Uniform Inference in High-Dimensional Generalized Additive Models

Philipp Bach, ***Sven Klaassen***[†],
Jannis Kueck and ***Martin Spindler***

University of Hamburg

Abstract: We develop a method for uniform valid confidence bands of a nonparametric component f_1 in the general additive model $Y = f_1(X_1) + \ldots + f_p(X_p) + \varepsilon$ in a high-dimensional setting. We employ sieve estimation and embed it in a high-dimensional Z-estimation framework allowing us to construct uniformly valid confidence bands for the first component f_1. As usual in high-dimensional settings where the number of regressors p may increase with sample, a sparsity assumption is critical for the analysis. We also run simulations studies which show that our proposed method gives reliable results concerning the estimation properties and coverage properties even in small samples. Finally, we illustrate our procedure with an empirical application demonstrating the implementation and the use of the proposed method in practice.

MSC 2010 subject classifications: Primary 62G08; Secondary 62-07, 41A15.
Keywords and phrases: General Additive Models, High-dimensional Setting, Z-estimation, Double Machine Learning, Lasso.

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 $\mathrm{E}[\varepsilon|X] = 0$. 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 (GAM):

$$Y = \alpha + f_1(X_1) + \ldots + f_p(X_p) + \varepsilon, \tag{1.1}$$

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 (1981), Stone (1985) and Hastie and Tibshirani (1990). Estimation and inference in the low-dimensional setting with fixed p has been analyzed widely in the literature. For

[*]Version April 2020.
[†]Corresponding author: *sven.klaassen@uni-hamburg.de*.

an introduction to GAMs we refer to the textbook treatments by Hastie and Tibshirani (1990) and Wood (2017). 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 non-zero. 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 (2004), Lin and Zhang (2006) and many others (Ravikumar et al., 2009; Meier et al., 2009; Huang et al., 2010; Koltchinskii and Yuan, 2010; Kato, 2012; Petersen et al., 2016; Lou et al., 2016). 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 (1989), Sun and Loader (1994), Fan and Zhang (2000), Claeskens and Keilegom (2003) and Zhang and Peng (2010). 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 (2015) 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. (2014b). 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. (2014) and Zhang and Zhang (2014). For a survey on post-selection inference in high-dimensional settings and generalization we refer to Chernozhukov et al. (2015b). We consider the same setting as Kozbur (2015), 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 (2015) 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. (2018) and Belloni et al. (2014a). Further, Kozbur (2015) relies on two high level assumptions on lasso estimation and variable selection (see Assumptions 9 and 10 in Kozbur (2015)) 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. (2016) use the so-called Debiasing approach in Zhang and Zhang (2014) 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 undersmoothed 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. (2016) 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. (2020) 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 fixed 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. (2020) 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. (2018). We also provide a theory driven choice of the penalization level. As in Gregory et al. (2016), Lu et al. (2020) 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 non-zero components $s = O(1)$ is bounded. In our setting, s may grow to infinity with increasing sample size. Nevertheless, their approach differs from ours in that they consider an ATLAS model, 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. (2016). The results show that the suggested method is able to perform valid simultaneous inference even in small 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.

1.1. Organization of the Paper

The paper is organized as follows. In Section 2, the setting is outlined. Section 3 introduces the estimation method. In Section 4, the main result is provided. A simulation study, highlighting the small sample properties and implementation of our proposed method, is presented in Section 5. Section 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 7. The Appendix includes additional technical material. In Appendix A, a general result for uniform inference about a high-dimensional linear functional is presented. Appendix B provides results regarding uniform lasso estimation rates in high-dimensions. Finally, computational details are presented in Appendix C.

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 $\mathbb{P}_n$ 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 non-zero 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 $P(a_n > C b_n) \leq \epsilon$ for all n.

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 $\mathrm{Var}(\varepsilon|X) \geq c$. Let the scalar response Y and features $X = (X_1, \ldots, X_p)$ take values in $\mathcal{Y}$ respectively $\mathcal{X} = (\mathcal{X}_1, \ldots, \mathcal{X}_p)$. 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 sample size. Assume that the approximations are given by

$$f_1(X_1) = \theta_0^T g(X_1) + b_1(X_1), \tag{2.1}$$

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

$$f_{-1}(X_{-1}) := \beta_0^T h(X_{-1}) + b_2(X_{-1}), \tag{2.2}$$

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

$$g_l(X_1) = (\gamma_0^{(l)})^T Z_{-l} + b_3^{(l)}(Z_{-l}) + \nu^{(l)} \tag{2.3}$$

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 (2.3) is used to construct an orthogonal score function for valid inference in a high-dimensional setting as in Chernozhukov et al. (2017). Estimating

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

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

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

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

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

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 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 $l = 1, \dots, d_1$.

Comment 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 $l = 1, \dots, d_1$, and, given further conditions mentioned in Section 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.

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 and solved for the target parameters, which yield an estimate for the first component $\hat{f}_1$. 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. The details are given in this section.

Let

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

and

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

for some vector

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

and

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

For each $l = 1, \dots, d_1$, 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

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

is defined as the solution of

$$\sup_{l=1,\dots,d_1} \left\{ \left|\mathbb{E}_n\left[\psi_l\big(W, \hat{\theta}_l, \hat{\eta}_l\big)\right]\right| - \inf_{\theta \in \Theta_l} \left|\mathbb{E}_n\left[\psi_l\big(W, \theta, \hat{\eta}_l\big)\right]\right| \right\} \leq \epsilon_n, \tag{3.1}$$

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

$$\hat{f}_1(\cdot) := \hat{\theta}_0^T g(\cdot). \tag{3.2}$$

Define the Jacobian matrix

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

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 $l = 1,\ldots,d_1$. 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 $\varepsilon\nu := (\varepsilon\nu^{(1)},\ldots,\varepsilon\nu^{(d_1)})$. 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

$$\hat{u}(x) := \hat{f}_1(x) + \frac{(g(x)^T\hat{\Sigma}_n g(x))^{1/2}c_\alpha}{\sqrt{n}}$$
$$\hat{l}(x) := \hat{f}_1(x) - \frac{(g(x)^T\hat{\Sigma}_n g(x))^{1/2}c_\alpha}{\sqrt{n}},$$

where c_α is a critical value determined by the following standard multiplier bootstrap method introduced in Chernozhukov et al. (2013). 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. Main Results

We now 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 (2.1) and (2.2) we need to choose an appropriate set of approximating functions. Let $\bar{d}_n := \max(d_1, d_2, n, e)$ and C a strictly positive constant independent of n and l. Additionally we set $t_1 := \sup_{x\in I}\|g(x)\|_0 \le d_1$. The following assumptions hold uniformly in $n \ge n_0, P \in \mathcal{P}_n$:

Assumption A. 1.

(i) It holds

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

and for all $\varepsilon > 0$

$$\log N(\varepsilon, g(I), \|\cdot\|_2) \le Ct_1 \log\left(\frac{A_n}{\varepsilon}\right).$$

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

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

Additionally, the approximation errors obey

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

and

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

(iii) We have

$$\sup_{\|\xi\|_2=1}\mathbb{E}\Big[(\xi^TZ)^2\big(b_1(X_1)+b_2(X_{-1})\big)^2\Big] \le C\mathbb{E}\Big[\big(b_1(X_1)+b_2(X_{-1})\big)^2\Big]$$

and

$$\sup_{\|\xi\|_2=1}\mathbb{E}\Big[(\xi^TZ)^2\big(b_3^{(l)}(Z_{-l})\big)^2\Big] \le C\mathbb{E}\Big[\big(b_3^{(l)}(Z_{-l})\big)^2\Big]$$

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

(iv) It holds

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

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. Van der Vaart and Wellner (1996)), 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*

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

(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

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

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 4.2. 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 4.1.

Theorem 1. *Given conditions 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 4.1. [**B-Splines**] 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

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

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

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

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

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

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., Belloni et al. (2018), Belloni et al. (2014a) 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. (2013). It guarantees the validity of multiplier bootstrap in our setting and allows us to construct uniformly valid confidence regions.

Comment 4.2. The sparsity condition in A.2(ii) restrict the number of nonzero regression coefficients $s = s_n$ in the Equations 2.1, 2.2 and 2.3. 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. (2016) and Lu et al. (2020) who restrict the number of relevant additive components in the GAM model 1.1. Our model also includes the approximate sparse setting due to the error terms b_1 and b_2 in 2.1 and 2.2. 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 2.1 (and $\beta_0^T h(X_{-1})$ for $f_{-1}(X_{-1})$ in 2.2) as in Gregory et al. (2016). 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).

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. (2016) and Meier et al. (2009). 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 1. We extend the initial setting in Gregory et al. (2016) 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 Gregory et al. (2016). 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 non-zero. 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. (2016) 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 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. (2016) and Meier et al. (2009).

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` (Chernozhukov et al., 2015a). Further details related to the implementation and parametrization in the simulation study can be found in Appendix C.

Table 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.

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 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, $I = [-2, 2]$.

The first two plots in Figure 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 1 show the empirical coverage as obtained for a sequence of values $x_{0,j}$ with $I = [-x_{0,j}, x_{0,j}]$ 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.

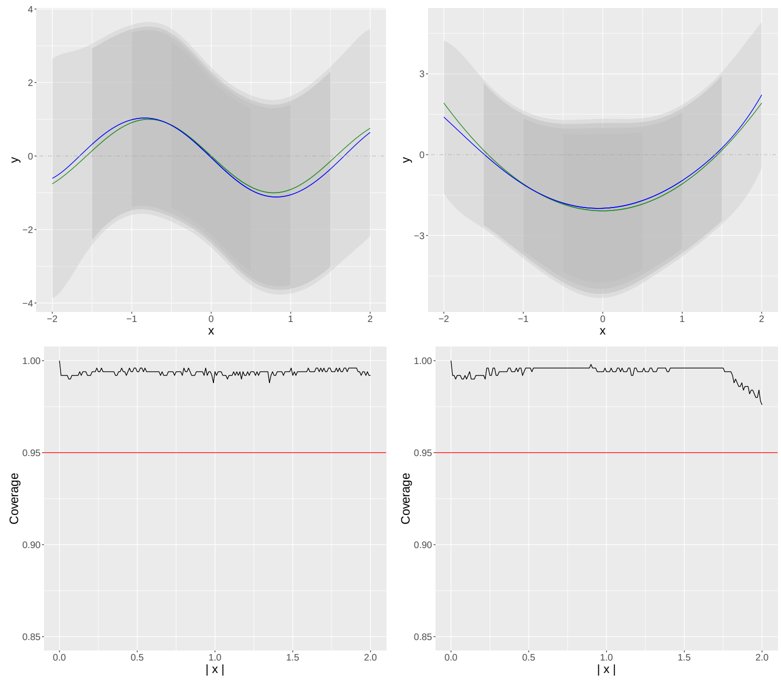

Fig 1: Simulation results for setting with $n = 100$ and $p = 150$. (Top) Gray shaded areas illustrate averaged 95%-confidence bands obtained in $R = 500$ repetitions for functions $f_1(x_1)$ and $f_2(x_2)$. Blue lines correspond to the estimated functions $\hat{f}_j(x_j)$ and green lines to the true 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 $f_2(x_2)$.

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 (1978) and later been reassessed in several studies, e.g., Kong and Xia (2012) and Doksum and Samarov (1995). The data set is available via the R package `mlbench` (Leisch and Dimitriadou, 2010; Newman et al., 1998). The data contain 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 3.

