SUPPLEMENTARY MATERIAL FOR
“TRANSFORMATION MODELS IN HIGH-DIMENSIONS”

By Sven Klaassen and Jannis Kück and Martin Spindler

University of Hamburg

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

\[
E_n[g(Z)] \equiv E_{P_n}[g(Z)] := \frac{1}{n} \sum_{i=1}^{n} g(Z_i)
\]

and

\[
G_n(g) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( g(Z_i) - E[g(Z_i)] \right).
\]

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

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

where the supremum is taken over all finitely discrete probability measures \(Q\) with \(0 < E_Q[|F|^r]^{1/r} < \infty\). For further definitions of entropy related terminology used in the paper, we refer to [9]. For any function \(\nu(\theta, u)\), we use the notation \(\dot{\nu}_{\theta^*}(u) := \partial \nu(\theta, u)/\partial \theta|_{\theta = \theta^*}\), respectively \(\nu_{\theta}(u^*) := \partial \nu(\theta, u)/\partial u|_{u = u^*}\).
APPENDIX A: PROOFS

Proof of Lemma 2.
We use the same argument as Neumeyer, Noh, Van Keilegom (2016) \[7\]. Define
\[
f^{(\theta)}(y|x) := \frac{1}{\sqrt{2\pi\sigma_\theta^2}} \exp \left( -\frac{(\Lambda_\theta(y) - x\beta_0)^2}{2\sigma_\theta^2} \right) \Lambda'_\theta(y).
\]
The expected Kullback-Leibler-Distance between \(f_{Y|X}\) and \(f^{(\theta)}\) is greater or equal to zero and equality only holds for the true parameter \(\theta_0\). Therefore, the following expression is minimized in \(\theta_0\)
\[
\int \int \log \left( \frac{f_{Y|X}(y|x)}{f^{(\theta)}(y|x)} \right) f_{Y|X}(y|x) dydF_X(x)
= \int \int \log(f_{Y|X}(y|x)) f_{Y|X}(y|x) dydF_X(x)
- \int \int \log(f^{(\theta)}(y|x)) f_{Y|X}(y|x) dydF_X(x).
\]
It follows that \(\mathbb{E}[\log(f^{(\theta)}(Y|X))]\) is maximized for the true parameter \(\theta = \theta_0\). Under the regularity conditions A1, A3 and A6, it holds
\[
\mathbb{E}[\psi((Y, X), \theta_0, h_0)] = \mathbb{E} \left[ \frac{\partial}{\partial \theta} \log(f^{(\theta)}(Y|X)) \big|_{\theta = \theta_0} \right]
= \frac{\partial}{\partial \theta} \mathbb{E}[\log(f^{(\theta)}(Y|X))] \big|_{\theta = \theta_0} = 0.
\]
Here, we used that for all \(\theta\)
\[
0 < c \leq \sigma_\theta^2 \quad \text{and} \quad \sigma_\theta^2 \leq \mathbb{E} \left[ \sup_{\theta \in \Theta} \varepsilon_\theta^2 \right] \leq C < \infty,
\]
which is shown in the proof of Theorem 5.

Proof of Lemma 3.
For the notation, we refer to Section 2.1. Let \(h = (h_1, h_2, h_3, h_4) \in \mathcal{H}'\) be arbitrary. First, we consider
\[
\partial_r \left( \Lambda_{\theta_0}(Y) - (m_{\theta_0}(X) + r(h_1(\theta_0, X) - m_{\theta_0}(X))) \right) \big|_{r=0}
\]
= m_{\theta_0}(X) - h_1(\theta_0, X)

and analogous

\[ \partial_r \left( \dot{\Lambda}_{\theta_0}(Y) - (\dot{m}_{\theta_0}(X) + r(h_3(\theta_0, X) - \dot{m}_{\theta_0}(X))) \right) \bigg|_{r=0} = \dot{m}_{\theta_0}(X) - h_3(\theta_0, X). \]

Additionally, we have

\[ \partial_r \left( \left( \sigma_{\theta_0}^2 + r(h_2(\theta_0) - \sigma_{\theta_0}^2) \right)^{-1} \right) \bigg|_{r=0} = -\frac{h_2(\theta_0) - \sigma_{\theta_0}^2}{\left( \sigma_{\theta_0}^2 \right)^2} \]

and

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

By the product rule, we obtain

\[
\mathbb{E} \left[ \partial_r I(\theta_0, \sigma_{\theta_0}^2 + r(h_2 - \sigma_{\theta_0}^2), \dot{\sigma}_{\theta_0}^2 + r(h_4 - \dot{\sigma}_{\theta_0}^2)) \bigg|_{r=0} | X \right]
\]

