dots,d_1} \|\gamma_0^{(l)}\|_2 \leq C.$$

(iii) There exists $1 \leq \rho \leq 2$ *such that*

$$\max_{j=1,\dots,d_1+d_2} \|Z_j\|_{\Psi_\rho} \leq C, \quad \|\varepsilon\|_{\Psi_\rho} \leq C.$$

(iv) It holds

$$\inf_{\|\xi\|_2=1} \mathbb{E}[(\xi^T Z)^2] \geq c, \quad \sup_{\|\xi\|_2=1} \mathbb{E}[(\xi^T Z)^4] \leq C$$

and the eigenvalues of the covariance matrix $\Sigma_{\varepsilon\nu}$ *are bounded from above and away from zero.*

(v) There exists a fixed $\bar{q} \geq 4$ *such that*

(a) $n^{\frac{1}{\bar{q}}} \frac{s^2 t_1^3 \log^{2+\frac{4}{\rho}}(\bar{d}_n) \log(A_n)}{n} = o(1)$,

(b) $n^{\frac{1}{\bar{q}}} \frac{\sup_{x \in I} \|g(x)\|_2^6 s t_1^4 \log(\bar{d}_n) \log^2(A_n)}{n} \left(\log^{\frac{2}{\rho}}(d_1) \vee s\sqrt{\frac{s \log(\bar{d}_n)}{n}} \right) = o(1)$,

(c) $n^{\frac{1}{\bar{q}}} \frac{t_1^{13} \log^{\frac{6}{\rho}}(d_1) \log^7(A_n)}{n} = o(1)$.

Assumptions A.2(i) and (ii) are regularity and sparsity conditions, where the number of nonzero regression coefficients $s = s_n$ is allowed to grow to infinity with increasing sample size. A detailed comment on the sparsity condition is given in Comment 4.4.2. Assumption A.2(iii) contains tail conditions on the approximating functions (and therefore on the original variables) as well as for the error term. Assumption A.2(iv) is a standard eigenvalue condition, which restricts the correlation between the basis elements (and therefore between the original variables). For example, if the conditional variance of $\nu^{(l)}$ is uniformly bounded away from zero, the second inequality of A.2(iv) holds. Finally, Assumption A.2(v) provides the growth conditions. These are given in general terms and depend on the choice of the approximation functions. Choosing B-Splines simplifies the growth conditions significantly as we discuss in Comment 4.4.1.

Theorem 9. *Given Assumptions A.1 and A.2, it holds that*

$$P\left(\hat{l}(x) \leq f_1(x) \leq \hat{u}(x), \forall x \in I\right) \to 1 - \alpha,$$

uniformly over $P \in \mathcal{P}_n$, *where* c_α *is a critical value determined by a multiplier bootstrap method.*

Comment 4.4.1. *An appropriate and common choice in series estimation are B-Splines. B-Splines are positive and local in the sense that* $g(x) \geq 0$ *and* $\sup_{x\in I} \|g(x)\|_0 \leq t_1$ *for every* x*, where* t_1 *is the degree of the spline. The* l_1*-norm of B-Splines is equal to 1, meaning*

$$\|g(x)\|_1 = \sum_{j=1}^{d_1} g_j(x) = 1$$

for every x *(partition of unity). Hence, Assumption A.1*(i) *is met with*

$$\frac{1}{\sqrt{t_1}} \leq \inf_{x\in I} \|g(x)\|_2^2 \leq \sup_{x\in I} \|g(x)\|_2^2 \leq 1 \quad \textit{and} \quad \sup_{x\in I} \sup_{l=1,\dots,d_1} |g_l(x)| \leq 1.$$

The covering numbers of $g(I)$ *is given by*

$$\begin{aligned}
\log N(\varepsilon, g(I), \|\cdot\|_2) &\leq \log\left(\sum_{j=1}^{d_1} N(\varepsilon, g^{(j)}(I), \|\cdot\|_2)\right) \\
&\leq t_1 \log\left(\left(\frac{6 d_1^{\frac{1}{t_1}} \sup_{x\in I} \|g(x)\|_2}{\varepsilon}\right)\right) \\
&\leq C \log\left(\frac{d_1}{\varepsilon}\right).
\end{aligned}$$

Choosing the degree of the B-Splines of order $t_1 = \log(n)$*, the growth rates in Assumption A.2*(v) *simplify to*

$$n^{\frac{1}{q}} \frac{s^2 \log^{2+\frac{4}{\rho}}(\bar{d}_n) \log(d_1)}{n} = o(1) \quad \textit{and} \quad n^{\frac{1}{q}} \frac{\log^{7+\frac{6}{\rho}}(d_1)}{n} = o(1).$$

It is worth to notice that in the first growth condition

$$n^{\frac{1}{q}} \frac{s^2 \log^{2+\frac{4}{\rho}}(\bar{d}_n) \log(d_1)}{n} = o(1)$$

both the total number of approximating functions d_1 *and* d_2*, and the number of relevant functions* s *may grow with the sample size in a balanced way. If* s *is bounded, the number of approximating functions can grow at an exponential rate with the sample size. This means that the set of approximating functions can be much larger than the sample size, only the number of relevant function* s *has to be smaller than the sample size. This situation is common for Lasso based estimators. Our growth condition is in line with other results in the literature, e.g., with Belloni et al. [12], Belloni et al. [10] and many others. The second growth condition ensures that*

$$n^{\frac{1}{q}} \frac{\log^{7+\frac{6}{\rho}}(d_1)}{n} = o(1)$$

and is in line with Chernozhukov et al. [30]. It guarantees the validity of multiplier bootstrap in our setting and allows us to construct uniformly valid confidence regions.

Comment 4.4.2. *The sparsity condition in A.2*(ii) *restricts the number of nonzero regression coefficients* $s = s_n$ *in the equations (4.2), (4.3) and (4.4). Through this, we especially assume that the regression function* f *can be approximated sufficiently well by only* s *relevant basis functions. Note that we do not directly control the number of relevant covariables but the number of approximating functions in total. This is another sparsity condition as in Gregory et al. [50] and Lu et al. [72] who restrict the number of relevant additive components in the GAM model (4.1). Our model also includes the approximate sparse setting due to the error terms* b_1 *and* b_2 *in (4.2) and (4.3). This is more flexible and more realistic for*

many applications. Furthermore, we do not define $\theta_0^T g(X_1)$ as the best projection of $f_1(X_1)$ in (4.2) (and $\beta_0^T h(X_{-1})$ for $f_{-1}(X_{-1})$ in (4.3)) as in Gregory et al. [50]. We only assume a sparse projection which is "close" to the best projection where the distance is measured with $\|\cdot\|_{P,2}$ as described in Assumption A.1(ii).

4.5 Simulation Results

To verify the theoretical guarantees of our estimator in practice, we perform a simulation study which is based on the settings in Gregory et al. [50] and Meier et al. [75]. We consider the finite sample performance of our estimator in a high-dimensional model of the form

$$y_i = \sum_{j=1}^{p} f_j(x_{i,j}) + \epsilon_{i,j}$$

with $i = 1, \ldots, n$, $j = 1, \ldots, p$. The definition of the functions $f_j(x_j)$, $j = 1, \ldots, j$, are presented in Table 4.1. We extend the initial setting in Gregory et al. [50] to allow for heteroscedasticity, i.e., we specify $\epsilon_j \sim N(0, \sigma_j)$ with $\sigma_j = \underline{\sigma} \cdot (1 + |x_j|)$ and $\underline{\sigma} = \sqrt{\frac{12}{67}}$. This value for $\underline{\sigma}$ ensures a signal-to-noise ratio that is comparable to the settings in [50]. Data sets are generated for scenarios with dimensions $n \in \{100, 1000\}$ and $p \in \{50, 150\}$. In all cases, sparsity is imposed by only allowing the first four components, $f_1, ..., f_4$, to be nonzero. The regressors X are marginally uniformly distributed on an interval $I = [-2.5, 2.5]$ with correlation matrix Σ with $\Sigma_{k,l} = 0.5^{|k-l|}$, $1 \leq k, l \leq p$, which corresponds to the setting in Gregory et al. [50] with the strongest correlation structure.

Component	Function
1	$f_1(x_1) = -\sin(2 \cdot x)$
2	$f_2(x_2) = x^2 - \frac{25}{12}$
3	$f_3(x_3) = x$
4	$f_4(x_4) = \exp(-x) - \frac{2}{5} \cdot \sinh(\frac{5}{2})$
$5, \ldots, p$	$f_j(x_j) = 0.$

Table 4.1: **Definition of the functions in the data generating processes that are used in the simulation study. Data generating processes are based on settings in Gregory et al. [50] and Meier et al. [75].**

In the simulation, we use the previously suggested estimator to generate predictions $\hat{f}_j(x_j)$ for the function $f_j(x_j)$ and construct simultaneous confidence bands that are defined by $\hat{l}_j(x_j)$ and $\hat{u}_j(x_j)$, accordingly. The functions $f_j(x_j)$ in the additive model are approximated using cubic B-splines. Variable selection is performed using post-Lasso with theory-based choice of the penalty level as implemented in the R package `hdm` [32]. Further details related to the implementation and parametrization in the simulation study can be found in Appendix 4.11.

Table 4.2 presents the empirical coverage achieved by the estimated simultaneous 95%-confidence bands in $R = 500$ repetitions as constructed over an interval of x_j $I = [-2, 2]$. A confidence band is considered to cover the function $f_j(x_j)$ if it contains the true function entirely, i.e., if for all values of $x_j \in I$ it holds that $\hat{l}_j(x_j) \leq f_j(x_j) \leq \hat{u}_j(x_j)$. The results serve as empirical evidence on the validity of the method. In most cases, the empirical coverage approaches 95% or is above the nominal level. This observation can be made even for the setting with more regressors than observations.

The first two plots in Figure 4.1 illustrate the averaged confidence bands as constructed for four different intervals of x_j, i.e., $I = [-x_0, x_0]$ with $x_0 = 0.5, 1.0, 1.5, 2$. It can be observed that as the interval I becomes wider, the width of the confidence bands increases, as well. The two plots at the bottom of Figure 4.1 show the empirical coverage as obtained for a sequence of values $x_{0,j}$ with $I = [-x_{0,j}, x_{0,j}]$

n	p	f_1	f_2	f_3	f_4	f_5
100	50	0.994	0.982	0.968	0.938	0.990
100	150	0.992	0.976	0.952	0.886	0.988
1000	50	0.998	0.980	0.962	0.848	1.000
1000	150	1.000	0.968	0.986	0.806	1.000

Table 4.2: **Simulation results. Coverage achieved by simultaneous 0.95%-confidence bands in $R = 500$ repetitions as generated over a range of values of x_j, $I = [-2, 2]$.**

with $x_{0,j} = 0.01, 0.02, \ldots, 2$. Whereas the coverage remains stable over a wide range of $x_{0,j}$ values, the coverage decreases slightly for larger $x_{0,j}$. This behavior arises due to boundary problems that are common in most nonparametric smoothing methods and explain the relatively low coverage achieved for f_4.

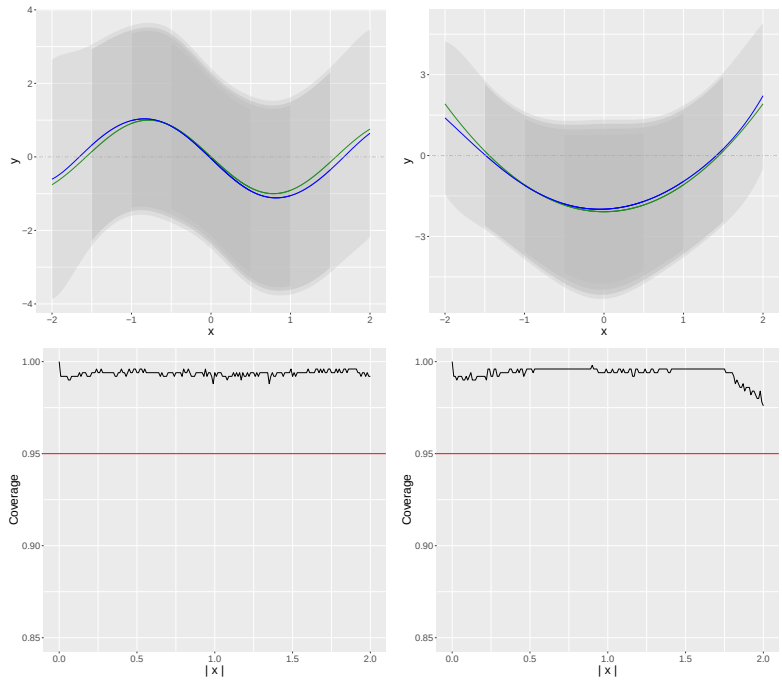


Figure 4.1: **Simulation results for the setting with $n = 100$ and $p = 150$. (Top) Gray shaded areas illustrate averaged 95%-confidence bands obtained in $R = 500$ repetitions for functions $f_1(x_1)$ and $f_2(x_2)$. Blue lines correspond to the estimated functions $\hat{f}_j(x_j)$ and green lines to the true functions $f_j(x_j)$. (Bottom) Empirical coverage achieved by confidence bands for a sequence of values $x_{0,j}$ with $I(x_j) = [-x_{0,j}, x_{0,j}]$ with $x_{j,0} = 0.01, 0.02, \ldots, 2$. Plots on the left refer to $f_1(x_1)$, plots on the right to $f_2(x_2)$.**

4.6 Empirical Application

As a real-data example, we apply our estimator to the Boston housing data that has been first used in Harrison Jr and Rubinfeld [52] and later been reassessed in several studies, e.g., Kong and Xia [63] and Doksum and Samarov [40]. The data set is available via the R package `mlbench` ([68, 80]). The data contains information on housing prices for $n = 506$ census tracts in Boston based on the 1970 census. We perform inference on the effect of 11 continuous variables on the dependent variable $MEDV$ which measures the median value of owner-occupied homes (in USD 1000's). A list of the explanatory variables is provided in Table 4.3.

$MEDV$	median value of owner-occupied homes in USD 1000's
$LSTAT$	percentage of lower status population
$CRIM$	per capita crime rate by town
NOX	nitric oxides
TAX	full-value property-tax rate per USD 10,000
AGE	proportion of owner-occupied units built prior to 1940
$DIST$	weighted distances to five Boston employment centers
RM	average number of rooms per dwelling
$INDUS$	proportion of non-retail business acres per town
ZN	proportion of residential land zoned for lots over 25,000 sq.ft
$BLACK$	$1000(B - 0.63)^2$ where B is the proportion of blacks by town
$PTRATIO$	pupil-teacher ratio by town
$CHAS$	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)

Table 4.3: **List of variables in the analysis of the Boston housing data.**

The implemented model is given by

$$\begin{aligned} MEDV_i = & f_1(\text{LSTAT}_i) + f_2(\text{CRIM}_i) + f_3(\text{NOX}_i) + f_4(\text{TAX}_i) + \\ & f_5(\text{AGE}_i) + f_6(\text{DIST}_i) + f_7(\text{RM}_i) + f_8(\text{INDUS}_i) + \\ & f_9(\text{ZN}_i) + f_{10}(\text{BLACK}_i) + f_{11}(\text{PTRATIO}_i) + \gamma \cdot \text{CHAS} + \epsilon_i. \end{aligned}$$

As in the simulation study, the functions $f_j(x_j)$ are approximated with cubic B-splines and variable selection is performed using post-Lasso with theory-based choice of the penalty term. The smoothing parameters $k = \{k_j, k_{-j}\}$ have been determined according to a heuristic cross-validation rule that is outlined in Appendix 4.11. The results illustrated in Figure 4.2 suggest nonlinear and significant effects for the variables LSTAT and RM that are generally in line with economic intuition and the findings in Kong and Xia [63] and Doksum and Samarov [40]. Whereas for small values of the LSTAT variable, i.e., the percentage of lower status of the population, the estimated effect $\hat{f}_1(\text{LSTAT})$ is positive, it decreases and, finally, becomes negative for higher values of LSTAT. The nonlinearities found for the variable RM suggest that the average number of rooms per dwelling impacts housing prices positively if the average number of rooms exceeds seven. The results for the other regressors that are presented in Appendix 4.11 point at nonlinear effects that are, however, not significant.

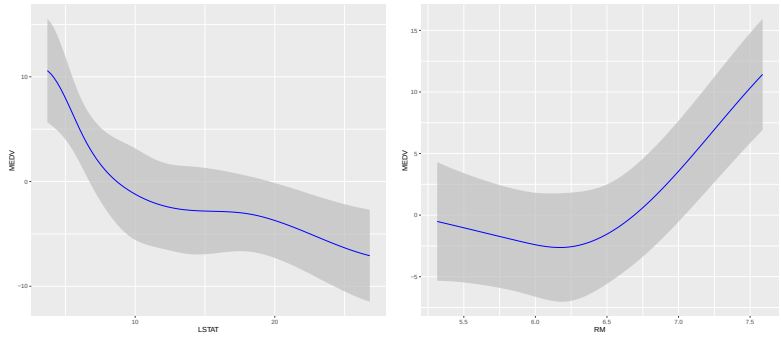


Figure 4.2: **Plots of $\hat{f}_1$(LSTAT) and $\hat{f}_7$(RM) with simultaneous 95%-confidence bands in the Boston housing data application.**

4.7 Conclusion

In this paper, we provide methodology for inference about a nonparametric component of an additively separable regression function in high-dimensions. We are able to construct uniformly valid confidence bands. Our work contributes to the double machine learning literature by extending this approach allowing to conduct valid inference about a linear functional of a high-dimensional target parameter. Our simulation studies show that our proposed method gives reliable results. We demonstrate the implementation and the use of the proposed method in practice by analyzing the well-known Boston housing data set. Our methodology suggests nonlinear and significant effects for the variables LSTAT and RM that denotes the percentage of lower status population and the average number of rooms per dwelling, respectively. This is in line with the economic intuition and the findings in the literature.

4.8 Proofs

Proof of Theorem 9.
We will prove that the Assumptions A.1 and A.2 imply the Assumptions B.1-B.5 stated in Appendix 4.9 and then the claim follows by applying Theorem 10. Without loss of generality, we assume $\min(d_1, n) \geq e$ to simplify notation.

Assumption B.1
Both conditions (i) and (ii) are directly assumed in A.1(i). Due to A.1(ii) and A.2(iv), it holds

$$\begin{aligned}\mathbb{E}\left[(\nu^{(l))})^2\right] &= \mathbb{E}\left[\left(g_l(X_1) - (\gamma_0^{(l)})^T Z_{-l} - b_3^{(l)}(Z_{-l})\right)^2\right] \\ &\leq C\left(\sup_{\|\xi\|_2=1} \mathbb{E}[(\xi^T Z)^2] + \mathbb{E}\left[\left(b_3^{(l)}(Z_{-l})\right)^2\right]\right) \\ &\lesssim C,\end{aligned}$$

where we used that $\|\gamma_0^{(l)}\|_2 \leq C$. It holds

$$\mathbb{E}\left[(\nu^{(l))})^2\right] \geq \mathrm{Var}(\nu^{(l)}|Z_{-l}) \geq c.$$

Since the eigenvalues of $\Sigma_{\varepsilon\nu}$ are bounded from above and away from zero,

$$\Sigma_n = J_0^{-1}\Sigma_{\varepsilon\nu}(J_0^{-1})^T \in \mathbb{R}^{d_1 \times d_1}$$

directly implies B.1(iii).

Assumption B.2
For each $l = 1, \ldots, d_1$, the moment condition holds

$$\begin{aligned}\mathbb{E}\left[\psi_l(W, \theta_{0,l}, \eta_{0,l})\right] &= \mathbb{E}\Big[\Big(Y - f(X)\Big)\Big(g_l(X_1) - (\gamma_0^{(l)})^T Z_{-l} - b_3^{(l)}(Z_{-l})\Big)\Big] \\ &= \mathbb{E}\big[\varepsilon\nu^{(l)}\big] \\ &= \mathbb{E}\big[\nu^{(l)} \underbrace{\mathbb{E}\left[\varepsilon|X\right]}_{=0}\big] \\ &= 0.\end{aligned}$$

For all $l = 1, \ldots, d_1$, define the convex set

$$\begin{aligned}T_l := \Big\{\eta = (\eta^{(1)}, \eta^{(2)}, \eta^{(3)}, \eta^{(4)})^T : &\eta^{(1)}, \eta^{(2)} \in \mathbb{R}^{d_1+d_2-1}, \\ &\eta^{(3)} \in \ell^\infty(\mathbb{R}^p), \eta^{(4)} \in \ell^\infty(\mathbb{R}^{d_1+d_2-1})\Big\}\end{aligned}$$

and endow T_l with the norm

$$\|\eta\|_e := \max\left\{\|\eta^{(1)}\|_2, \|\eta^{(2)}\|_2, \|\eta^{(3)}(X)\|_{P,2}, \|\eta^{(4)}(Z_{-l})\|_{P,2}\right\}.$$

Further, let $\tau_n := \sqrt{\frac{s\log(\bar{d}_n)}{n}}$ and define the corresponding nuisance realization set

$$\mathcal{T}_l := \Big\{\eta \in T_l : \eta^{(3)} \equiv 0, \eta^{(4)} \equiv 0, \|\eta^{(1)}\|_0 \vee \|\eta^{(2)}\|_0 \leq Cs,$$

$$\left.\begin{aligned}\|\eta^{(1)}-\beta_0^{(l)}\|_2 \vee \|\eta^{(2)}-\gamma_0^{(l)}\|_2 \leq C\tau_n,\\ \|\eta^{(1)}-\beta_0^{(l)}\|_1 \vee \|\eta^{(2)}-\gamma_0^{(l)}\|_1 \leq C\sqrt{s}\tau_n\end{aligned}\right\} \cup \{\eta_{0,l}\}$$

for a sufficiently large constant C. For arbitrary random variables X and Y, it holds

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

Due to Assumption A.2(iii), this implies

$$\begin{aligned}\max_{l=1,\dots,d_1}\|\nu^{(l)}\|_{\Psi_\rho} &= \max_{l=1,\dots,d_1}\|g_l(X_1)-\mathbb{E}[g_l(X_1)|Z_{-l}]\|_{\Psi_\rho}\\ &\leq \max_{l=1,\dots,d_1}\|g_l(X_1)\|_{\Psi_\rho} + \max_{l=1,\dots,d_1}\|\mathbb{E}[g_l(X_1)|Z_{-l}]\|_{\Psi_\rho}\\ &\lesssim C.\end{aligned}$$

Therefore, we are able to bound the q-th moments of the maxima by

$$\begin{aligned}\mathbb{E}\left[\max_{l=1,\dots,d_1}|\nu^{(l)}|^q\right]^{\frac{1}{q}} &= \|\max_{l=1,\dots,d_1}|\nu^{(l)}|\|_{P,q}\\ &\leq q!\|\max_{l=1,\dots,d_1}|\nu^{(l)}|\|_{\Psi_1}\\ &\leq q!\log^{\frac{1}{\rho}-1}(2)\|\max_{l=1,\dots,d_1}|\nu^{(l)}|\|_{\Psi_1}\\ &\leq Cq!\log^{\frac{1}{\rho}-1}(2)\log^{\frac{1}{\rho}}(1+d_1)\max_{l=1,\dots,d_1}\|\nu^{(l)}|\|_{\Psi_p}\\ &\leq C\log^{\frac{1}{\rho}}(d_1),\end{aligned}$$

where C does depend on q and ρ but not on n. For $\mathcal{F} := \{\varepsilon\nu^{(l)} : l=1,\dots,d_1\}$, it holds

$$\begin{aligned}\mathcal{S}_n :&= \mathbb{E}\left[\sup_{l=1,\dots,d_1}\left|\sqrt{n}\mathbb{E}_n\left[\psi_l(W,\theta_{0,l},\eta_{0,l})\right]\right|\right]\\ &= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\mathbb{G}_n(f)\right]\end{aligned}$$

and the envelope $\sup_{f\in\mathcal{F}}|f|$ satisfies

$$\begin{aligned}\|\max_{l=1,\dots,d_1}\varepsilon\nu^{(l)}\|_{P,q} &\leq \|\varepsilon\|_{P,2q}\|\max_{l=1,\dots,d_1}\nu^{(l)}\|_{P,2q}\\ &\leq C\log^{\frac{1}{\rho}}(d_1).\end{aligned}$$

We can apply Lemma P.2 from Belloni et al. [12] with $|\mathcal{F}| = d_1$ to obtain

$$\mathcal{S}_n \leq C\log^{\frac{1}{2}}(d_1) + C\log^{\frac{1}{2}}(d_1)\left(n^{\frac{2}{q}}\frac{\log^{\frac{2}{\rho}+1}(d_1)}{n}\right)^{1/2} \lesssim \log^{\frac{1}{2}}(d_1),$$

due to A.2(v)(a). Finally, Assumption A.2(i) implies B.2(i). Assumption B.2(ii) holds since, for all $l=1,\dots,d_1$, the map $(\theta_l,\eta_l)\mapsto\psi_l(X,\theta_l,\eta_l)$ is twice continuously Gateaux-differentiable on $\Theta_l\times\mathcal{T}_l$, which directly implies the differentiability of the map $(\theta_l,\eta_l)\mapsto\mathbb{E}[\psi_l(X,\theta_l,\eta_l)]$. Additionally, for every