MEDV	median value of owner-occupied homes in USD 1000's
LSTAT	percentage of lower status of the 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 centres
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 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 C. The results illustrated in Figure 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 (2012) and Doksum and Samarov (1995). 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 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 C point at nonlinear effects that are, however, not significant.

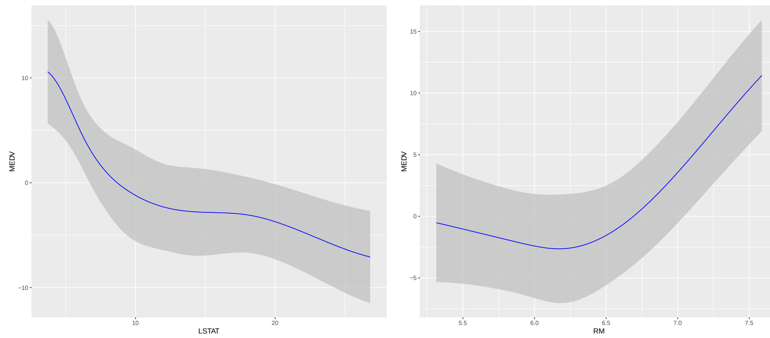

Fig 2: Plots of $\hat{f}_1$(LSTAT) and $\hat{f}_7$(RM) with simultaneous 95%-confidence bands in the Boston housing data application.

7. Proofs

Proof of Theorem 1.
We will prove that the Assumptions A.1 and A.2 imply the Assumptions B.1-B.5 stated in Appendix A and then the claim follows by applying Theorem 2. 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

$$\mathbb{E}\left[(\nu^{(l)})^2\right] \geq \text{Var}(\nu^{(l)}|Z_{-l}) \geq c.$$

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 $l = 1, \dots, d_1$, 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 $l = 1, \dots, d_1$, 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

$$\|\eta\|_e := \max\Big\{\|\eta^{(1)}\|_2, \|\eta^{(2)}\|_2, \|\eta^{(3)}(X)\|_{P,2}, \|\eta^{(4)}(Z_{-l})\|_{P,2}\Big\}.$$

Further, let $\tau_n := \sqrt{\frac{s\log(\bar{d}_n)}{n}}$ and define the corresponding nuisance realization set

$$\begin{aligned}\mathcal{T}_l := \Big\{\eta \in T_l : &\eta^{(3)} \equiv 0, \eta^{(4)} \equiv 0, \|\eta^{(1)}\|_0 \vee \|\eta^{(2)}\|_0 \leq Cs, \\ &\|\eta^{(1)} - \beta_0^{(l)}\|_2 \vee \|\eta^{(2)} - \gamma_0^{(l)}\|_2 \leq C\tau_n, \\ &\|\eta^{(1)} - \beta_0^{(l)}\|_1 \vee \|\eta^{(2)} - \gamma_0^{(l)}\|_1 \leq C\sqrt{s}\tau_n\Big\} \cup \{\eta_{0,l}\}\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}\left[\max_{l=1,\dots,d_1} |\nu^{(l)}|^q\right]^{\frac{1}{q}} &= \|\max_{l=1,\dots,d_1} |\nu^{(l)}|\|_{P,q} \\
&\le q!\|\max_{l=1,\dots,d_1} |\nu^{(l)}|\|_{\Psi_1} \\
&\le q!\log^{\frac{1}{\rho}-1}(2)\|\max_{l=1,\dots,d_1} |\nu^{(l)}|\|_{\Psi_1} \\
&\le Cq!\log^{\frac{1}{\rho}-1}(2)\log^{\frac{1}{\rho}}(1+d_1)\max_{l=1,\dots,d_1} \|\nu^{(l)}|\|_{\Psi_p} \\
&\le C\log^{\frac{1}{\rho}}(d_1)
\end{aligned}$$

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

$$\begin{aligned}
\|\max_{l=1,\dots,d_1} \varepsilon\nu^{(l)}\|_{P,q} &\le \|\varepsilon\|_{P,2q}\|\max_{l=1,\dots,d_1} \nu^{(l)}\|_{P,2q} \\
&\le C\log^{\frac{1}{\rho}}(d_1).
\end{aligned}$$

We can apply Lemma P.2 from Belloni et al. (2018) with $|\mathcal{F}| = d_1$ to obtain

$$\mathcal{S}_n \le C\log^{\frac{1}{2}}(d_1) + C\log^{\frac{1}{2}}(d_1)\left(n^{\frac{2}{q}}\frac{\log^{\frac{2}{\rho}+1}(d_1)}{n}\right)^{1/2} \lesssim \log^{\frac{1}{2}}(d_1),$$

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

$$\begin{aligned}
D_{l,0}[\eta,\eta_{0,l}] &:= \partial_t\big\{\mathbb{E}[\psi_l(W,\theta_{0,l},\eta_{0,l}+t(\eta-\eta_{0,l}))]\big\}\big|_{t=0} \\
&= \mathbb{E}\big[\partial_t\big\{\psi_l(W,\theta_{0,l},\eta_{0,l}+t(\eta-\eta_{0,l}))\big\}\big]\big|_{t=0} \\
&= \mathbb{E}\Bigg[\partial_t\Bigg\{\Big(Y-\theta_{0,l}g_l(X_1)-\big(\eta_{0,l}^{(1)}+t(\eta^{(1)}-\eta_{0,l}^{(1)})\big)^T Z_{-l} \\
&\quad -\big(\eta_{0,l}^{(3)}(X)+t(\eta^{(3)}(X)-\eta_{0,l}^{(3)}(X))\big)\Big) \\
&\quad \Big(g_l(X_1)-\big(\eta_{0,l}^{(2)}+t(\eta^{(2)}-\eta_{0,l}^{(2)})\big)^T Z_{-l}
\end{aligned}$$

$$
\begin{aligned}
&\quad - \left(\eta_{0,l}^{(4)}(Z_{-l}) + t(\eta^{(4)}(Z_{-l}) - \eta_{0,l}^{(4)}(Z_{-l}))\right)\Big\}\Big]\Big|_{t=0} \\
&= \mathbb{E}\left[\varepsilon(\eta_{0,l}^{(2)} - \eta^{(2)})^T Z_{-l}\right] + \mathbb{E}\left[\nu^{(l)}(\eta_{0,l}^{(1)} - \eta^{(1)})^T Z_{-l}\right] \\
&\quad + \mathbb{E}\left[\varepsilon\left(\eta_{0,l}^{(4)}(Z_{-l}) - \eta^{(4)}(Z_{-l})\right)\right] + \mathbb{E}\left[\nu^{(l)}\left(\eta_{0,l}^{(3)}(X) - \eta^{(3)}(X)\right)\right]
\end{aligned}
$$

with

$$
\mathbb{E}\left[\varepsilon(\eta_{0,l}^{(2)} - \eta^{(2)})^T Z_{-l}\right] = \mathbb{E}\left[((\eta_{0,l}^{(2)} - \eta^{(2)})^T Z_{-l}\mathbb{E}[\varepsilon|X]\right] = 0,
$$

$$
\mathbb{E}\left[\nu^{(l)}(\eta_{0,l}^{(1)} - \eta^{(1)})^T Z_{-l}\right] = \mathbb{E}\left[(\eta_{0,l}^{(1)} - \eta^{(1)})^T Z_{-l}\mathbb{E}[\nu^{(l)}|Z_{-l}]\right] = 0,
$$

$$
\mathbb{E}\left[\varepsilon\left(\eta_{0,l}^{(4)}(Z_{-l}) - \eta^{(4)}(Z_{-l})\right)\right] = \mathbb{E}\left[\left(\eta_{0,l}^{(4)}(Z_{-l}) - \eta^{(4)}(Z_{-l})\right)\mathbb{E}[\varepsilon|X]\right] = 0
$$

and

$$
\mathbb{E}\left[\nu^{(l)}\left(\eta_{0,l}^{(3)}(X) - \eta^{(3)}(X)\right)\right] = \mathbb{E}\left[\nu^{(l)}\big(b_1(X_1) + b_2(X_{-1})\big)\right] \le C\delta_n n^{-1/2}
$$

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 $t \in [0,1), l = 1,\dots,d_1, \theta_l \in \Theta_l, \eta_l \in \mathcal{T}_l \setminus \{\eta_{0,l}\}$ we have

$$
\begin{aligned}
&\mathbb{E}\left[(\psi_l(W,\theta_l,\eta_l) - \psi_l(W,\theta_{0,l},\eta_{0,l}))^2\right] \\
&= \mathbb{E}\left[(\psi_l(W,\theta_l,\eta_l) - \psi_l(W,\theta_{0,l},\eta_l) + \psi_l(W,\theta_{0,l},\eta_l) - \psi_l(W,\theta_{0,l},\eta_{0,l}))^2\right] \\
&\le C\bigg(\mathbb{E}\left[(\psi_l(W,\theta_l,\eta_l) - \psi_l(W,\theta_{0,l},\eta_l))^2\right] \\
&\qquad \vee \mathbb{E}\left[(\psi_l(W,\theta_{0,l},\eta_l) - \psi_l(W,\theta_{0,l},\eta_{0,l}))^2\right]\bigg)
\end{aligned}
$$

with

$$
\begin{aligned}
&\mathbb{E}\left[(\psi_l(W,\theta_l,\eta_l) - \psi_l(W,\theta_{0,l},\eta_l))^2\right] \\
&= |\theta_l - \theta_{0,l}|^2\mathbb{E}\left[\left(g_l(X_1)(g_l(X_1) - (\eta_l^{(2)})^T Z_{-l}) - \eta_l^{(4)}(Z_{-l})\right)^2\right] \\
&\le C|\theta_l - \theta_{0,l}|^2\left(\mathbb{E}\left[g_l(X_1)^4\right]\mathbb{E}\left[\left(g_l(X_1) - (\eta_l^{(2)})^T Z_{-l} - \eta_l^{(4)}(Z_{-l})\right)^4\right]\right)^{\frac{1}{2}} \\
&\le C|\theta_l - \theta_{0,l}|^2
\end{aligned}
$$

due to Assumption A.2(ii), (iv) and the definition of $\mathcal{T}_l$. With similar arguments we obtain