\[
= \mathbb{E} \left[ \frac{h_4(\theta_0) - \dot{\sigma}_{\theta_0}^2}{2\sigma_{\theta_0}^2} - \frac{h_2(\theta_0) - \sigma_{\theta_0}^2}{2 \left( \sigma_{\theta_0}^2 \right)^2} \bigg| X \right]
\]

\[
= \frac{h_4(\theta_0) - \dot{\sigma}_{\theta_0}^2}{2\sigma_{\theta_0}^2} - \frac{h_2(\theta_0) - \sigma_{\theta_0}^2}{2 \left( \sigma_{\theta_0}^2 \right)^2}.
\]

\[
\mathbb{E} \left[ \partial_r II(\theta_0, m_{\theta} + r(h_1 - m_{\theta}), \sigma_{\theta}^2 + r(h_2 - \sigma_{\theta}^2), \dot{m}_{\theta} + r(h_1 - \dot{m}_{\theta})) \bigg|_{r=0} | X \right]
\]

\[
= \mathbb{E} \left[ -\frac{h_2(\theta_0) - \sigma_{\theta_0}^2}{\left( \sigma_{\theta_0}^2 \right)^2} \left( \Lambda_{\theta_0}(Y) - m_{\theta_0}(X) \right) \left( \dot{\Lambda}_{\theta_0}(Y) - \dot{m}_{\theta_0}(X) \right) \bigg| X \right]
\]

\[
+ \mathbb{E} \left[ \frac{1}{\sigma_{\theta_0}^2} \left( m_{\theta_0}(X) - h_1(\theta_0, X) \right) \left( \dot{\Lambda}_{\theta_0}(Y) - \dot{m}_{\theta_0}(X) \right) \bigg| X \right]
\]

\[
+ \mathbb{E} \left[ \frac{1}{\sigma_{\theta_0}^2} \left( \Lambda_{\theta_0}(Y) - m_{\theta_0}(X) \right) \left( \dot{m}_{\theta_0}(X) - h_3(\theta_0, X) \right) \bigg| X \right]
\]

\[
= -\frac{h_2(\theta_0) - \sigma_{\theta_0}^2}{\left( \sigma_{\theta_0}^2 \right)^2} \mathbb{E} \left[ \varepsilon_{\theta_0} \dot{\varepsilon}_{\theta_0} \bigg| X \right] + \frac{m_{\theta_0}(X) - h_1(\theta_0, X)}{\sigma_{\theta_0}^2} \mathbb{E} \left[ \dot{\varepsilon}_{\theta_0} \bigg| X \right] \bigg|_{\sigma_{\theta_0}^2}.
\]
\[
\frac{\dot{m}_{\theta_0}(X) - h_3(\theta_0, X)}{\sigma_{\theta_0}^2} = \left(\frac{\sigma_{\theta_0}^2}{\sigma_{\theta_0}^2}\right) \mathbb{E}[\varepsilon_{\theta_0} | X] = 0
\]

and

\[
\mathbb{E} \left[ \partial_0^3 \left( \theta_0, m_\theta + r(h_1 - m_\theta), \sigma_0^2 + r(h_2 - \sigma_0^2), \dot{\sigma}_0^2 + r(h_4 - \dot{\sigma}_0^2) \right) | r=0 | X \right]
\]

\[
= \mathbb{E} \left[ \frac{h_4(\theta_0) - \dot{\sigma}_0^2}{2 (\sigma_{\theta_0}^2)^2} \left( \Lambda_{\theta_0}(Y) - m_{\theta_0}(X) \right)^2 | X \right]
\]

\[- \mathbb{E} \left[ \frac{\dot{\sigma}_0^2}{(\sigma_{\theta_0}^2)^2} \left( \Lambda_{\theta_0}(Y) - m_{\theta_0}(X) \right)^2 | X \right]
\]

\[+ \mathbb{E} \left[ \frac{\dot{\sigma}_0^2}{(\sigma_{\theta_0}^2)^2} \left( \Lambda_{\theta_0}(Y) - m_{\theta_0}(X) \right)^2 | X \right]
\]

\[= \frac{h_4(\theta_0) - \dot{\sigma}_0^2}{2 (\sigma_{\theta_0}^2)^2} \mathbb{E} \left[ (\Lambda_{\theta_0}(Y) - m_{\theta_0}(X))^2 | X \right]
\]

\[- \frac{\dot{\sigma}_0^2}{(\sigma_{\theta_0}^2)^2} \mathbb{E} \left[ (\Lambda_{\theta_0}(Y) - m_{\theta_0}(X))^2 | X \right]
\]

\[+ \frac{\dot{\sigma}_0^2}{(\sigma_{\theta_0}^2)^2} \left( m_{\theta_0}(X) - h_1(\theta_0, X) \right) \mathbb{E} \left[ \varepsilon_{\theta_0} | X \right]
\]

\[= \frac{h_4(\theta_0) - \dot{\sigma}_0^2}{2 \sigma_{\theta_0}^2} - \frac{\dot{\sigma}_0^2}{(\sigma_{\theta_0}^2)^2} h_2(\theta_0) - \frac{\sigma_{\theta_0}^2}{(\sigma_{\theta_0}^2)^2} \cdot \hat{\mathbb{E}}_{\theta_0} \left[ \varepsilon_{\theta_0} | X \right]
\]