$\eta \in \mathcal{T}_l \setminus \{\eta_{0,l}\}$, we have

$$\begin{aligned}
D_{l,0}[\eta, \eta_{0,l}] &:= \partial_t \big\{ \mathbb{E}[\psi_l(W, \theta_{0,l}, \eta_{0,l} + t(\eta - \eta_{0,l}))] \big\}\big|_{t=0} \\
&= \mathbb{E}\big[\partial_t \big\{ \psi_l(W, \theta_{0,l}, \eta_{0,l} + t(\eta - \eta_{0,l})) \big\}\big]\big|_{t=0} \\
&= \mathbb{E}\Big[\partial_t \Big\{ \Big(Y - \theta_{0,l} g_l(X_1) - \big(\eta_{0,l}^{(1)} + t(\eta^{(1)} - \eta_{0,l}^{(1)})\big)^T Z_{-l} \\
&\quad - \big(\eta_{0,l}^{(3)}(X) + t(\eta^{(3)}(X) - \eta_{0,l}^{(3)}(X))\big)\Big) \\
&\quad \Big(g_l(X_1) - \big(\eta_{0,l}^{(2)} + t(\eta^{(2)} - \eta_{0,l}^{(2)})\big)^T Z_{-l} \\
&\quad - \big(\eta_{0,l}^{(4)}(Z_{-l}) + t(\eta^{(4)}(Z_{-l}) - \eta_{0,l}^{(4)}(Z_{-l}))\big)\Big) \Big\}\Big]\Big|_{t=0} \\
&= \mathbb{E}\left[\varepsilon(\eta_{0,l}^{(2)} - \eta^{(2)})^T Z_{-l}\right] + \mathbb{E}\left[\nu^{(l)}(\eta_{0,l}^{(1)} - \eta^{(1)})^T Z_{-l}\right] \\
&\quad + \mathbb{E}\left[\varepsilon\left(\eta_{0,l}^{(4)}(Z_{-l}) - \eta^{(4)}(Z_{-l})\right)\right] + \mathbb{E}\left[\nu^{(l)}\left(\eta_{0,l}^{(3)}(X) - \eta^{(3)}(X)\right)\right]
\end{aligned}$$

with

$$\begin{aligned}
\mathbb{E}\left[\varepsilon(\eta_{0,l}^{(2)} - \eta^{(2)})^T Z_{-l}\right] &= \mathbb{E}\left[((\eta_{0,l}^{(2)} - \eta^{(2)})^T Z_{-l}\mathbb{E}[\varepsilon|X]\right] = 0, \\
\mathbb{E}\left[\nu^{(l)}(\eta_{0,l}^{(1)} - \eta^{(1)})^T Z_{-l}\right] &= \mathbb{E}\left[(\eta_{0,l}^{(1)} - \eta^{(1)})^T Z_{-l}\mathbb{E}[\nu^{(l)}|Z_{-l}]\right] = 0, \\
\mathbb{E}\left[\varepsilon\left(\eta_{0,l}^{(4)}(Z_{-l}) - \eta^{(4)}(Z_{-l})\right)\right] &= \mathbb{E}\left[\left(\eta_{0,l}^{(4)}(Z_{-l}) - \eta^{(4)}(Z_{-l})\right)\mathbb{E}[\varepsilon|X]\right] = 0
\end{aligned}$$

and

$$\mathbb{E}\left[\nu^{(l)}\left(\eta_{0,l}^{(3)}(X) - \eta^{(3)}(X)\right)\right] = \mathbb{E}\left[\nu^{(l)}\big(b_1(X_1) + b_2(X_{-1})\big)\right] \leq C\delta_n n^{-1/2}$$

due to Assumption A.1 with $\delta_n = o\big(t_1^{-\frac{3}{2}} \log^{-\frac{1}{2}}(A_n)\big)$. Due to the linearity of the score and the moment condition, it holds

$$\mathbb{E}[\psi_l(W, \theta_l, \eta_{0,l})] = J_{0,l}(\theta_l - \theta_{0,l})$$

and, due to

$$|J_{0,l}| = \mathbb{E}\left[(\nu^{(l)})^2\right],$$

Assumption B.2(iv) is satisfied.
For all $t \in [0, 1)$, $l = 1, \ldots, d_1$, $\theta_l \in \Theta_l$ and $\eta_l \in \mathcal{T}_l \setminus \{\eta_{0,l}\}$, we have

$$\begin{aligned}
&\mathbb{E}\left[(\psi_l(W, \theta_l, \eta_l) - \psi_l(W, \theta_{0,l}, \eta_{0,l}))^2\right] \\
&= \mathbb{E}\left[(\psi_l(W, \theta_l, \eta_l) - \psi_l(W, \theta_{0,l}, \eta_l) + \psi_l(W, \theta_{0,l}, \eta_l) - \psi_l(W, \theta_{0,l}, \eta_{0,l}))^2\right] \\
&\leq C\Big(\mathbb{E}\left[(\psi_l(W, \theta_l, \eta_l) - \psi_l(W, \theta_{0,l}, \eta_l))^2\right] \\
&\qquad \vee \mathbb{E}\left[(\psi_l(W, \theta_{0,l}, \eta_l) - \psi_l(W, \theta_{0,l}, \eta_{0,l}))^2\right]\Big)
\end{aligned}$$

with

$$\begin{aligned}
&\mathbb{E}\left[\left(\psi_l(W,\theta_l,\eta_l)-\psi_l(W,\theta_{0,l},\eta_l)\right)^2\right]\\
&=|\theta_l-\theta_{0,l}|^2\mathbb{E}\left[\left(g_l(X_1)(g_l(X_1)-(\eta_l^{(2)})^TZ_{-l})-\eta_l^{(4)}(Z_{-l})\right)^2\right]\\
&\leq C|\theta_l-\theta_{0,l}|^2\left(\mathbb{E}\left[g_l(X_1)^4\right]\mathbb{E}\left[\left(g_l(X_1)-(\eta_l^{(2)})^TZ_{-l}-\eta_l^{(4)}(Z_{-l})\right)^4\right]\right)^{\frac{1}{2}}\\
&\leq C|\theta_l-\theta_{0,l}|^2
\end{aligned}$$

due to Assumption A.2(ii), (iv) and the definition of $\mathcal{T}_l$. By similar arguments, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\left(\psi_l(W,\theta_{0,l},\eta_l)-\psi_l(W,\theta_{0,l},\eta_{0,l})\right)^2\right]\\
&=\mathbb{E}\Bigg[\Bigg(\left(Y-\theta_{0,l}g_l(X_1)-(\eta_l^{(1)})^TZ_{-l}-\eta_l^{(3)}(X)\right)\left(g_l(X_1)-(\eta_l^{(2)})^TZ_{-l}-\eta_l^{(4)}(Z_{-l})\right)\\
&\quad-\left(Y-\theta_{0,l}g_l(X_1)-(\eta_{0,l}^{(1)})^TZ_{-l}-\eta_{0,l}^{(3)}(X)\right)\left(g_l(X_1)-(\eta_{0,l}^{(2)})^TZ_{-l}-\eta_{0,l}^{(4)}(Z_{-l})\right)\Bigg)^2\Bigg]\\
&=\mathbb{E}\Bigg[\Bigg(\left(Y-\theta_{0,l}g_l(X_1)-(\eta_l^{(1)})^TZ_{-l}-\eta_l^{(3)}(X)\right)\\
&\quad\cdot\left((\eta_{0,l}^{(2)}-\eta_l^{(2)})^TZ_{-l}+\eta_{0,l}^{(4)}(Z_{-l})-\eta_l^{(4)}(Z_{-l})\right)\\
&\quad+\left(g_l(X_1)-(\eta_{0,l}^{(2)})^TZ_{-l}-\eta_{0,l}^{(4)}(Z_{-l})\right)\\
&\quad\cdot\left((\eta_{0,l}^{(1)}-\eta_l^{(1)})^TZ_{-l}+\eta_{0,l}^{(3)}(X)-\eta_l^{(3)}(X)\right)\Bigg)^2\Bigg]\\
&\leq C\left(\|\eta_{0,l}^{(2)}-\eta_l^{(2)}\|_2\vee\|\eta_{0,l}^{(1)}-\eta_l^{(1)}\|_2\vee\|\eta_{0,l}^{(3)}(X)\|_{P,2}\vee\|\eta_{0,l}^{(4)}(Z_{-l})\|_{P,2}\right)^2\\
&=C\|\eta_{0,l}-\eta_l\|_e^2,
\end{aligned}$$

where we used the definition of $\mathcal{T}_l$, A.1(iii) and

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

Therefore, Assumption B.2$(v)(a)$ holds with $\omega=2$ since it is straightforward to show Assumption B.2(v) for $\eta_l=\eta_{0,l}$. It holds

$$\begin{aligned}
&\left|\partial_t\mathbb{E}\Big[\psi_l(W,\theta_l,\eta_{0,l}+t(\eta_l-\eta_{0,l}))\Big]\right|\\
&=\Bigg|\mathbb{E}\Bigg[\partial_t\Bigg\{\Big(Y-\theta_{0,l}g_l(X_1)-\big(\eta_{0,l}^{(1)}+t(\eta_l^{(1)}-\eta_{0,l}^{(1)})\big)^TZ_{-l}\\
&\quad-\big(\eta_{0,l}^{(3)}(X)+t(\eta_l^{(3)}(X)-\eta_{0,l}^{(3)}(X))\big)\Big)\\
&\quad\cdot\Big(g_l(X_1)-\big(\eta_{0,l}^{(2)}+t(\eta_l^{(2)}-\eta_{0,l}^{(2)})\big)^TZ_{-l}\\
&\quad-\big(\eta_{0,l}^{(4)}(Z_{-l})+t(\eta_l^{(4)}(Z_{-l})-\eta_{0,l}^{(4)}(Z_{-l}))\big)\Big)\Bigg\}\Bigg]\Bigg|\\
&=\Bigg|\mathbb{E}\Bigg[\Big(Y-\theta_{0,l}g_l(X_1)-(\eta_{0,l}^{(1)}+t(\eta_l^{(1)}-\eta_{0,l}^{(1)}))^TZ_{-l}\\
&\quad-(\eta_{0,l}^{(3)}(X)+t(\eta_l^{(3)}(X)-\eta_{0,l}^{(3)}(X)))\Big)
\end{aligned}$$

$$
\begin{aligned}
&\cdot \Big((\eta_{0,l}^{(2)} - \eta_l^{(2)})^T Z_{-l} + \eta_{0,l}^{(4)}(Z_{-l}) - \eta_l^{(4)}(Z_{-l})\Big) \\
&+ \Big(g_l(X_1) - (\eta_{0,l}^{(2)} + t(\eta_l^{(2)} - \eta_{0,l}^{(2)}))^T Z_{-l} \\
&- \big(\eta_{0,l}^{(4)}(Z_{-l}) + t(\eta_l^{(4)}(Z_{-l}) - \eta_{0,l}^{(4)}(Z_{-l}))\big)\Big) \\
&\cdot \Big((\eta_{0,l}^{(1)} - \eta_l^{(1)})^T Z_{-l} + \eta_{0,l}^{(3)}(X) - \eta_l^{(3)}(X)\Big)\Big]\Big| \\
&= |I_{1,1} + I_{1,2} + I_{1,3} + I_{1,4}|
\end{aligned}
$$

with

$$
\begin{aligned}
I_{1,1} &= \mathbb{E}\Big[\Big(Y - \theta_{0,l} g_l(X_1) - (\eta_{0,l}^{(1)} + t(\eta_l^{(1)} - \eta_{0,l}^{(1)}))^T Z_{-l} \\
&\quad - (\eta_{0,l}^{(3)}(X) + t(\eta_l^{(3)}(X) - \eta_{0,l}^{(3)}(X)))\Big)\Big((\eta_{0,l}^{(2)} - \eta_l^{(2)})^T Z_{-l}\Big)\Big] \\
&\leq C\|\eta_{0,l}^{(2)} - \eta_l^{(2)}\|_2, \\
I_{1,2} &= \mathbb{E}\Big[\Big(Y - \theta_{0,l} g_l(X_1) - (\eta_{0,l}^{(1)} + t(\eta_l^{(1)} - \eta_{0,l}^{(1)}))^T Z_{-l} \\
&\quad - (\eta_{0,l}^{(3)}(X) + t(\eta_l^{(3)}(X) - \eta_{0,l}^{(3)}(X)))\Big)\Big(\eta_{0,l}^{(4)}(Z_{-l})\Big)\Big] \\
&\leq C\|\eta_{0,l}^{(4)}(X)\|_{P,2}, \\
I_{1,3} &= \mathbb{E}\Big[\Big(g_l(X_1) - (\eta_{0,l}^{(2)} + t(\eta_l^{(2)} - \eta_{0,l}^{(2)}))^T Z_{-l} \\
&\quad - \big(\eta_{0,l}^{(4)}(Z_{-l}) + t(\eta_l^{(4)}(Z_{-l}) - \eta_{0,l}^{(4)}(Z_{-l}))\big)\Big)\Big((\eta_{0,l}^{(1)} - \eta_l^{(1)})^T Z_{-l}\Big)\Big] \\
&\leq C\|\eta_{0,l}^{(1)} - \eta_l^{(1)}\|_2, \\
I_{1,4} &= \mathbb{E}\Big[\Big(g_l(X_1) - (\eta_{0,l}^{(2)} + t(\eta_l^{(2)} - \eta_{0,l}^{(2)}))^T Z_{-l} \\
&\quad - \big(\eta_{0,l}^{(4)}(Z_{-l}) + t(\eta_l^{(4)}(Z_{-l}) - \eta_{0,l}^{(4)}(Z_{-l}))\big)\Big)\Big(\eta_{0,l}^{(3)}(X)\Big)\Big] \\
&\leq C\|\eta_{0,l}^{(3)}(X)\|_{P,2}.
\end{aligned}
$$

This implies Assumption B.2$(v)(b)$ with $B_{1n} = C$. Finally, to obtain Assumption B.2$(v)(c)$ with $B_{2n} = C$, we note that

$$
\begin{aligned}
&\partial_t^2 \mathbb{E}\left[\psi_l(W, \theta_{0,l} + t(\theta_l - \theta_{0,l}), \eta_{0,l} + t(\eta_l - \eta_{0,l}))\right] \\
&= \partial_t \mathbb{E}\Big[\Big(Y - \big(\theta_{0,l} + t(\theta_l - \theta_{0,l})\big) g_l(X_1) - \big(\eta_{0,l}^{(1)} + t(\eta_l^{(1)} - \eta_{0,l}^{(1)})\big)^T Z_{-l} \\
&\quad - \big(\eta_{0,l}^{(3)}(X) + t(\eta_l^{(3)}(X) - \eta_{0,l}^{(3)}(X))\big)\Big) \\
&\quad \cdot \Big((\eta_{0,l}^{(2)} - \eta_l^{(2)})^T Z_{-l} + \eta_{0,l}^{(4)}(Z_{-l})\Big) \\
&\quad + \Big(g_l(X_1) - \big(\eta_{0,l}^{(2)} + t(\eta_l^{(2)} - \eta_{0,l}^{(2)})\big)^T Z_{-l} \\
&\quad - \big(\eta_{0,l}^{(4)}(Z_{-l}) + t(\eta_l^{(4)}(Z_{-l}) - \eta_{0,l}^{(4)}(Z_{-l}))\big)\Big) \\
&\quad \cdot \Big((\theta_{0,l} - \theta_l)) g_l(X_1) + (\eta_{0,l}^{(1)} - \eta_l^{(1)})^T Z_{-l} + \eta_{0,l}^{(3)}(X)\Big)\Big] \\
&= 2\mathbb{E}\Big[\Big((\theta_{0,l} - \theta_l) g_l(X_1) + (\eta_{0,l}^{(1)} - \eta_l^{(1)})^T Z_{-l} + \eta_{0,l}^{(3)}(X)\Big)
\end{aligned}
$$

$$\cdot \Big((\eta_{0,l}^{(2)} - \eta_l^{(2)}))^T Z_{-l} + \eta_{0,l}^{(4)}(Z_{-l}) \Big) \Big]$$
$$\leq C \left(|\theta_{0,l} - \theta_l|^2 \vee \|\eta_{0,l} - \eta_l\|_e^2 \right)$$

using the same arguments as above.

Assumption B.3

Note that the Assumptions B.3(ii) and (iii) both hold by the construction of $\mathcal{T}_l$ and by the Assumptions A.1(ii) and A.2(ii). The main part to verify Assumption B.3 is to show that the estimates of the nuisance function are contained in the nuisance realization set with high probability. We will rely on uniform Lasso estimation results stated in Appendix 4.10. Therefore, we have to check the Assumptions C.1(i) to (v). Due to Assumption A.2(iii), it holds

$$\max_{j=1,\dots,d_1+d_2} \|Z_j\|_{\Psi_\rho} \leq C \text{ and } \max_{l=1,\dots,d_1} \|\nu^{(l)}\|_{\Psi_\rho} \leq C,$$

which are the tail conditions in Assumption C.1(i) for the auxiliary regressions. Assumption C.1(ii) is directly implied by Assumption A.2(iv) and

$$\min_{l=1,\dots,d_1} \min_{j \neq l} \mathbb{E}\big[(\nu^{(l)})^2 Z_{-l,j}^2\big] = \min_{l=1,\dots,d_1} \min_{j \neq l} \mathbb{E}\big[Z_{-l,j}^2 \underbrace{\mathbb{E}[(\nu^{(l)})^2 | Z_{-l}]}_{=\mathrm{Var}(\nu^{(l)}|Z_{-l}) \geq c} \big] \geq c.$$

Additionally, the uniform sparsity condition in Assumption C.1(iii) holds by Assumption A.2(ii) and the growth condition in Assumption C.1(iv) by Assumption A.2(v)(a). Finally, the condition on the approximation error in Assumption C.1(v) holds due to A.1(ii). Therefore,

$$\hat{\eta}_l^{(2)} \in \mathcal{T}_l \quad \text{for all } l = 1, \dots, d_1$$

with probability $1 - o(1)$. To estimate $\eta_{0,l}^{(1)}$, we run a Lasso regression of Y on Z. By analogous arguments, it holds

$$\|\beta_0^{(l)} - \hat{\beta}^{(l)}\|_0 \leq \|\hat{\theta}\|_0 + \|\hat{\beta}\|_0 \leq Cs,$$
$$\|\beta_0^{(l)} - \hat{\beta}^{(l)}\|_2 \leq \sqrt{\|\theta - \hat{\theta}\|_2^2 + \|\beta_0 - \hat{\beta}\|_2^2} \leq C\sqrt{\frac{s \log(\bar{d}_n)}{n}},$$
$$\|\beta_0^{(l)} - \hat{\beta}^{(l)}\|_1 \leq \|\theta - \hat{\theta}\|_1 + \|\beta_0 - \hat{\beta}\|_1 \leq C\sqrt{\frac{s^2 \log(\bar{d}_n)}{n}}$$

with probability $1 - o(1)$ using Assumptions A.1(ii), A.2(ii)-(v) and

$$\min_{l=1,\dots,d_1+d_2} \mathbb{E}\big[\varepsilon^2 Z_l^2\big] = \min_{l=1,\dots,d_1+d_2} \mathbb{E}\big[Z_l^2 \underbrace{\mathbb{E}[\varepsilon^2 | X]}_{=\mathrm{Var}(\varepsilon|X) \geq c} \big] \geq c.$$

This directly implies that with probability $1 - o(1)$ the nuisance realization set $\mathcal{T}_l$ contains $\hat{\eta}_l^{(1)}$ for all $l = 1, \dots, d_1$.

Combining the results above with $\hat{\eta}^{(3)} \equiv 0$ and $\hat{\eta}^{(4)} \equiv 0$, we obtain Assumption B.3(i). Define

$$\mathcal{F}_1 := \big\{ \psi_l(\cdot, \theta_l, \eta_l) : l = 1, \dots, d_1, \theta_l \in \Theta_l, \eta_l \in \mathcal{T}_l \big\}.$$

To bound the complexity of $\mathcal{F}_1$, we exclude the true nuisance function (the true nuisance function is the only element of $\mathcal{T}_l$ with a nonzero approximation error)

$$\mathcal{F}_{1,1} := \left\{\psi_l(\cdot, \theta_l, \eta_l) : l = 1, \dots, d_1, \theta_l \in \Theta_l, \eta_l \in \mathcal{T}_l \setminus \{\eta_0^{(l)}\}\right\} \subseteq \mathcal{F}_{1,1}^{(1)} \mathcal{F}_{1,1}^{(2)}$$

with

$$\begin{aligned}
\mathcal{F}_{1,1}^{(1)} &:= \left\{W \mapsto Y - \theta_l g_l(X_1) - (\eta_l^{(1)})^T Z_{-l} : l = 1, \dots, d_1, \theta_l \in \Theta_l, \eta_l \in \mathcal{T}_l \setminus \{\eta_0^{(l)}\}\right\}, \\
\mathcal{F}_{1,1}^{(2)} &:= \left\{W \mapsto g_l(X_1) - (\eta_l^{(2)})^T Z_{-l} : l = 1, \dots, d_1, \theta_l \in \Theta_l, \eta_l \in \mathcal{T}_l \setminus \{\eta_0^{(l)}\}\right\}.
\end{aligned}$$

Note that the envelope $F_{1,1}^{(1)}$ of $\mathcal{F}_{1,1}^{(1)}$ satisfies

$$\begin{aligned}
\|F_{1,1}^{(1)}\|_{P,2q} &\leq \Bigg\| \sup_{l=1,\dots,d_1} \sup_{\theta_l \in \Theta_l, \|\eta_{0,l}^{(1)} - \eta_l^{(l)}\|_1 \leq C\sqrt{s}\tau_n} \Big(|\varepsilon| + |\eta_0^{(3)}(X)| \\
&\quad + |(\theta_{0,l} - \theta_l) g_l(X_1)| + |(\eta_{0,l}^{(1)} - \eta_l^{(1)})^T Z_{-l}| \Big) \Bigg\|_{P,2q} \\
&\lesssim \|\varepsilon\|_{P,2q} + \|\eta_0^{(3)}(X)\|_{P,2q} + \| \sup_{l=1,\dots,d_1} g_l(X_1)\|_{P,2q} \\
&\quad + \sqrt{s}\tau_n \| \sup_{j=1,\dots,d_1+d_2} Z_j \|_{P,2q} \\
&\lesssim C + \log^{\frac{1}{\rho}}(d_1) + \sqrt{s}\tau_n \log^{\frac{1}{\rho}}(d_1 + d_2) \\
&\lesssim \log^{\frac{1}{\rho}}(d_1)
\end{aligned}$$

due to A.1(ii), A.2(v) and analogously

$$\|F_{1,1}^{(2)}\|_{P,2q} \lesssim \log^{\frac{1}{\rho}}(d_1),$$

where we assumed $d_1 \geq 2$ without loss of generality. Next, note that due to Lemma 2.6.15 from Vaart and Wellner [94] the set

$$\mathcal{G}_{1,1} := \left\{Z \mapsto \xi^T Z : \xi \in \mathbb{R}^{d_1+d_2+1}, \|\xi\|_0 \leq Cs, \|\xi\|_2 \leq C\right\}$$

is a union over $\binom{d_1+d_2+1}{Cs}$ VC-subgraph classes $\mathcal{G}_{1,1,k}$ with VC indices less or equal to $Cs + 2$. Therefore, $\mathcal{F}_{1,1}^{(1)}$ and $\mathcal{F}_{1,1}^{(2)}$ are unions over $\binom{d_1+d_2+1}{Cs}$ and $\binom{d_1+d_2}{Cs}$ VC-subgraph classes, respectively, which combined with Theorem 2.6.7 from Vaart and Wellner [94] implies

$$\sup_Q \log N(\varepsilon \|F_{1,1}^{(1)}\|_{Q,2}, \mathcal{F}_{1,1}^{(1)}, \|\cdot\|_{Q,2}) \lesssim s \log\left(\frac{d_1 + d_2}{\varepsilon}\right)$$

and

$$\sup_Q \log N(\varepsilon \|F_{1,1}^{(2)}\|_{Q,2}, \mathcal{F}_{1,1}^{(2)}, \|\cdot\|_{Q,2}) \lesssim s \log\left(\frac{d_1 + d_2}{\varepsilon}\right).$$

Using basic calculations, we obtain

$$\sup_Q \log N(\varepsilon \|F_{1,1}^{(1)} \mathcal{F}_{1,1}^{(2)}|_{Q,2}, \mathcal{F}_{1,1}, \|\cdot\|_{Q,2}) \lesssim s \log\left(\frac{d_1 + d_2}{\varepsilon}\right),$$

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

$$\|F_{1,1}\|_{P,q} \leq \|F_{1,1}^{(1)}\|_{P,2q} \|F_{1,1}^{(2)}\|_{P,2q} \lesssim \log^{\frac{2}{\rho}}(d_1).$$

Define

$$\mathcal{F}_{1,2} := \left\{ \psi_l(\cdot, \theta_l, \eta_{0,l}) : l = 1, \ldots, d_1, \theta_l \in \Theta_l \right\}$$

and with an analogous argument we obtain

$$\sup_Q \log N(\varepsilon \|F_{1,2}\|_{Q,2}, \mathcal{F}_{1,2}, \|\cdot\|_{Q,2}) \lesssim \log\left(\frac{d_1}{\varepsilon}\right),$$

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

$$\|F_{1,2}\|_{P,q} \lesssim \log^{\frac{2}{\rho}}(d_1).$$

Combining the results above, we obtain

$$\sup_Q \log N(\varepsilon \|F_1\|_{Q,2}, \mathcal{F}_1, \|\cdot\|_{Q,2}) \lesssim s \log\left(\frac{d_1 + d_2}{\varepsilon}\right),$$

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

$$\|F_1\|_{P,q} \lesssim \log^{\frac{2}{\rho}}(d_1).$$

Therefore, Assumption B.3(iv) holds with $\upsilon_n \lesssim s$, $a_n = d_1 \vee d_2$ and $K_n \lesssim \log^{\frac{2}{\rho}}(d_1)$. For all $f \in \mathcal{F}_1$, we have

$$\mathbb{E}[f^2]^{\frac{1}{2}} \lesssim \sup_{\|\xi\|_2 = 1} \mathbb{E}[(\xi^T Z)^4]^{\frac{1}{2}} \lesssim C$$

and, for each $l = 1, \ldots, d_1$,

$$\begin{aligned}
&\mathbb{E}\left[\psi_l(W, \theta_l, \eta_l)^2\right]^{\frac{1}{2}} \\
&= \mathbb{E}\left[\left(Y - \theta_l g_l(X_1) - (\eta^{(1)})^T Z_{-l} - \eta^{(3)}(X)\right)^2 \left(g_l(X_1) - (\eta^{(2)})^T Z_{-l} - \eta^{(4)}(Z_{-l})\right)^2\right]^{\frac{1}{2}} \\
&= \mathbb{E}\Big[\left(g_l(X_1) - (\eta^{(2)})^T Z_{-l} - \eta^{(4)}(Z_{-l})\right)^2 \\
&\quad \cdot \underbrace{\mathbb{E}\left[\left(Y - \theta_l g_l(X_1) - (\eta^{(1)})^T Z_{-l} - \eta^{(3)}(X)\right)^2 | X\right]}_{\geq Var(\varepsilon|X) \geq c} \Big]^{\frac{1}{2}} \\
&\geq c
\end{aligned}$$

due to Assumption A.2(iv). This implies Assumption B.3(v). Assumption B.3$(vi)(a)$ holds by the definition of τ_n and $\upsilon_n \lesssim s$. For the next growth condition, we note

$$\begin{aligned}
&(B_{1n}\tau_n + \mathcal{S}_n \log(n)/\sqrt{n})^{\omega/2} (\upsilon_n \log(a_n))^{1/2} + n^{-1/2+1/q} \upsilon_n K_n \log(a_n) \\
&\lesssim (\tau_n + \log^{\frac{1}{2}}(d_1) \log(n)/\sqrt{n})(s \log(a_n))^{1/2} + n^{-1/2+1/q} s \log^{\frac{2}{\rho}}(d_1) \log(a_n) \\
&\lesssim \left(n^{\frac{2}{q}} \frac{s^2 \log^{2+\frac{4}{\rho}}(\bar{d}_n)}{n}\right)^{\frac{1}{2}} \\
&\lesssim \delta_n
\end{aligned}$$

with $\delta_n = o\big(t_1^{-\frac{3}{2}} \log^{-\frac{1}{2}}(A_n)\big)$ due to Assumption A.2$(v)(a)$ and analogously

$$n^{1/2} B_{1n}^2 B_{2n}^2 \tau_n^2 \lesssim n^{1/2} \tau_n^2 = \sqrt{\frac{s^2 \log^2(\bar{d}_n)}{n}} \lesssim \delta_n,$$

since q can be chosen arbitrarily large.