$$
\begin{aligned}
&\mathbb{E}\Big[\big(\psi_l(W,\theta_{0,l},\eta_l)-\psi_l(W,\theta_{0,l},\eta_{0,l})\big)^2\Big]\\
&=\mathbb{E}\Bigg[\bigg(\Big(Y-\theta_{0,l}g_l(X_1)-(\eta_l^{(1)})^TZ_{-l}-\eta_l^{(3)}(X)\Big)\Big(g_l(X_1)-(\eta_l^{(2)})^TZ_{-l}-\eta_l^{(4)}(Z_{-l})\Big)\\
&\qquad-\Big(Y-\theta_{0,l}g_l(X_1)-(\eta_{0,l}^{(1)})^TZ_{-l}-\eta_{0,l}^{(3)}(X)\Big)\Big(g_l(X_1)-(\eta_{0,l}^{(2)})^TZ_{-l}-\eta_{0,l}^{(4)}(Z_{-l})\Big)\bigg)^2\Bigg]\\
&=\mathbb{E}\Bigg[\bigg(\Big(Y-\theta_{0,l}g_l(X_1)-(\eta_l^{(1)})^TZ_{-l}-\eta_l^{(3)}(X)\Big)\\
&\qquad\cdot\Big((\eta_{0,l}^{(2)}-\eta_l^{(2)})^TZ_{-l}+\eta_{0,l}^{(4)}(Z_{-l})-\eta_l^{(4)}(Z_{-l})\Big)\\
&\qquad+\Big(g_l(X_1)-(\eta_{0,l}^{(2)})^TZ_{-l}-\eta_{0,l}^{(4)}(Z_{-l})\Big)\\
&\qquad\cdot\Big((\eta_{0,l}^{(1)}-\eta_l^{(1)})^TZ_{-l}+\eta_{0,l}^{(3)}(X)-\eta_l^{(3)}(X)\Big)\bigg)^2\Bigg]\\
&\le C\bigg(\|\eta_{0,l}^{(2)}-\eta_l^{(2)}\|_2\vee\|\eta_{0,l}^{(1)}-\eta_l^{(1)}\|_2\vee\|\eta_{0,l}^{(3)}(X)\|_{P,2}\vee\|\eta_{0,l}^{(4)}(Z_{-l})\|_{P,2}\bigg)^2\\
&=C\|\eta_{0,l}-\eta_l\|_e^2
\end{aligned}
$$

where we used the definition of $\mathcal{T}_l$, A.1(iii) and

$$\sup_{\|\xi\|_2=1}\mathbb{E}[(\xi^TZ)^4]\le C.$$

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

$$
\begin{aligned}
&\Big|\partial_t\mathbb{E}\Big[\psi_l(W,\theta_l,\eta_{0,l}+t(\eta_l-\eta_{0,l}))\Big]\Big|\\
&=\Bigg|\mathbb{E}\Bigg[\partial_t\bigg\{\Big(Y-\theta_{0,l}g_l(X_1)-\big(\eta_{0,l}^{(1)}+t(\eta_l^{(1)}-\eta_{0,l}^{(1)})\big)^TZ_{-l}\\
&\qquad-\big(\eta_{0,l}^{(3)}(X)+t(\eta_l^{(3)}(X)-\eta_{0,l}^{(3)}(X))\big)\Big)\\
&\qquad\cdot\Big(g_l(X_1)-\big(\eta_{0,l}^{(2)}+t(\eta_l^{(2)}-\eta_{0,l}^{(2)})\big)^TZ_{-l}\\
&\qquad-\big(\eta_{0,l}^{(4)}(Z_{-l})+t(\eta_l^{(4)}(Z_{-l})-\eta_{0,l}^{(4)}(Z_{-l}))\big)\Big)\bigg\}\Bigg]\Bigg|\\
&=\Bigg|\mathbb{E}\Bigg[\Big(Y-\theta_{0,l}g_l(X_1)-(\eta_{0,l}^{(1)}+t(\eta_l^{(1)}-\eta_{0,l}^{(1)}))^TZ_{-l}\\
&\qquad-(\eta_{0,l}^{(3)}(X)+t(\eta_l^{(3)}(X)-\eta_{0,l}^{(3)}(X)))\Big)\\
&\qquad\cdot\Big((\eta_{0,l}^{(2)}-\eta_l^{(2)}))^TZ_{-l}+\eta_{0,l}^{(4)}(Z_{-l})-\eta_l^{(4)}(Z_{-l})\Big)
\end{aligned}
$$

$$
\begin{aligned}
&+ \Big(g_l(X_1) - (\eta_{0,l}^{(2)} + t(\eta_l^{(2)} - \eta_{0,l}^{(2)}))^T Z_{-l} \\
&- \big(\eta_{0,l}^{(4)}(Z_{-l}) + t(\eta_l^{(4)}(Z_{-l}) - \eta_{0,l}^{(4)}(Z_{-l}))\big)\Big) \\
&\cdot \Big((\eta_{0,l}^{(1)} - \eta_l^{(1)})^T Z_{-l} + \eta_{0,l}^{(3)}(X) - \eta_l^{(3)}(X)\Big)\Big]\Big| \\
= &|I_{1,1} + I_{1,2} + I_{1,3} + I_{1,4}|,
\end{aligned}
$$

with

$$
\begin{aligned}
I_{1,1} &= \mathbb{E}\Big[\Big(Y - \theta_{0,l} g_l(X_1) - (\eta_{0,l}^{(1)} + t(\eta_l^{(1)} - \eta_{0,l}^{(1)}))^T Z_{-l} \\
&\quad - (\eta_{0,l}^{(3)}(X) + t(\eta_l^{(3)}(X) - \eta_{0,l}^{(3)}(X)))\Big)\Big((\eta_{0,l}^{(2)} - \eta_l^{(2)}))^T Z_{-l}\Big)\Big] \\
&\leq C\|\eta_{0,l}^{(2)} - \eta_l^{(2)}\|_2, \\
I_{1,2} &= \mathbb{E}\Big[\Big(Y - \theta_{0,l} g_l(X_1) - (\eta_{0,l}^{(1)} + t(\eta_l^{(1)} - \eta_{0,l}^{(1)}))^T Z_{-l} \\
&\quad - (\eta_{0,l}^{(3)}(X) + t(\eta_l^{(3)}(X) - \eta_{0,l}^{(3)}(X)))\Big)\Big(\eta_{0,l}^{(4)}(Z_{-l})\Big)\Big] \\
&\leq C\|\eta_{0,l}^{(4)}(X)\|_{P,2},
\end{aligned}
$$

$$
\begin{aligned}
I_{1,3} &= \mathbb{E}\Big[\Big(g_l(X_1) - (\eta_{0,l}^{(2)} + t(\eta_l^{(2)} - \eta_{0,l}^{(2)}))^T Z_{-l} \\
&\quad - \big(\eta_{0,l}^{(4)}(Z_{-l}) + t(\eta_l^{(4)}(Z_{-l}) - \eta_{0,l}^{(4)}(Z_{-l}))\big)\Big)\Big((\eta_{0,l}^{(1)} - \eta_l^{(1)})^T Z_{-l}\Big)\Big] \\
&\leq C\|\eta_{0,l}^{(1)} - \eta_l^{(1)}\|_2, \\
I_{1,4} &= \mathbb{E}\Big[\Big(g_l(X_1) - (\eta_{0,l}^{(2)} + t(\eta_l^{(2)} - \eta_{0,l}^{(2)}))^T Z_{-l} \\
&\quad - \big(\eta_{0,l}^{(4)}(Z_{-l}) + t(\eta_l^{(4)}(Z_{-l}) - \eta_{0,l}^{(4)}(Z_{-l}))\big)\Big)\Big(\eta_{0,l}^{(3)}(X)\Big)\Big] \\
&\leq C\|\eta_{0,l}^{(3)}(X)\|_{P,2}.
\end{aligned}
$$

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

$$
\begin{aligned}
&\partial_t^2 \mathbb{E}\left[\psi_l(W, \theta_{0,l} + t(\theta_l - \theta_{0,l}), \eta_{0,l} + t(\eta_l - \eta_{0,l}))\right] \\
= &\partial_t \mathbb{E}\Big[\Big(Y - \big(\theta_{0,l} + t(\theta_l - \theta_{0,l})\big) g_l(X_1) - \big(\eta_{0,l}^{(1)} + t(\eta_l^{(1)} - \eta_{0,l}^{(1)})\big)^T Z_{-l} \\
&- \big(\eta_{0,l}^{(3)}(X) + t(\eta_l^{(3)}(X) - \eta_{0,l}^{(3)}(X))\big)\Big) \\
&\cdot \Big((\eta_{0,l}^{(2)} - \eta_l^{(2)}))^T Z_{-l} + \eta_{0,l}^{(4)}(Z_{-l})\Big)
\end{aligned}
$$

$$
\begin{aligned}
&+\Big(g_l(X_1)-\big(\eta_{0,l}^{(2)}+t(\eta_l^{(2)}-\eta_{0,l}^{(2)})\big)^T Z_{-l}\\
&-\big(\eta_{0,l}^{(4)}(Z_{-l})+t(\eta_l^{(4)}(Z_{-l})-\eta_{0,l}^{(4)}(Z_{-l}))\big)\Big)\\
&\cdot\Big((\theta_{0,l}-\theta_l))g_l(X_1)+(\eta_{0,l}^{(1)}-\eta_l^{(1)})^T Z_{-l}+\eta_{0,l}^{(3)}(X)\Big)\Big]\\
=2\mathbb{E}\Big[&\Big((\theta_{0,l}-\theta_l)g_l(X_1)+(\eta_{0,l}^{(1)}-\eta_l^{(1)})^T Z_{-l}+\eta_{0,l}^{(3)}(X)\Big)\\
&\cdot\Big((\eta_{0,l}^{(2)}-\eta_l^{(2)}))^T Z_{-l}+\eta_{0,l}^{(4)}(Z_{-l})\Big)\Big]\\
\leq& C\left(|\theta_{0,l}-\theta_l|^2\vee\|\eta_{0,l}-\eta_l\|_e^2\right)
\end{aligned}
$$

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 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 B. 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 one lasso regression of Y on Z. With analogous arguments it holds

$$
\begin{aligned}
&\|\beta_0^{(l)}-\hat{\beta}^{(l)}\|_0\leq\|\hat{\theta}\|_0+\|\hat{\beta}\|_0\leq Cs,\\
&\|\beta_0^{(l)}-\hat{\beta}^{(l)}\|_2\leq\sqrt{\|\theta-\hat{\theta}\|_2^2+\|\beta_0-\hat{\beta}\|_2^2}\leq C\sqrt{\frac{s\log(\bar{d}_n)}{n}},\\
&\|\beta_0^{(l)}-\hat{\beta}^{(l)}\|_1\leq\|\theta-\hat{\theta}\|_1+\|\beta_0-\hat{\beta}\|_1\leq C\sqrt{\frac{s^2\log(\bar{d}_n)}{n}}
\end{aligned}
$$

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)

$$\mathcal{F}_{1,1} := \big\{\psi_l(\cdot, \theta_l, \eta_l) : l = 1, \dots, d_1, \theta_l \in \Theta_l, \eta_l \in \mathcal{T}_l \setminus \{\eta_0^{(l)}\}\big\} \subseteq \mathcal{F}_{1,1}^{(1)}\mathcal{F}_{1,1}^{(2)}$$

with

$$\mathcal{F}_{1,1}^{(1)} := \big\{W \mapsto Y - \theta_l g_l(X_1) - (\eta_l^{(1)})^T Z_{-l} : l = 1, \dots, d_1, \theta_l \in \Theta_l, \eta_l \in \mathcal{T}_l \setminus \{\eta_0^{(l)}\}\big\}$$
$$\mathcal{F}_{1,1}^{(2)} := \big\{W \mapsto g_l(X_1) - (\eta_l^{(2)})^T Z_{-l} : l = 1, \dots, d_1, \theta_l \in \Theta_l, \eta_l \in \mathcal{T}_l \setminus \{\eta_0^{(l)}\}\big\}.$$