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

\[D_0[h - h_0] = \partial_r \left\{ \mathbb{E} \left[ \psi((Y, X), \theta_0, h_0 + r(h - h_0)) \right] \right\} \bigg|_{r=0}
\]
and using Markov’s inequality we obtain

$$\mathbb{E} \left[ \partial_r \psi \left( (Y, X), \theta_0, h_0 + r(h - h_0) \right) \bigg| r=0 \right]$$

$$= \mathbb{E} \left[ \partial_r \psi \left( (Y, X), \theta_0, h_0 + r(h - h_0) \right) \bigg| X \right]$$

$$= \mathbb{E} \left[ - \mathbb{E} \left[ \partial_r I(\theta_0, \sigma_0^2 + r(h_2 - \sigma_0^2), \dot{\sigma}_0 + r(h_4 - \dot{\sigma}_0^2)) \bigg| r=0 \right] X \right]$$

$$- \mathbb{E} \left[ \partial_r II(\theta_0, m_\theta + r(h_1 - m_\theta), \sigma_0^2 + r(h_2 - \sigma_0^2), \dot{m}_\theta + r(h_1 - \dot{m}_\theta)) \bigg| r=0 \right] + \mathbb{E} \left[ \partial_r III(\theta_0, m_\theta + r(h_1 - m_\theta), \sigma_0^2 + r(h_2 - \sigma_0^2), \dot{m}_\theta + r(h_1 - \dot{m}_\theta)) \bigg| r=0 \right]$$

$$= \mathbb{E} \left[ - \mathbb{E} \left[ \partial_r I(\theta_0, \sigma_0^2 + r(h_2 - \sigma_0^2), \dot{\sigma}_0 + r(h_4 - \dot{\sigma}_0^2)) \bigg| r=0 \right] X \right]$$

$$- \mathbb{E} \left[ \partial_r II(\theta_0, m_\theta + r(h_1 - m_\theta), \dot{m}_\theta + r(h_1 - \dot{m}_\theta)) \bigg| r=0 \right] + \mathbb{E} \left[ \partial_r III(\theta_0, m_\theta + r(h_1 - m_\theta), \dot{m}_\theta + r(h_1 - \dot{m}_\theta)) \bigg| r=0 \right]$$

$$= 0,$$

where we used \( \dot{\sigma}_0^2 = 2\mathbb{E}[\varepsilon_0, \dot{\varepsilon}_0] \) in the last step.

**Proof of Theorem 1.**

The Assumptions A1-A8 directly imply the conditions in Theorem 5 for the model (2.1) and (2.2) except for Assumption B7. We need to show that the empirical eigenvalues converge to the restricted sparse eigenvalues defined in Assumption A8. By Lemma P.1 in [2], we have

$$\mathbb{E} \left[ \sup_{\|\delta\|_0 \leq s \log(n), \|\delta\|=1} \left| \|X^T \delta\|_{\hat{P}_n,2}^2 - \|X^T \delta\|_{\hat{P}_n,2}^2 \right| \right]$$

$$\leq C \left( \frac{s \log^2(n) \log(p)}{n} + \sqrt{\frac{s \log^2(n) \log(p)}{n}} \right) \leq C \sqrt{\frac{s \log^2(n) \log(p)}{n}}$$

and using Markov’s inequality we obtain

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

with probability \( 1 - o(1) \). This implies Assumption B7 for \( n \) large enough since the restricted sparse eigenvalues are bounded away from zero and
above. The statement follows with $\tilde{\delta}_n = O\left(\sqrt{n^{-1/2} s \log(p \vee n)}\right) = o(1)$ by the growth condition A2.

**Proof of Theorem 2.**
As shown in the proof of Theorem 5, we have

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

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

and with an analogous argument

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

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

$$\sup_{\theta \in \Theta} |\hat{\sigma}^2_{\theta} - \sigma^2_{\theta}|$$

$$= \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \left( \Lambda_{\theta}(Y_i) - X_i^T \hat{\beta}_{\theta} \right)^2 - E[\varepsilon_{\theta,i}^2] \right|$$

$$= \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \left( \varepsilon_{\theta,i}^2 - E[\varepsilon_{\theta,i}^2] \right) - \frac{2}{n} \sum_{i=1}^{n} \varepsilon_{\theta,i} X_i^T (\hat{\beta}_{\theta} - \beta_{\theta}) \right.$$ 