Assumption B.4$(i) - (ii)$
Define

$$\mathcal{F}_0 := \{\psi_x(\cdot) : x \in I\},$$

where $\psi_x(\cdot) := (g(x)^T \Sigma_n g(x))^{-1/2} g(x)^T J_0^{-1} \psi(\cdot, \theta_0, \eta_0)$. We note that for any $q > 0$ the envelope F_0 of $\mathcal{F}_0$ satisfies

$$\begin{aligned}
\|F_0\|_{P,q} &= \mathbb{E}\left[\sup_{x \in I} \left|(g(x)^T \Sigma_n g(x))^{-1/2} g(x)^T J_0^{-1} \psi(W, \theta_0, \eta_0)\right|^q\right]^{\frac{1}{q}} \\
&\lesssim \mathbb{E}\left[\sup_{x \in I} \left|g(x)^T J_0^{-1} \psi(W, \theta_0, \eta_0)\right|^q\right]^{\frac{1}{q}} \\
&= \mathbb{E}\left[\sup_{x \in I} \left|\sum_{l=1}^{d_1} g_l(x) J_{0,l}^{-1} \psi_l(W, \theta_{0,l}, \eta_{0,l})\right|^q\right]^{\frac{1}{q}} \\
&\lesssim \mathbb{E}\left[\sup_{x \in I} \left|\sum_{l=1}^{d_1} g_l(x) \varepsilon \nu^{(l)}\right|^q\right]^{\frac{1}{q}} \\
&\lesssim t_1 \mathbb{E}\left[\sup_{l=1,\ldots,d_1} \left|\varepsilon \nu^{(l)}\right|^q\right]^{\frac{1}{q}} \\
&\lesssim t_1 \log^{\frac{1}{\rho}}(d_1).
\end{aligned}$$

By using the same argument as above, we directly obtain B.4(ii) with

$$L_n \lesssim t_1^3 \log^{\frac{3}{\rho}}(d_1).$$

Therefore, we can find a larger envelope $\tilde{F}_0$ with

$$\|\tilde{F}_0\|_{P,q} \lesssim t_1^3 \log^{\frac{3}{\rho}}(d_1).$$

To bound the entropy of $\mathcal{F}_0$, we note that

$$\begin{aligned}
&\big\|\psi_x(W) - \psi_{\tilde{x}}(W)\big\|_{P,2} \\
&= \Big\|(g(x)^T \Sigma_n g(x))^{-1/2} \sum_{l=1}^{d_1} g_l(x) \mathbb{E}[(\nu^{(l)})^2]^{-1} \psi_l(W, \theta_{0,l}, \eta_{0,l}) \\
&\quad - (g(\tilde{x})^T \Sigma_n g(\tilde{x}))^{-1/2} \sum_{l=1}^{d_1} g_l(\tilde{x}) \mathbb{E}[(\nu^{(l)})^2]^{-1} \psi_l(W, \theta_{0,l}, \eta_{0,l})\Big\|_{P,2} \\
&\leq |(g(x)^T \Sigma_n g(x))^{-1/2} - (g(\tilde{x})^T \Sigma_n g(\tilde{x}))^{-1/2}| \\
&\quad \cdot \Big\| \sum_{l=1}^{d_1} g_l(x) \mathbb{E}[(\nu^{(l)})^2]^{-1} \psi_l(W, \theta_{0,l}, \eta_{0,l})\Big\|_{P,2}
\end{aligned}$$

$$
\begin{aligned}
&+ (g(\tilde{x})^T \Sigma_n g(\tilde{x}))^{-1/2} \Big\| \sum_{l=1}^{d_1} \big(g_l(x) - g_l(\tilde{x})\big) \mathbb{E}[(\nu^{(l)})^2]^{-1} \psi_l(W, \theta_{0,l}, \eta_{0,l}) \Big\|_{P,2} \\
&= |(g(x)^T \Sigma_n g(x))^{-1/2} - (g(\tilde{x})^T \Sigma_n g(\tilde{x}))^{-1/2}| \Big\| g(x)^T J_0^{-1} \psi(W, \theta_{0,l}, \eta_{0,l}) \Big\|_{P,2} \\
&\quad + (g(\tilde{x})^T \Sigma_n g(\tilde{x}))^{-1/2} \Big\| \big(g(x) - g(\tilde{x})\big)^T J_0^{-1} \psi(W, \theta_{0,l}, \eta_{0,l}) \Big\|_{P,2} \\
&\lesssim |(g(x)^T \Sigma_n g(x))^{-1/2} - (g(\tilde{x})^T \Sigma_n g(\tilde{x}))^{-1/2}| \sup_{x \in I} \|g(x)\|_2 \\
&\quad + \|g(x) - g(\tilde{x})\|_2
\end{aligned}
$$

due to the sub-multiplicativity of the spectral norm and the bounded eigenvalues.

Additionally, it holds

$$
\begin{aligned}
&|(g(x)^T \Sigma_n g(x))^{-1/2} - (g(\tilde{x})^T \Sigma_n g(\tilde{x}))^{-1/2}| \\
\lesssim& \left| \left(\frac{g(\tilde{x})^T \Sigma_n g(\tilde{x})}{g(x)^T \Sigma_n g(x)} \right)^{1/2} - 1 \right| \\
\lesssim& |g(\tilde{x})^T \Sigma_n g(\tilde{x}) - g(x)^T \Sigma_n g(x)| \\
=& |(g(x) - g(\tilde{x}))^T \Sigma_n (g(x) + g(\tilde{x}))| \\
\leq& |\langle \Sigma_n (g(x) - g(\tilde{x})), (g(x) + g(\tilde{x})) \rangle| \\
\lesssim& \|g(x) - g(\tilde{x})\|_2 \sup_x \|g(x)\|_2,
\end{aligned}
$$

which implies

$$
\big\| \psi_x(W) - \psi_{\tilde{x}}(W) \big\|_{P,2} \lesssim \|g(x) - g(\tilde{x})\|_2 \sup_x \|g(x)\|_2^2.
$$

Using the same argument as in Theorem 2.7.11 from Vaart and Wellner [94], we obtain

$$
\begin{aligned}
&\sup_Q \log N(\varepsilon \|\tilde{F}_0\|_{Q,2}, \mathcal{F}_0, \|\cdot\|_{Q,2}) \\
\lesssim& \sup_Q \log N \left(\left(\frac{\varepsilon t_1^3 \log^{\frac{3}{\rho}}(d_1)}{\sup_x \|g(x)\|_2^2} \right) \sup_x \|g(x)\|_2^2, \mathcal{F}_0, \|\cdot\|_{Q,2} \right) \\
\leq& \log N \left(\left(\frac{\varepsilon t_1^3 \log^{\frac{3}{\rho}}(d_1)}{\sup_x \|g(x)\|_2^2} \right), g(I), \|\cdot\|_2 \right) \\
\lesssim& t_1 \log \left(\frac{A_n}{\varepsilon} \right).
\end{aligned}
$$

Therefore, Assumption B.4(i) is satisfied with $\varrho_n = t_1$.

Assumption B.5

Next, we want to prove that with probability $1 - o(1)$ it holds

$$
\sup_{l=1,\ldots,d_1} |\hat{J}_l - J_{0,l}| = o(1),
$$

where $\hat{J}_l = \mathbb{E}_n[-g_l(X_1)(g_l(X_1) - (\hat{\eta}_l^{(2)})^T Z_{-l})]$.

It holds

$$\begin{aligned}|\hat{J}_l - J_{0,l}| \leq |\hat{J}_l - \mathbb{E}[-g_l(X_1)(g_l(X_1) - (\hat{\eta}_l^{(2)})^T Z_{-l})]| \\ + |\mathbb{E}[-g_l(X_1)(g_l(X_1) - (\hat{\eta}_l^{(2)})^T Z_{-l})] + J_{0,l}|\end{aligned}$$

with

$$\begin{aligned}&|\mathbb{E}[-g_l(X_1)(g_l(X_1) - (\hat{\eta}_l^{(2)})^T Z_{-l})] + J_{0,l}| \\ &\leq |\mathbb{E}[g_l(X_1)(\hat{\eta}_l^{(2)} - \eta_{0,l}^{(2)})^T Z_{-l})]| + |\mathbb{E}[g_l(X_1)\eta_{0,l}^{(4)}(Z_{-l})]| \\ &\lesssim \tau_n.\end{aligned}$$

Let

$$\begin{aligned}\tilde{\mathcal{G}}_1 := \Big\{ X \mapsto -g_l(X_1)(g_l(X_1) - (\eta_l^{(2)})^T Z_{-l}) : l = 1, \dots, d_1, \|\eta_l^{(2)}\|_0 \leq Cs, \\ \|\eta_l^{(2)} - \eta_{0,l}^{(2)}\|_2 \leq C\tau_n, \|\eta^{(2)} - \eta_{0,l}^{(2)}\|_1 \leq C\sqrt{s}\tau_n \Big\}.\end{aligned}$$

The envelope $\tilde{G}_1$ of $\tilde{\mathcal{G}}_1$ satisfies

$$\begin{aligned}\mathbb{E}[\tilde{G}_1^q]^{\frac{1}{q}} &\leq \mathbb{E}\left[\sup_{l=1,\dots,d_1} \sup_{\eta^{(2)}:\|\eta_l^{(2)} - \eta_{0,l}^{(2)}\|_2 \leq C\sqrt{s}\tau_n} |g_l(X_1)|^q |(g_l(X_1) - (\eta_l^{(2)})^T Z_{-l})|^q\right]^{\frac{1}{q}} \\ &\leq \| \sup_{l=1,\dots,d_1} g_l(X_1)\|_{P,2q} \\ &\quad \cdot \mathbb{E}\left[\sup_{l=1,\dots,d_1} \sup_{\eta^{(2)}:\|\eta_l^{(2)} - \eta_{0,l}^{(2)}\|_2 \leq C\sqrt{s}\tau_n} |(g_l(X_1) - (\eta_l^{(2)})^T Z_{-l})|^{2q}\right]^{\frac{1}{2q}} \\ &\lesssim \log^{\frac{1}{\rho}}(d_1)\Big(\| \sup_{l=1,\dots,d_1} \nu^{(l)}\|_{P,2q} \vee \| \sup_{l=1,\dots,d_1} b_3^{(l)}(Z_{-l})\|_{P,2q} \\ &\quad \vee \mathbb{E}\left[\sup_{l=1,\dots,d_1} \sup_{\eta^{(2)}:\|\eta_l^{(2)} - \eta_{0,l}^{(2)}\|_2 \leq C\sqrt{s}\tau_n} (\eta_{0,l}^{(2)} - \eta_l^{(2)})^T Z_{-l})^{2q}\right]^{\frac{1}{2q}}\Big) \\ &\lesssim \log^{\frac{1}{\rho}}(d_1)\left(\log^{\frac{1}{\rho}}(d_1) \vee \sqrt{s}\tau_n \log^{\frac{1}{\rho}}(d_1 + d_2)\right) \\ &\lesssim \log^{\frac{2}{\rho}}(d_1)\end{aligned}$$

and with the same arguments as above we obtain

$$\sup_Q \log N(\varepsilon \|\tilde{G}_1\|_{Q,2}, \tilde{\mathcal{G}}_1, \|\cdot\|_{Q,2}) \lesssim s \log\left(\frac{d_1 + d_2}{\varepsilon}\right).$$

Therefore, by using Lemma P.2 from Belloni et al. [12], it holds

$$\begin{aligned}\sup_{l=1,\dots,d_1} |\hat{J}_l - J_{0,l}| &\lesssim \sup_{f \in \tilde{\mathcal{G}}_1} |\mathbb{E}_n[f(X)] - \mathbb{E}[f(X)]| + \tau_n \\ &\lesssim K\left(\sqrt{\frac{s\log(\bar{d}_n)}{n}} + n^{\frac{1}{q}} \frac{s\log^{\frac{2}{\rho}}(d_1)\log(\bar{d}_n)}{n}\right) + \tau_n\end{aligned}$$

with probability $1 - o(1)$.

Next, we want to bound the restricted eigenvalues of $\hat{\Sigma}_{\varepsilon\nu}$ with high probability by showing

$$\sup_{\|v\|_2=1,\|v\|_0\leq t_1} |v^T\big(\hat{\Sigma}_{\varepsilon\nu} - \Sigma_{\varepsilon\nu}\big)v| \lesssim u_n \tag{4.7}$$

with

$$u_n \lesssim t_1 \left(n^{\frac{1}{q}} \log^{\frac{2}{\rho}}(d_1)\tau_n^2 \vee s\tau_n^3\right)^{\frac{1}{2}}$$

for a suitable $\tilde{q} > \bar{q}$. Define $\xi_i := \varepsilon_i\nu_i$, $\hat{\xi}_i := \hat{\varepsilon}_i\hat{\nu}_i$ and observe that

$$\begin{aligned}
&\hat{\Sigma}_{\varepsilon\nu} - \Sigma_{\varepsilon\nu} \\
&= \frac{1}{n}\sum_{i=1}^n \hat{\xi}_i\hat{\xi}_i^T - \mathbb{E}[\xi_i\xi_i^T] \\
&= \frac{1}{n}\sum_{i=1}^n \xi_i\xi_i^T - \mathbb{E}[\xi_i\xi_i^T] \\
&\quad + \frac{1}{n}\sum_{i=1}^n \xi_i\big(\hat{\xi}_i - \xi_i\big)^T + \frac{1}{n}\sum_{i=1}^n \big(\hat{\xi}_i - \xi_i\big)\xi_i^T + \frac{1}{n}\sum_{i=1}^n \big(\hat{\xi}_i - \xi_i\big)\big(\hat{\xi}_i - \xi_i\big)^T.
\end{aligned}$$

Using the Lemma Q.1 from Belloni et al. [12], we can bound the first part.
Due to the tail conditions on ε and ν, we obtain

$$\begin{aligned}
\left(\mathbb{E}\left[\max_{1\leq i\leq n} \|\varepsilon_i\nu_i\|_\infty^2\right]\right)^{1/2} &\leq \left(\mathbb{E}\left[\max_{1\leq i\leq n} \|\varepsilon_i\|^4\right]\mathbb{E}\left[\max_{1\leq i\leq n} \|\nu_i\|_\infty^4\right]\right)^{1/4} \\
&\lesssim n^{\frac{2}{q}} \log^{\frac{1}{\rho}}(d_1)
\end{aligned}$$

for an arbitrary but fixed $q \geq 4$. Then, Lemma Q.1 implies

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\|v\|_2=1,\|v\|_0\leq t_1} \left|v^T\Big(\frac{1}{n}\sum_{i=1}^n \xi_i\xi_i^T - \mathbb{E}[\xi_i\xi_i^T]\Big)v\right|\right] \\
&= \mathbb{E}\left[\sup_{\|v\|_2=1,\|v\|_0\leq t_1} \left|\mathbb{E}_n\Big[\big(v^T\xi_i\big)^2 - \mathbb{E}\big[\big(v^T\xi_i\big)^2\big]\Big]\right|\right] \\
&\lesssim \tilde{\delta}_n^2 + \tilde{\delta}_n
\end{aligned}$$

with

$$\begin{aligned}
\tilde{\delta}_n &\lesssim \left(n^{\frac{4}{q}} \log^{\frac{2}{\rho}}(d_1) t_1 \log^2(t_1)\log(d_1)\log(n)n^{-1}\right)^{\frac{1}{2}} \\
&\lesssim \left(n^{\frac{5}{q}} \frac{t_1 \log^{1+\frac{2}{\rho}}(d_1)}{n}\right)^{\frac{1}{2}}
\end{aligned}$$

and

$$\frac{\tilde{\delta}_n^2}{u_n^2} \lesssim \left(n^{\frac{1}{\tilde{q}}-\frac{5}{q}} t_1 s\right)^{-1} = o(1)$$

for $q > 5\tilde{q}$. Using Markov's inequality, we directly obtain

$$\sup_{\|v\|_2=1,\|v\|_0\leq t_1} \left|v^T\Big(\frac{1}{n}\sum_{i=1}^n \xi_i\xi_i^T - \mathbb{E}[\xi_i\xi_i^T]\Big)v\right| \lesssim u_n$$

with probability $1 - o(1)$. Note that applying the results on covariance estimation from Chen et al. [28] instead would lead to comparable growth rates. Further, with probability $1 - o(1)$, it holds

$$\sup_{l=1,\dots,d_1} |\hat{\theta}_l - \theta_{0,l}| \lesssim \tau_n$$

due to Appendix A from Belloni et al. [12]. Define

$$\tilde{\mathcal{G}}_2^2 := \{(\psi_l(\cdot, \theta_l, \eta_l) - \psi_l(\cdot, \theta_{0,l}, \eta_{0,l}))^2 : \ l = 1, \dots, d_1, |\theta_l - \theta_{0,l}| \leq C\tau_n, \eta_l \in \mathcal{T}_l \setminus \{\eta_{0,l}\}\}$$

with

$$\sup_Q \log N(\varepsilon \|\tilde{G}_2^2\|_{Q,2}, \tilde{\mathcal{G}}_2^2, \|\cdot\|_{Q,2}) \lesssim s \log\left(\frac{d_1 + d_2}{\varepsilon}\right).$$

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

$$\tilde{G}_2^2 = \sup_{l=1,\dots,d_1} \sup_{\theta_l : |\theta_l - \theta_{0,l}| \leq C\tau_n, \eta_l \in \mathcal{T}_l} \left(\psi_l(W, \theta_l, \eta_l) - \psi_l(W, \theta_{0,l}, \eta_{0,l})\right)^2$$

and

$$\begin{aligned}
&\|\tilde{G}_2^2\|_{P,q} \\
&\lesssim \Big\| \sup_{l, \theta_l, \eta_l^{(2)}, \eta_l^{(4)}} \Big((\theta_{0,l} - \theta_l) g_l(X_1) \big(g_l(X_1) - (\eta_l^{(2)})^T Z_{-l} - \eta_l^{(4)}(Z_{-l}) \big) \Big)^2 \Big\|_{P,q} \\
&\quad + \Big\| \sup_{l, \eta_l} \Big(\big(Y - \theta_{0,l} g_l(X_1) - (\eta_l^{(1)})^T Z_{-l} - \eta_l^{(3)}(X) \big) \big((\eta_{0,l}^{(2)} - \eta_l^{(2)})^T Z_{-l} + \eta_{0,l}^{(4)}(Z_{-l}) - \eta_l^{(4)}(Z_{-l}) \big) \Big)^2 \Big\|_{P,q} \\
&\quad + \Big\| \sup_{l, \eta_l^{(1)}, \eta_l^{(3)}} \Big(\big(g_l(X_1) - (\eta_{0,l}^{(2)})^T Z_{-l} - \eta_{0,l}^{(4)}(Z_{-l}) \big) \big((\eta_{0,l}^{(1)} - \eta_l^{(1)})^T Z_{-l} + \eta_{0,l}^{(3)}(X) - \eta_l^{(3)}(X) \big) \Big)^2 \Big\|_{P,q} \\
&=: T_1 + T_2 + T_3.
\end{aligned}$$

It holds

$$\begin{aligned}
T_1 &\lesssim \tau_n^2 \Big\| \sup_{l, \eta_l^{(2)}, \eta_l^{(4)}} \Big(g_l(X_1) \big(g_l(X_1) - (\eta_l^{(2)})^T Z_{-l} - \eta_l^{(4)}(Z_{-l}) \big) \Big)^2 \Big\|_{P,q} \\
&\leq \tau_n^2 \| \sup_l (g_l(X_1))^2 \|_{P,2q} \Big\| \sup_{l, \eta_l^{(2)}, \eta_l^{(4)}} \Big(g_l(X_1) - (\eta_l^{(2)})^T Z_{-l} - \eta_l^{(4)}(Z_{-l}) \Big)^2 \Big\|_{P,2q} \\
&\lesssim \tau_n^2 \log^{\frac{4}{\rho}}(d_1),
\end{aligned}$$

$$\begin{aligned}
T_2 &\leq \Big\| \sup_{l, \eta_l^{(1)}, \eta_l^{(3)}} \Big(Y - \theta_{0,l} g_l(X_1) - (\eta_l^{(1)})^T Z_{-l} - \eta_l^{(3)}(X) \Big)^2 \Big\|_{P,2q} \\
&\qquad \Big\| \sup_{l, \eta_l^{(2)}, \eta_l^{(4)}} \Big((\eta_{0,l}^{(2)} - \eta_l^{(2)})^T Z_{-l} + \eta_{0,l}^{(4)}(Z_{-l}) - \eta_l^{(4)}(Z_{-l}) \Big)^2 \Big\|_{P,2q} \\
&\lesssim s\tau_n^2 \Big\| \sup_l \|Z_{-l}\|_\infty^2 \Big\|_{P,2q} + \log^{\frac{2}{\rho}}(d_1) \\
&\lesssim s\tau_n^2 \log^{\frac{2}{\rho}}(d_1 + d_2) + \log^{\frac{2}{\rho}}(d_1)
\end{aligned}$$

and

$$\begin{aligned}
T_3 &\leq \|\sup_l (\nu^{(l)})^2\|_{P,2q} \Big\| \sup_{l,\eta_l^{(1)},\eta_l^{(3)}} \Big(\eta_{0,l}^{(1)} - \eta_l^{(1)})^T Z_{-l} + \eta_{0,l}^{(3)}(X) - \eta_l^{(3)}(X)\Big)^2 \Big\|_{P,2q} \\
&\lesssim \log^{\frac{2}{\rho}}(d_1)\Big(s\tau_n^2 \Big\| \sup_l \|Z_{-l}\|_\infty^2 \Big\|_{P,2q} + 1\Big) \\
&\lesssim \log^{\frac{2}{\rho}}(d_1)\Big(s\tau_n^2 \log^{\frac{2}{\rho}}(d_1+d_2) + 1\Big).
\end{aligned}$$

By using an analogous argument as above, we obtain

$$\begin{aligned}
\tilde{\sigma} :&= \sup_{f\in\tilde{\mathcal{G}}_2^2} \mathbb{E}\left[f(X)^2\right]^{\frac{1}{2}} \\
&= \sup_{l=1,\dots,d_1} \sup_{\theta_l : |\theta_l - \theta_{0,l}| \leq C\tau_n, \eta_l \in \mathcal{T}_l} \mathbb{E}\left[(\psi_l(W,\theta_l,\eta_l) - \psi_l(W,\theta_{0,l},\eta_{0,l}))^4\right]^{\frac{1}{2}} \\
&\lesssim \frac{s^2\log(d_1\vee d_2)}{n}.
\end{aligned}$$

Again, we can apply Lemma P.2 from Belloni et al. [12] to obtain

$$\begin{aligned}
\sup_{f\in\tilde{\mathcal{G}}_2^2} |\mathbb{E}_n[f(X)] - \mathbb{E}[f(X)]| &\leq K\left(\tilde{\sigma}\sqrt{\frac{s\log(\bar{d}_n)}{n}} + n^{\frac{1}{q}}\|\tilde{G}_2^2\|_{P,q}\frac{s\log(\bar{d}_n)}{n}\right) \\
&\lesssim s\tau_n^3 \vee n^{\frac{1}{q}}\log^{\frac{2}{\rho}}(d_1)\tau_n^2
\end{aligned}$$

with probability $1 - o(1)$. Note that we have already shown Assumption B.2$(v)(a)$ which implies

$$\sup_{f\in\tilde{\mathcal{G}}_2^2} \mathbb{E}[f(X)] \leq C\left(|\theta_l - \theta_{0,l}|^2 \vee \|\eta_{0,l} - \eta_l\|_e^2\right) \lesssim \tau_n^2.$$

This implies

$$\sup_{l=1,\dots,d_1} \mathbb{E}_n\left[\left(\hat{\varepsilon}_i\hat{\nu}_i^{(l)} - \varepsilon_i\nu_i^{(l)}\right)^2\right] \leq \sup_{f\in\tilde{\mathcal{G}}_2^2} \mathbb{E}_n[f(X)] \lesssim n^{\frac{1}{q}}\log^{\frac{2}{\rho}}(d_1)\tau_n^2 \vee s\tau_n^3$$

and, with an analogous argument, we obtain

$$\sup_{l=1,\dots,d_1} \mathbb{E}_n\left[\left(\varepsilon_i\nu_i^{(l)}\right)^2\right] \lesssim 1.$$