Note that the envelope $F_{1,1}^{(1)}$ of $\mathcal{F}_{1,1}^{(1)}$ satisfies

$$\begin{aligned}
\|F_{1,1}^{(1)}\|_{P,2q} \leq \bigg\| &\sup_{l=1,\dots,d_1} \sup_{\theta_l \in \Theta_l, \|\eta_{0,l}^{(1)} - \eta_l^{(l)}\|_1 \leq C\sqrt{s}\tau_n} \Big(|\varepsilon| + |\eta_0^{(3)}(X)| \\
&+ |(\theta_{0,l} - \theta_l) g_l(X_1)| + |(\eta_{0,l}^{(1)} - \eta_l^{(1)})^T Z_{-l}|\Big)\bigg\|_{P,2q} \\
&\lesssim \|\varepsilon\|_{P,2q} + \|\eta_0^{(3)}(X)\|_{P,2q} + \|\sup_{l=1,\dots,d_1} g_l(X_1)\|_{P,2q} \\
&\quad + \sqrt{s}\tau_n \|\sup_{j=1,\dots,d_1+d_2} Z_j\|_{P,2q} \\
&\lesssim C + \log^{\frac{1}{\rho}}(d_1) + \sqrt{s}\tau_n \log^{\frac{1}{\rho}}(d_1 + d_2) \\
&\lesssim \log^{\frac{1}{\rho}}(d_1)
\end{aligned}$$

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 Van der Vaart and Wellner (1996) the set

$$\mathcal{G}_{1,1} := \big\{Z \mapsto \xi^T Z : \xi \in \mathbb{R}^{d_1+d_2+1}, \|\xi\|_0 \leq Cs, \|\xi\|_2 \leq C\big\}$$

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 $\binom{d_1+d_2+1}{Cs}$ respectively

$\binom{d_1+d_2}{Cs}$ VC-subgraph classes, which combined with Theorem 2.6.7 from Van der Vaart and Wellner (1996) 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

$$\mathcal{F}_{1,2} := \left\{\psi_l(\cdot,\theta_l,\eta_{0,l}) : l=1,\dots,d_1, \theta_l \in \Theta_l\right\}$$

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 $l = 1, \dots, d_1$

$$\begin{aligned}
&\mathbb{E}\left[\psi_l(W,\theta_l,\eta_l)^2\right]^{\frac{1}{2}} \\
&= \mathbb{E}\left[\left(Y-\theta_l g_l(X_1)-(\eta^{(1)})^T Z_{-l}-\eta^{(3)}(X)\right)^2\left(g_l(X_1)-(\eta^{(2)})^T Z_{-l}-\eta^{(4)}(Z_{-l})\right)^2\right]^{\frac{1}{2}} \\
&= \mathbb{E}\Big[\left(g_l(X_1)-(\eta^{(2)})^T Z_{-l}-\eta^{(4)}(Z_{-l})\right)^2 \\
&\qquad \cdot \underbrace{\mathbb{E}\left[\left(Y-\theta_l g_l(X_1)-(\eta^{(1)})^T Z_{-l}-\eta^{(3)}(X)\right)^2|X\right]}_{\geq Var(\varepsilon|X)\geq c}\Big]^{\frac{1}{2}} \\
&\geq c
\end{aligned}$$

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

$$n^{1/2}B_{1n}^2B_{2n}^2\tau_n^2 \lesssim n^{1/2}\tau_n^2 = \sqrt{\frac{s^2\log^2(\bar{d}_n)}{n}} \lesssim \delta_n,$$

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

$$\begin{aligned}
\|F_0\|_{P,q} &= \mathbb{E}\left[\sup_{x\in I}\left|(g(x)^T\Sigma_n g(x))^{-1/2}g(x)^T J_0^{-1}\psi(W,\theta_0,\eta_0)\right|^q\right]^{\frac{1}{q}} \\
&\lesssim \mathbb{E}\left[\sup_{x\in I}\left|g(x)^T J_0^{-1}\psi(W,\theta_0,\eta_0)\right|^q\right]^{\frac{1}{q}} \\
&= \mathbb{E}\left[\sup_{x\in I}\left|\sum_{l=1}^{d_1} g_l(x)J_{0,l}^{-1}\psi_l(W,\theta_{0,l},\eta_{0,l})\right|^q\right]^{\frac{1}{q}}
\end{aligned}$$

$$\lesssim \mathbb{E}\left[\sup_{x\in I}\left|\sum_{l=1}^{d_1} g_l(x)\varepsilon\nu^{(l)}\right|^q\right]^{\frac{1}{q}}$$
$$\lesssim t_1\mathbb{E}\left[\sup_{l=1,\dots,d_1}\left|\varepsilon\nu^{(l)}\right|^q\right]^{\frac{1}{q}}$$
$$\lesssim t_1\log^{\frac{1}{\rho}}(d_1).$$

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

$$\begin{aligned}
&\left\|\psi_x(W)-\psi_{\tilde{x}}(W)\right\|_{P,2}\\
&=\Big\|(g(x)^T\Sigma_n g(x))^{-1/2}\sum_{l=1}^{d_1} g_l(x)\mathbb{E}[(\nu^{(l)})^2]^{-1}\psi_l(W,\theta_{0,l},\eta_{0,l})\\
&\quad-(g(\tilde{x})^T\Sigma_n g(\tilde{x}))^{-1/2}\sum_{l=1}^{d_1} g_l(\tilde{x})\mathbb{E}[(\nu^{(l)})^2]^{-1}\psi_l(W,\theta_{0,l},\eta_{0,l})\Big\|_{P,2}\\
&\le |(g(x)^T\Sigma_n g(x))^{-1/2}-(g(\tilde{x})^T\Sigma_n g(\tilde{x}))^{-1/2}|\\
&\quad\cdot\Big\|\sum_{l=1}^{d_1} g_l(x)\mathbb{E}[(\nu^{(l)})^2]^{-1}\psi_l(W,\theta_{0,l},\eta_{0,l})\Big\|_{P,2}\\
&\quad+(g(\tilde{x})^T\Sigma_n g(\tilde{x}))^{-1/2}\Big\|\sum_{l=1}^{d_1}\big(g_l(x)-g_l(\tilde{x})\big)\mathbb{E}[(\nu^{(l)})^2]^{-1}\psi_l(W,\theta_{0,l},\eta_{0,l})\Big\|_{P,2}\\
&=|(g(x)^T\Sigma_n g(x))^{-1/2}-(g(\tilde{x})^T\Sigma_n g(\tilde{x}))^{-1/2}|\Big\|g(x)^T J_0^{-1}\psi(W,\theta_{0,l},\eta_{0,l})\Big\|_{P,2}\\
&\quad+(g(\tilde{x})^T\Sigma_n g(\tilde{x}))^{-1/2}\Big\|\big(g(x)-g(\tilde{x})\big)^T J_0^{-1}\psi(W,\theta_{0,l},\eta_{0,l})\Big\|_{P,2}\\
&\lesssim |(g(x)^T\Sigma_n g(x))^{-1/2}-(g(\tilde{x})^T\Sigma_n g(\tilde{x}))^{-1/2}|\sup_{x\in I}\|g(x)\|_2\\
&\quad+\|g(x)-g(\tilde{x})\|_2
\end{aligned}$$

due to the sub-multiplicativity of the spectral norm and the bounded eigenvalues.

Additionally, it holds

$$\begin{aligned}
&|(g(x)^T\Sigma_n g(x))^{-1/2} - (g(\tilde{x})^T\Sigma_n g(\tilde{x}))^{-1/2}| \\
\lesssim &\left| \left(\frac{g(\tilde{x})^T\Sigma_n g(\tilde{x})}{g(x)^T\Sigma_n g(x)} \right)^{1/2} - 1 \right| \\
\lesssim &|g(\tilde{x})^T\Sigma_n g(\tilde{x}) - g(x)^T\Sigma_n g(x)| \\
= &|(g(x) - g(\tilde{x}))^T\Sigma_n (g(x) + g(\tilde{x}))| \\
\le &|\langle \Sigma_n (g(x) - g(\tilde{x})), (g(x) + g(\tilde{x}))\rangle| \\
\lesssim &\|g(x) - g(\tilde{x})\|_2 \sup_x \|g(x)\|_2
\end{aligned}$$

which implies

$$\left\|\psi_x(W) - \psi_{\tilde{x}}(W)\right\|_{P,2} \lesssim \|g(x) - g(\tilde{x})\|_2 \sup_x \|g(x)\|_2^2.$$

Using the same argument as in Theorem 2.7.11 from Van der Vaart and Wellner (1996) we obtain

$$\begin{aligned}
&\sup_Q \log N(\varepsilon\|\tilde{F}_0\|_{Q,2}, \mathcal{F}_0, \|\cdot\|_{Q,2}) \\
\lesssim &\sup_Q \log N\left(\left(\frac{\varepsilon t_1^3 \log^{\frac{3}{\rho}}(d_1)}{\sup_x \|g(x)\|_2^2} \right) \sup_x \|g(x)\|_2^2, \mathcal{F}_0, \|\cdot\|_{Q,2} \right) \\
\le &\log N\left(\left(\frac{\varepsilon t_1^3 \log^{\frac{3}{\rho}}(d_1)}{\sup_x \|g(x)\|_2^2} \right), g(I), \|\cdot\|_2 \right) \\
\lesssim &t_1 \log\left(\frac{A_n}{\varepsilon}\right).
\end{aligned}$$

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

$$\sup_{l=1,\dots,d_1} |\hat{J}_l - J_{0,l}| = o(1),$$

where $\hat{J}_l = \mathbb{E}_n[-g_l(X_1)(g_l(X_1) - (\hat{\eta}_l^{(2)})^T Z_{-l})]$. It holds

$$\begin{aligned}
|\hat{J}_l - J_{0,l}| \le &|\hat{J}_l - \mathbb{E}[-g_l(X_1)(g_l(X_1) - (\hat{\eta}_l^{(2)})^T Z_{-l})]| \\
&+ |\mathbb{E}[-g_l(X_1)(g_l(X_1) - (\hat{\eta}_l^{(2)})^T Z_{-l})] + J_{0,l}|
\end{aligned}$$

with

$$\begin{aligned}
&|\mathbb{E}[-g_l(X_1)(g_l(X_1) - (\hat{\eta}_l^{(2)})^T Z_{-l})] + J_{0,l}| \\
\le &|\mathbb{E}[g_l(X_1)(\hat{\eta}_l^{(2)} - \eta_{0,l}^{(2)})^T Z_{-l})]| + |\mathbb{E}[g_l(X_1)\eta_{0,l}^{(4)}(Z_{-l})]| \\
\lesssim &\tau_n.
\end{aligned}$$

Let

$$\tilde{\mathcal{G}}_1 := \Big\{ X \mapsto -g_l(X_1)(g_l(X_1) - (\eta_l^{(2)})^T Z_{-l}) : l = 1, \dots, d_1, \|\eta_l^{(2)}\|_0 \le Cs, \\ \|\eta_l^{(2)} - \eta_{0,l}^{(2)}\|_2 \le C\tau_n, \|\eta^{(2)} - \eta_{0,l}^{(2)}\|_1 \le C\sqrt{s}\tau_n \Big\}.$$