$$+ \frac{1}{n} \sum_{i=1}^{n} \left( X_i^T (\hat{\beta}_{\theta} - \beta_{\theta}) \right)^2 \right|$$

$$\leq \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \left( \varepsilon_{\theta,i}^2 - E[\varepsilon_{\theta,i}^2] \right) \right| + \sup_{\theta \in \Theta} \left| \frac{2}{n} \sum_{i=1}^{n} \varepsilon_{\theta,i} X_i^T (\hat{\beta}_{\theta} - \beta_{\theta}) \right|$$

$$+ \sup_{\theta \in \Theta} \left| X_i^T (\hat{\beta}_{\theta} - \beta_{\theta}) \right| \leq 2 \sup_{\theta \in \Theta} \left| X_i^T (\hat{\beta}_{\theta} - \beta_{\theta}) \right| \left| E_{n,2} \right| \left( \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{\theta,i}^2 \right)$$

$$= O\left( s \log(p \vee n) \right)$$
\[
\leq 2 \sup_{\theta \in \Theta} ||X^T (\hat{\beta}_\theta - \beta_\theta)||_{\mathbb{P}_{n,2}} \sup_{\theta \in \Theta} \left( \frac{1}{n} \sum_{i=1}^{n} \tilde{\varepsilon}_{\theta,i}^2 + O\left( \frac{s \log(p \vee n)}{n} \right) + O(\log(n)n^{-1/2}) \right)
\]
\[
= O\left( \max \left( \sqrt{\frac{s \log(p \vee n)}{n}}, \frac{\log(n)}{n^{1/2}} \right) \right) \leq \tilde{\delta}_n n^{-\frac{1}{4}}
\]

for a sequence \( \tilde{\delta}_n = O\left( \max \left( \sqrt{n^{-1/2}}s \log(p \vee n), n^{-1/4} \log(n) \right) \right) \) \( n \), due to the growth condition A2. By the same argument, we obtain with probability \( 1 - o(1) \)

\[
\sup_{\theta \in \Theta} |\hat{\sigma}_\theta^2 - \hat{\sigma}_\theta^2| = \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} (\tilde{\varepsilon}_{\theta,i} \hat{\varepsilon}_{\theta,i} - \varepsilon_{\theta,i} \varepsilon_{\theta,i}) - \frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_{\theta,i} (\hat{m}_\theta(X_i) - m_\theta(X_i)) \right|
\]

\[
\leq \frac{1}{n} \sum_{i=1}^{n} \left( \hat{m}_\theta(X_i) - m_\theta(X_i) \right)^2 \left( \frac{1}{n} \sum_{i=1}^{n} \left( \hat{m}_\theta(X_i) - m_\theta(X_i) \right)^2 \right)^{1/2}
\]

\[
\leq \frac{1}{n} \sum_{i=1}^{n} \left( \hat{m}_\theta(X_i) - m_\theta(X_i) \right)^2 + 2 \sup_{\theta \in \Theta} \left[ \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{\theta,i} (\hat{m}_\theta(X_i) - m_\theta(X_i)) \right]
\]

\[
\leq 2 \sup_{\theta \in \Theta} \left[ ||X^T (\hat{\beta}_\theta - \beta_\theta)||_{\mathbb{P}_{n,2}} \sqrt{\frac{1}{n} \sum_{i=1}^{n} \tilde{\varepsilon}_{\theta,i}^2} \right]
\]

\[
+ 2 \sup_{\theta \in \Theta} \left[ \frac{1}{n} \sum_{i=1}^{n} \tilde{\varepsilon}_{\theta,i} (\hat{m}_\theta(X_i) - m_\theta(X_i)) \right]
\]
\[ + \left( \sup_{\theta \in \Theta} \left\| (\hat{m}_\theta(X_i) - m_\theta(X_i)) \right\|_{P_n,2}^2 \right) \sup_{\theta \in \Theta} \left\| (\hat{m}_\theta(X_i) - \hat{m}_\theta(X_i)) \right\|^2_{P_n,2} \]

\[ = O \left( \frac{s \log(p \vee n)}{n} \right) \]

\[ \leq 2 \sup_{\theta \in \Theta} \| X^T (\hat{\beta}_\theta - \hat{\beta}_\theta) \|_{P_n,2} \sup_{\theta \in \Theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \epsilon_{\theta,i}^2 \right\} \]

\[ = O \left( \frac{s \log(p \vee n)}{n} \right) \]

\[ + 2 \sup_{\theta \in \Theta} \| X^T (\hat{\beta}_\theta - \beta_\theta) \|_{P_n,2} \sup_{\theta \in \Theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \epsilon_{\theta,i}^2 \right\} + O \left( \frac{s \log(p \vee n)}{n} \right) \]

\[ = O \left( \frac{s \log(p \vee n)}{n} \right) \]

\[ + O(\log(n)n^{-1/2}) \]

and therefore

\[ \sup_{\theta \in \Theta} |\hat{\sigma}_\theta^2 - \hat{\sigma}_\theta^2| = O \left( \max \left( \sqrt{\frac{s \log(p \vee n)}{n}}, \frac{\log(n)}{n^{1/2}} \right) \right) \leq \tilde{\delta}_n n^{-1/4}. \]

**Proof of Theorem 3.**

The strategy of the proof is similar to the proof of Theorem 1 from Belloni et al. (2014) [4]. Let \( C, C_1 \) and \( C_2 \) denote generic positive constants that may differ in each appearance, but do not depend on the sequence \( P \in \mathcal{P}_n \).