Therefore, it holds

$$\begin{aligned}
&\sup_{\|v\|_2=1,\|v\|_0\leq t_1} |v^T\frac{1}{n}\sum_{i=1}^n \xi_i\big(\hat{\xi}_i - \xi_i\big)^T v| \\
=\ &\sup_{\|v\|_2=1,\|v\|_0\leq t_1} |\mathbb{E}_n\left[v^T\xi_i\big(\hat{\xi}_i - \xi_i\big)^T v\right]| \\
\leq\ &\sup_{\|v\|_2=1,\|v\|_0\leq t_1} \left|\left(\mathbb{E}_n\left[\big(v^T\xi_i\big)^2\right]\mathbb{E}_n\left[\Big(v^T\big(\hat{\xi}_i - \xi_i\big)\Big)^2\right]\right)^{\frac{1}{2}}\right| \\
\lesssim\ &\sup_{\|v\|_2=1,\|v\|_0\leq t_1} \left|\left(\mathbb{E}_n\left[\Big(v^T\big(\hat{\xi}_i - \xi_i\big)\Big)^2\right]\right)^{\frac{1}{2}}\right| \\
=\ &\sup_{\|v\|_2=1,\|v\|_0\leq t_1} \left(\sum_{k=1}^{d_1}\sum_{l=1}^{d_1} v_k v_l \mathbb{E}_n\left[(\hat{\varepsilon}_i\hat{\nu}_i^{(k)} - \varepsilon_i\nu_i^{(k)})(\hat{\varepsilon}_i\hat{\nu}_i^{(l)} - \varepsilon_i\nu_i^{(l)})\right]\right)^{\frac{1}{2}}
\end{aligned}$$

$$\lesssim t_1 \sup_{l=1,\dots,d_1} \mathbb{E}_n \left[(\hat{\varepsilon}_i \hat{\nu}_i^{(l)} - \varepsilon_i \nu_i^{(l)})^2 \right]^{\frac{1}{2}}$$
$$\lesssim t_1 \left(n^{\frac{1}{q}} \log^{\frac{2}{\rho}}(d_1) \tau_n^2 \vee s\tau_n^3 \right)^{\frac{1}{2}}$$

and

$$\sup_{\|v\|_2=1, \|v\|_0 \le t_1} |v^T \frac{1}{n} \sum_{i=1}^n \big(\hat{\xi}_i - \xi_i\big)\big(\hat{\xi}_i - \xi_i\big)^T v| \lesssim t_1^2 \left(n^{\frac{1}{q}} \log^{\frac{2}{\rho}}(d_1) \tau_n^2 \vee s\tau_n^3 \right)$$

with probability $1 - o(1)$. Combining the steps above, this implies (4.7) if $u_n = o(1)$ which is ensured by the growth conditions. Next, note that for every sparse vector $w \in \mathbb{R}^{d_1}$ ($\|w\|_0 \le t_1$) there exists a corresponding matrix

$$M_w \in \mathbb{R}^{d_1 \times d_1} : (M_w)_{k,l} = \begin{cases} 1 \text{ if } w_k \neq 0 \wedge w_l \neq 0 \\ 0 \text{ else} \end{cases}$$

such that

$$w^T(\hat{\Sigma}_{\varepsilon\nu} - \Sigma_{\varepsilon\nu})w = w^T \left(M_w \odot (\Sigma_n - \hat{\Sigma}_n) \right) w.$$

Due to (4.7), it holds

$$\sup_{\|w\|_0 \le t_1} \sup_{\|v\|_2=1} \left| v^T \left(M_w \odot (\hat{\Sigma}_{\varepsilon\nu} - \Sigma_{\varepsilon\nu}) \right) v \right| \le \sup_{\|v\|_2=1, \|v\|_0 \le t_1} \left| v^T (\hat{\Sigma}_{\varepsilon\nu} - \Sigma_{\varepsilon\nu}) v \right| \lesssim u_n,$$

which implies

$$\sup_{\|w\|_0 \le t_1} \|M_w \odot (\hat{\Sigma}_{\varepsilon\nu} - \Sigma_{\varepsilon\nu})\|_2 \lesssim u_n$$

and

$$\sup_{\|w\|_0 \le t_1} \|M_w \odot \hat{\Sigma}_{\varepsilon\nu}\|_2 \lesssim 1$$

due to Assumption A.2(iv). For $v \in \mathbb{R}^{d_1}$, this can be used to show

$$\sup_{\|v\|_2=1, \|v\|_0 \le t_1} |v^T \big(\hat{\Sigma}_n - \Sigma_n\big) v| \lesssim u_n \tag{4.8}$$

with probability $1 - o(1)$ which can be interpreted as an upper bound for the sparse eigenvalues of $\hat{\Sigma}_n - \Sigma_n$. It holds

$$\begin{aligned} \hat{\Sigma}_n - \Sigma_n &= \hat{J}^{-1} \hat{\Sigma}_{\varepsilon\nu} (\hat{J}^{-1})^T - J_0^{-1} \Sigma_{\varepsilon\nu} (J_0^{-1})^T \\ &= \hat{J}^{-1} \hat{\Sigma}_{\varepsilon\nu} \big(\hat{J}^{-1} - J_0^{-1}\big)^T + \big(\hat{J}^{-1} - J_0^{-1}\big) \hat{\Sigma}_{\varepsilon\nu} (J_0^{-1})^T \\ &\quad + J_0^{-1} \big(\hat{\Sigma}_{\varepsilon\nu} - \Sigma_{\varepsilon\nu}\big) \big(J_0^{-1}\big)^T. \end{aligned}$$

Note that

$$\begin{aligned} &\sup_{\|v\|_2=1, \|v\|_0 \le t_1} |v^T \hat{J}^{-1} \hat{\Sigma}_{\varepsilon\nu} \big(\hat{J}^{-1} - J_0^{-1}\big)^T v| \\ = &\sup_{\|v\|_2=1, \|v\|_0 \le t_1} |v^T \hat{J}^{-1} \left(M_v \odot \hat{\Sigma}_{\varepsilon\nu} \right) \big(\hat{J}^{-1} - J_0^{-1}\big)^T v| \end{aligned}$$

$$\leq \left\|\hat{J}^{-1}\right\|_2 \sup_{\|w\|_0\leq t_1} \left\|\left(M_w \odot \hat{\Sigma}_{\varepsilon\nu}\right)\right\|_2 \left\|(\hat{J}^{-1} - J_0^{-1})^T\right\|_2$$
$$\lesssim n^{\frac{1}{q}} \frac{s \log^{\frac{2}{\rho}}(d_1) \log(\bar{d}_n)}{n} + \tau_n$$

due to the sub-multiplicative spectral norm. The same bound holds for the second term. The third term can be bounded by

$$\sup_{\|v\|_2=1, \|v\|_0\leq t_1} |v^T J_0^{-1}\big(\hat{\Sigma}_{\varepsilon\nu} - \Sigma_{\varepsilon\nu}\big)\big(J_0^{-1}\big)^T v| \lesssim u_n.$$

This implies (4.8). We finally obtain

$$\begin{aligned}\sup_{x\in I}\left|\frac{(g(x)^T\hat{\Sigma}_n g(x))^{1/2}}{(g(x)^T\Sigma_n g(x))^{1/2}} - 1\right| &\lesssim \sup_{x\in I}\left|g(x)^T\big(\hat{\Sigma}_n - \Sigma_n\big)g(x)\right| \\ &\leq \sup_{x\in I}\|g(x)\|_2^2 \sup_{\|v\|_2=1,\|v\|_0\leq t_1} |v^T\big(\hat{\Sigma}_n - \Sigma_n\big)v| \\ &\lesssim \sup_{x\in I}\|g(x)\|_2^2 u_n\end{aligned}$$

with probability $1 - o(1)$ and $\epsilon_n \lesssim \sup_{x\in I}\|g(x)\|_2^2 u_n$ which is the first part of Assumption B.5.

Assumption B.4$(iii) - (iv)$
Define

$$\sigma_x := (g(x)^T\Sigma_n g(x))^{1/2}, \quad \hat{\sigma}_x := (g(x)^T\hat{\Sigma}_n g(x))^{1/2}$$

and

$$\hat{\mathcal{F}}_0 := \{\psi_x(\cdot) - \hat{\psi}_x(\cdot) : x \in I\}$$

with $\hat{\psi}_x(\cdot) := \hat{\sigma}_x^{-1} g(x)^T \hat{J}_0^{-1}\psi(\cdot, \hat{\theta}, \hat{\eta})$. For every x and $\tilde{x}$, it holds

$$\begin{aligned}&\|\psi_x(W) - \hat{\psi}_x(W) - (\psi_{\tilde{x}}(W) - \hat{\psi}_{\tilde{x}}(W))\|_{\mathbb{P}_n,2} \\ =&\Big\|\sigma_x^{-1} g(x)^T J_0^{-1}\psi(W,\theta_0,\eta_0) - \sigma_{\tilde{x}}^{-1} g(\tilde{x})^T J_0^{-1}\psi(W,\theta_0,\eta_0) \\ &- \big(\hat{\sigma}_x^{-1} g(x)^T \hat{J}^{-1}\psi(W,\hat{\theta},\hat{\eta}) - \hat{\sigma}_{\tilde{x}}^{-1} g(\tilde{x})^T \hat{J}^{-1}\psi(W,\hat{\theta},\hat{\eta})\big)\Big\|_{\mathbb{P}_n,2} \\ =&\Big\|\sum_{l=1}^{d_1}(\sigma_x^{-1} g_l(x) - \sigma_{\tilde{x}}^{-1} g_l(\tilde{x}))J_{0,l}^{-1}\psi_l(W,\theta_{0,l},\eta_{0,l}) \\ &- \sum_{l=1}^{d_1}(\hat{\sigma}_x^{-1} g_l(x) - \hat{\sigma}_{\tilde{x}}^{-1} g_l(\tilde{x}))\hat{J}_l^{-1}\psi_l(W,\hat{\theta}_l,\hat{\eta}_l)\Big\|_{\mathbb{P}_n,2} \\ \leq&\Big\|\sum_{l=1}^{d_1}(\sigma_x^{-1} g_l(x) - \sigma_{\tilde{x}}^{-1} g_l(\tilde{x}))\Big(J_{0,l}^{-1} - \hat{J}_l^{-1}\Big)\psi_l(W,\theta_{0,l},\eta_{0,l})\Big\|_{\mathbb{P}_n,2} \\ &+ \Big\|\sum_{l=1}^{d_1}(\sigma_x^{-1} g_l(x) - \sigma_{\tilde{x}}^{-1} g_l(\tilde{x}))\hat{J}_l^{-1}\Big(\psi_l(W,\theta_{0,l},\eta_{0,l}) - \psi_l(W,\hat{\theta}_l,\hat{\eta}_l)\Big)\Big\|_{\mathbb{P}_n,2} \\ &+ \Big\|\sum_{l=1}^{d_1}\Big((\sigma_x^{-1} g_l(x) - \sigma_{\tilde{x}}^{-1} g_l(\tilde{x})) - (\hat{\sigma}_x^{-1} g_l(x) - \hat{\sigma}_{\tilde{x}}^{-1} g_l(\tilde{x}))\Big)\hat{J}_l^{-1}\psi_l(W,\hat{\theta}_l,\hat{\eta}_l)\Big\|_{\mathbb{P}_n,2} \\ =:&I_{4,1} + I_{4,2} + I_{4,3}.\end{aligned}$$

We obtain

$$\begin{aligned}
I_{4,1} &= \Big\| \sum_{l=1}^{d_1} (\sigma_x^{-1} g_l(x) - \sigma_{\tilde{x}}^{-1} g_l(\tilde{x})) \Big(J_{0,l}^{-1} - \hat{J}_l^{-1} \Big) \psi_l(W, \theta_{0,l}, \eta_{0,l}) \Big\|_{\mathbb{P}_n,2} \\
&\leq \sigma_x^{-1} \Big\| (g(x) - g(\tilde{x}))^T \Big(J_0^{-1} - \hat{J}^{-1} \Big) \psi(W, \theta_0, \eta_0) \Big\|_{\mathbb{P}_n,2} \\
&\quad + |\sigma_x^{-1} - \sigma_{\tilde{x}}^{-1}| \Big\| g(\tilde{x})^T \Big(J_0^{-1} - \hat{J}^{-1} \Big) \psi(W, \theta_0, \eta_0) \Big\|_{\mathbb{P}_n,2} \\
&\lesssim \|g(x) - g(\tilde{x})\|_2 \sup_{\|v\|_2=1, \|v\|_0 \leq 2t_1} \Big\| v^T \Big(J_0^{-1} - \hat{J}^{-1} \Big) \psi(W, \theta_0, \eta_0) \Big\|_{\mathbb{P}_n,2} \\
&\quad + \|g(x) - g(\tilde{x})\|_2 \sup_{x \in I} \|g(x)\|_2^2 \sup_{\|v\|_2=1, \|v\|_0 \leq t_1} \Big\| v^T \Big(J_0^{-1} - \hat{J}^{-1} \Big) \psi(W, \theta_0, \eta_0) \Big\|_{\mathbb{P}_n,2} \\
&\lesssim \|g(x) - g(\tilde{x})\|_2 \sup_{x \in I} \|g(x)\|_2^2 u_n,
\end{aligned}$$

where we used that

$$\begin{aligned}
&\sup_{\|v\|_2=1, \|v\|_0 \leq t_1} \Big\| v^T \Big(J_0^{-1} - \hat{J}^{-1} \Big) \psi(W, \theta_0, \eta_0) \Big\|_{\mathbb{P}_n,2}^2 \\
&= \sup_{\|v\|_2=1, \|v\|_0 \leq t_1} \Big| v^T \Big(J_0^{-1} - \hat{J}^{-1} \Big) \frac{1}{n} \sum_{i=1}^n \xi_i \xi_i^T \Big(J_0^{-1} - \hat{J}^{-1} \Big)^T v \Big| \\
&\leq \Big\| J_0^{-1} - \hat{J}^{-1} \Big\|_2^2 \sup_{\|v\|_0 \leq t_1} \Big\| M_v \odot \Big(\frac{1}{n} \sum_{i=1}^n \xi_i \xi_i^T \Big) \Big\|_2^2 \\
&\lesssim u_n^2.
\end{aligned}$$

Analogously, we obtain

$$\begin{aligned}
I_{4,2} &= \Big\| \sum_{l=1}^{d_1} (\sigma_x^{-1} g_l(x) - \sigma_{\tilde{x}}^{-1} g_l(\tilde{x})) \hat{J}_l^{-1} \Big(\psi_l(W, \theta_{0,l}, \eta_{0,l}) - \psi_l(W, \hat{\theta}_l, \hat{\eta}_l) \Big) \Big\|_{\mathbb{P}_n,2} \\
&\leq \sigma_x^{-1} \Big\| (g(x) - g(\tilde{x}))^T \hat{J}^{-1} \Big(\psi(W, \theta_0, \eta_0) - \psi(W, \hat{\theta}, \hat{\eta}) \Big) \Big\|_{\mathbb{P}_n,2} \\
&\quad + |\sigma_x^{-1} - \sigma_{\tilde{x}}^{-1}| \Big\| g(\tilde{x})^T \hat{J}^{-1} \Big(\psi(W, \theta_0, \eta_0) - \psi(W, \hat{\theta}, \hat{\eta}) \Big) \Big\|_{\mathbb{P}_n,2} \\
&\lesssim \|g(x) - g(\tilde{x})\|_2 \sup_{\|v\|_2=1, \|v\|_0 \leq 2t_1} \Big\| v^T \hat{J}^{-1} \Big(\psi(W, \theta_0, \eta_0) - \psi(W, \hat{\theta}, \hat{\eta}) \Big) \Big\|_{\mathbb{P}_n,2} \\
&\quad + \|g(x) - g(\tilde{x})\|_2 \sup_{x \in I} \|g(x)\|_2^2 \sup_{\|v\|_2=1, \|v\|_0 \leq t_1} \Big\| v^T \hat{J}^{-1} \Big(\psi(W, \theta_0, \eta_0) - \psi(W, \hat{\theta}, \hat{\eta}) \Big) \Big\|_{\mathbb{P}_n,2} \\
&\lesssim \|g(x) - g(\tilde{x})\|_2 \sup_{x \in I} \|g(x)\|_2^2 u_n.
\end{aligned}$$

It holds

$$\begin{aligned}
I_{4,3} &= \Big\| \sum_{l=1}^{d_1} \Big((\sigma_x^{-1} g_l(x) - \sigma_{\tilde{x}}^{-1} g_l(\tilde{x})) - (\hat{\sigma}_x^{-1} g_l(x) - \hat{\sigma}_{\tilde{x}}^{-1} g_l(\tilde{x})) \Big) \hat{J}_l^{-1} \psi_l(W, \hat{\theta}_l, \hat{\eta}_l) \Big\|_{\mathbb{P}_n,2} \\
&\leq \big| \sigma_x^{-1} - \hat{\sigma}_x^{-1} \big| \Big\| (g(x) - g(\tilde{x}))^T \hat{J}^{-1} \psi(W, \hat{\theta}, \hat{\eta}) \Big\|_{\mathbb{P}_n,2} \\
&\quad + \big| (\sigma_x^{-1} - \hat{\sigma}_x^{-1}) - (\sigma_{\tilde{x}}^{-1} - \hat{\sigma}_{\tilde{x}}^{-1}) \big| \Big\| g(\tilde{x})^T \hat{J}^{-1} \psi(W, \hat{\theta}, \hat{\eta}) \Big\|_{\mathbb{P}_n,2}.
\end{aligned}$$

Note that

$$\begin{aligned}
&\big|(\sigma_x^{-1} - \hat{\sigma}_x^{-1}) - (\sigma_{\tilde{x}}^{-1} - \hat{\sigma}_{\tilde{x}}^{-1})\big| \\
&= \Big|\frac{1}{\sigma_x\sigma_{\tilde{x}}}(\sigma_{\tilde{x}} - \sigma_x) - \frac{1}{\hat{\sigma}_x\hat{\sigma}_{\tilde{x}}}(\hat{\sigma}_{\tilde{x}} - \hat{\sigma}_x)\Big| \\
&= \frac{1}{\hat{\sigma}_x\hat{\sigma}_{\tilde{x}}}\Big|\frac{\hat{\sigma}_x\hat{\sigma}_{\tilde{x}}}{\sigma_x\sigma_{\tilde{x}}}(\sigma_{\tilde{x}} - \sigma_x) - (\hat{\sigma}_{\tilde{x}} - \hat{\sigma}_x)\Big| \\
&\lesssim \big|(\sigma_{\tilde{x}} - \sigma_x) - (\hat{\sigma}_{\tilde{x}} - \hat{\sigma}_x)\big| + \Big|\frac{\hat{\sigma}_x\hat{\sigma}_{\tilde{x}}}{\sigma_x\sigma_{\tilde{x}}} - 1\Big|\big|\sigma_{\tilde{x}} - \sigma_x\big|
\end{aligned}$$

with

$$\begin{aligned}
\Big|\frac{\hat{\sigma}_x\hat{\sigma}_{\tilde{x}}}{\sigma_x\sigma_{\tilde{x}}} - 1\Big|\big|\sigma_{\tilde{x}} - \sigma_x\big| &\leq \Big(\Big|\frac{\hat{\sigma}_x}{\sigma_x} - 1\Big|\frac{\hat{\sigma}_{\tilde{x}}}{\sigma_{\tilde{x}}} + \Big|\frac{\hat{\sigma}_{\tilde{x}}}{\sigma_{\tilde{x}}} - 1\Big|\Big)\big|\sigma_{\tilde{x}} - \sigma_x\big| \\
&\lesssim \epsilon_n\frac{1}{\sigma_x}\big|\sigma_{\tilde{x}}^2 - \sigma_x^2\big| \\
&\lesssim \epsilon_n\|g(x) - g(\tilde{x})\|_2 \sup_x \|g(x)\|_2
\end{aligned}$$

uniformly over $x \in I$ with probability $1 - o(1)$ and

$$\begin{aligned}
&\big|(\sigma_{\tilde{x}} - \sigma_x) - (\hat{\sigma}_{\tilde{x}} - \hat{\sigma}_x)\big| \\
&\leq \frac{1}{(\hat{\sigma}_{\tilde{x}} + \hat{\sigma}_x)}\big|(\sigma_{\tilde{x}}^2 - \sigma_x^2) - (\hat{\sigma}_{\tilde{x}}^2 - \hat{\sigma}_x^2)\big| + \Big|\Big(\frac{1}{(\sigma_{\tilde{x}} + \sigma_x)} - \frac{1}{(\hat{\sigma}_{\tilde{x}} + \hat{\sigma}_x)}\Big)(\sigma_{\tilde{x}}^2 - \sigma_x^2)\Big| \\
&\lesssim \big|(\sigma_{\tilde{x}}^2 - \sigma_x^2) - (\hat{\sigma}_{\tilde{x}}^2 - \hat{\sigma}_x^2)\big| + \Big|\frac{(\hat{\sigma}_{\tilde{x}} + \hat{\sigma}_x)}{(\sigma_{\tilde{x}} + \sigma_x)} - 1\Big|\big|\sigma_{\tilde{x}}^2 - \sigma_x^2\big|.
\end{aligned}$$

Using an analogous argument as in the verification of Assumption B.5, we obtain

$$\begin{aligned}
|(\sigma_x^2 - \hat{\sigma}_x^2) - (\sigma_{\tilde{x}}^2 - \hat{\sigma}_{\tilde{x}}^2)| &= |(g(x) - g(\tilde{x}))^T(\Sigma_n - \hat{\Sigma}_n)(g(x) + g(\tilde{x}))| \\
&\leq \|(\Sigma_n - \hat{\Sigma}_n)(g(x) - g(\tilde{x}))\|_2 \sup_{x\in I}\|g(x)\|_2 \\
&\lesssim \|g(x) - g(\tilde{x})\|_2 u_n \sup_{x\in I}\|g(x)\|_2
\end{aligned}$$

with probability $1 - o(1)$, where the last inequality holds due the order of the sparse eigenvalues in (4.8). Additionally,

$$\begin{aligned}
\Big|\frac{(\hat{\sigma}_{\tilde{x}} + \hat{\sigma}_x)}{(\sigma_{\tilde{x}} + \sigma_x)} - 1\Big|\big|\sigma_{\tilde{x}}^2 - \sigma_x^2\big| &\leq \sup_{x\in I}\Big|\frac{\hat{\sigma}_x}{\sigma_x} - 1\Big|\big|\sigma_{\tilde{x}}^2 - \sigma_x^2\big| \\
&\lesssim \epsilon_n\|g(x) - g(\tilde{x})\|_2 \sup_{x\in I}\|g(x)\|_2
\end{aligned}$$

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

$$\begin{aligned}
I_{4,3} &\lesssim \epsilon_n\|g(x) - g(\tilde{x})\|_2 \sup_{\|v\|_2=1,\|v\|_0\leq 2t_1}\Big\|v^T\hat{J}^{-1}\psi(W,\hat{\theta},\hat{\eta})\Big\|_{\mathbb{P}_n,2} \\
&\quad + (\epsilon_n \vee u_n)\|g(x) - g(\tilde{x})\|_2 \sup_{x\in I}\|g(x)\|_2^2 \sup_{\|v\|_2=1,\|v\|_0\leq t_1}\Big\|v^T\hat{J}^{-1}\psi(W,\hat{\theta},\hat{\eta})\Big\|_{\mathbb{P}_n,2} \\
&\lesssim \|g(x) - g(\tilde{x})\|_2\epsilon_n \sup_{x\in I}\|g(x)\|_2^2.
\end{aligned}$$

Combining the steps above, we obtain

$$\|\psi_x(W) - \hat{\psi}_x(W) - (\psi_{\tilde{x}}(W) - \hat{\psi}_{\tilde{x}}(W))\|_{\mathbb{P}_n,2} \leq \|g(x) - g(\tilde{x})\|_2\|\hat{F}_0\|_{\mathbb{P}_n,2}$$

with

$$\|\hat{F}_0\|_{\mathbb{P}_n,2} \lesssim \epsilon_n \sup_{x\in I} \|g(x)\|_2^2 = o(1)$$

due to the growth condition in Assumption A.2$(v)(b)$ as shown below. Using the same argument as in Theorem 2.7.11 from Vaart and Wellner [94], we obtain with probability $1 - o(1)$

$$\begin{aligned}\log N(\varepsilon, \hat{\mathcal{F}}_0, \|\cdot\|_{\mathbb{P}_n,2}) &\leq \log N(\varepsilon \|\hat{F}_0\|_{\mathbb{P}_n,2}, \hat{\mathcal{F}}_0, \|\cdot\|_{\mathbb{P}_n,2}) \\ &\leq \log N(\varepsilon, g(I), \|\cdot\|_2) \\ &\leq \bar{\varrho}_n \log\left(\frac{\bar{A}_n}{\varepsilon}\right)\end{aligned}$$

with $\bar{\varrho}_n = t_1$ and $\bar{A}_n \lesssim A_n$. Additionally, it holds

$$\begin{aligned}&\|\psi_x(W) - \hat{\psi}_x(W)\|_{\mathbb{P}_n,2} \\ &= \left\|\sigma_x^{-1} g(x)^T J_0^{-1} \psi(W, \theta_0, \eta_0) - \hat{\sigma}_x^{-1} g(x)^T \hat{J}^{-1} \psi(W, \hat{\theta}, \hat{\eta})\right\|_{\mathbb{P}_n,2} \\ &\leq \sigma_x^{-1} \left\|g(x)^T \left(J_0^{-1} - \hat{J}^{-1}\right) \psi(W, \theta_0, \eta_0)\right\|_{\mathbb{P}_n,2} \\ &\quad + \sigma_x^{-1} \left\|g(x)^T \hat{J}^{-1} \Big(\psi(W, \theta_0, \eta_0) - \psi(W, \hat{\theta}, \hat{\eta})\Big)\right\|_{\mathbb{P}_n,2} \\ &\quad + |\sigma_x^{-1} - \hat{\sigma}_x^{-1}| \left\|g(x)^T \hat{J}^{-1} \psi(W, \hat{\theta}, \hat{\eta})\right\|_{\mathbb{P}_n,2} \\ &\lesssim \sup_{x\in I} \|g(x)\|_2 (u_n \vee \epsilon_n) \\ &\lesssim \sup_{x\in I} \|g(x)\|_2 \epsilon_n\end{aligned}$$

with an analogous argument as above. Therefore, B.4(iii) holds with

$$\bar{\delta}_n \lesssim \sup_{x\in I} \|g(x)\|_2 \epsilon_n.$$

To complete the proof, we verify all growth conditions from Assumptions B.4 and B.5. As shown in the verification of B.3(vi), it holds

$$t_1^2 \delta_n^2 \varrho_n \log(A_n) = \delta_n^2 t_1^3 \log(A_n) = o(1).$$

Additionally, it holds

$$n^{-\frac{1}{7}} L_n^{\frac{2}{7}} \varrho_n \log(A_n) = \frac{t_1^{\frac{13}{7}} \log^{\frac{6}{7\rho}}(d_1) \log(A_n)}{n^{\frac{1}{7}}} = o(1)$$

and

$$n^{\frac{2}{3q} - \frac{1}{3}} L_n^{\frac{2}{3}} \varrho_n \log(A_n) = n^{\frac{2}{3q}} \frac{t_1^3 \log^{\frac{2}{\rho}}(d_1) \log(A_n)}{n^{\frac{1}{3}}} = o(1)$$

for q large enough due to the growth condition in Assumption A.2$(v)(c)$. Note that

$$\varepsilon_n \varrho_n \log(A_n) = \varepsilon_n t_1 \log(A_n) \lesssim \bar{\delta}_n t_1 \log(A_n).$$

Hence, we need to show that

$$\bar{\delta}_n^2 \bar{\varrho}_n \varrho_n \log(\bar{A}_n) \log(A_n) = \bar{\delta}_n^2 t_1^2 \log^2(A_n) = o(1).$$

It holds

$$\begin{aligned}
\bar{\delta}_n^2 t_1^2 \log^2(A_n) &\lesssim u_n^2 \sup_{x \in I} \|g(x)\|_2^6 t_1^2 \log^2(A_n) \\
&\lesssim \left(n^{\frac{1}{q}} \log^{\frac{2}{\rho}}(d_1) \tau_n^2 \vee s\tau_n^3 \right) \sup_{x \in I} \|g(x)\|_2^6 t_1^4 \log^2(A_n) \\
&= o(1)
\end{aligned}$$

due to Assumption A.2$(v)(b)$. This completes the proof.