The envelope $\tilde{G}_1$ of $\tilde{\mathcal{G}}_1$ satisfies

$$\begin{aligned}
\mathbb{E}[\tilde{G}_1^q]^{\frac{1}{q}} &\le \mathbb{E}\left[\sup_{l=1,\dots,d_1} \sup_{\eta^{(2)}:\|\eta_l^{(2)} - \eta_{0,l}^{(2)}\|_2 \le C\sqrt{s}\tau_n} |g_l(X_1)|^q |(g_l(X_1) - (\eta_l^{(2)})^T Z_{-l})|^q \right]^{\frac{1}{q}} \\
&\le \| \sup_{l=1,\dots,d_1} g_l(X_1) \|_{P,2q} \\
&\quad \cdot \mathbb{E}\left[\sup_{l=1,\dots,d_1} \sup_{\eta^{(2)}:\|\eta_l^{(2)} - \eta_{0,l}^{(2)}\|_2 \le C\sqrt{s}\tau_n} |(g_l(X_1) - (\eta_l^{(2)})^T Z_{-l})|^{2q} \right]^{\frac{1}{2q}} \\
&\lesssim \log^{\frac{1}{\rho}}(d_1) \Bigg(\| \sup_{l=1,\dots,d_1} \nu^{(l)} \|_{P,2q} \vee \| \sup_{l=1,\dots,d_1} b_3^{(l)}(Z_{-l}) \|_{P,2q} \\
&\quad \vee \mathbb{E}\Big[\sup_{l=1,\dots,d_1} \sup_{\eta^{(2)}:\|\eta_l^{(2)} - \eta_{0,l}^{(2)}\|_2 \le C\sqrt{s}\tau_n} (\eta_{0,l}^{(2)} - \eta_l^{(2)})^T Z_{-l})^{2q} \Big]^{\frac{1}{2q}} \Bigg) \\
&\lesssim \log^{\frac{1}{\rho}}(d_1) \left(\log^{\frac{1}{\rho}}(d_1) \vee \sqrt{s}\tau_n \log^{\frac{1}{\rho}}(d_1 + d_2) \right) \\
&\lesssim \log^{\frac{2}{\rho}}(d_1)
\end{aligned}$$

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. (2018) 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

$$\sup_{\|v\|_2 = 1, \|v\|_0 \le t_1} |v^T (\hat{\Sigma}_{\varepsilon\nu} - \Sigma_{\varepsilon\nu}) v| \lesssim u_n \tag{7.1}$$

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

$$\begin{aligned}
&\hat{\Sigma}_{\varepsilon\nu} - \Sigma_{\varepsilon\nu} \\
=& \frac{1}{n}\sum_{i=1}^n \hat{\xi}_i \hat{\xi}_i^T - \mathbb{E}[\xi_i \xi_i^T] \\
=& \frac{1}{n}\sum_{i=1}^n \xi_i \xi_i^T - \mathbb{E}[\xi_i \xi_i^T] \\
&+ \frac{1}{n}\sum_{i=1}^n \xi_i \big(\hat{\xi}_i - \xi_i\big)^T + \frac{1}{n}\sum_{i=1}^n \big(\hat{\xi}_i - \xi_i\big)\xi_i^T + \frac{1}{n}\sum_{i=1}^n \big(\hat{\xi}_i - \xi_i\big)\big(\hat{\xi}_i - \xi_i\big)^T.
\end{aligned}$$

Using the Lemma Q.1 from Belloni et al. (2018) we can bound the first part. Due to the tail conditions on ε and ν we obtain

$$\begin{aligned}
\left(\mathbb{E}\left[\max_{1\le i\le n} \|\varepsilon_i \nu_i\|_\infty^2\right]\right)^{1/2} &\le \left(\mathbb{E}\left[\max_{1\le i\le n} \|\varepsilon_i\|^4\right]\mathbb{E}\left[\max_{1\le i\le n} \|\nu_i\|_\infty^4\right]\right)^{1/4} \\
&\lesssim n^{\frac{2}{q}} \log^{\frac{1}{\rho}}(d_1)
\end{aligned}$$

for an arbitrary but fixed $q \ge 4$. Then Lemma Q.1 implies

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\|v\|_2=1,\|v\|_0\le t_1} \left| v^T\Big(\frac{1}{n}\sum_{i=1}^n \xi_i\xi_i^T - \mathbb{E}[\xi_i\xi_i^T]\Big)v\right|\right] \\
=& \mathbb{E}\left[\sup_{\|v\|_2=1,\|v\|_0\le t_1} \left| \mathbb{E}_n\Big[\big(v^T\xi_i\big)^2 - \mathbb{E}\big[\big(v^T\xi_i\big)^2\big]\Big]\right|\right] \\
\lesssim& \tilde{\delta}_n^2 + \tilde{\delta}_n
\end{aligned}$$

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

$$\sup_{\|v\|_2=1,\|v\|_0\le t_1} \left| v^T\Big(\frac{1}{n}\sum_{i=1}^n \xi_i\xi_i^T - \mathbb{E}[\xi_i\xi_i^T]\Big)v\right| \lesssim u_n$$

with probability $1 - o(1)$. Note that by applying the results on covariance estimation from Chen et al. (2012) instead would lead to comparable growth rates. With probability $1 - o(1)$ it holds

$$\sup_{l=1,\ldots,d_1} |\hat{\theta}_l - \theta_{0,l}| \lesssim \tau_n$$

due to Appendix A from Belloni et al. (2018). Define

$$\begin{aligned}\tilde{\mathcal{G}}_2^2 := \big\{(\psi_l(\cdot,\theta_l,\eta_l) - \psi_l(\cdot,\theta_{0,l},\eta_{0,l}))^2 : & l = 1,\ldots,d_1, |\theta_l - \theta_{0,l}| \le C\tau_n \\ & \eta_l \in \mathcal{T}_l \setminus \{\eta_{0,l}\}\big\},\end{aligned}$$

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

$$\tilde{G}_2^2 = \sup_{l=1,\ldots,d_1} \sup_{\theta_l:|\theta_l-\theta_{0,l}|\le C\tau_n, \eta_l\in\mathcal{T}_l} \big(\psi_l(W,\theta_l,\eta_l) - \psi_l(W,\theta_{0,l},\eta_{0,l})\big)^2$$

and

$$\begin{aligned}
&\|\tilde{G}_2^2\|_{P,q} \\
\lesssim & \Big\| \sup_{l,\theta_l,\eta_l^{(2)},\eta_l^{(4)}} \Big((\theta_{0,l}-\theta_l) g_l(X_1)\big(g_l(X_1) - (\eta_l^{(2)})^T Z_{-l} - \eta_l^{(4)}(Z_{-l})\big)\Big)^2 \Big\|_{P,q} \\
& + \Big\| \sup_{l,\eta_l} \Big(\big(Y - \theta_{0,l} g_l(X_1) - (\eta_l^{(1)})^T Z_{-l} - \eta_l^{(3)}(X)\big) \\
& \quad \big((\eta_{0,l}^{(2)} - \eta_l^{(2)})^T Z_{-l} + \eta_{0,l}^{(4)}(Z_{-l}) - \eta_l^{(4)}(Z_{-l})\big)\Big)^2 \Big\|_{P,q} \\
& + \Big\| \sup_{l,\eta_l^{(1)},\eta_l^{(3)}} \Big(\big(g_l(X_1) - (\eta_{0,l}^{(2)})^T Z_{-l} - \eta_{0,l}^{(4)}(Z_{-l})\big) \\
& \quad \big((\eta_{0,l}^{(1)} - \eta_l^{(1)})^T Z_{-l} + \eta_{0,l}^{(3)}(X) - \eta_l^{(3)}(X)\big)\Big)^2 \Big\|_{P,q} \\
=: & T_1 + T_2 + T_3.
\end{aligned}$$

It holds

$$\begin{aligned}
T_1 &\lesssim \tau_n^2 \Big\| \sup_{l,\eta_l^{(2)},\eta_l^{(4)}} \Big(g_l(X_1)\big(g_l(X_1) - (\eta_l^{(2)})^T Z_{-l} - \eta_l^{(4)}(Z_{-l})\big)\Big)^2 \Big\|_{P,q} \\
&\le \tau_n^2 \|\sup_l (g_l(X_1))^2\|_{P,2q} \Big\| \sup_{l,\eta_l^{(2)},\eta_l^{(4)}} \Big(g_l(X_1) - (\eta_l^{(2)})^T Z_{-l} - \eta_l^{(4)}(Z_{-l})\Big)^2 \Big\|_{P,2q} \\
&\lesssim \tau_n^2 \log^{\frac{4}{\rho}}(d_1),
\end{aligned}$$

$$
\begin{aligned}
T_2 &\leq \Big\| \sup_{l,\eta_l^{(1)},\eta_l^{(3)}} \Big(Y - \theta_{0,l} g_l(X_1) - (\eta_l^{(1)})^T Z_{-l} - \eta_l^{(3)}(X)\Big)^2 \Big\|_{P,2q} \\
&\quad \Big\| \sup_{l,\eta_l^{(2)},\eta_l^{(4)}} \Big((\eta_{0,l}^{(2)} - \eta_l^{(2)})^T Z_{-l} + \eta_{0,l}^{(4)}(Z_{-l}) - \eta_l^{(4)}(Z_{-l})\Big)^2 \Big\|_{P,2q} \\
&\lesssim s\tau_n^2 \Big\| \sup_l \|Z_{-l}\|_\infty^2 \Big\|_{P,2q} + \log^{\frac{2}{\rho}}(d_1) \\
&\lesssim s\tau_n^2 \log^{\frac{2}{\rho}}(d_1 + d_2) + \log^{\frac{2}{\rho}}(d_1)
\end{aligned}
$$

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

$$
\begin{aligned}
\tilde{\sigma} &:= \sup_{f \in \tilde{\mathcal{G}}_2^2} \mathbb{E}\left[f(X)^2\right]^{\frac{1}{2}} \\
&= \sup_{l=1,\dots,d_1} \sup_{\theta_l : |\theta_l - \theta_{0,l}| \leq C\tau_n, \eta_l \in \mathcal{T}_l} \mathbb{E}\left[\left(\psi_l(W, \theta_l, \eta_l) - \psi_l(W, \theta_{0,l}, \eta_{0,l})\right)^4 \right]^{\frac{1}{2}} \\
&\lesssim \frac{s^2 \log(d_1 \vee d_2)}{n}.
\end{aligned}
$$

Again, we can apply Lemma P.2 from Belloni et al. (2018) 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

$$
\begin{aligned}
\sup_{f \in \tilde{\mathcal{G}}_2^2} \mathbb{E}[f(X)] &\leq C \left(|\theta_l - \theta_{0,l}|^2 \vee \|\eta_{0,l} - \eta_l\|_e^2 \right) \\
&\lesssim \tau_n^2.
\end{aligned}
$$

Combined 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

$$\begin{aligned}
&\sup_{\|v\|_2=1,\|v\|_0\le t_1} |v^T \frac{1}{n}\sum_{i=1}^n \xi_i\big(\hat{\xi}_i-\xi_i\big)^T v| \\
=& \sup_{\|v\|_2=1,\|v\|_0\le t_1} |\mathbb{E}_n\left[v^T\xi_i\big(\hat{\xi}_i-\xi_i\big)^T v\right]| \\
\le& \sup_{\|v\|_2=1,\|v\|_0\le t_1} \left|\left(\mathbb{E}_n\left[\big(v^T\xi_i\big)^2\right]\mathbb{E}_n\left[\left(v^T\big(\hat{\xi}_i-\xi_i\big)\right)^2\right]\right)^{\frac{1}{2}}\right| \\
\lesssim& \sup_{\|v\|_2=1,\|v\|_0\le t_1} \left|\left(\mathbb{E}_n\left[\left(v^T\big(\hat{\xi}_i-\xi_i\big)\right)^2\right]\right)^{\frac{1}{2}}\right| \\
=& \sup_{\|v\|_2=1,\|v\|_0\le t_1} \left(\sum_{k=1}^{d_1}\sum_{l=1}^{d_1} v_k v_l \mathbb{E}_n\left[(\hat{\varepsilon}_i\hat{\nu}_i^{(k)}-\varepsilon_i\nu_i^{(k)})(\hat{\varepsilon}_i\hat{\nu}_i^{(l)}-\varepsilon_i\nu_i^{(l)})\right]\right)^{\frac{1}{2}} \\
\lesssim& \, t_1 \sup_{l=1,\dots,d_1} \mathbb{E}_n\left[(\hat{\varepsilon}_i\hat{\nu}_i^{(l)}-\varepsilon_i\nu_i^{(l)})^2\right]^{\frac{1}{2}} \\
\lesssim& \, t_1\left(n^{\frac{1}{q}}\log^{\frac{2}{\rho}}(d_1)\tau_n^2 \vee s\tau_n^3\right)^{\frac{1}{2}}
\end{aligned}$$