For every \( \theta \in \Theta \), the set \( \hat{H}_1(\theta) \) consists of unions of \( p \) choose \( Cs \) sets, where the set of indices \( \{ i \in \{1, \ldots, p\} : \beta_i \neq 0 \} \) has cardinality not more than \( Cs \) and therefore is a subset of a vector space with dimension \( Cs \). It follows that \( \hat{H}_1(\theta) \) consists of unions of \( p \) choose \( Cs \) VC classes \( \hat{H}_{1,k}(\theta) \) with VC indices less or equal to \( Cs + 2 \) (Lemma 2.6.15, Van der Vaart and Wellner (1996)) [9]. Using Theorem 2.6.7 in Van der Vaart and Wellner (1996), we obtain

\[ \sup_{Q} \log N(\varepsilon \| \hat{H}_1 \|_{Q,2}, \hat{H}_1(\theta), L_2(Q)) \]

\[ \leq \sup_{Q} \log \left( \sum_{k=1}^{p} N(\varepsilon \| \hat{H}_1 \|_{Q,2}, \hat{H}_{1,k}(\theta), L_2(Q)) \right) \]
\[
\sup_Q \log \left( \left( \frac{p}{C_S} \right)^{K(CS + 2)(16e)^{Cs + 1}} \left( \frac{1}{\varepsilon} \right)^{2Cs + 2} \right) \\
\leq \left( \frac{p}{C_S} \right)^{Cs}
\leq \log \left( \left( \frac{e \cdot p}{C_S} \right)^{Cs} K(CS + 2)(16e)^{Cs + 1} \left( \frac{1}{\varepsilon} \right)^{2Cs + 2} \right)
\leq Cs \log \left( \frac{p}{\varepsilon} \right)
\]

with \( C \) being independent from \( \theta \). Since

\[
\sup_{h_1(\theta) \in \tilde{H}_1(\theta)} |h_1(\theta, x)| \leq \sup_{\beta: \|\beta_0 - \beta_\theta\|_1 \leq \delta_n \sqrt{n}^{-1/4}} |x^T \tilde{\beta}| \\
\leq \sup_{\beta: \|\beta_0 - \beta_\theta\|_1 \leq \delta_n \sqrt{n}^{-1/4}} |x^T \tilde{\beta} - x^T \beta_\theta| + |x^T \beta_\theta| \\
\leq KC + E[F_\Lambda | X = x] =: \tilde{H}_1(x),
\]

the envelope \( \tilde{H}_1 \) can be chosen independent from \( \theta \). Here and in the following, we omit the dependence from \( Y \) in \( F_\Lambda \equiv F_\Lambda(Y) \) to simplify notation. With the same argument we obtain

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

with envelope \( \tilde{H}_3(x) := KC + E[\tilde{F}_\Lambda | X = x] \).

Next, we consider

\[
\tilde{H}_4(\theta) := \left\{ c \in \mathbb{R} \mid |c - \hat{\sigma}^2_\theta| \leq \delta_n n^{-1/4} \right\} \subseteq \left[ \hat{\sigma}^2_\theta - Cn^{-1/4}, \hat{\sigma}^2_\theta + Cn^{-1/4} \right]
\subseteq \left[ -(c + Cn^{-1/4}), (c + Cn^{-1/4}) \right],
\]

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

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

for all \( \theta \in \Theta \) with envelope \( \tilde{H}_4 = c + C \) and \( C \) independent from \( \theta \). Remark that \( 0 < c_1 = \inf_{\theta \in \Theta} \sigma^2_\theta \) and \( c_2 = \sup_{\theta \in \Theta} \sigma^2_\theta < \infty \) due to the Assumptions
A4–A5. For $n$ sufficiently large, we find a $c_3$ with $0 < c_3 \leq c_1 - Cn^{-1/4}$. Therefore, we can define

$$
\tilde{H}_2(\theta) := \left\{ \frac{1}{h_2(\theta)} | \tilde{h}_2(\theta) \in \tilde{H}_2(\theta) \right\} 
$$

$$
\subseteq \left\{ \frac{1}{c} | c - \sigma_\theta^2 | \leq Cn^{-1/4} \right\} 
$$