□

Appendix

4.9 Uniformly Valid Confidence Bands

As in Belloni et al. [12], we consider the problem of estimating the set of parameters $\theta_{0,l}$, $l = 1, \dots, d_1$, in the moment condition model

$$\mathbb{E}[\psi_l(W, \theta_{0,l}, \eta_{0,l})] = 0, \qquad l = 1, \dots, d_1, \tag{4.9}$$

where W is a random variable, ψ_l a known score function, $\theta_{0,l} \in \Theta_l$ a scalar of interest, and $\eta_{0,l} \in T_l$ a high-dimensional nuisance parameter. T_l is a convex set in a normed space equipped with a norm $\|\cdot\|_e$. Let $\mathcal{T}_l$ be some subset of T_l which contains the nuisance estimate $\hat{\eta}_l$ with high probability. Belloni et al. [12] provide an appropriate estimator $\hat{\theta}_l$ and are able to construct simultaneous confidence bands for $(\theta_{0,l})_{l=1,\dots,d_1}$, where d_1 may increase with the sample size n. In this section, we are particularly interested in the linear functional

$$G(x) = \sum_{l=1}^{d_1} \theta_{0,l} g_l(x),$$

where $(g_l)_{l=1,\dots,d_1}$ is a given set of functions with

$$g_l : I \subseteq \mathbb{R} \to \mathbb{R}, \qquad l = 1, \dots, d_1.$$

We assume that the score functions ψ_l are constructed to satisfy the near-orthogonality condition, namely

$$D_{l,0}[\eta, \eta_{0,l}] := \partial_t \big\{ \mathbb{E}[\psi_l(W, \theta_{0,l}, \eta_{0,l} + t(\eta - \eta_{0,l}))] \big\}\big|_{t=0} \lesssim \delta_n n^{-1/2}, \tag{4.10}$$

where ∂_t denotes the derivative with respect to t and $(\delta_n)_{n \geq 1}$ a sequence of positive constants converging to zero. We aim to construct uniform valid confidence bands for the target function $G(x)$, namely

$$P(\hat{l}(x) \leq G(x) \leq \hat{u}(x), \forall x \in I) \to 1 - \alpha.$$

Let $\hat{\eta}_l$ be an estimator of the nuisance function. The estimator $\hat{\theta}_0$ of the target parameter

$$\theta_0 = (\theta_{0,1}, \dots, \theta_{0,d_1})^T$$

is defined as the solution of

$$\sup_{l=1,\dots,d_1} \left\{ \left| \mathbb{E}_n \Big[\psi_l\big(W, \hat{\theta}_l, \hat{\eta}_l\big) \Big] \right| - \inf_{\theta \in \Theta_l} \left| \mathbb{E}_n \Big[\psi_l\big(W, \theta, \hat{\eta}_l\big) \Big] \right| \right\} \leq \epsilon_n, \tag{4.11}$$

where $\epsilon_n = o\left(\delta_n n^{-1/2}\right)$ is the numerical tolerance. Let

$$g(x) = (g_1(x), \dots, g_{d_1}(x))^T \in \mathbb{R}^{d_1 \times 1}$$

and

$$\psi(W, \theta, \eta) = (\psi_1(W, \theta, \eta), \dots, \psi_{d_1}(W, \theta, \eta))^T \in \mathbb{R}^{d_1 \times 1}.$$

Define the Jacobian matrix

$$J_0 := \frac{\partial}{\partial \theta}\mathbb{E}[\psi(W,\theta,\eta_0)]\Big|_{\theta=\theta_0} = \operatorname{diag}\left(J_{0,1},\ldots,J_{0,d_1}\right) \in \mathbb{R}^{d_1\times d_1}$$

and the approximate covariance matrix

$$\Sigma_n := J_0^{-1}\mathbb{E}\big[\psi(W,\theta_0,\eta_0)\psi(W,\theta_0,\eta_0)^T\big](J_0^{-1})^T \in \mathbb{R}^{d_1\times d_1}.$$

Additionally, define

$$\mathcal{S}_n := \mathbb{E}\left[\sup_{l=1,\ldots,d_1}\left|\sqrt{n}\mathbb{E}_n\left[\psi_l(W,\theta_{0,l},\eta_{0,l})\right]\right|\right]$$

and

$$t_1 := \sup_{x\in I}\|g(x)\|_0.$$

The definition of t_1 is helpful if the functions g_l, $l=1,\ldots,d_1$, are local in the sense that for any point x in I there are at most $t_1 \ll d_1$ nonzero functions. Now, we state the conditions needed for the uniformly valid confidence bands.

Assumption B. 1. *It holds*

(i) $\inf_{x\in I}\|g(x)\|_2^2 \geq c > 0$,

(ii) $\sup_{x\in I}\sup_{l=1,\ldots,d_1}|g_l(x)| \leq C < \infty$,

(iii) The eigenvalues from Σ_n *are uniformly bounded from above and away from zero.*

Since the proof of our main result in this section relies on the techniques in Belloni et al. [12], we try formulate the following conditions as similar as possible to make the use of their methodology transparent.

Assumption B. 2. *For all* $n \geq n_0, P \in \mathcal{P}_n$ *and* $l \in \{1,\ldots,d_1\}$*, the following conditions hold:*

(i) The true parameter value $\theta_{0,l}$ *obeys (4.9), and* Θ_l *contains a ball of radius* $C_0 n^{-1/2}\mathcal{S}_n \log(n)$ *centered at* $\theta_{0,l}$.

(ii) The map $(\theta_l,\eta_l) \mapsto \mathbb{E}[\psi_l(W,\theta_l,\eta_l)]$ *is twice continuously Gateaux-differentiable on* $\Theta_l \times \mathcal{T}_l$.

(iii) The score function ψ_l *obeys the near orthogonality condition (4.10) for the set* $\mathcal{T}_l \subset T_l$.

(iv) For all $\theta_l \in \Theta_l$, $|\mathbb{E}[\psi_l(W,\theta_l,\eta_{0,l})]| \geq 2^{-1}|J_{0,l}(\theta_l-\theta_{0,l})| \wedge c_0$, *where* $J_{0,l}$ *satisfies* $c_0 \leq |J_{0,l}| \leq C_0$.

(v) For all $r \in [0,1), \theta_l \in \Theta_l$ *and* $\eta_l \in \mathcal{T}_l$*, it holds*

(a) $\mathbb{E}[(\psi_l(W,\theta_l,\eta_l)-\psi_l(W,\theta_{0,l},\eta_{0,l}))^2] \leq C_0(|\theta_l-\theta_{0,l}| \vee \|\eta_l-\eta_{0,l}\|_e)^\omega$,

(b) $|\partial_r\mathbb{E}[\psi_l(W,\theta_l,\eta_{0,l}+r(\eta_l-\eta_{0,l}))]| \leq B_{1n}\|\eta_l-\eta_{0,l}\|_e$,

(c) $|\partial_r^2\mathbb{E}[\psi_l(W,\theta_{0,l}+r(\theta_l-\theta_{0,l}),\eta_{0,l}+r(\eta_l-\eta_{0,l}))]| \leq B_{2n}(|\theta_l-\theta_{0,l}|^2 \vee \|\eta_l-\eta_{0,l}\|_e^2)$.

Note that the notation $\mathbb{E}$ abbreviates $\mathbb{E}_P$. For a detailed discussion of the ideas and intuitions of these and the following assumptions, we refer to Belloni et al. [12].
Let $(\Delta_n)_{n\geq 1}$ and $(\tau_n)_{n\geq 1}$ be some sequences of positive constants converging to zero. Also, let $(a_n)_{n\geq 1}$, $(\upsilon_n)_{n\geq 1}$, and $(K_n)_{n\geq 1}$ be some sequences of positive constants, possibly growing to infinity, where $a_n \geq n \vee K_n$ and $\upsilon \geq 1$ for all $n \geq 1$. Finally, let $q \geq 2$ be some constant.

Assumption B. 3. *For all $n \geq n_0$ and $P \in \mathcal{P}_n$, the following conditions hold:*

(i) With probability at least $1 - \Delta_n$, we have $\hat{\eta}_l \in \mathcal{T}_l$ for all $l = 1, \ldots, d_1$.

(ii) For all $l = 1, \ldots, d_1$ and $\eta_l \in \mathcal{T}_l$, it holds $\|\eta_l - \eta_{0,l}\|_e \leq \tau_n$.

(iii) For all $l = 1, \ldots, d_1$, we have $\eta_{0,l} \in \mathcal{T}_l$.

(iv) The function class $\mathcal{F}_1 = \{\psi_l(\cdot, \theta_l, \eta_l) : l = 1, \ldots, d_1, \theta_l \in \Theta_l, \eta_l \in \mathcal{T}_l\}$ is suitably measurable and its uniform entropy numbers obey

$$\sup_Q \log N(\epsilon \|F_1\|_{Q,2}, \mathcal{F}_1, \|\cdot\|_{Q,2}) \leq \upsilon_n \log(a_n/\epsilon), \quad \text{for all } 0 < \epsilon \leq 1,$$

where F_1 is a measurable envelope for $\mathcal{F}_1$ that satisfies $\|F_1\|_{P,q} \leq K_n$.

(v) For all $f \in \mathcal{F}_1$, we have $c_0 \leq \|f\|_{P,2} \leq C_0$.

(vi) The complexity characteristics a_n and υ_n satisfy

(a) $(\upsilon_n \log(a_n)/n)^{1/2} \leq C_0 \tau_n$,

(b) $(B_{1n}\tau_n + \mathcal{S}_n \log(n)/\sqrt{n})^{\omega/2} (\upsilon_n \log(a_n))^{1/2} + n^{-1/2+1/q} \upsilon_n K_n \log(a_n) \leq C_0 \delta_n$,

(c) $n^{1/2} B_{1n}^2 B_{2n}^2 \tau_n^2 \leq C_0 \delta_n$.

Whereas the Assumptions B.2 and B.3 are identical to the Assumptions 2.1 and 2.2 from Belloni et al. [12], the analogs to their Assumptions 2.3 and 2.4 need modifications to fit our setting constructing a uniformly valid confidence band for the linear functional $G(x)$. In this context, define

$$\psi_x(\cdot) := (g(x)^T \Sigma_n g(x))^{-1/2} g(x)^T J_0^{-1} \psi(\cdot, \theta_0, \eta_0)$$

and the corresponding plug-in estimator

$$\hat{\psi}_x(\cdot) := (g(x)^T \hat{\Sigma}_n g(x))^{-1/2} g(x)^T \hat{J}_0^{-1} \psi(\cdot, \hat{\theta}_0, \hat{\eta}_0).$$

Let $(\bar{\delta}_n)_{n \geq 1}$ be a sequence of positive constants converging to zero. Also, let $(\varrho_n)_{n \geq 1}$, $(\bar{\varrho}_n)_{n \geq 1}$, $(A_n)_{n \geq 1}$, $(\bar{A}_n)_{n \geq 1}$, and $(L_n)_{n \geq 1}$ be some sequences of positive constants, possibly growing to infinity, where $\varrho \geq 1$, $A_n \geq n$, and $\bar{A}_n \geq n$ for all $n \geq 1$. In addition, assume that $q > 4$.

Assumption B. 4. *For all $n \geq n_0$ and $P \in \mathcal{P}_n$, the following conditions hold:*

(i) The function class $\mathcal{F}_0 = \{\psi_x(\cdot) : x \in I\}$ is suitably measurable and its uniform entropy numbers obey

$$\sup_Q \log N(\varepsilon \|F_0\|_{Q,2}, \mathcal{F}_0, \|\cdot\|_{Q,2}) \leq \varrho_n \log(A_n/\varepsilon), \quad \text{for all } 0 < \epsilon \leq 1,$$

where F_0 is a measurable envelope for $\mathcal{F}_0$ that satisfies $\|F_0\|_{P,q} \leq L_n$.

(ii) For all $f \in \mathcal{F}_0$ and $k = 3, 4$, we have $\mathbb{E}[|f(W)|^k] \leq C_0 L_n^{k-2}$.

(iii) The function class $\hat{\mathcal{F}}_0 = \{\psi_x(\cdot) - \hat{\psi}_x(\cdot) : x \in I\}$ satisfies with probability $1 - \Delta_n$:

$$\log N(\varepsilon, \hat{\mathcal{F}}_0, \|\cdot\|_{\mathbb{P}_n,2}) \leq \bar{\varrho}_n \log(\bar{A}_n/\varepsilon), \quad \text{for all } 0 < \epsilon \leq 1$$

and $\|f\|_{\mathbb{P}_n,2} \leq \bar{\delta}_n$ for all $f \in \hat{\mathcal{F}}_0$.

(iv) $t_1^2\delta_n^2\varrho_n\log(A_n)=o(1)$, $L_n^{2/7}\varrho_n\log(A_n)=o(n^{1/7})$ *and* $L_n^{2/3}\varrho_n\log(A_n)=o(n^{1/3-2/(3q)})$.

Additionally, we need to be able to estimate the variance of the linear functional sufficiently well. Let $\hat{\Sigma}_n$ be an estimator of Σ_n.

Assumption B. 5. *For all* $n\geq n_0$ *and* $P\in\mathcal{P}_n$*, it holds*

$$P\left(\sup_{x\in I}\left|\frac{(g(x)^T\hat{\Sigma}_n g(x))^{1/2}}{(g(x)^T\Sigma_n g(x))^{1/2}}-1\right|>\varepsilon_n\right)\leq\Delta_n,$$

where $\varepsilon_n\varrho_n\log(A_n)=o(1)$ *and* $\bar{\delta}_n^2\bar{\varrho}_n\varrho_n\log(\bar{A}_n)\log(A_n)=o(1)$.

As in Chernozhukov et al. [30], we employ the Gaussian multiplier bootstrap method to estimate the respective quantiles. Let

$$\hat{\mathcal{G}}=\left(\hat{\mathcal{G}}_x\right)_{x\in I}=\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n\xi_i\hat{\psi}_x(W_i)\right)_{x\in I},$$

where $(\xi_i)_{i=1}^n$ are independent standard normal random variables (especially independent from $(W_i)_{i=1}^n$). Define the multiplier bootstrap critical value c_α as the $(1-\alpha)$-quantile of the conditional distribution of $\sup_{x\in I}|\hat{\mathcal{G}}_x|$ given $(W_i)_{i=1}^n$.

Theorem 10. *Define*

$$\hat{u}(x):=\hat{G}(x)+\frac{(g(x)'\hat{\Sigma}_n g(x))^{1/2}c_\alpha}{\sqrt{n}},$$
$$\hat{l}(x):=\hat{G}(x)-\frac{(g(x)'\hat{\Sigma}_n g(x))^{1/2}c_\alpha}{\sqrt{n}}$$

with $\hat{G}(x)=g(x)^T\hat{\theta}_0$*. Given the Assumptions B.1-B.5, it holds*

$$P\left(\hat{l}(x)\leq G(x)\leq\hat{u}(x),\forall x\in I\right)\to 1-\alpha$$

uniformly over $P\in\mathcal{P}_n$.

Proof. Since Theorem 2.1 in Belloni et al. [12] is not directly applicable to our problem, we have to modify the proof to obtain a uniform Bahadur representation. We want to prove that

$$\sup_{x\in I}\left|\sqrt{n}(g(x)^T\Sigma_n g(x))^{-1/2}g(x)^T\left(\hat{\theta}-\theta_0\right)\right|=\sup_{x\in I}\left|\mathbb{G}_n(\psi_x)\right|+O_P(t_1\delta_n).\tag{4.12}$$

Assumptions B.2 and B.3 contain Assumptions 2.1 and 2.2 from Belloni et al. [12] which enables us to use parts of their results. Therefore, it holds

$$\sup_{l=1,\dots,d_1}\left|J_{0,l}^{-1}\sqrt{n}\mathbb{E}_n\left[\psi_l(W,\theta_{0,l},\eta_{0,l})\right]+\sqrt{n}\left(\hat{\theta}_l-\theta_{0,l}\right)\right|=O_P(\delta_n).$$

Using Assumption B.1, this implies

$$\begin{aligned}&\sup_{x\in I}\left|\sqrt{n}\mathbb{E}_n\left[g(x)^TJ_0^{-1}\psi(W,\theta_0,\eta_0)\right]+\sqrt{n}g(x)^T\left(\hat{\theta}-\theta_0\right)\right|\\ =\;&\sup_{x\in I}\left|\sum_{j=1}^{d_1}g_l(x)\left(J_{0,l}^{-1}\sqrt{n}\mathbb{E}_n\left[\psi_l(W,\theta_{0,l},\eta_{0,l})\right]+\sqrt{n}(\hat{\theta}_l-\theta_{0,l})\right)\right|\end{aligned}$$

$$\leq t_1 \underbrace{\sup_{x\in I}\sup_{l=1,\dots,d_1}|g_l(x)|}_{\leq C}\sup_{l=1,\dots,d_1}\left|J_{0,l}^{-1}\sqrt{n}\mathbb{E}_n\left[\psi_l(W,\theta_{0,l},\eta_{0,l})\right]+\sqrt{n}\left(\hat{\theta}_l-\theta_{0,l}\right)\right|$$
$$= O_p(t_1\delta_n).$$

Since the minimal eigenvalue of Σ_n is uniformly bounded away from zero, it follows that $g(x)^T\Sigma_n g(x)$ is uniformly bounded away from zero as long as $\|g(x)\|_2^2$ is uniformly bounded away from zero due to Assumption B.1. This implies (4.12).
Due to Assumption B.5, it holds

$$P\left(\sup_{x\in I}\left|\frac{(g(x)^T\hat{\Sigma}_n g(x))^{1/2}}{(g(x)^T\Sigma_n g(x))^{1/2}}-1\right|>\varepsilon_n\right)\leq\Delta_n$$

with $\Delta_n = o(1)$, which is an analogous version of the Assumption 2.4 from Belloni et al. [12]. Therefore, given the Assumptions B.2-B.5, the proofs of Corollary 2.1 and Corollary 2.2 from Belloni et al. [12] can be applied implying the stated theorem. □

4.10 Uniform Nuisance Function Estimation

To establish uniform estimation properties of the nuisance function, we rely on uniform estimation results from Klaassen et al. [60]. Consider the following linear regression model

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

with centered regressors and $a_r(X_r)$ accounts for an approximation error. The errors ε_r are assumed to satisfy $\mathbb{E}[\varepsilon_r|X_r]=0$ for each $r=1,\dots,d$.
The true parameter obeys

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

We show that the Lasso and post-Lasso estimators have sufficiently fast uniform estimation rates if the vector β_r is sparse for all $r=1,\dots,d$. Due to the approximation error $a_r(X_r)$, the sparsity assumption is quite mild and contains an approximate sparse setting. 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 $\Psi_\rho(x)=\exp(x^\rho)-1$.

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 fixed $m \geq 0$. Define the post-regularized weighted least squares (post-Lasso) estimator as

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

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

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 $n \geq n_0$ and $P \in \mathcal{P}_n$:

Assumption C. 1.

(i) There exists $1 \leq \rho \leq 2$ *such that*

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

(ii) For all $r = 1, \ldots, d_n$, *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

$$\min_{j=1,\ldots,p} \mathbb{E}[\epsilon_r^2 X_{r,j}^2] \geq c > 0.$$

(iii) The coefficients obey

$$\max_{r=1,\ldots,d} \|\beta_r\|_0 \leq s.$$

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

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

(v) The approximation error obeys

$$\max_{r=1,\ldots,d} \|a_r(X_r)\|_{P,2} \leq C \sqrt{\frac{s \log(a_n)}{n}}$$

and

$$\max_{r=1,\ldots,d} (\mathbb{E}_n[(a_r(X_r))^2] - E[(a_r(X_r))^2]) \leq C \frac{s \log(a_n)}{n}$$

with probability $1 - o(1)$.

Theorem 11. *Under Assumption C.1, the Lasso estimator $\hat{\beta}_r$ obeys uniformly over all $P \in \mathcal{P}_n$ with probability $1 - o(1)$*

$$\max_{r=1,\dots,d} \|\hat{\beta}_r - \beta_r\|_2 \leq C\sqrt{\frac{s\log(a_n)}{n}}, \tag{4.13}$$

$$\max_{r=1,\dots,d} \|\hat{\beta}_r - \beta_r\|_1 \leq C\sqrt{\frac{s^2\log(a_n)}{n}} \tag{4.14}$$

with

$$\max_{r=1,\dots,d} \|\hat{\beta}_r\|_0 \leq Cs. \tag{4.15}$$

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

$$\max_{r=1,\dots,d} \|\tilde{\beta}_r - \beta_r\|_2 \leq C\sqrt{\frac{s\log(a_n)}{n}}, \tag{4.16}$$

$$\max_{r=1,\dots,d} \|\tilde{\beta}_r - \beta_r\|_1 \leq C\sqrt{\frac{s^2\log(a_n)}{n}}. \tag{4.17}$$

Proof of Theorem 11.

In the following, 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 that there exists a sequence $(b_n)_{n\geq 1}$ with $|a_n| \leq b_n$, which is independent of $P \in \mathcal{P}_n$ and $b_n \to 0$. Finally, the notation $a_n \lesssim_P b_n$ means that, for any $\epsilon > 0$, there exists a C such that, uniformly over all n, we have $P_P(a_n > Cb_n) \leq \epsilon$.

Due to Assumption C.1(i), 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 Vaart and Wellner [94]. Now, we essentially modify the proof of Theorem 4.2 from Belloni et al. [12] to fit our setting and keep the notation as similar as possible. Let $\mathcal{U} = \{1, \dots, d\}$ and

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

for all $r = 1, \dots, d$. The approximation error $a_r(X_r)$ is estimated 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

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

First, we verify the Condition WL from Belloni et al. [12]. 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), we note 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 $t \in (0, 1/2)$, it holds

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

with

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

uniformly over all $j = 1, \dots, p$ and $r = 1, \dots, d$ by Assumption C.1(i) and C.1(iv). Further, it holds

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

for all $j = 1, \dots, p$ and $r = 1, \dots, d$ by Assumption C.1(i) and

$$\mathbb{E}\left[S_{r,j}^2\right] = \mathbb{E}\left[\varepsilon_r^2 X_{r,j}^2\right] \geq c$$

by Assumption C.1(ii) which implies Condition $WL(ii)$. Note that Condition $WL(iii)$ simplifies to

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

with probability $1 - \Delta_n$. We use the Maximal Inequality, see for example Lemma $P.2$ from Belloni et al. [12]. Let $\mathcal{W} = (\mathcal{Y}, \mathcal{X})$ with $Y = (Y_1, \dots, Y_d) \in \mathcal{Y}$ and $X = (X_1, \dots, X_d) \in \mathcal{X}$, respectively. Define

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

with

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

Note 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

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

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, it holds $|\mathcal{F}| = dp$ 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 $P.2$ from Belloni et al. [12], 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 in Assumption C.1(iv). We proceed by verifying Assumption $M.1$ in Belloni et al. [12]. The function $\beta \mapsto M_r(Y_r, X_r, \beta)$ is convex which is the first requirement of Assumption $M.1$. We now proceed with a simplified version of the proof of $K.1$ from Belloni et al. [12]. To show Assumption $M.1$ (a), note that for all $\delta \in \mathbb{R}^p$

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

for all $r = 1, \dots, d$ due to Assumption C.1(v). Further, we have

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

where

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

and

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

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

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

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

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

and additionally

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

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

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

where the last inequality holds due to Assumption C.1(ii) and an application of the Maximal Inequality. Also uniformly over $r = 1, \dots, d$, $j = 1, \dots, p$ and for an arbitrary $q > 0$, it holds

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

with

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

By Maximal Inequality, we obtain with probability $1 - o(1)$ for a sufficiently large $q' > 0$

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

since

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

Uniformly over r, we conclude

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

Therefore, Assumption $M.1(b)$ 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 obtain

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

due to Lemma $M.4$ in Belloni et al. [12]. 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 C.1(iv). We apply Lemma $Q.1$ in Belloni et al. [12] 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 the growth condition in Assumption C.1(iv), it holds

$$\delta_n = o(1),$$

which implies

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

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

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

Let the restricted eigenvalues be defined as

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

where $\Delta_{2\tilde{c},r} := \{\delta : \|\delta_{T_r}^c\|_1 \leq 2\tilde{c}\|\delta_{T_r}\|_1\}$. By the argument given in Bickel et al. [13], it holds

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

$$\gtrsim \left(\min_{\|\delta\|_0 \leq l_n s} \frac{\|\delta X_r\|_{\mathbb{P}_n,2}^2}{\|\delta\|_2^2} \right)^{1/2} - 2n^{\frac{1}{q}-\frac{1}{\bar{q}}} \left(\max_{\|\delta\|_0 \leq l_n s} \frac{\|\delta X_r\|_{\mathbb{P}_n,2}^2}{\|\delta\|_2^2} \right)^{1/2}$$
$$\gtrsim 1$$

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 using the uniformly bounded penalty loading from above and away from zero, we obtain

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

by Lemma $M.1$ from Belloni et al. [12]. To show Assumption $M.1(b)$ 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

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

with probability $1 - o(1)$. 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 $M.1(b)$ holds for $\hat{\Psi}_{r,m}$ for some $\Delta_n = o(1)$, $l \gtrsim 1$ and $L \lesssim 1$. Consequently, we obtain

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

and

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

with probability $1 - o(1)$ due to Lemma $M.1$ in Belloni et al. [12]. Uniformly over all $r = 1, \ldots, d$, it holds

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

$$\leq \|(\hat{\beta}_r - \beta_r)X_r\|_{\mathbb{P}_n,2}\|\delta X_r\|_{\mathbb{P}_n,2} \leq L_n\|\delta X_r\|_{\mathbb{P}_n,2}$$

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

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

are uniformly bounded from above, Lemma $M.2$ from Belloni et al. [12] implies

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

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

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

with probability $1-o(1)$. We now proceed by using Lemma $M.3$ in Belloni et al. [12]. We obtain uniformly over all $r = 1, \ldots, d$

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

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

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

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

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

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

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

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

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

This completes the proof. □

4.11 Computational Details

4.11.1 Computation and Infrastructure

The simulation study has been run on a x86_64_redhat_linux-gnu (64-bit) (CentOS Linux 7 (Core)) cluster using R version `3.5.3 (2019-03-11)`. All Lasso estimations are performed using the R package `hdm`, version `0.3.1` by [32] which can be downloaded from CRAN. The construction of B-splines is based on the R package `splines`. The R code is available upon request.