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 implies (7.1) 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

$$M_w \in \mathbb{R}^{d_1\times d_1} : (M_w)_{k,l} = \begin{cases} 1 \text{ if } w_k \neq 0 \wedge w_l \neq 0 \\ 0 \text{ else,} \end{cases}$$

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 (7.1) 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). This can be used to show for $v \in \mathbb{R}^{d_1}$

$$\sup_{\|v\|_2=1, \|v\|_0 \le t_1} |v^T \big(\hat{\Sigma}_n - \Sigma_n\big) v| \lesssim u_n \tag{7.2}$$

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

$$\begin{aligned}&\sup_{\|v\|_2=1, \|v\|_0 \le t_1} |v^T \hat{J}^{-1}\hat{\Sigma}_{\varepsilon\nu}\big(\hat{J}^{-1} - J_0^{-1}\big)^T v| \\ =& \sup_{\|v\|_2=1, \|v\|_0 \le t_1} |v^T \hat{J}^{-1}\left(M_v \odot \hat{\Sigma}_{\varepsilon\nu}\right)\big(\hat{J}^{-1} - J_0^{-1}\big)^T v| \\ \le& \left\|\hat{J}^{-1}\right\|_2 \sup_{\|w\|_0 \le t_1} \left\|\left(M_w \odot \hat{\Sigma}_{\varepsilon\nu}\right)\right\|_2 \left\|\big(\hat{J}^{-1} - J_0^{-1}\big)^T\right\|_2 \\ \lesssim& \, n^{\frac{1}{q}} \frac{s \log^{\frac{2}{\rho}}(d_1)\log(\bar{d}_n)}{n} + \tau_n\end{aligned}$$

due to the sub-multiplicative spectral norm and an analogous argument holds for the second term. The third term can be bounded by

$$\sup_{\|v\|_2=1, \|v\|_0 \le t_1} |v^T J_0^{-1}\big(\hat{\Sigma}_{\varepsilon\nu} - \Sigma_{\varepsilon\nu}\big)\big(J_0^{-1}\big)^T v| \lesssim u_n.$$

This implies (7.2). We finally obtain

$$\begin{aligned}\sup_{x \in I}\left|\frac{(g(x)^T\hat{\Sigma}_n g(x))^{1/2}}{(g(x)^T\Sigma_n g(x))^{1/2}} - 1\right| &\lesssim \sup_{x\in I}\left|g(x)^T\big(\hat{\Sigma}_n - \Sigma_n\big)g(x)\right| \\ &\le \sup_{x\in I}\|g(x)\|_2^2 \sup_{\|v\|_2=1, \|v\|_0 \le t_1} |v^T\big(\hat{\Sigma}_n - \Sigma_n\big)v| \\ &\lesssim \sup_{x\in I}\|g(x)\|_2^2 u_n\end{aligned}$$

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

$$\sigma_x := (g(x)^T\Sigma_n g(x))^{1/2},$$
$$\hat\sigma_x := (g(x)^T\hat\Sigma_n g(x))^{1/2}$$

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

$$\begin{aligned}
&\|\psi_x(W)-\hat\psi_x(W)-(\psi_{\tilde x}(W)-\hat\psi_{\tilde x}(W))\|_{\mathbb{P}_n,2}\\
=&\Big\|\sigma_x^{-1}g(x)^TJ_0^{-1}\psi(W,\theta_0,\eta_0)-\sigma_{\tilde x}^{-1}g(\tilde x)^TJ_0^{-1}\psi(W,\theta_0,\eta_0)\\
&-\big(\hat\sigma_x^{-1}g(x)^T\hat J^{-1}\psi(W,\hat\theta,\hat\eta)-\hat\sigma_{\tilde x}^{-1}g(\tilde x)^T\hat J^{-1}\psi(W,\hat\theta,\hat\eta)\big)\Big\|_{\mathbb{P}_n,2}\\
=&\Big\|\sum_{l=1}^{d_1}(\sigma_x^{-1}g_l(x)-\sigma_{\tilde x}^{-1}g_l(\tilde x))J_{0,l}^{-1}\psi_l(W,\theta_{0,l},\eta_{0,l})\\
&-\sum_{l=1}^{d_1}(\hat\sigma_x^{-1}g_l(x)-\hat\sigma_{\tilde x}^{-1}g_l(\tilde x))\hat J_l^{-1}\psi_l(W,\hat\theta_l,\hat\eta_l)\Big\|_{\mathbb{P}_n,2}\\
\le&\Big\|\sum_{l=1}^{d_1}(\sigma_x^{-1}g_l(x)-\sigma_{\tilde x}^{-1}g_l(\tilde x))\Big(J_{0,l}^{-1}-\hat J_l^{-1}\Big)\psi_l(W,\theta_{0,l},\eta_{0,l})\Big\|_{\mathbb{P}_n,2}\\
&+\Big\|\sum_{l=1}^{d_1}(\sigma_x^{-1}g_l(x)-\sigma_{\tilde x}^{-1}g_l(\tilde x))\hat J_l^{-1}\Big(\psi_l(W,\theta_{0,l},\eta_{0,l})-\psi_l(W,\hat\theta_l,\hat\eta_l)\Big)\Big\|_{\mathbb{P}_n,2}\\
&+\Big\|\sum_{l=1}^{d_1}\Big((\sigma_x^{-1}g_l(x)-\sigma_{\tilde x}^{-1}g_l(\tilde x))-(\hat\sigma_x^{-1}g_l(x)-\hat\sigma_{\tilde x}^{-1}g_l(\tilde x))\Big)\hat J_l^{-1}\psi_l(W,\hat\theta_l,\hat\eta_l)\Big\|_{\mathbb{P}_n,2}\\
=:&I_{4,1}+I_{4,2}+I_{4,3}.
\end{aligned}$$

We obtain

$$\begin{aligned}
I_{4,1} &= \Big\| \sum_{l=1}^{d_1} (\sigma_x^{-1} g_l(x) - \sigma_{\tilde{x}}^{-1} g_l(\tilde{x})) \Big(J_{0,l}^{-1} - \hat{J}_l^{-1} \Big) \psi_l(W, \theta_{0,l}, \eta_{0,l}) \Big\|_{\mathbb{P}_n,2} \\
&\le \sigma_x^{-1} \Big\| (g(x) - g(\tilde{x}))^T \Big(J_0^{-1} - \hat{J}^{-1} \Big) \psi(W, \theta_0, \eta_0) \Big\|_{\mathbb{P}_n,2} \\
&\quad + |\sigma_x^{-1} - \sigma_{\tilde{x}}^{-1}| \Big\| g(\tilde{x})^T \Big(J_0^{-1} - \hat{J}^{-1} \Big) \psi(W, \theta_0, \eta_0) \Big\|_{\mathbb{P}_n,2} \\
&\lesssim \|g(x) - g(\tilde{x})\|_2 \sup_{\|v\|_2=1, \|v\|_0 \le 2t_1} \Big\| v^T \Big(J_0^{-1} - \hat{J}^{-1} \Big) \psi(W, \theta_0, \eta_0) \Big\|_{\mathbb{P}_n,2} \\
&\quad + \|g(x) - g(\tilde{x})\|_2 \sup_{x \in I} \|g(x)\|_2^2 \sup_{\|v\|_2=1, \|v\|_0 \le t_1} \Big\| v^T \Big(J_0^{-1} - \hat{J}^{-1} \Big) \psi(W, \theta_0, \eta_0) \Big\|_{\mathbb{P}_n,2} \\
&\lesssim \|g(x) - g(\tilde{x})\|_2 \sup_{x \in I} \|g(x)\|_2^2 u_n,
\end{aligned}$$

where we used that

$$\begin{aligned}
& \sup_{\|v\|_2=1, \|v\|_0 \le t_1} \Big\| v^T \Big(J_0^{-1} - \hat{J}^{-1} \Big) \psi(W, \theta_0, \eta_0) \Big\|_{\mathbb{P}_n,2}^2 \\
= & \sup_{\|v\|_2=1, \|v\|_0 \le t_1} \Big| v^T \Big(J_0^{-1} - \hat{J}^{-1} \Big) \frac{1}{n} \sum_{i=1}^n \xi_i \xi_i^T \Big(J_0^{-1} - \hat{J}^{-1} \Big)^T v \Big| \\
\le & \Big\| J_0^{-1} - \hat{J}^{-1} \Big\|_2^2 \sup_{\|v\|_0 \le t_1} \Big\| M_v \odot \left(\frac{1}{n} \sum_{i=1}^n \xi_i \xi_i^T \right) \Big\|_2^2 \\
\lesssim & \; u_n^2.
\end{aligned}$$

Analogously, we obtain

$$\begin{aligned}
I_{4,2} &= \Big\| \sum_{l=1}^{d_1} (\sigma_x^{-1} g_l(x) - \sigma_{\tilde{x}}^{-1} g_l(\tilde{x})) \hat{J}_l^{-1} \Big(\psi_l(W, \theta_{0,l}, \eta_{0,l}) - \psi_l(W, \hat{\theta}_l, \hat{\eta}_l) \Big) \Big\|_{\mathbb{P}_n,2} \\
&\le \sigma_x^{-1} \Big\| (g(x) - g(\tilde{x}))^T \hat{J}^{-1} \Big(\psi(W, \theta_0, \eta_0) - \psi(W, \hat{\theta}, \hat{\eta}) \Big) \Big\|_{\mathbb{P}_n,2} \\
&\quad + |\sigma_x^{-1} - \sigma_{\tilde{x}}^{-1}| \Big\| g(\tilde{x})^T \hat{J}^{-1} \Big(\psi(W, \theta_0, \eta_0) - \psi(W, \hat{\theta}, \hat{\eta}) \Big) \Big\|_{\mathbb{P}_n,2} \\
&\lesssim \|g(x) - g(\tilde{x})\|_2 \sup_{\|v\|_2=1, \|v\|_0 \le 2t_1} \Big\| v^T \hat{J}^{-1} \Big(\psi(W, \theta_0, \eta_0) - \psi(W, \hat{\theta}, \hat{\eta}) \Big) \Big\|_{\mathbb{P}_n,2} \\
&\quad + \|g(x) - g(\tilde{x})\|_2 \sup_{x \in I} \|g(x)\|_2^2 \sup_{\|v\|_2=1, \|v\|_0 \le t_1} \Big\| v^T \hat{J}^{-1} \Big(\psi(W, \theta_0, \eta_0) - \psi(W, \hat{\theta}, \hat{\eta}) \Big) \Big\|_{\mathbb{P}_n,2} \\
&\lesssim \|g(x) - g(\tilde{x})\|_2 \sup_{x \in I} \|g(x)\|_2^2 u_n.
\end{aligned}$$