$$
= \left\{ \frac{1}{c} | c - \sigma_\theta^2 | \leq \frac{1}{|c\sigma_\theta^2|} Cn^{-1/4} \right\} 
$$

$$
\subseteq \left\{ \frac{1}{c} | c - \sigma_\theta^2 | \leq C^*n^{-1/4} \right\} 
$$

$$
= \left\{ \frac{c}{c - 1/|\sigma_\theta^2|} \leq C^*n^{-1/4} \right\} 
$$

$$
= \left[ \frac{1}{\sigma_\theta^2} - C^*n^{-1/4}, 1/\sigma_\theta^2 + C^*n^{-1/4} \right] 
$$

$$
\subseteq \left[ \frac{1}{c_2} - C^*n^{-1/4}, 1/c_2 + C^*n^{-1/4} \right] 
$$

with $C^* = \frac{C}{c_3c_1}$. Analogously, we obtain for all $\theta \in \Theta$

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

with envelope $\tilde{H}_2 = 1/c_2 + C^*$ and $C$ independent from $\theta$. Let us define

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

$$
II(\theta, \tilde{H}_1, \tilde{H}_2, \tilde{H}_3) := \left\{ (y,x) \mapsto -h_2(\theta) (\Lambda_\theta(y) - h_1(\theta,x)) (\Lambda_\theta(y) - h_3(\theta,x)) 
| h_1(\theta) \in \tilde{H}_1(\theta), h_2(\theta) \in \tilde{H}_2(\theta), h_3(\theta) \in \tilde{H}_3(\theta) \right\} 
$$

and

$$
III(\theta, \tilde{H}_1, \tilde{H}_2, \tilde{H}_4) := \left\{ (y,x) \mapsto \frac{1}{2} h_2^2(\theta) h_4(\theta) (\Lambda_\theta(y) - h_1(\theta,x))^2 
| h_1(\theta) \in \tilde{H}_1(\theta), h_2(\theta) \in \tilde{H}_2(\theta), h_3(\theta) \in \tilde{H}_3(\theta) \right\}. 
$$
By Lemma L.1 in the supplement to [3], we have

\[ \log N \left( \epsilon \| 1/2 \tilde{H}_2 \tilde{H}_4 \|_{Q, 2}, I(\theta, \tilde{H}_2, \tilde{H}_4), L_2(\Omega) \right) \]

\[ \leq \log N \left( \frac{\epsilon}{4} \| \tilde{H}_2 \|_{Q, 2}, \tilde{H}_2(\theta), L_2(\Omega) \right) + \log N \left( \frac{\epsilon}{4} \| \tilde{H}_4 \|_{Q, 2}, \tilde{H}_4(\theta), L_2(\Omega) \right) \]

\[ \leq 2 \log \left( \frac{C}{\epsilon} \right). \]

By Assumption A3, we obtain

\[ \log N \left( \epsilon \| \tilde{H}_2(F_\Lambda + \tilde{H}_1)(\hat{F}_\Lambda + \tilde{H}_3) \|_{Q, 2}, II(\theta, \tilde{H}_1, \tilde{H}_2, \tilde{H}_3), L_2(\Omega) \right) \]

\[ \leq \log N \left( \frac{\epsilon}{2} \| \tilde{H}_2(\theta) \|_{Q, 2}, \tilde{H}_2(\theta), L_2(\Omega) \right) \]

\[ + \log N \left( \frac{\epsilon}{4} \| (F_\Lambda + \tilde{H}_1) \|_{Q, 2}, \mathcal{F}_\Lambda - \tilde{H}_1(\theta), L_2(\Omega) \right) \]

\[ + \log N \left( \frac{\epsilon}{4} \| (\hat{F}_\Lambda + \tilde{H}_3) \|_{Q, 2}, \hat{F}_\Lambda - \tilde{H}_3(\theta), L_2(\Omega) \right) \]

\[ \leq \log \left( \frac{2C}{\epsilon} \right) + \log N \left( \frac{\epsilon}{8} \| F_\Lambda \|_{Q, 2}, \mathcal{F}_\Lambda, L_2(\Omega) \right) \]

\[ + \log N \left( \frac{\epsilon}{8} \| \hat{F}_\Lambda \|_{Q, 2}, \hat{F}_\Lambda, L_2(\Omega) \right) \]

\[ + \log N \left( \frac{\epsilon}{8} \| \tilde{H}_1 \|_{Q, 2}, \tilde{H}_1(\theta), L_2(\Omega) \right) \]

\[ + \log N \left( \frac{\epsilon}{8} \| \tilde{H}_3 \|_{Q, 2}, \tilde{H}_3(\theta), L_2(\Omega) \right) \]

\[ \leq \log \left( \frac{2C}{\epsilon} \right) + C'_\Lambda \log(8C'_\Lambda/\epsilon) + C'_\Lambda \log(8C'_\Lambda/\epsilon) + C_s \log \left( \frac{8p}{\epsilon} \right) \]