4.11.2 Simulation Study: Smoothing Parameters in B-splines

Table 4.4 presents the corresponding smoothing parameters $\{k_j, k_{-j}\}$ of the cubic B-splines that are used in the simulation study. k_j denotes the degrees of freedom chosen to approximate the function $f_j(x_j)$ and k_{-j} is chosen for all other functions.

n	p	f_1	f_2	f_3	f_4	f_5
100	50	$\{7,4\}$	$\{6,4\}$	$\{7,4\}$	$\{5,4\}$	$\{7,4\}$
100	150	$\{7,4\}$	$\{6,4\}$	$\{6,4\}$	$\{5,4\}$	$\{5,4\}$
1000	50	$\{7,4\}$	$\{6,5\}$	$\{5,4\}$	$\{5,4\}$	$\{5,4\}$
1000	150	$\{7,4\}$	$\{6,5\}$	$\{7,4\}$	$\{5,5\}$	$\{4,4\}$

Table 4.4: **Smoothing parameters used in the simulation study in Table 4.2.**

4.11.3 Empirical Application: Cross-Validation Procedure

The choice of the degrees of freedom parameter k for the construction of B-splines in the empirical application is based on a heuristic cross-validation which exploits the additive structure of the model. Let $k = \{k_j, k_{-j}\}$ be the degrees of freedom with k_j specifying the smoothing parameters for $f_j(x_j)$ and k_{-j} denoting the parameter for all other functions $f_{-j}(x_{-j})$. To explicitly address the dependence of the fitted function on the chosen degrees of freedom parameter, we use a notation $\hat{f}_j(x_j, k_j)$ which leads to the model

$$y_i = f_j(x_{i,j}, k_j) + f_{-j}(x_{i,-j}, k_{-j}) + \epsilon_i.$$

Then, the heuristic rule for choosing k proceeds as follows:

For $j = 1, ..., p$,

1. set up a grid of values for k_{-j},
2. perform a 5-fold cross-validated search for an optimal k_j over a grid of values $\underline{k}_j, ..., \overline{k}_j$, i.e., fit the regression
$$y_i = f_j(x_{i,j}, k_j) + f_{-j}(x_{i,-j}, k_{-j}) + \epsilon_i$$
and compute $MSE_{CV}(k_j, k_{-j})$, where $MSE_{CV}(k_j, k_{-j})$ is the cross-validated mean squared error in prediction provided values k_j and k_{-j}.
3. find the optimal value of k_j^* which minimizes MSE_{CV} over all values of k_{-j}.

We experimented with different settings and repeated the procedure multiple times. The resulting parameters are listed in Table 4.5.

NOX	11
$CRIM$	6
ZN	3
$INDUS$	6
RM	6
AGE	5
$DIST$	9
TAX	5
$PTRATIO$	11
$BLACK$	5
$LSTAT$	7

Table 4.5: **Smoothing parameters used in the empirical application.**

4.11.4 Empirical Application: Additional Plots for Explanatory Variables

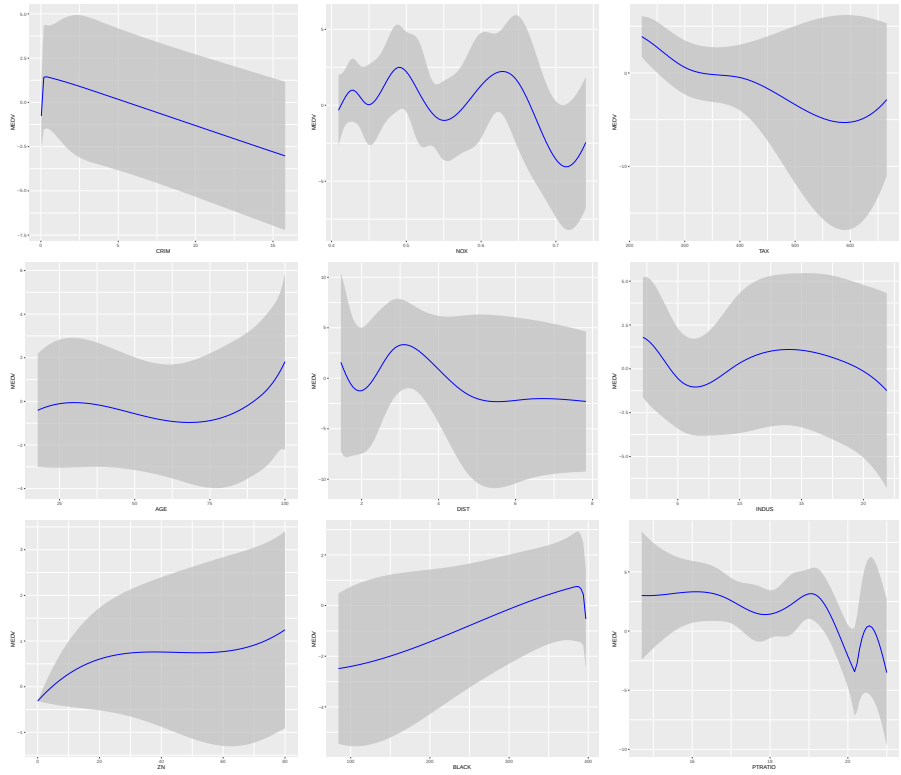


Figure 4.3: **Additional plots of the effect of the explanatory variables on the dependent variable $MEDV$ with simultaneous 95%-confidence bands in the Boston housing data application.**

Chapter 5

Uniform Inference in High-Dimensional Gaussian Graphical Models

5.1 Introduction

We provide methodology and theory for uniform inference on high-dimensional graphical models with the number of target parameters being potentially 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 is normally distributed, every edge corresponds to a nonzero entry in the inverse covariance matrix (Lauritzen [66]).

In the last decade, significant progress has been made on the estimation of a large precision matrix in order to analyze the dependence structure of a high-dimensional normally 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, e.g., in Lam and Fan [65], Rothman et al. [86], Ravikumar et al. [84] and Yuan and Lin [104]. The second approach, first introduced by Meinshausen and Bühlmann [76], 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 [103], Cai et al. [24] and Sun and Zhang [92].

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 [58] and Ren et al. [85]. Both approaches lead to an estimate that is element-wise asymptotically normally distributed 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 and under a random design. In particular, our results rely on approximate sparsity instead of row sparsity which restricts the number of nonzero 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 framework 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. [9], Belloni et al. [12] and Chernozhukov et al. [35]. 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. [12] to ensure sufficient fast convergence rate of the Lasso estimator under approximate sparsity conditions. Moreover, we provide auxiliary results for the square-root Lasso estimator establishing new uniform estimation rates and sparsity guarantees in a random design under approximate sparsity conditions. Square-root Lasso is very popular in graphical models but these results might be of independent interest for related problems in high-dimensional linear models.
Chang et al. [27] consider testing for high-dimensional parameters of the precision matrix similar to our setting, in particular conducting inference for a growing number of parameters in a high-dimensional setting. Our setting and analysis differs from them in several ways. First, Chang propose a biased correction estimate for the parameter of interest, while we use the Z-estimation framework which does not require the bias correction step. Second, Chang relies on results from Bühlmann and Van De Geer [22] to estimate the nuisance parameter. We explicitly derive uniform convergence results for Lasso and square-root Lasso in a random design setting. Thus, we provide a feasible nuisance parameter estimate and show how it can be implemented. The choice of the penalization parameter is both theory-grounded and also feasible in empirical studies. Third, our assumptions are tailored for Gaussian graphical models and hence are more structured. Finally, we allow for approximate sparsity instead of strict sparsity which is more realistic for many applications. In a simulation study, we show that our proposed method gives reliable results even in this challenging setting.

Plan of this Paper

The rest of this paper is organized as follows. First, we introduce the technical notation that is used in the paper. In Section 5.2, we formally define the setting and embed the problem of support discovery in Gaussian graphical models into a general Z-estimation problem with a high-dimensional nuisance function. In Section 5.3, 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 5.4. Section 5.5 provides implementation details and shows how our estimation procedure can be modified by cross-fitting to improve small sample properties. Section 5.6 provides a simulation study on the proposed method. We conclude in Section 5.7. The supplementary material includes additional technical material. The proof of our main theorem is provided in Appendix 5.8. The uniform nuisance function estimation is discussed

in Appendix 5.9. Appendix 5.9.1 formally discusses conditions for the uniform convergence rates of the Lasso estimator. Finally, Appendix 5.9.2 provides auxiliary results for the square-root Lasso estimator.

Notation

Throughout the paper, we consider a random element X from some common probability space $(\Omega, \mathcal{A}, P)$. We denote by $P \in \mathcal{P}_n$ a probability measure out of a large class of probability measures, which may vary with the sample size (since the model is allowed to change with n), and by $\mathbb{P}_n$ the empirical probability measure. $\|\cdot\|_{P,q}$ denotes the $L^q(P)$-norm. Additionally, let $\mathbb{E}$ and $\mathbb{E}_n$ be the expectation with respect to P and $\mathbb{P}_n$, respectively. $\mathbb{G}_n(\cdot)$ denotes the empirical process

$$\mathbb{G}_n(f) := \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X_i)]\right)$$

for a class of suitably measurable functions $\mathcal{F} : \mathcal{X} \to \mathbb{R}$.
Further, $\|v\|_1 = \sum_{l=1}^{p} |v_l|$ denotes the ℓ_1-norm, $\|v\|_2 = \sqrt{v^T v}$ the ℓ_2-norm and $\|v\|_0$ equals the number of nonzero components of a vector $v \in \mathbb{R}^p$. We define $v_{-l} := (v_1, \ldots, v_{l-1}, v_{l+1} \ldots, v_p)^T \in \mathbb{R}^{p-1}$ for any $1 \leq l \leq p$. $\|v\|_\infty = \sup_{l=1,\ldots,p} |v_l|$ denotes the sup-norm. Let c and C be positive constants independent of n which value may change at each appearance. The notation $a_n \lesssim b_n$ means $a_n \leq C b_n$ for all n and some C. Additionally, $a_n = o(1)$ denotes that there exists a sequence $(b_n)_{\geq 1}$ of positive numbers such that $|a_n| \leq b_n$ for all n, where b_n is independent of $P \in \mathcal{P}_n$ for all n and b_n converges to zero. Finally, the notation $a_n \lesssim_P b_n$ means that, for any $\epsilon > 0$, there exists a C such that, uniformly over all n, we have $P_P(a_n > C b_n) \leq \epsilon$.

5.2 Setting

Let

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

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

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

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 $\Phi_0 = \Sigma_X^{-1}$, see e.g., Janková and Van De Geer [58]. Hence,

$$\beta_k^{(j)} = 0 \Leftrightarrow \beta_j^{(k)} = 0 \Leftrightarrow X_j \perp X_k | X_{-\{j,k\}} \tag{5.1}$$

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 the sample size n. In the following, the dependence on n is omitted to simplify the notation. In order to test whether the variables X_{j_r} and X_{k_r} are conditionally independent with $m_r = (j_r, k_r)$ for all $r \in \{1, \ldots, d\}$, 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 5.2.1. *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

$$\mathbb{E}\left[\psi_{m_r}\left(X, \theta_{m_r}, \eta_{m_r}\right)\right] = 0 \tag{5.2}$$

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

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

for all $\tilde{\eta}$ in an appropriate set, where ∂_t denotes the derivative 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 12.

5.3 Estimation

Let $X^{(i)}$, $i = 1, \dots, n$, 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

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

is defined as the solution of the empirical version of the moment condition

$$\sup_{r=1,\dots,d} \left\{ \left| \mathbb{E}_n \left[\psi_{m_r}\left(X, \hat{\theta}_{m_r}, \hat{\eta}_{m_r}\right) \right] \right| - \inf_{\theta \in \Theta_{m_r}} \left| \mathbb{E}_n \left[\psi_{m_r}\left(X, \theta, \hat{\eta}_{m_r}\right) \right] \right| \right\} \leq \epsilon_n, \tag{5.3}$$

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 slowly converging to zero but at least at a polynomial rate in n (cf. proof of Theorem 12).

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

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

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

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 5.3.1 below. The number of relevant variables $s_n \equiv s$ captured by the regression coefficient $\beta^{(1,m_r)}$ and $\gamma^{(1,m_r)}$, respectively, can grow with the sample size. The coefficients $\beta^{(2,m_r)}$ and $\gamma^{(2,m_r)}$, respectively, are the approximate sparse parts of the true regression coefficients. The misspecification of the 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 number of tested edges becomes more restrictive. This growth condition ensures that $\log(d) = o(n^{1/9})$ and is in line with Chernozhukov et al. [30]. It guarantees the validity of multiplier bootstrap in our setting and allows us to construct uniformly valid confidence regions. 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 13 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 5.3.1. *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 condition in A1 for testing on an edge (j,k) is satisfied if each node j and k is only sparsely connected to all other nodes.

5.4 Main Results

We are able to construct uniformly valid confidence intervals for a growing number of hypothesis $d = d_n$ by applying new results regarding confidence regions for many parameters from Belloni et al. [12]. To approximate the limit process of $\sup_{r=1,\dots,d} \hat{\theta}_{m_r}$, we employ the Gaussian multiplier bootstrap approach. In this context, we define

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

and the corresponding plug-in estimators

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

for $r = 1, \dots, d$. Further, let

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

and we define 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 normally distributed random variables which are independent from $\left(X^{(i)}\right)_{i=1}^n$. We define c_α as the conditional $(1-\alpha)$-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 12.
Under Assumptions A1-A4 with probability $1 - o(1)$ *uniformly in* $P \in \mathcal{P}_n$*, the estimator* $\hat{\theta}$ *in (5.3) obeys*

$$P\left(\hat{\theta}_{m_r} - \frac{c_\alpha \hat{\sigma}_{m_r}}{\sqrt{n}} \leq \theta_{m_r} \leq \hat{\theta}_{m_r} + \frac{c_\alpha \hat{\sigma}_{m_r}}{\sqrt{n}}, r = 1, \dots, d\right) \to 1 - \alpha. \tag{5.4}$$

Using Theorem 12 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 5.4.1. *Theorem 12 provides critical regions of the form*

$$\sup_{r=1,\dots,d} \left|\sqrt{n} \frac{\hat{\theta}_{m_r}}{\hat{\sigma}_{m_r}}\right| > c_{1-\alpha}. \tag{5.5}$$

Alternatively, we can reject the null hypothesis if

$$\sup_{r=1,\dots,d} \left|\sqrt{n} \frac{\hat{\theta}_{m_r}}{\hat{\sigma}_{m_r}}\right| < c_{\frac{\alpha}{2}} \quad or \quad \sup_{r=1,\dots,d} \left|\sqrt{n} \frac{\hat{\theta}_{m_r}}{\hat{\sigma}_{m_r}}\right| > c_{1-\frac{\alpha}{2}}. \tag{5.6}$$

The confidence region (5.6) 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 in Vershynin [99] and therefore might have smaller volume. In future research, we plan to address the challenging problem analyzing if there is an optimal test that delivers the smallest confidence region. In this paper, we compare the empirical performance of the two confidence regions. It is worth to notice that both of the regions (5.5) and (5.6) are based on Gaussian approximation and multiplier bootstrap for maxima of sums of high-dimensional random vectors in Chernozhukov et al. [30]. The central limit theorem and bootstrap in high-dimension introduced in Chernozhukov et al. [34] extend this result 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 $1-\alpha$. In this context, let us 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 rejects the null hypothesis if

$$\sup_{r=1,\ldots,d} \left| \hat{\theta}^*_{m_r}(S, exp) \right| > c^*_{1-\alpha} \tag{5.7}$$

has level α since the constructed confidence regions correspond to S-sparsely convex sets, see Chernozhukov et al. [34]. 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

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

5.5 Notes on the Implementation

We have implemented a function that estimates the target coefficients

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

corresponding to the considered set of potential edges

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

by the proposed method described in Section 5.3. It can be used to perform hypothesis tests with asymptotic level α based on the different confidence regions described in Comment 5.4.1. The nuisance function can be estimated by Lasso, post-Lasso or square-root Lasso.

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. [35]. 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 3 cross-fitting

1) Take a K-fold random partition $(I_k)_{k=1}^K$ of observation indices $[n] = \{1, \dots, n\}$ such that the size of each fold I_k is N. Also, for each $k \in [K] = \{1, \dots, K\}$, define $I_k^c := \{1, \dots, N\} \setminus I_k$.

2) For each $k \in [K]$ and $r = 1, \dots, d$, construct an estimator $\hat{\eta}_{k,m_r} = \hat{\eta}_{m_r}\big((X_i)_{i \in I_k^c}\big)$ by Lasso/ post-Lasso or square-root Lasso.

3) For each $k \in [K]$, construct an estimator $\hat{\theta}_k = (\hat{\theta}_{k,m_1}, \dots, \hat{\theta}_{k,m_d})$ as in (5.3):

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

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

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

with

$$\hat{J}_{m_r}^K = -\frac{1}{K} \sum_{k=1}^K (X_k (X_k - \hat{\eta}_{k,m_r}^{(2)} X_{-m_r})),$$

$$\hat{\sigma}_{m_r}^K = \sqrt{(\hat{J}_{m_r}^K)^{-2} \frac{1}{K} \sum_{k=1}^K \Big(\psi_{m_r}^2 (X, \hat{\theta}_{m_r}^K, \hat{\eta}_{k,m_r}) \Big)}.$$

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

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

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

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

The confidence region above corresponds to (5.5). Confidence regions corresponding to (5.6) or (5.7) can be constructed in an analogous way.

5.6 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* [107]. 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

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

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 $1 - \alpha = 0.95$.

5.6.1 Simulation Settings

In our simulation study, we estimate the correlation structure of four different designs that are described in the following. In Example 3, we explicitly allow for approximate sparsity instead of strict sparsity which is more realistic for many applications.

Example 1: Random Graph

Each pair of off-diagonal elements of the covariance matrix of the first $p-1$ regressors is randomly set to nonzero 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

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

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

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

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 nonzero 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 test the tuples

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

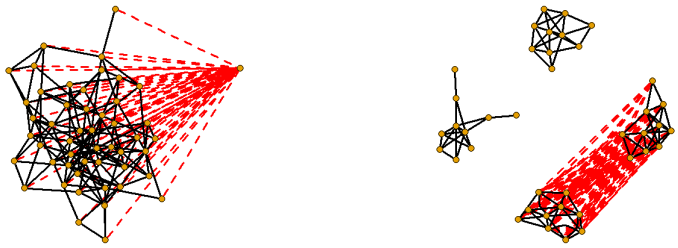


Figure 5.1: **Examples of a Random Graph (left) and a Cluster Graph (right).**
The edges of the graph are colored in black and the edges contained in the hypothesis in red.

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 matrix 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. Again, we test the hypothesis whether the last regressor is independent from all other regressors that corresponds to

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

Example 4: Independent Graph

By setting

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

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

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

5.6.2 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 (5.5), whereas Confidence Intervall II is based on the approximation of the sphere in (5.6). 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, which has got a mean absolute deviation of 1.875%, 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 5.2 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 5.1: **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
random	20	19	0.969	0.925	0.956	0.951	0.932	0.947
	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
cluster	20	25	0.972	0.958	0.973	0.914	0.936	0.914
	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
approx	20	19	0.958	0.958	0.958	0.965	0.965	0.965
	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
indepent	5	10	0.951	0.951	0.951	0.951	0.951	0.951
	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 5.2: **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 5.3: **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 5.4: **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 5.5: **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
	20	19	0.914	0.897	0.918	0.922	0.921	0.923
random	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
	20	25	0.885	0.882	0.888	0.900	0.896	0.901
cluster	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
	20	19	0.931	0.930	0.931	0.938	0.937	0.938
approx	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
	5	10	0.923	0.922	0.923	0.943	0.942	0.943
indepent	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 5.6: **Simulation results for S=5, exp=2 and 3-fold.**

5.7 Conclusion

In this paper, we provide results for uniform inference on high-dimensional graphical models with the number of target parameters being possibly much larger than the sample size. This is in particular important when certain structures of a causal model should be recovered. As the square-root Lasso estimator is very popular for the estimation of graphical models, we provide uniform estimation rates and sparsity guarantees of the square-root Lasso estimator under a random design and approximate sparsity. These results might be of independent interest for related problems. We show that our proposed method has very good small sample properties in simulation studies. Although the estimation of graphical models has been considered as very challenging, as the number of parameters to estimate is often large compared to the sample size, our results from the simulation studies are very encouraging.

Appendix

5.8 Proof of Theorem 12

Proof. Let $m_r = (j,k)$ be an arbitrary tuple in $\mathcal{M}$. First, we remark that

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

due to the Assumptions A3 and A4. Let us 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 we 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 $C > 0$.
We verify the Assumptions 2.1-2.4 in Belloni et al. [12] to apply Corollary 2.2 of this paper. 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 $\mathcal{F} = \{\varepsilon^{(m_r)}\nu^{(m_r)} | r = 1, \ldots, d\}$ and $\mathbb{G}_n(f) := \sqrt{n}|\mathbb{E}_n[f] - \mathbb{E}[f]|$. By the same arguments as in the beginning of the proof of Theorem 13, we conclude that the envelope $\sup_{f\in\mathcal{F}} |f|$ of $\mathcal{F}$ fulfills

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

since the error terms are normally distributed. Using Lemma P.2 (Maximal Inequality I) in Belloni et al. [12] with $|\mathcal{F}| = d$, we have

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

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

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

Therefore, Assumptions 2.1 (ii) and 2.1 (iii) hold. We notice 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

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

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

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

due to Assumption A3, Assumption 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 above using

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

due to the normal design. Combining these results gives us Assumption 2.1 (v) (a).
We conclude 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 $B_{2n} = C$, notice that

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

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 13 from Section 5.9 hold which implies Assumption 2.2 (i). Notice that Assumption B1 and Assumption B4 are satisfied with $\rho = 2$. Condition A1 implies

Assumption 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 $r = 1, \dots, d$ and $l \in \{1, \dots, p\} \setminus \{j\}$, 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 Assumption A4, this implies Assumption B2. Therefore, we are able to estimate the nuisance parameters at a sufficiently fast rate.
Define

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

For now, we exclude the true nuisance parameter to bound the covering entropy of $\mathcal{F}_1$ and define

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

with

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

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

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

and with an analogous argument

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

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

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

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, Vaart and Wellner [94]). 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 Vaart and Wellner [94], we obtain

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

where K is an universal constant and with an analogous argument

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

Using basic calculations on covering entropies (Lemma N.1 in Appendix N, Belloni et al. [9]), we can bound the covering entropy of the class $\mathcal{F}_{1,1}$ by

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

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

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

Additionally, define

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

By 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

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

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

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

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). 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 normally distributed random variables whose variances are 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 in 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, it holds

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

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

and

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

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

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

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 $\sigma_{m_r}^2 := J_{m_r}^{-2}\mathbb{E}[\psi_{m_r}^2(X, \theta_{m_r}, \eta_{m_r})]$. By the Cauchy-Schwarz inequality, for any $q > 0$, the envelope F_0 for $\mathcal{F}_0$ satisfies

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

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

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

for all $< \varepsilon \le 1$. Therefore, Assumption 2.3 (i) is satisfied with $\varrho_n = 1$ and $A_n = d \vee n$. Since the errors are centered normally 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 $\mathbb{E}[f^4] \le C$ for all $f \in \mathcal{F}_0$ which gives us Assumption 2.3 (ii). The growth conditions from Corollary 2.1 are satisfied due to Assumption A2. Notice that

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

and

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

Thus, 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

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

with

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

Further, let us define

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

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

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

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

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

By the Maximal Inequality, 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

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

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

due to Assumption 2.1 (a) and Assumption 2.2 (v). Note 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 A from Belloni et al. [12]. Since

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

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

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

due to $\left\| \left(F^{(1)}_{1,1}\right)^4 \right\|_{P,q} \lesssim \log^2(a_n)$ and $\left\| \left(F^{(2)}_{1,1}\right)^4 \right\|_{P,q} \lesssim \log^2(a_n)$. 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 $1 \lesssim \sigma^2_{m_r} \lesssim 1$ due to Assumption 2.1 (iv) and Assumption 2.2 (v), we have

$$
\begin{aligned}
\left| \frac{\hat{\sigma}_{m_r}}{\sigma_{m_r}} - 1 \right| &\leq \left| \frac{\hat{\sigma}^2_{m_r}}{\sigma^2_{m_r}} - 1 \right| \\
&\lesssim \left| \hat{\sigma}^2_{m_r} - \sigma^2_{m_r} \right| \\
&\leq \left| \hat{J}^{-2}_{m_r} - J^{-2}_{m_r} \right| \mathbb{E}_n[\psi^2_{m_r}(X, \hat{\theta}_{m_r}, \hat{\eta}_{m_r})]
\end{aligned}
$$

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

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

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

hold 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, we note that uniformly over all $r = 1, \dots, d$

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

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

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

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

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

It holds

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

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 $r = 1, \dots, d$. To bound the second term, define the class

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

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 number obeys

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

and the envelope

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

satisfies

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

with

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

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

and

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

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

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

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

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

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 verify that the growth conditions of Corollary 2.2 in Belloni et al. [12] hold. This completes the proof. □

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

5.9.1 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 fixed $m \geq 0$. Define the post-regularized weighted least squares estimator (post-Lasso) as

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

The penalty loadings $\hat{\Psi}_{r,m} = \text{diag}(\{\hat{l}_{r,j,m}, j = 1, \ldots, p\})$ 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 4 penalty loadings

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 $n \geq n_0$ and $P \in \mathcal{P}_n$:

Assumptions **B1**-**B4**.