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 (7.2). 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

$$
\begin{aligned}
I_{4,3} &\lesssim \epsilon_n \|g(x) - g(\tilde{x})\|_2 \sup_{\|v\|_2=1, \|v\|_0 \le 2t_1} \left\| v^T \hat{J}^{-1} \psi(W, \hat{\theta}, \hat{\eta}) \right\|_{\mathbb{P}_n,2} \\
&\quad + (\epsilon_n \vee u_n) \|g(x) - g(\tilde{x})\|_2 \sup_{x \in I} \|g(x)\|_2^2 \sup_{\|v\|_2=1, \|v\|_0 \le t_1} \left\| v^T \hat{J}^{-1} \psi(W, \hat{\theta}, \hat{\eta}) \right\|_{\mathbb{P}_n,2} \\
&\lesssim \|g(x) - g(\tilde{x})\|_2 \epsilon_n \sup_{x \in I} \|g(x)\|_2^2.
\end{aligned}
$$

Combining the steps above we obtain

$$
\|\psi_x(W) - \hat{\psi}_x(W) - (\psi_{\tilde{x}}(W) - \hat{\psi}_{\tilde{x}}(W))\|_{\mathbb{P}_n,2} \le \|g(x) - g(\tilde{x})\|_2 \|\hat{F}_0\|_{\mathbb{P}_n,2}
$$

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 Theorem 2.7.11 from Van der Vaart and Wellner (1996) we obtain with probability $1 - o(1)$

$$
\begin{aligned}
\log N(\varepsilon, \hat{\mathcal{F}}_0, \|\cdot\|_{\mathbb{P}_n,2}) &\le \log N(\varepsilon \|\hat{F}_0\|_{\mathbb{P}_n,2}, \hat{\mathcal{F}}_0, \|\cdot\|_{\mathbb{P}_n,2}) \\
&\le \log N(\varepsilon, g(I), \|\cdot\|_2) \\
&\le \bar{\varrho}_n \log\left(\frac{\bar{A}_n}{\varepsilon}\right)
\end{aligned}
$$

with $\bar{\varrho}_n = t_1$ and $\bar{A}_n \lesssim A_n$. Additionally, it holds

$$
\begin{aligned}
&\|\psi_x(W) - \hat{\psi}_x(W)\|_{\mathbb{P}_n,2} \\
&= \left\| \sigma_x^{-1} g(x)^T J_0^{-1} \psi(W, \theta_0, \eta_0) - \hat{\sigma}_x^{-1} g(x)^T \hat{J}^{-1} \psi(W, \hat{\theta}, \hat{\eta}) \right\|_{\mathbb{P}_n,2} \\
&\le \sigma_x^{-1} \left\| g(x)^T \left(J_0^{-1} - \hat{J}^{-1} \right) \psi(W, \theta_0, \eta_0) \right\|_{\mathbb{P}_n,2} \\
&\quad + \sigma_x^{-1} \left\| g(x)^T \hat{J}^{-1} \left(\psi(W, \theta_0, \eta_0) - \psi(W, \hat{\theta}, \hat{\eta}) \right) \right\|_{\mathbb{P}_n,2} \\
&\quad + |\sigma_x^{-1} - \hat{\sigma}_x^{-1}| \left\| g(x)^T \hat{J}^{-1} \psi(W, \hat{\theta}, \hat{\eta}) \right\|_{\mathbb{P}_n,2} \\
&\lesssim \sup_{x \in I} \|g(x)\|_2 (u_n \vee \epsilon_n) \\
&\lesssim \sup_{x \in I} \|g(x)\|_2 \epsilon_n
\end{aligned}
$$

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,

$$n^{-\frac{1}{7}} L_n^{\frac{2}{7}} \varrho_n \log(A_n) = \frac{t_1^{\frac{13}{7}} \log^{\frac{6}{7\rho}}(d_1) \log(A_n)}{n^{\frac{1}{7}}} = o(1)$$

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 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$(v)(b)$.

■

Appendix A Uniformly valid confidence bands

As in Belloni et al. (2018), we consider the problem of estimating the set of parameters $\theta_{0,l}$ for $l = 1, \dots, d_1$ in the moment condition model,

$$\mathbb{E}[\psi_l(W, \theta_{0,l}, \eta_{0,l})] = 0, \qquad l = 1, \dots, d_1, \tag{A.1}$$

where W is a random variable, ψ_l a known score function, $\theta_{0,l} \in \Theta_l$ a scalar of interest, and $\eta_{0,l} \in T_l$ is a high-dimensional nuisance parameter, where 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. (2018) 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 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

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

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 $\hat{\eta}_l = \left(\hat{\eta}_l^{(1)}, \hat{\eta}_l^{(2)}\right)$ 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

$$\sup_{l=1,\dots,d_1} \left\{ \left| \mathbb{E}_n \left[\psi_l(W, \hat{\theta}_l, \hat{\eta}_l) \right] \right| - \inf_{\theta \in \Theta_l} \left| \mathbb{E}_n \left[\psi_l(W, \theta, \hat{\eta}_l) \right] \right| \right\} \leq \epsilon_n, \tag{A.3}$$

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 converging to zero. 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

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

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

$$\mathcal{S}_n := \mathbb{E}\left[\sup_{l=1,\dots,d_1} \left| \sqrt{n} \mathbb{E}_n \left[\psi_l(W, \theta_{0,l}, \eta_{0,l}) \right] \right| \right]$$

and

$$t_1 := \sup_{x \in I} \|g(x)\|_0.$$

The definition of t_1 is helpful if the functions g_l, $l = 1, \dots, d_1$ are local in the sense that for any point x in I there are at most $t_1 \ll d_1$ non-zero functions. 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. (2018), 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 (A.1), 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 (A.2) 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 $\eta_l \in \mathcal{T}_l$*
 (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 about the ideas and intuitions of these and the following assumptions see Belloni et al. (2018).
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 $\eta_l \in \mathcal{T}_l, \|\eta_l - \eta_{0,l}\|_e \leq \tau_n$.*
(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. (2018) 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* $P \in \mathcal{P}_n$,

$$P\left(\sup_{x\in I}\left|\frac{(g(x)^T \hat{\Sigma}_n g(x))^{1/2}}{(g(x)^T \Sigma_n g(x))^{1/2}} - 1\right| > \varepsilon_n\right) \leq \Delta_n,$$

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. (2013) we employ the Gaussian multiplier bootstrap method to estimate the needed 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 2. *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 Assumptions B.1 - B.5 it holds*

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

uniformly over $P \in \mathcal{P}_n$*.*

Proof. Since Theorem 2.1 in Belloni et al. (2018) is not directly applicable to our problem we have to modify the proof to obtain a uniform Bahadur representation. We want to prove that

$$\sup_{x\in I}\left|\sqrt{n}(g(x)^T\Sigma_n g(x))^{-1/2}g(x)^T\left(\hat{\theta}-\theta_0\right)\right| = \sup_{x\in I}\left|\mathbb{G}_n(\psi_x)\right| + O_P(t_1\delta_n). \quad \text{(A.4)}$$

Assumptions B.2 and B.3 contain Assumptions 2.1 and 2.2 from Belloni et al. (2018) 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

$$\begin{aligned}
&\sup_{x\in I}\left|\sqrt{n}\mathbb{E}_n\left[g(x)^TJ_0^{-1}\psi(W,\theta_0,\eta_0)\right] + \sqrt{n}g(x)^T\left(\hat{\theta}-\theta_0\right)\right| \\
&= \sup_{x\in I}\left|\sum_{j=1}^{d_1}g_l(x)\left(J_{0,l}^{-1}\sqrt{n}\mathbb{E}_n\left[\psi_l(W,\theta_{0,l},\eta_{0,l})\right] + \sqrt{n}(\hat{\theta}_l-\theta_{0,l})\right)\right| \\
&\le t_1\underbrace{\sup_{x\in I}\sup_{l=1,\dots,d_1}|g_l(x)|}_{\le C}\sup_{l=1,\dots,d_1}\left|J_{0,l}^{-1}\sqrt{n}\mathbb{E}_n\left[\psi_l(W,\theta_{0,l},\eta_{0,l})\right] + \sqrt{n}\left(\hat{\theta}_l-\theta_{0,l}\right)\right| \\
&= O_p(t_1\delta_n).
\end{aligned}$$

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 (A.4). Due to Assumption B.5 it holds

$$P\left(\sup_{x\in I}\left|\frac{(g(x)^T\hat{\Sigma}_n g(x))^{1/2}}{(g(x)^T\Sigma_n g(x))^{1/2}}-1\right|>\varepsilon_n\right)\le\Delta_n,$$

with $\Delta_n = o(1)$, which is an analogous version of the Assumption 2.4 from Belloni et al. (2018). Therefore, given the Assumptions B.2 - B.5, the proofs of Corollary 2.1 and 2.2 from Belloni et al. (2018) can be applied implying the stated theorem. ■

Appendix B Uniform nuisance function estimation

To establish uniform estimation properties of the nuisance function we rely on uniform estimation results from Klaassen et al. (2018). 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 $r=1,\ldots,d$.
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 lasso estimators have sufficiently fast uniform estimation rates if the vector β_r is sparse for all $r=1,\ldots,d$. 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

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

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

B.1 Uniform lasso estimation

Define the weighted lasso estimator

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

with the penalty level

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

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

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

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

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

Assumption C. 1.
(i) There exists $1 \leq \rho \leq 2$ such that

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

(ii) For all $r = 1, \dots, d_n$, it holds

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

and

$$\min_{j=1,\dots,p} \mathbb{E}[\epsilon_r^2 X_{r,j}^2] \geq c > 0.$$

(iii) The coefficients obey

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

(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

$$\max_{r=1,\dots,d} \|a_r(X_r)\|_{P,2} \leq C\sqrt{\frac{s\log(a_n)}{n}}$$

and

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

with probability $1 - o(1)$*.*

Theorem 3. *Under condition C.1 the lasso estimator* $\hat{\beta}_r$ *obeys uniformly over all* $P \in \mathcal{P}_n$ *with probability* $1 - o(1)$

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

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

with

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

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

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

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

Proof of Theorem 3.

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 there exists sequence $(b_n)_{n\geq 1}$ with $|a_n| \leq b_n$, b_n is independent of $P \in \mathcal{P}_n$ for all n and $b_n \to 0$. Finally, the notation $a_n \lesssim_P b_n$ means that for any $\epsilon > 0$, there exists C such that uniformly over all n we have $P_P(a_n > Cb_n) \leq \epsilon$.