\[ + C_s \log \left( \frac{8p}{\epsilon} \right) \]

\[ \leq C_1 s \log \left( \frac{C_2 p}{\epsilon} \right) \]

and with an analogous argument

\[ \log N \left( \epsilon \| \frac{1}{2} \tilde{H}_2^2 \tilde{H}_4^2 (F_\Lambda + \tilde{H}_1)^2 \|_{Q, 2}, III(\theta, \tilde{H}_1, \tilde{H}_2, \tilde{H}_4), L_2(\Omega) \right) \]

\[ \leq C_1 s \log \left( \frac{C_2 p}{\epsilon} \right). \]

Since

\[ \Psi(\theta) = I(\theta, \tilde{H}_2, \tilde{H}_4) + II(\theta, \tilde{H}_1, \tilde{H}_2, \tilde{H}_3) + III(\theta, \tilde{H}_1, \tilde{H}_2, \tilde{H}_4) + c_0, \]
we can define the envelope
\[ \tilde{\psi}(Y, X) := \frac{1}{2} \tilde{H}_2 \tilde{H}_4 + \tilde{H}_2 (F_A + \tilde{H}_3) (F_A + \tilde{H}_3) + \frac{1}{2} \tilde{H}_2^2 (F_A + \tilde{H}_1)^2 + J_{\tilde{A}}, \]
which is independent from \( \theta \) with
\[ \mathbb{E} \left[ \left( \tilde{\psi}(Y, X) \right)^2 \right] = \mathbb{E} \left[ \left( \frac{1}{2} \tilde{H}_2 \tilde{H}_4 + \tilde{H}_2 (F_A + \tilde{H}_3) (F_A + \tilde{H}_3) + \frac{1}{2} \tilde{H}_2^2 (F_A + \tilde{H}_1)^2 + J_{\tilde{A}} \right)^2 \right] \leq C < \infty, \]
where we used Assumption A3 and A4. Additionally, by using
\[ N(\varepsilon \|J_{\tilde{A}}\|_{Q,2}, c_\theta, L_2(Q)) = 1 \]
for all \( \theta \in \Theta \) and Lemma L.1 in the supplement to [3], we obtain
\[ \sup_Q \log N(\varepsilon \|\tilde{\psi}\|_{Q,2}, \Psi(\theta), L_2(Q)) \leq C_1 s \log \left( \frac{C_2 (p \vee n)}{\varepsilon} \right), \]
where the supremum is taken over all finitely discrete probability measures \( Q \) with \( \mathbb{E}_Q \left[ (\tilde{\psi}(Y, X))^2 \right] < \infty. \)

**Proof of Theorem 4.**

We demonstrate that the Conditions C1-C7 from Theorem 6 are satisfied. Most of these conditions are already proven in the preceding theorems. The Condition C1 is shown in Lemma 2. Due to Theorem 1 and Theorem 2, Condition C3 is satisfied with \( \mathcal{H} \) and \( \mathcal{H}(\theta) \) as defined in Section 3.4. Condition C5 is proved in Theorem 3. Again, choosing \( \mathcal{H}' = \mathcal{H} \) as defined in Section 3.4, the conditions in Lemma 3 hold where we used (B.3) and the envelope in C5 which implies C4. Since the Conditions C2 and C7 are the same as A11 and A10, we need to verify C6. Due to Theorem 1 and Theorem 2, choosing \( \rho_n = \delta_n n^{-1/4} \) for a suitable sequence \( \delta_n = o(1) \), we have
\[ \sup_{\theta \in \Theta, \tilde{h} \in \mathcal{H}(\theta)} \|\mathbb{E}[\tilde{\psi}((Y, X), \theta, h_0(\theta))] - \mathbb{E}[\psi((Y, X), \theta, \tilde{h}(\theta))]\| \]
\[ \leq \sup_{\theta \in \Theta, \tilde{h} \in \mathcal{H}(\theta)} \mathbb{E} \left[ \left( \psi((Y, X), \theta, \tilde{h}(\theta), X) - \psi((Y, X), \theta, h_0(\theta), X) \right)^2 \right]^{1/2} \]
\[ \leq \sup_{\theta \in \Theta, \tilde{h} \in \mathcal{H}(\theta)} C \mathbb{E} \left[ \|\tilde{h}(\theta, X) - h_0(\theta, X)\|_2 \right]^{1/2} \]
\[ \leq C \rho_n, \]

where we used Assumptions A9 (ii) and

\[
\mathbb{E} \left[ \| \hat{h}(\theta, X) - h_0(\theta, X) \|_2^2 \right] = \mathbb{E} \left[ (\hat{h}_1(\theta, X) - m_\theta(X))^2 \right] + \mathbb{E} \left[ (\hat{h}_2(\theta) - \sigma_\theta^2)^2 \right] \\
+ \mathbb{E} \left[ (\hat{h}_3(\theta, X) - m_\theta(X))^2 \right] + \mathbb{E} \left[ (\hat{h}_4(\theta) - \sigma_\theta^2)^2 \right] \\
\leq C \rho_n^2.
\]