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

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

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

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

and

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

B3 (Uniform approximate sparsity)
The coefficients obey

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

and

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

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 13. *Under the Assumptions B1-B4, the Lasso estimator $\hat{\beta}_r$ obeys uniformly over all $P \in \mathcal{P}_n$ with probability $1 - o(1)$*

$$\max_{r=1,\dots,d} \|\hat{\beta}_r - \beta_r^{(1)}\|_2 \leq C\sqrt{\frac{s \log(a_n)}{n}}, \tag{5.8}$$

$$\max_{r=1,\dots,d} \|\hat{\beta}_r - \beta_r^{(1)}\|_1 \leq C\sqrt{\frac{s^2 \log(a_n)}{n}} \tag{5.9}$$

with

$$\max_{r=1,\dots,d} \|\hat{\beta}_r\|_0 \leq Cs. \tag{5.10}$$

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

$$\max_{r=1,\dots,d} \|\tilde{\beta}_r - \beta_r^{(1)}\|_2 \leq C\sqrt{\frac{s \log(a_n)}{n}}, \tag{5.11}$$

$$\max_{r=1,\dots,d} \|\tilde{\beta}_r - \beta_r^{(1)}\|_1 \leq C\sqrt{\frac{s^2 \log(a_n)}{n}}. \tag{5.12}$$

5.9.2 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 defined 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

$$\lambda = c' \sqrt{n} \Phi^{-1}\big(1 - \gamma/(2pd)\big), \tag{5.13}$$

where $1 - \gamma$ is a confidence level associated with the probability of the event (5.14), and $c' > c$ is a slack constant. The first part of the analysis is to control the event

$$\frac{\lambda}{n} \geq c \max_{r=1,\dots,d} \|S_r\|_\infty, \tag{5.14}$$

where

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

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

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

The following conditions and Lemma 4 are similar to the condition WL and Lemma M.4 in Belloni et al. [12]. 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 (5.14) with high probability.

Lemma 4. *Suppose that Condition* ***WL*** *holds. In addition, suppose that* λ *satisfies (5.13) 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 13 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 5.9.2, we additionally assumed independence between the regressors and the error terms. This eliminates the need to estimate the penalty loadings.

Theorem 14. *Suppose that the Assumptions B1-B4 hold. In addition, suppose that* λ *satisfies (5.13) for some* $c' > c$ *and* $\gamma = \gamma_n \in [1/n, 1/\log(n)]$*. Then, with probability at least* $1 - o(1)$*, we have*

$$\max_{r=1,\dots,d} \|\hat{\beta}_r - \beta_r^{(1)}\|_2 \leq C\sqrt{\frac{s\log(a_n)}{n}}, \tag{5.15}$$

$$\max_{r=1,\dots,d} \|\hat{\beta}_r - \beta_r^{(1)}\|_1 \leq C\sqrt{\frac{s^2\log(a_n)}{n}} \tag{5.16}$$

with

$$\max_{r=1,\dots,d} \|\hat{\beta}_r\|_0 \leq Cs. \tag{5.17}$$

5.9.3 Proofs

Proof of Theorem 13.
As in the previous proof, we use C for a strictly positive constant, independent of n, which value may differ 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 a sequence $(b_n)_{n\geq 1}$ with $|a_n| \leq b_n$, where 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 a C such that, uniformly over all n, we have $P_P(a_n > Cb_n) \leq \epsilon$.

Due to Assumption 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 Vaart and Wellner [94]. Now, we modify the proof of Theorem 4.2 from Belloni et al. [12] to fit our setting, but we keep the notation as similar as possible. Let us define $\mathcal{U} = \{1, \dots, d\}$ 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 B3, we estimate the nuisance parameter a_r with $\hat{a}_r \equiv 0$. Define

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

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

and

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

First, we verify Condition WL from Belloni et al. [12]. 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), notice that

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

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

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

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). Condition WL (iii) simplifies 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 $1 - \Delta_n$. Let $\mathcal{W} = (\mathcal{Y}, \mathcal{X})$ with $Y = (Y_1, \ldots, Y_d) \in \mathcal{Y}$ and $X = (X_1, \ldots, X_d) \in \mathcal{X}$, respectively. Define

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

with

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

We notice that

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

Since we have

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

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 $P.2$ from Belloni et al. [12], we obtain with probability not less than $1-o(1)$

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

by the growth condition in B4. We proceed by verifying Assumption $M.1$ in Belloni et al. [12]. The function $\beta\mapsto M_r(Y_r,X_r,\beta)$ is convex which is the first requirement of Assumption $M.1$. Define

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

with envelope

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

Note that

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

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

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

we can use Lemma $P.2$ from Belloni et al. [12] to obtain with probability not less than $1-o(1)$

$$\begin{aligned}
&\max_{r=1,\ldots,d}|(\mathbb{E}_n-\mathbb{E})[(\beta_r^{(2)}X_r)^2]| \\
&= n^{-1/2}\sup_{g\in\mathcal{G}}|\mathbb{G}_n(g)|
\end{aligned}$$

$$\begin{aligned}
&\lesssim C\left(\sqrt{\frac{\log(a_n)\max_{r=1,\dots,d}\|\beta_r^{(2)}\|_1^4}{n}}+n^{-1+1/q}\max_{r=1,\dots,d}\|\beta_r^{(2)}\|_1^2\log^{1+\frac{2}{\rho}}(a_n)\right)\\
&\lesssim C\left(\sqrt{\frac{\log(a_n)}{n}}\sqrt{\frac{s^2\log(a_n)}{n}}+\frac{s\log(a_n)}{n}\sqrt{n^{2/q}\frac{\log^{1+\frac{4}{\rho}}(a_n)}{n}}\right)\\
&\lesssim \frac{s\log(a_n)}{n}
\end{aligned}$$

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

$$\begin{aligned}
\max_{r=1,\dots,d}\mathbb{E}_n[(\beta_r^{(2)}X_r)^2] &\leq \max_{r=1,\dots,d}|(\mathbb{E}_n-\mathbb{E})[(\beta_r^{(2)}X_r)^2]|+\max_{r=1,\dots,d}\mathbb{E}[(\beta_r^{(2)}X_r)^2]\\
&\lesssim_P \frac{s\log(a_n)}{n}. \qquad (5.18)
\end{aligned}$$

To show Assumption $M.1$ (a), note that

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

for all $\delta\in\mathbb{R}^p$ and for all $r=1,\dots,d$. Further, we have

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

where

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

and

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

with $\sqrt{w_r}=1/4$. This gives us Assumption $M.1$ (c) with $\Delta_n=0$ and $\bar{q}_{A_r}=\infty$. Since Condition WL (ii) and WL (iii) hold, we conclude 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

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

For now, we suppose that $m = 0$ in Algorithm 4. Uniformly over $r = 1, \dots, d$ and $j = 1, \dots, p$, we have

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

where the last inequality holds due to Assumption B2 and an application of the Maximal Inequality. Also uniformly over $r = 1, \dots, d$ and $j = 1, \dots, p$, it holds

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

for an arbitrary $q > 0$, where

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

By Maximal Inequality, it holds

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

with probability $1 - o(1)$ for a sufficiently large $q' > 0$ 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 $M.1(b)$ 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

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

due to Lemma $M.4$ from Belloni et al. [12]. 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 $\tilde{q}$ defined in Assumption B4.

We apply Lemma $Q.1$ in Belloni et al. [12] 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 Assumption 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 $r = 1, \ldots, d$. Define $T_r := \text{supp}(\beta_r^{(1)})$ and

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

Let the restricted eigenvalues be defined as

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

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. [13], 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

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

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

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

by Lemma $M.1$ from Belloni et al. [12].

To establish Assumption $M.1(b)$ 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

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

with probability $1-o(1)$. We notice that

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

as shown in (5.18). Using the triangle inequality, we obtain with probability $1-o(1)$

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

This implies

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

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

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

and

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

with probability $1-o(1)$ due to Lemma $M.1$ from Belloni et al. [12]. Uniformly over all $r=1,\dots,d$, it holds

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

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

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

are uniformly bounded from above, Lemma $M.2$ from Belloni et al. [12] directly implies

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

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

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

with probability $1 - o(1)$. Now, we proceed by using Lemma $M.3$ from Belloni et al. [12]. Uniformly over all $r = 1, \ldots, d$, it holds

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

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

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

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

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

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

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

and that the minimum sparse eigenvalues are uniformly bounded away from zero. By the same argument as above, it holds

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

This completes the proof. □

Proof of Lemma 4.
We rely upon the proof of Lemma M.4 from Belloni et al. [12]. Since the regressors are standardized for all $j = 1, \ldots, p$ and independent from the error terms for all $r = 1, \ldots, d$, notice that

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

Due to Condition **WL**(iii), it holds

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

for a 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 Condition **WL**(ii). Next, for each $j = 1,\dots,p$ and $r = 1,\dots,d$, we apply Lemma P.1 from Belloni et al. [12] 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, Condition **WL**(i) and Condition **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

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

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

Proof of Theorem 14.
The proof relies upon the proof of Lemma M.1. from Belloni et al. [12]. At first, we show that Condition **WL** is fulfilled. Condition **WL**(i), Condition **WL**(ii) and the first part of Condition **WL**(iii) have been verified in the proof of Theorem 13. Hence, we need to show

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

with probability converging to one.
Let $\mathcal{W} = (\mathcal{Y}, \mathcal{X})$ with $Y = (Y_1, \dots, Y_d) \in \mathcal{Y}$ and $X = (X_1, \dots, X_d) \in \mathcal{X}$, respectively.
Define $\mathcal{F} := \{f_r | r = 1, \dots, d\}$ 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, notice that

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

where we used the same argument as in the beginning of the proof of Theorem 13. Due to Assumption B1, the second moments of the error terms are uniformly bounded. 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 P.2 from Belloni et al. [12] which implies

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

with probability $1 - o(1)$. 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

$$\hat{Q}_r^{1/2}(\hat{\beta}_r) - \hat{Q}_r^{1/2}(\beta_r^{(1)}) \leq \frac{\lambda}{n} \left(\|\delta_{r,T_r}\|_1 - \|\delta_{r,T_r^c}\|_1 \right) \tag{5.19}$$

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

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

with probability $1 - o(1)$. For a sequence $C_n \lesssim \sqrt{\frac{s \log(a_n)}{n}}$ independent from r, it holds

$$|\delta_r \hat{S}_r| \leq |\delta_r S_r| + |\delta_r (\hat{S}_r - S_r)|$$

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

To obtain the last inequality, notice 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

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

uniformly converges towards zero with probability $1 - o(1)$ 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 the proof of Theorem 13. This implies

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

by an analogous argument as above. Hence, it holds

$$\hat{Q}_r^{1/2}(\hat{\beta}_r) - \hat{Q}_r^{1/2}(\beta_r^{(1)}) \geq \delta_r \hat{S}_r \gtrsim -\|\delta_r\|_1 \frac{\lambda}{nc} - C_n \|\delta_r X_r\|_{\mathbb{P}_n,2} \tag{5.20}$$

with probability $1 - o(1)$. Combining the inequalities (5.19) and (5.20), we obtain

$$\begin{aligned} &- \|\delta_r\|_1 \frac{\lambda}{nc} - C_n \|\delta_r X_r\|_{\mathbb{P}_n,2} \lesssim_P \frac{\lambda}{n} \left(\|\delta_{r,T_r}\|_1 - \|\delta_{r,T_r^c}\|_1 \right) \\ \iff &\|\delta_{r,T_r^c}\|_1 \lesssim_P \underbrace{\frac{c+1}{c-1}}_{:=\tilde{c}} \|\delta_{r,T_r}\|_1 + \frac{n}{\lambda} \frac{c}{c-1} C_n \|\delta_r X_r\|_{\mathbb{P}_n,2}. \end{aligned} \tag{5.21}$$

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

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

due to the Hölder inequality. Due to Lemma $Q.1$ in Belloni et al. [12] with $K \lesssim n^{1/\bar{q}} \log^{\frac{1}{\rho}}(a_n)$ for a suitable $\bar{q} > \tilde{q}$, $k \lesssim s$ and

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

by Assumption 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

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

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

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

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

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 $1 - o(1)$, it holds

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

and we obtain

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

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 $||\delta_{r,T_r^c}||_1 > 2\tilde{c}||\delta_{r,T_r}||_1$), (5.21) 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

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

since $\tilde{c} \geq 1$. Additionally, (5.21) 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

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

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

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

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, we notice 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 $r = 1, \ldots, d$. Now, we show that

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

It holds

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

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

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

due to

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

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

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

for every j with $|\hat{\beta}_{r,j}| > 0$. Let $\hat{T}_r := \operatorname{supp}(\hat{\beta}_r)$ and $|\hat{T}_r| := \hat{s}_r$. We obtain

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

where we used

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

Hence, it holds

$$\begin{aligned}\hat{s}_r \leq& \left(\frac{2CnC_n}{\lambda(1-1/c)}\right)^2 \sup_{\|\delta\|_2=1, \|\delta\|_0 \leq \hat{s}_r} \|\delta X_r\|^2_{\mathbb{P}_n,2} \\ \leq& \underbrace{\left(\frac{2CnC_n}{\lambda(1-1/c)}\right)}_{:=L}^2 \phi_{max}(\hat{s}_r, r) \lesssim s\phi_{max}(\hat{s}_r, r) \end{aligned} \tag{5.22}$$

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

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

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, Belloni and Chernozhukov [6]), for any integer $k \geq 0$ and constant $l \geq 0$, we have

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

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

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

that violates the condition that $M \in \mathcal{M}_r$. Therefore, we have $\hat{s}_r \leq M$. Using (5.22), 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 $1 - o(1)$. □

Chapter 6

General Conclusion and Outlook

In this chapter, some concluding remarks on this dissertation and an outlook on future research is given. This work has highlighted how machine learning methods can be used to perform valid inference in high-dimensional settings. For a detailed discussion of the previous chapters, I refer to the conclusions of each paper at the end of each chapter. Here, a high-level overview of the main results of the dissertation is given and the main contributions to the literature are pointed out. The papers that are presented in this dissertation make the double machine learning approach applicable to a wider range of problems in high-dimensions by extending and adjusting the underlying proofs.

In Chapter 2, we derived results for the asymptotic distribution of the estimated treatment effects using L_2-Boosting. We achieved this in a high-dimensional linear model with many controls as well as in an instrumental variable model with potentially many instruments. As mentioned in the introduction, the estimation and the variable selection in both models often rely on the Lasso estimator. We managed to proof that L_2-Boosting algorithms are a competitive alternative to Lasso due to the comparable estimation rates. This makes L_2-Boosting attractive for applied work, in which one is interested in estimating a treatment effect after variable selection. In Chapter 3, we constructed an estimator for the transformation parameter in a high-dimensional transformation model and proved the asymptotic normality of the estimator. In high-dimensional transformation models, the nuisance parameter depends on the target parameter. We provide new results regarding inference in a general Z-estimation framework under a different set of entropy conditions where such a dependency is explicitly allowed. As such a dependency occurs in many statistical models, these results are of independent interest for other high-dimensional problems with the same underlying structure. Chapter 4 provided a methodology for uniform valid confidence bands of a nonparametric component in the generalized additive model in high-dimensions. In this setting, one is particularly interested in the linear functional of a high-dimensional target parameter. This is a nontrivial extension of the double machine learning approach that only provides valid inference for the target parameter itself. So far generalized additive models are frequently used in empirical work when the number of covariates is small. Through our work however, the reach of this tool is extended to also cover settings where the number of available covariates is large compared to the sample size. In Chapter 5, we presented results for uniform inference in high-dimensional Gaussian graphical models. We showed how recent methodology can be applied to conduct valid inference in situations where the number of target parameters is potentially much larger than the sample size. This allows to estimate the dependencies within a large set of variables and to recover causal structures in complex data sets. Further, as explained in Chapter 1, uniform estimation rates for the nuisance parameter are crucial to conduct valid inference. In this context, we established uniform estimation rates and sparsity guarantees of the Lasso estimator and of the square-root Lasso estimator in a random design under approximate sparsity conditions. These results are of independent interest for other high-dimensional linear regression

problems. The papers that are collected in this dissertation emphasize the power and the potential of the double machine approach. With this dissertation, I would like to contribute that the double machine learning approach finds more application in practice and, thus, help to avoid flawed conclusions that might arise from naive approaches in high-dimensions.

The findings of each paper presented in this dissertation pave the way for future research to enhance the field of double machine learning further. In Chapter 2, we showed that L_2-Boosting algorithms are a competitive alternative to Lasso. However, a more detailed comparison to Lasso is required to understand in which situations Boosting is superior to Lasso and vice versa. This could be achieved by extensive simulation studies or a detailed theoretical analysis of L_2-Boosting. In contrast to L_1-penalized methods like Lasso, the estimation properties of Boosting algorithms for regression in high-dimensions are not widely discussed in the literature yet. Further, transformation models in high-dimensions can be extended by considering another identification strategy for the true transformation parameter. In many applications, the response variable is transformed to achieve homogeneity of the error terms. In a research project, I am working on an adjustment of the proposed transformation model, where the true transformation ensures that the zero conditional mean assumption for the error term holds. In addition, transformation models can be applied to analyze duration data. In this context, I am working on an adjustment of the proposed transformation model to link a failure time with a high-dimensional vector of covariates in an empirical application that analyzes US credit data. In Chapter 4 and Chapter 5, new methodologies and key theoretical insights are provided. A natural extension for future work is to increasingly use these methodologies in empirical applications. For instance, I am working on an application of the high-dimensional Gaussian graphical model to large data sets from biology and finance in order to analyze correlation networks. I hope that future research in empirical applications will greatly benefit from the new methodologies provided by this work.

Bibliography

[1] T. Amemiya and J. L. Powell. "A comparison of the Box-Cox maximum likelihood estimator and the non-linear two-stage least squares estimator". In: *Journal of Econometrics* 17.3 (1981), pp. 351–381.

[2] D. Andrews. "A note on the selection of data transformations". In: *Biometrika* 58.2 (1971), pp. 249–254.

[3] P. Assouad. "Densité et dimension". In: *Annales de l'Institut Fourier*. Vol. 33. 3. 1983, pp. 233–282.

[4] A. Atkinson. "Testing transformations to normality". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1973), pp. 473–479.

[5] A. Belloni, D. Chen, V. Chernozhukov, and C. Hansen. "Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain". In: *Econometrica* 80.6 (2012), pp. 2369–2429. URL: http://dx.doi.org/10.3982/ECTA9626.

[6] A. Belloni and V. Chernozhukov. "Least squares after model selection in high-dimensional sparse models". In: *Bernoulli* 19.2 (2013), pp. 521–547. URL: http://dx.doi.org/10.3150/11-BEJ410.

[7] A. Belloni, V. Chernozhukov, et al. "ℓ1-penalized quantile regression in high-dimensional sparse models". In: *The Annals of Statistics* 39.1 (2011), pp. 82–130.

[8] A. Belloni, V. Chernozukov, and C. Hansen. "Inference on Treatment Effects after Selection among High-Dimensional Controls". In: *The Review of Economic Studies* 81.2 (287) (2014), pp. 608–650. URL: http://www.jstor.org/stable/43551575.

[9] A. Belloni, V. Chernozhukov, and K. Kato. *Uniform post selection inference for LAD regression and other Z-estimation problems*. Tech. rep. cemmap working paper, Centre for Microdata Methods and Practice, 2014.

[10] A. Belloni, V. Chernozhukov, and K. Kato. "Uniform post-selection inference for least absolute deviation regression and other Z-estimation problems". In: *Biometrika* 102.1 (2014), 77–94. URL: http://dx.doi.org/10.1093/biomet/asu056.

[11] A. Belloni, V. Chernozhukov, I. Fernández-Val, and C. Hansen. "Program Evaluation and Causal Inference With High-Dimensional Data". In: *Econometrica* 85.1 (2017), pp. 233–298.

[12] A. Belloni, V. Chernozhukov, D. Chetverikov, and Y. Wei. "Uniformly valid post-regularization confidence regions for many functional parameters in z-estimation framework". In: *Annals of statistics* 46.6B (2018), p. 3643.

[13] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. "Simultaneous analysis of Lasso and Dantzig selector". In: *Annals of Statistics* 37.4 (2009), pp. 1705–1732.

[14] P. J. Bickel and K. A. Doksum. "An Analysis of Transformations Revisited". In: *Journal of the American Statistical Association* 76.374 (1981), pp. 296–311.

[15] A. Bloniarz, H. Liu, C.-H. Zhang, J. S. Sekhon, and B. Yu. "Lasso adjustments of treatment effect estimates in randomized experiments". In: *Proceedings of the National Academy of Sciences* (2016). eprint: http://www.pnas.org/content/early/2016/06/30/1510506113.full.pdf. URL: http://www.pnas.org/content/early/2016/06/30/1510506113.abstract.

[16] G. E. P. Box and D. R. Cox. "An analysis of transformations". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1964), pp. 211–252.

[17] L. Breiman. "Bagging predictors". In: *Machine learning* 24.2 (1996), pp. 123–140.

[18] L. Breiman. "Arcing Classifiers". In: *The Annals of Statistics* 26.3 (1998), pp. 801–824.

[19] L. Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[20] P. Bühlmann. "Boosting for High-Dimensional Linear Models". In: *The Annals of Statistics* 34.2 (2006), pp. 559–583.

[21] P. Bühlmann and T. Hothorn. "Boosting Algorithms: Regularization, Prediction and Model Fitting". In: *Statistical Science* 22.4 (2007). with discussion, pp. 477–505. URL: http://dx.doi.org/10.1214/07-STS242.

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

[23] P. Bühlmann and B. Yu. "Boosting with the L_2 Loss: Regression and Classification". In: *Journal of the American Statistical Association* 98.462 (2003), pp. 324–339. URL: http://www.jstor.org/stable/30045243.

[24] T. Cai, W. Liu, and X. Luo. "A constrained l1 minimization approach to sparse precision matrix estimation". In: *Journal of the American Statistical Association* 106.494 (2011), pp. 594–607.

[25] E. Candes, T. Tao, et al. "The Dantzig selector: Statistical estimation when p is much larger than n". In: *The Annals of Statistics* 35.6 (2007), pp. 2313–2351.

[26] R. J. Carroll. "A robust method for testing transformations to achieve approximate normality". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1980), pp. 71–78.

[27] J. Chang, Y. Qiu, Q. Yao, and T. Zou. "Confidence regions for entries of a large precision matrix". In: *Journal of Econometrics* 206.1 (2018), pp. 57 –82. URL: http://www.sciencedirect.com/science/article/pii/S0304407618300782.

[28] R. Y. Chen, A. Gittens, and J. A. Tropp. "The masked sample covariance estimator: an analysis using matrix concentration inequalities". In: *Information and Inference: A Journal of the IMA* 1.1 (2012), pp. 2–20.

[29] X. Chen and H. White. "Improved rates and asymptotic normality for nonparametric neural network estimators". In: *IEEE Transactions on Information Theory* 45.2 (1999), pp. 682–691.

[30] V. Chernozhukov, D. Chetverikov, K. Kato, et al. "Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors". In: *The Annals of Statistics* 41.6 (2013), pp. 2786–2819.

[31] V. Chernozhukov, D. Chetverikov, K. Kato, et al. "Gaussian approximation of suprema of empirical processes". In: *The Annals of Statistics* 42.4 (2014), pp. 1564–1597.

[32] V. Chernozhukov, C. Hansen, and M. Spindler. *hdm: High-Dimensional Metrics*. R package version 0.1. 2015.

[33] V. Chernozhukov, C. Hansen, and M. Spindler. "Valid Post-Selection and Post-Regularization Inference: An Elementary, General Approach". In: *Annual Review of Economics* 7.1 (2015), pp. 649–688. eprint: https://doi.org/10.1146/annurev-economics-012315-015826. URL: https://doi.org/10.1146/annurev-economics-012315-015826.

[34] V. Chernozhukov, D. Chetverikov, K. Kato, et al. "Central limit theorems and bootstrap in high dimensions". In: *The Annals of Probability* 45.4 (2017), pp. 2309–2352.

[35] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. "Double/debiased machine learning for treatment and structural parameters". In: *The Econometrics Journal* 21.1 (2018), pp. C1–C68. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/ectj.12097.

[36] G. Claeskens and I. Keilegom. "Bootstrap confidence bands for regression curves and their derivatives". In: *Annals of Statistics* 31 (2003).

[37] A. F. Connors, T. Speroff, N. V. Dawson, C. Thomas, F. E. Harrell, D. Wagner, N. Desbiens, L. Goldman, A. W. Wu, R. M. Califf, et al. "The effectiveness of right heart catheterization in the initial care of critically III patients". In: *Jama* 276.11 (1996), pp. 889–897.