Due to Assumption C.1(i) we can bound the q-th moments of the maxima of the regressors uniformly by

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

where C does depend on q and ρ but not on n. For the norm inequalities we refer to Van der Vaart and Wellner (1996). Now, we essentially modify the proof from Theorem 4.2 from Belloni et al. (2018) 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

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

First, we verify the Condition WL from Belloni et al. (2018). Since $N_n = d$ we have $N(\varepsilon,\mathcal{U},d_\mathcal{U}) \le N_n$ 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

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

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. (2018). Let $\mathcal{W} = (\mathcal{Y}, \mathcal{X})$ with $Y = (Y_1, \dots, Y_d) \in \mathcal{Y}$ and $X = (X_1, \dots, X_d) \in \mathcal{X}$. Define

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

with

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

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. (2018) 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. (2018). 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 proof of $K.1$ from Belloni et al. (2018). 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 C.1(v). Further, we have

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

where

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

and

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

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 have with probability $1 - o(1)$

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

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

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

and additionally

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

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

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

where the last inequality holds due to 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

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

with

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

By Maximal Inequality, we obtain with probability $1 - o(1)$ for a sufficiently large $q' > 0$

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

since

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

We conclude

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

uniformly over r. Therefore, Assumption $M.1(b)$ 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

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

due to Lemma $M.4$ in Belloni et al. (2018). 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. (2018) 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

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

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

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

Let the restricted eigenvalues be definied as

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

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

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

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

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

and the uniformly bounded penalty loading from above and away from zero, we obtain

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

by Lemma $M.1$ from Belloni et al. (2018). 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

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

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

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

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 obtain

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

and

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

with probability $1 - o(1)$ due to Lemma $M.1$ in Belloni et al. (2018). Uniformly over all $r = 1, \dots, d$, it holds

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

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

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

are uniformly bounded from above, Lemma $M.2$ from Belloni et al. (2018) implies

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

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

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

with probability $1 - o(1)$. We now proceed by using Lemma $M.3$ in Belloni et al. (2018). We obtain uniformly over all $r = 1, \dots, d$

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

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

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

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

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

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

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

and that the minimum sparse eigenvalues are uniformly bounded away from zero. With the same argument as above we 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 finally completes the proof. ∎

Appendix C Computational Details

C.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 Chernozhukov et al. (2015a) which can be downloaded from CRAN. Construction of B-splines is based on the R package `splines`. R code is available upon request.

C.2 Simulation Study: Smoothing Parameters in B-splines

Table 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

Smoothing parameters $\{k_j, k_{-j}\}$ corresponding to simulation results in Table 2.

C.3 Empirical Application: Cross-Validation Procedure for Choice of Smoothing Parameter

The choice of the degrees of freedom parameter k for 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

- For $j = 1, ..., p$,
  (i) Set up a grid of values for k_{-j},
  (ii) 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 $k_{0,-j}$.
  (iii) 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 5.

NOX	11
$CRIM$	6
ZN	3
$INDUS$	6
RM	6
AGE	5
$DIST$	9
TAX	5
$PTRATIO$	11
$BLACK$	5
$LSTAT$	7

TABLE 5
Smoothing parameters used in empirical application.

C.4 Empirical Application: Additional Plots for Explanatory Variables

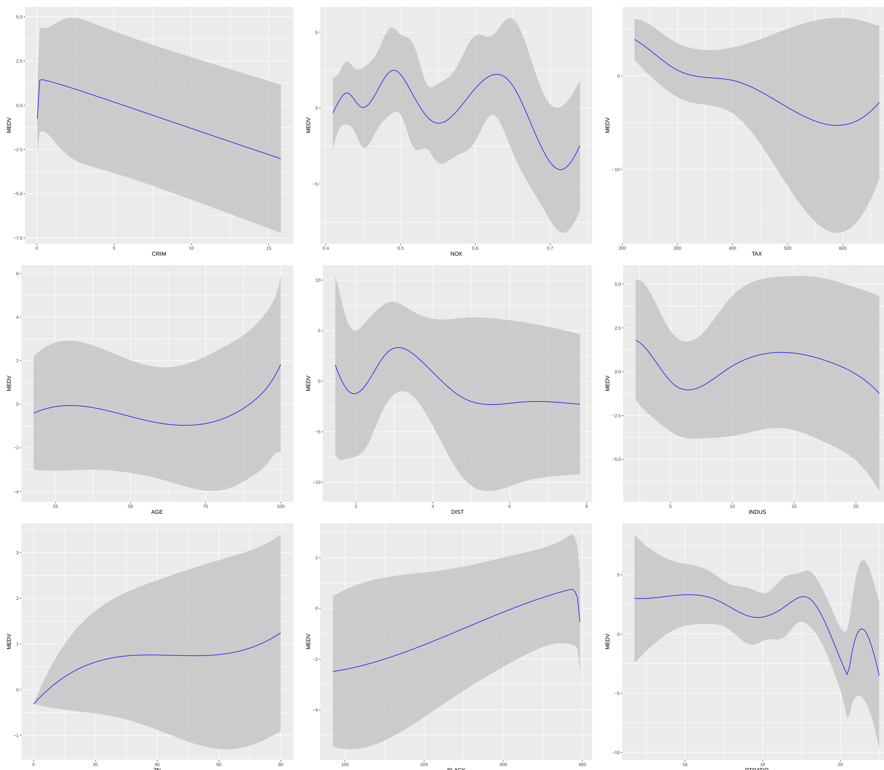

Fig 3: Additional plots of the effect of the explanatory variables on the dependent variable $MEDV$ with simultaneous 95%-confidence bands in the Boston housing data application.

References

Alexandre Belloni, Victor Chernozhukov, and Kengo Kato. Uniform post-selection inference for least absolute deviation regression and other z-estimation problems. *Biometrika*, 102(1):77–94, Dec 2014a. ISSN 1464-3510. . URL http://dx.doi.org/10.1093/biomet/asu056.

Alexandre Belloni, Victor Chernozukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2 (287)):608–650, 2014b. ISSN 00346527, 1467937X. URL http://www.jstor.org/stable/43551575.

Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, and Ying Wei. Uniformly valid post-regularization confidence regions for many functional parameters in z-estimation framework. *Annals of Statistics*, 46(6B):3643–3675, 2018.

Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

Richard Y. Chen, Alex Gittens, and Joel A. Tropp. The masked sample covariance estimator: an analysis using matrix concentration inequalities. *Information and Inference: A Journal of the IMA*, 1(1):2–20, 2012.

Victor Chernozhukov, Denis Chetverikov, Kengo Kato, et al. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819, 2013.

Victor Chernozhukov, Christian Hansen, and Martin Spindler. *hdm: High-Dimensional Metrics*, 2015a. R package version 0.1.

Victor Chernozhukov, Christian Hansen, and Martin Spindler. Valid post-selection and post-regularization inference: An elementary, general approach. *Annual Review of Economics*, 7(1):649–688, 2015b. . URL https://doi.org/10.1146/annurev-economics-012315-015826.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 2017.

Gerda Claeskens and Ingrid Keilegom. Bootstrap confidence bands for regression curves and their derivatives. *Annals of Statistics*, 31, 2003. .

Kjell Doksum and Alexander Samarov. Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression. *The Annals of Statistics*, 23(5):1443–1473, 1995.

Jianqing Fan and Wenyang Zhang. Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scandinavian Journal of Statistics*, 27(4):715–731, 2000. ISSN 03036898, 14679469. URL http://www.jstor.org/stable/4616637.

Jerome H. Friedman and Werner Stuetzle. Projection pursuit regression. *Journal of the American Statistical Association*, 76(376):817–823, 1981. .

Karl Gregory, Enno Mammen, and Martin Wahl. Statistical inference in sparse high-dimensional additive models. 2016. URL https://arxiv.org/abs/

1603.07632.

David Harrison Jr and Daniel L Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1):81–102, 1978.

Trevor Hastie and Robert Tibshirani. *Generalized Additive Models*, volume 43. Chapman and Hall, Ltd., London, 1990.

Wolfgang Härdle. Asymptotic maximal deviation of M-smoothers. *Journal of Multivariate Analysis*, 29(2):163–179, May 1989. URL https://ideas.repec.org/a/eee/jmvana/v29y1989i2p163-179.html.

Jian Huang, Joel Horowitz, and Fengrong Wei. Variable selection in nonparametric additive models. *Annals of Statistics*, 38:2282–2313, 08 2010. .

Kengo Kato. Two-step estimation of high dimensional additive models. 2012. URL https://arxiv.org/abs/1207.5313.

Sven Klaassen, Jannis Kück, Martin Spindler, and Victor Chernozhukov. Uniform inference in high-dimensional gaussian graphical models. *arXiv preprint arXiv:1808.10532*, 2018.

Vladimir Koltchinskii and Ming Yuan. Sparsity in multiple kernel learning. *Annals of Statistics*, 38(6):3660–3695, 12 2010. . URL https://doi.org/10.1214/10-AOS825.

Efang Kong and Yingcun Xia. A single-index quantile regression model and its estimation. *Econometric Theory*, 28(4):730–768, 2012.

Damian Kozbur. Inference in additively separable models with a high dimensional conditioning set. *SSRN Electronic Journal*, 03 2015. .

Friedrich Leisch and Evgenia Dimitriadou. *mlbench: Machine learning benchmark problems*, 2010. R package version 2.1-1.

Yi Lin and Hao Helen Zhang. Component selection and smoothing in multivariate nonparametric regression. *Annals of Statistics*, 34(5):2272–2297, 10 2006. . URL https://doi.org/10.1214/009053606000000722.

Yin Lou, Jacob Bien, Rich Caruana, and Johannes Gehrke. Sparse partially linear additive models. *Journal of Computational and Graphical Statistics*, 25(4):1126–1140, 2016. .

Junwei Lu, Mladen Kolar, and Han Liu. Kernel meets sieve: Post-regularization confidence bands for sparse additive model. *Journal of the American Statistical Association*, 0(ja):1–16, 2020. . URL https://doi.org/10.1080/01621459.2019.1689984.

Lukas Meier, Sara Van de Geer, and Peter Bühlmann. High-dimensional additive modeling. *Annals of Statistics*, 37(6B):3779–3821, 2009.

David J. Newman, SCLB Hettich, Cason L. Blake, and Christopher J. Merz. Uci repository of machine learning databases, 1998. URL http://www.ics.uci.edu/~mlearn/MLRepository.html.

Ashley Petersen, Daniela Witten, and Noah Simon. Fused lasso additive model. *Journal of Computational and Graphical Statistics*, 25(4):1005–1025, 2016. . PMID: 28239246.

Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030, 2009. . URL https://rss.onlinelibrary.

wiley.com/doi/abs/10.1111/j.1467-9868.2009.00718.x.

Sylvain Sardy and Paul Tseng. Amlet, ramlet, and gamlet: Automatic nonlinear fitting of additive models, robust and generalized, with wavelets. *Journal of Computational and Graphical Statistics*, 13(2):283–309, 2004. ISSN 10618600. URL http://www.jstor.org/stable/1391177.

Charles J. Stone. Additive regression and other nonparametric models. *Annals of Statistics*, 13(2):689–705, 06 1985. . URL https://doi.org/10.1214/aos/1176349548.

Jiayang Sun and Clive R. Loader. Simultaneous confidence bands for linear regression and smoothing. *Annals of Statistics*, 22(3):1328–1345, 09 1994. . URL https://doi.org/10.1214/aos/1176325631.

Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42(3):1166–1202, 06 2014. . URL https://doi.org/10.1214/14-AOS1221.

Aad W. Van der Vaart and Jon A. Wellner. Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer, 1996.

Simon N. Wood. *Generalized additive models: an introduction with R*. CRC press, 2017.

Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014. . URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12026.

Wenyang Zhang and Heng Peng. Simultaneous confidence band and hypothesis test in generalised varying-coefficient models. *Journal of Multivariate Analysis*, 101(7):1656 – 1680, 2010. ISSN 0047-259X. . URL http://www.sciencedirect.com/science/article/pii/S0047259X10000539.