[38] D. R. Cox. "Some Problems Connected with Statistical Inference". In: *Ann. Math. Statist.* 29.2 (June 1958), pp. 357–372. URL: https://doi.org/10.1214/aoms/1177706618.

[39] J. E. Dalen. "The pulmonary artery catheter—friend, foe, or accomplice?" In: *Jama* 286.3 (2001), pp. 348–350.

[40] K. Doksum and A. Samarov. "Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression". In: *The Annals of Statistics* 23.5 (1995), pp. 1443–1473.

[41] N. R. Draper and D. R. Cox. "On distributions and their transformation to normality". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1969), pp. 472–476.

[42] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, et al. "Least angle regression". In: *The Annals of statistics* 32.2 (2004), pp. 407–499.

[43] J. Fan and W. Zhang. "Simultaneous Confidence Bands and Hypothesis Testing in Varying-Coefficient Models". In: *Scandinavian Journal of Statistics* 27.4 (2000), pp. 715–731. URL: http://www.jstor.org/stable/4616637.

[44] C. Feng, H. Wang, N. Lu, and X. M. Tu. "Log transformation: application and interpretation in biomedical research". In: *Statistics in medicine* 32.2 (2013), pp. 230–239.

[45] Y. Freund and R. E. Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting". In: *Journal of computer and system sciences* 55.1 (1997), pp. 119–139.

[46] J. H. Friedman, T. Hastie, and R. Tibshirani. "Additive Logistic Regression: A Statistical View of Boosting". In: *The Annals of Statistics* 28 (2000). with discussion, pp. 337–407. URL: http://projecteuclid.org/euclid.aos/1016218223.

[47] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York, 2001.

[48] J. H. Friedman. "Greedy Function Approximation: A Gradient Boosting Machine". English. In: *The Annals of Statistics* 29.5 (2001), pp. 1189–1232. URL: http://www.jstor.org/stable/2699986.

[49] J. H. Friedman and W. Stuetzle. "Projection Pursuit Regression". In: *Journal of the American Statistical Association* 76.376 (1981), pp. 817–823.

[50] K. Gregory, E. Mammen, and M. Wahl. "Statistical inference in sparse high-dimensional additive models". In: *arXiv preprint arXiv:1603.07632* (2016).

[51] C. Hansen, J. Hausman, and W. Newey. "Estimation with many instrumental variables". In: *Journal of Business & Economic Statistics* 26.4 (2008), pp. 398–422.

[52] D. Harrison Jr and D. L. Rubinfeld. "Hedonic housing prices and the demand for clean air". In: *Journal of Environmental Economics and Management* 5.1 (1978), pp. 81–102.

[53] S. Harvey, D. A. Harrison, M. Singer, J. Ashcroft, C. M. Jones, D. Elbourne, W. Brampton, D. Williams, D. Young, K. Rowan, et al. "Assessment of the clinical effectiveness of pulmonary artery catheters in management of patients in intensive care (PAC-Man): a randomised controlled trial". In: *Lancet* 366.9484 (2005), pp. 472–477.

[54] T. Hastie and R. Tibshirani. *Generalized Additive Models*. Vol. 43. Chapman and Hall, Ltd., London, 1990.

[55] W. Härdle. "Asymptotic maximal deviation of M-smoothers". In: *Journal of Multivariate Analysis* 29.2 (1989), pp. 163–179. URL: https://ideas.repec.org/a/eee/jmvana/v29y1989i2p163-179.html.

[56] J. Huang, J. L. Horowitz, and F. Wei. "Variable selection in nonparametric additive models". In: *Annals of statistics* 38.4 (2010), p. 2282.

[57] G. W. Imbens and D. B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. New York, NY, USA: Cambridge University Press, 2015.

[58] J. Janková and S. Van De Geer. "Honest confidence regions and optimality in high-dimensional precision matrix estimation". In: *Test* 26.1 (2017), pp. 143–162.

[59] K. Kato. "Two-step estimation of high dimensional additive models". In: *arXiv preprint* (2012). URL: https://arxiv.org/abs/1207.5313.

[60] S. Klaassen, J. Kueck, M. Spindler, and V. Chernozhukov. "Uniform Inference in High-Dimensional Gaussian Graphical Models". In: *arXiv preprint arXiv:1808.10532* (2018).

[61] N. Kloodt and N. Neumeyer. "Specification tests in semiparametric transformation models". In: *ArXiv e-prints* (2017). arXiv: 1709.06855 [stat.ME].

[62] V. Koltchinskii and M. Yuan. "Sparsity in multiple kernel learning". In: *Annals of Statistics* 38.6 (2010), pp. 3660–3695. URL: https://doi.org/10.1214/10-AOS825.

[63] E. Kong and Y. Xia. "A single-index quantile regression model and its estimation". In: *Econometric Theory* 28.4 (2012), pp. 730–768.

[64] D. Kozbur. "Inference in Additively Separable Models With a High Dimensional Conditioning Set". In: *SSRN Electronic Journal* (2015).

[65] C. Lam and J. Fan. "Sparsistency and rates of convergence in large covariance matrix estimation". In: *Annals of statistics* 37.6B (2009), p. 4254.

[66] S. L. Lauritzen. *Graphical models*. Vol. 17. Clarendon Press, 1996.

[67] H. Leeb and B. M. Pötscher. "Model Selection and Inference: Facts and Fiction". English. In: *Econometric Theory* 21.1 (2005), pp. 21–59. URL: http://www.jstor.org/stable/3533623.

[68] F. Leisch and E. Dimitriadou. *mlbench: Machine learning benchmark problems*. R package version 2.1-1. 2010.

[69] Y. Lin and H. H. Zhang. "Component selection and smoothing in multivariate nonparametric regression". In: *Annals of Statistics* 34.5 (2006), pp. 2272–2297. URL: https://doi.org/10.1214/009053606000000722.

[70] O. Linton, S. Sperlich, and I. Van Keilegom. "Estimation of a semiparametric transformation model". In: *The Annals of Statistics* (2008), pp. 686–718.

[71] Y. Lou, J. Bien, R. Caruana, and J. Gehrke. "Sparse Partially Linear Additive Models". In: *Journal of Computational and Graphical Statistics* 25.4 (2016), pp. 1126–1140.

[72] J. Lu, M. Kolar, and H. Liu. "Kernel Meets Sieve: Post-Regularization Confidence Bands for Sparse Additive Model". In: *Journal of the American Statistical Association* (2020), pp. 1–16. URL: https://doi.org/10.1080/01621459.2019.1689984.

[73] Y. Luo and M. Spindler. "High-Dimensional $L2$ Boosting: Rate of Convergence". In: *arXiv preprint arXiv:1602.08927* (2016).

[74] W. G. Manning and J. Mullahy. "Estimating log models: to transform or not to transform?" In: *Journal of health economics* 20.4 (2001), pp. 461–494.

[75] L. Meier, S. Van de Geer, and P. Bühlmann. "High-dimensional additive modeling". In: *The Annals of Statistics* 37.6B (2009), pp. 3779–3821.

[76] N. Meinshausen and P. Bühlmann. "High-dimensional graphs and variable selection with the lasso". In: *The Annals of Statistics* (2006), pp. 1436–1462.

[77] H. L. Nelson and C. Granger. "Experience with using the Box-Cox transformation when forecasting economic time series". In: *Journal of Econometrics* 10.1 (1979), pp. 57 –69. URL: http://www.sciencedirect.com/science/article/pii/0304407679900642.

[78] N. Neumeyer, H. Noh, and I. Van Keilegom. "Heteroscedastic semiparametric transformation models: estimation and testing for validity". In: *Statistica Sinica* 26 (2016), pp. 925–954.

[79] W. K. Newey. "The asymptotic variance of semiparametric estimators". In: *Econometrica: Journal of the Econometric Society* (1994), pp. 1349–1382.

[80] D. J. Newman, S. Hettich, C. L. Blake, and C. J. Merz. *UCI Repository of machine learning databases*. 1998. URL: http://www.ics.uci.edu/~mlearn/MLRepository.html.

[81] A. Petersen, D. Witten, and N. Simon. "Fused Lasso Additive Model". In: *Journal of Computational and Graphical Statistics* 25.4 (2016). PMID: 28239246, pp. 1005–1025.

[82] A. J. Quiroz, M. Nakamura, and F. J. Pérez. "Estimation of a multivariate Box-Cox transformation to elliptical symmetry via the empirical characteristic function". In: *Annals of the Institute of Statistical Mathematics* 48.4 (1996), pp. 687–709.

[83] P. Ravikumar, J. Lafferty, H. Liu, and L. Wasserman. "Sparse additive models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71.5 (2009), pp. 1009–1030. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2009.00718.x. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2009.00718.x.

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

[85] Z. Ren, T. Sun, C.-H. Zhang, H. H. Zhou, et al. "Asymptotic normality and optimalities in estimation of large Gaussian graphical models". In: *The Annals of Statistics* 43.3 (2015), pp. 991–1026.

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

[87] S. Ruggles, K. Genadek, R. Goeken, J. Grover, and M. Sobek. "Integrated Public Use Microdata Series: Version 6.0 [dataset]". In: *Minneapolis: University of Minnesota* (2015.). URL: http://doi.org/10.18128/D010.V6.0.

[88] R. M. Sakia. "The Box-Cox Transformation Technique: A Review". In: *Journal of the Royal Statistical Society. Series D (The Statistician)* 41.2 (1992), pp. 169–178.

[89] S. Sardy and P. Tseng. "AMlet, RAMlet, and GAMlet: Automatic Nonlinear Fitting of Additive Models, Robust and Generalized, with Wavelets". In: *Journal of Computational and Graphical Statistics* 13.2 (2004), pp. 283–309. URL: http://www.jstor.org/stable/1391177.

[90] C. J. Stone. "Additive Regression and Other Nonparametric Models". In: *Annals of Statistics* 13.2 (1985), pp. 689–705. URL: https://doi.org/10.1214/aos/1176349548.

[91] J. Sun and C. R. Loader. "Simultaneous Confidence Bands for Linear Regression and Smoothing". In: *Annals of Statistics* 22.3 (1994), pp. 1328–1345. URL: https://doi.org/10.1214/aos/1176325631.

[92] T. Sun and C.-H. Zhang. "Sparse matrix inversion with scaled lasso". In: *The Journal of Machine Learning Research* 14.1 (2013), pp. 3385–3418.

[93] R. Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.

[94] A. Van der Vaart and J. Wellner. *Weak convergence and empirical processes*. 1996.

[95] S. Van De Geer, P. Bühlmann, et al. "On the conditions used to prove oracle results for the Lasso". In: *Electronic Journal of Statistics* 3 (2009), pp. 1360–1392.

[96] S. Van De Geer, P. Bühlmann, Y. Ritov, R. Dezeure, et al. "On asymptotically optimal confidence regions and tests for high-dimensional models". In: *The Annals of Statistics* 42.3 (2014), pp. 1166–1202.

[97] A. Vanhems and I. Van Keilegom. "Estimation of a Semiparametric Transformation Model in the Presence of Endogeneity". In: *Econometric Theory* (2018), pp. 1–38.

[98] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York: Springer, 2002. URL: http://www.stats.ox.ac.uk/pub/MASS4.

[99] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge university press, 2018.

[100] S. Wager and G. Walther. "Adaptive concentration of regression trees, with application to random forests". In: *arXiv preprint arXiv:1503.06388* (2015).

[101] S. N. Wood. *Generalized additive models: an introduction with R*. CRC press, 2017.

[102] I.-K. Yeo and R. A. Johnson. "A New Family of Power Transformations to Improve Normality or Symmetry". In: *Biometrika* 87.4 (2000), pp. 954–959.

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

[104] M. Yuan and Y. Lin. "Model selection and estimation in the Gaussian graphical model". In: *Biometrika* 94.1 (2007), pp. 19–35.

[105] C.-H. Zhang and S. S. Zhang. "Confidence intervals for low dimensional parameters in high dimensional linear models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76.1 (2014), pp. 217–242. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12026.

[106] W. Zhang and H. Peng. "Simultaneous confidence band and hypothesis test in generalised varying-coefficient models". In: *Journal of Multivariate Analysis* 101.7 (2010), pp. 1656 –1680. URL: http://www.sciencedirect.com/science/article/pii/S0047259X10000539.

[107] T. Zhao, H. Liu, K. Roeder, J. Lafferty, and L. Wasserman. "The huge package for high-dimensional undirected graph estimation in R". In: *Journal of Machine Learning Research* 13.1 (2012), pp. 1059–1062.

Appendices

A.1 Statement of Personal Contribution Pursuant to §6(4) PromO

Estimation and Inference of Treatment Effects with L_2-Boosting in High-Dimensional Settings

Jannis Kueck, Ye Luo and Martin Spindler

- Performed literature review
- Constructed proofs
- Revised simulation study and empirical application
- Interpreted results
- Prepared manuscript

Transformation Models in High-Dimensions

Sven Klaassen, Jannis Kueck and Martin Spindler

- Conducted literature review
- Conceptualized statistical model
- Defined research question
- Constructed proofs
- Implemented algorithm and simulation study
- Visualized and interpreted results
- Prepared manuscript
- Presented paper at the research seminar of the Deep Data Lab, UCI (2019)
- Collected feedback and revised manuscript

Uniform Inference in High-Dimensional Generalized Additive Models

Philipp Bach, Sven Klaassen, Jannis Kueck and Martin Spindler

- Conducted literature review
- Conceptualized statistical model
- Constructed proofs
- Interpreted results
- Prepared manuscript
- Presented paper at School of Economics and Political Science, University of St.Gallen (2020)

Uniform Inference in High-Dimensional Gaussian Graphical Models

Victor Chernozhukov, Sven Klaassen, Jannis Kueck and Martin Spindler

- Performed literature review
- Constructed proofs
- Implemented algorithm and simulation
- Visualized and interpreted results
- Prepared manuscript
- Presented paper at Workshop "Machine Learning in Economics and Econometrics", Munich, Max Planck Society (2018), Data Science Summer School, Paris, École Polytechnique (2018), Conference "Statistics of Machine Learning", Prague, Charles University (2019)
- Collected feedback and revised manuscript

A.2 Short Summaries of Papers Pursuant to §6(6) PromO

Short summaries in English language

Estimation and Inference of Treatment Effects with L_2-Boosting in High-Dimensional Settings (Chapter 2)

Boosting algorithms are very popular in machine learning and have proven to be particularly useful for prediction and variable selection. Nevertheless, in many applications, one is interested in inference about treatment effects or policy variables in a high-dimensional setting. As rich data sets become more and more available containing many controls or instrumental variables, variable selection is increasingly challenging for empirical researchers. We provide a methodology for valid inference about a treatment effect when post- or orthogonal L_2-Boosting is used for the variable selection. This methodology is applied in a high-dimensional linear model with many controls and in an instrumental variable model with potentially many instruments. We present simulation results for the proposed methodology as well as an empirical application. The application is based on the so-called "PAC-man" study, which has analyzed the effectiveness of a pulmonary artery catheter in a randomized control trial. We confirm that the treatment effect is not significantly different from zero.

Transformation Models in High-Dimensions (Chapter 3)

Transformation models are a very important tool for applied statisticians and econometricians. In many applications, the dependent variable is transformed so that homogeneity or normal distribution of the error holds. We analyze transformation models in a high-dimensional setting, where the set of potential covariates is large. Our proposed model builds on a high-dimensional linear model and combines it with a parametric transformation of the response variable. We propose an estimator for the transformation parameter and show that it is asymptotically normally distributed by using an orthogonalized moment condition where the nuisance functions depend on the target parameter. We provide general results regarding inference in Z-estimation frameworks where we explicitly allow for such a dependency. These results are of independent interest for general Z-estimation problems with the same underlying structure. A common practice in labor economics is to first transform wages with the log function and then to further process since wage data is non-negative and often highly skewed. In an empirical application, we test if this transformation holds for American Community Survey (ACS) data from the United States. We conclude that the log transformation is rejected on a 5% significance level. In a simulation study, we show that our estimator works well even in small samples.

Uniform Inference in High-Dimensional Generalized Additive Models (Chapter 4)

Generalized additive models $Y = f_1(X_1) + \ldots + f_p(X_p) + \varepsilon$ are very popular in statistics. As the estimation of a nonparametric regression function $f(X_1, \ldots, X_p)$ is practically infeasible when p is large, these models impose an additive structure of the regression function. We develop a methodology for the estimation of the nonparametric component f_1 in a high-dimensional setting, where the number of regressors p may increase with the sample size. As usual in high-dimensions, a sparsity assumption is crucial for the analysis. We employ sieve estimation and embed it in a high-dimensional Z-estimation framework which allows us to construct uniformly valid confidence bands for the function f_1. We also run simulation studies which show that our proposed method gives reliable results concerning the estimation and coverage properties even in small samples. Finally, we demonstrate the use of the proposed method empirically by analyzing the well-known Boston housing data set. Our methodology suggests nonlinear and significant effects on the median value of owner-occupied homes for the variables LSTAT and RM,

that denotes the percentage of lower status population and the average number of rooms per dwelling, respectively. This is in line with the economic intuition and the findings in the literature.

Uniform Inference in High-Dimensional Gaussian Graphical Models (Chapter 5)

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 being possibly much larger than the 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. We do not aim to estimate the precision matrix but we focus on quantifying the uncertainty of recovering its support by providing a significance test for a set of potential edges. 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 Lasso 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 comprehensive simulation studies that our procedure has good small sample properties.

Kurzfassungen in deutscher Sprache

Schätzung und Inferenz von Behandlungseffekten mit L_2-Boosting in hochdimensionalen Situationen (Kapitel 2)

Boosting-Algorithmen sind sehr beliebte Methoden des maschinellen Lernens und haben sich für die Vorhersage und Variablenauswahl als äußerst nützlich erwiesen. Nichtsdestotrotz ist man bei vielen Anwendungen in hochdimensionalen Situationen daran interessiert, kausale Rückschlüsse über Behandlungseffekte oder politische Maßnahmen zu ziehen. Da immer größere Datensätze zur Verfügung stehen, die eine Vielzahl von Kontroll- oder Instrumentalvariablen enthalten, wird die Variablenauswahl zunehmend herausfordernd für empirische Forscher. Wir liefern eine Methodik zur validen Inferenz über Behandlungseffekte, wenn post- oder orthogonal L_2-Boosting für die Variablenauswahl verwendet wird. Diese Methodik wird in einem hochdimensionalen linearen Modell mit vielen Kontrollvariablen und in einem Modell mit potenziell vielen Instrumentalvariablen angewandt. Wir präsentieren sowohl Simulationsergebnisse für die vorgeschlagene Methode als auch eine empirische Anwendung. Die Anwendung basiert auf der sogenannten „PAC-man“-Studie, welche die Effektivität des Pulmonalarterienkatheters in einer randomisierten Fallstudie untersucht hat. Wir bestätigen in unserer Analyse, dass der Behandlungseffekt nicht signifikant von null verschieden ist.

Transformationsmodelle in hoher Dimension (Kapitel 3)

Transformationsmodelle sind ein wichtiges Werkzeug für angewandte Statistiker und Ökonomen. In vielen Anwendungen wird die abhängige Variable transformiert, um Homogenität und/oder Normalität der Fehlerterme zu erzeugen. Wir analysieren Transformationsmodelle in hochdimensionalen Situationen, in denen die Anzahl an potenziellen Kovariablen groß ist. Unser entwickeltes Modell baut auf einem hochdimensionalen linearen Modell auf und kombiniert es mit einer parametrischen Transformation der abhängigen Variable durch eine gegebene Familie von streng monoton steigenden Funktionen. Basierend auf einer orthogonalisierten Momentenbedingung leiten wir einen Schätzer für den Transformationsparameter her und zeigen, dass dieser asymptotisch normalverteilt ist. Dabei ist hervorzuheben, dass die unbekannten Störfunktionen von dem Zielparameter abhängen. Wir leiten allgemeine Ergebnisse für Inferenz in Z-Schätzungsproblemen her, in denen wir eine solche Abhängigkeit ausdrücklich erlauben. Diese Ergebnisse sind von unabhängigem Interesse für allgemeine Z-Schätzungsprobleme mit der gleichen zugrundeliegenden Struktur. Eine gängige Praxis in der Arbeitsökonomie ist es, Löhne mit dem Logarithmus zu transformieren und dann weiter zu analysieren, da Löhne nicht negativ und oft stark rechtsschief sind. In einer empirischen Studie untersuchen wir, ob die logarithmische Transformation für American Community Survey (ACS) Daten aus den USA geeignet ist. Wir kommen zu dem Schluss, dass der Logarithmus zu einem Signifikanzniveau von 5 % abgelehnt wird. Unsere Simulationsstudie zeigt, dass der vorgeschlagene Schätzer auch in kleinen Stichproben gute Ergebnisse liefert.

Gleichmäßige Inferenz in hochdimensionalen verallgemeinerten additiven Modellen (Kapitel 4)

Verallgemeinerte additive Modelle $Y = f_1(X_1) + \ldots + f_p(X_p) + \varepsilon$ sind in der Statistik sehr beliebt. Da die Schätzung einer nichtparametrischen Regressionsfunktion $f(X_1, \ldots, X_p)$ praktisch nicht möglich ist, wenn p groß ist, nehmen diese Modelle eine additive Struktur der Regressionsfunktion an. Wir entwickeln eine Methode für die Schätzung der nichtparametrischen Komponente f_1 in hochdimensionalen Situationen, in denen die Anzahl der Kovariablen p mit der Stichprobengröße steigen kann. Wie üblich in hoher Dimension, ist eine sogenannte Spärlichkeitsannahme („sparsity“-Annahme) für die Analyse entscheidend. Wir verwenden die Sieve-Schätzung und betten das Modell in ein hochdimensionales Z-Schätzungsproblem

ein, um gleichmäßig valide Konfidenzbänder für die Funktion f_1 zu konstruieren. Wir führen Simulationsstudien durch, die zeigen, dass unsere Schätzmethode selbst für kleine Stichproben zuverlässige Ergebnisse bezüglich der Schätzung und den Überdeckungswahrscheinlichkeiten liefert. Abschließend demonstrieren wir den Nutzen unserer Methodik empirisch, indem wir den bekannten „Boston housing“ Datensatz analysieren. Unsere Methodik legt nichtlineare und signifikante Effekte auf Medianwert von Eigenheimen für die Variablen LSTAT und RM nahe, die den Prozentsatz der Bevölkerung mit niedrigem Status in der Wohngegend bzw. die durchschnittliche Anzahl von Zimmern pro Wohnung angeben. Dies steht im Einklang mit der ökonomischen Intuition und den Erkenntnissen aus der Fachliteratur.

Gleichmäßige Inferenz in hochdimensionalen Gaußschen graphischen Modellen (Kapitel 5)

Grafische Modelle sind zu einem beliebten Werkzeug zur Darstellung von Abhängigkeiten einer großen Menge an Variablen geworden und sind der Schlüssel zur Darstellung kausaler Strukturen. Wir liefern Ergebnisse für gleichmäßige Inferenz in hochdimensionalen grafischen Modellen, wobei die Anzahl der Zielparameter möglicherweise bedeutend größer ist als die Stichprobengröße. Dies ist vor allem dann von Bedeutung, wenn bestimmte Merkmale oder Strukturen eines kausalen Modells untersucht werden sollen. Unsere Ergebnisse zeigen, wie in hochdimensionalen Situationen graphische Modelle mit modernen Methoden des maschinellen Lernens in komplexen Datensätzen geschätzt und untersucht werden können. Wir versuchen dabei nicht die gesamte Präzisionsmatrix zu schätzen, sondern konzentrieren uns auf die Quantifizierung der Unsicherheit bei der Identifizierung des Trägers, indem wir einen Signifikanztest für eine große Menge an potenziellen Kanten durchführen. Um gleichmäßige Konfidenzregionen für eine hohe Anzahl an Zielparametern zu konstruieren, sind ausreichend schnelle Schätzraten der Störfunktionen von entscheidender Bedeutung. In diesem Zusammenhang leiten wir gleichmäßige Schätzraten und Spärlichkeitsgarantien für den Square-Root-Lasso Schätzer in einem zufälligen Design unter approximativen Spärlichkeitsbedingungen her, die für verwandte Probleme in hoher Dimension von unabhängigem Interesse sein könnten. Ebenfalls zeigen wir in umfangreichen Simulationsstudien, dass unsere Methodik selbst bei kleinen Stichprobenumfängen gute Schätzeigenschaften aufweist.

A.3 List of Publications Pursuant to §6 (6) PromO

Journal article	**Publication status**
Kueck, J., Luo Y., and Spindler, M. (2020). Estimation and Inference of Treatment Effects with L_2-Boosting in High-Dimensional Settings.	Revised and resubmit at *Journal of Econometrics*
Klaassen, S., Kueck, J., and Spindler, M. (2017). Transformation models in high-dimensions.	Revised and resubmit at *Journal of Business & Economic Statistics*
Klaassen, S., Kueck, J., Spindler, M. and Chernozhukov, V. (2018). Uniform inference in high-dimensional Gaussian graphical models.	Reject and resubmit at *Biometrika*
Bach, P., Klaassen, S., Kueck, J. and Spindler, M. (2020). Uniform Inference in High-Dimensional Generalized Additive Models.	*Working Paper*

Affidavit

Hiermit erkläre ich, Jannis Malte Kück, an Eides statt, dass ich die Dissertation mit dem Titel

Advances in Machine Learning: Valid Inference about High-Dimensional Parameters

selbständig – und bei einer Zusammenarbeit mit anderen Wissenschaftlerinnen und Wissenschaftlern gemäß den beigefügten Darstellungen nach §6 Abs. 4 der Promotionsordnung der Fakultät der Betriebswirtschaft vom 9. Juli 2014 – verfasst habe und keine anderen als die von mir angegebenen Hilfsmittel benutzt habe. Die den herangezogenen Werken wörtlich oder sinngemäß entnommenen Stellen sind als solche gekennzeichnet.

Ich versichere, dass ich keine kommerzielle Promotionsberatung in Anspruch genommen habe und die Arbeit nicht schon in einem früheren Promotionsverfahren im In- oder Ausland angenommen oder als ungenügend beurteilt worden ist.