The last inequality follows from the properties of \( \hat{H} \) and Condition A8. We have

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

and

\[
\mathbb{E} \left[ (\hat{h}_2(\theta) - \sigma_\theta^2)^2 \right] \leq C \rho_n^2
\]

due to the bounded empirical sparse eigenvalue. The same holds for the two remaining terms with an analogous argument. Therefore, Condition C6 (i) holds. In the following, we take the supremum over all \( \theta \) with \( |\theta - \theta_0| \leq C \rho_n \) and \( \hat{h} \in \hat{H}(\theta) \), meaning

\[
\sup_{\theta : |\theta - \theta_0| \leq C \rho_n, \hat{h} \in \hat{H}(\theta)} \equiv \sup_{|\theta - \theta_0| \leq C \rho_n, \hat{h} \in \hat{H}(\theta)}.
\]

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

\[
\sup \mathbb{E} \left[ \left( \psi((Y, X), \theta, \hat{h}(\theta, X)) - \psi((Y, X), \theta_0, h_0(\theta_0, X)) \right)^2 \right]^{1/2} \\
= \sup \mathbb{E} \left[ \left( \psi((Y, X), \theta, \hat{h}(\theta, X)) - \psi((Y, X), \theta, h_0(\theta, X)) \right) + \psi((Y, X), \theta, h_0(\theta, X)) - \psi((Y, X), \theta_0, h_0(\theta_0, X)) \right]^2 \right]^{1/2} \\
\leq \sup \mathbb{E} \left[ \left( \psi((Y, X), \theta, \hat{h}(\theta, X)) - \psi((Y, X), \theta, h_0(\theta, X)) \right)^2 \right]
\]
\begin{align*}
&+ \left( \psi((Y, X), \theta, h_0(\theta, X)) - \psi((Y, X), \theta_0, h_0(\theta_0, X)) \right)^2 \\
&+ 2 \left( \psi((Y, X), \theta, \tilde{h}(\theta, X)) - \psi((Y, X), \theta, h_0(\theta, X)) \right) \\
&\left( \psi((Y, X), \theta, h_0(\theta, X)) - \psi((Y, X), \theta_0, h_0(\theta_0, X)) \right)^{1/2} \\
&\leq \sup C \left( \mathbb{E} \left[ \|\tilde{h}(\theta, X) - h_0(\theta, X)\|_2^2 \right] + |\theta - \theta_0|^2 \\
&+ |\theta - \theta_0| \sqrt{\mathbb{E} \left[ \|\tilde{h}(\theta, X) - h_0(\theta, X)\|_2^2 \right]} \right)^{1/2} \\
&\leq C \rho_n.
\end{align*}

Due to the growth condition in A2, we have

r_n s^{1/2} \log \left( \frac{p \vee n}{r_n} \right)^{1/2} + n^{-1/2} + \frac{3}{4} s \log \left( \frac{p \vee n}{r_n} \right) \\
\leq \delta_n^{-1} \left( r_n s^{1/2} \log(p \vee n)^{1/2} + n^{-1/2} + \frac{3}{4} s \log(p \vee n) \right) \\
\leq \delta_n^{-1} \left( n^{-1/4} s^{1/2} \log(p \vee n)^{1/2} + n^{-1/2} + \frac{1}{4} s \log(p \vee n) \right) = o(1)

for a sequence \( \delta_n \) approaching zero at a polynomial speed in \( n \). Hence, Assumption C6 (ii) follows. Condition C6 (iii) follows directly from A9 (iii):

\begin{align*}
&\sup_{r \in (0, 1)} \sup_{\theta : |\theta - \theta_0| \leq C \rho_n, \tilde{h} \in \tilde{H}(\theta)} \left| \frac{\partial^2 }{\partial \theta^2} \mathbb{E} \left[ \psi((Y, X), \theta_0 + r(\theta - \theta_0), h_0 + r(\tilde{h} - h_0)) \right] \right| \\
&\leq C \rho_n^2 = o(n^{-1/2}).
\end{align*}

PROOF OF LEMMA 1.

REMARK A.1. The proof for Box-Cox-Transformations is from [9], who refer to [8]. It heavily relies on the properties of the dual density from [1]. We give a detailed version of the proof of [8] and extend the idea to the class of derivatives and Yeo-Johnson Power Transformations.

Since adding a single function to a class of functions can increase the VC
index at most by one, we exclude the parameter \( \theta = 0 \) from the proof and restrict the class to
\[
\mathcal{F}_1' = \{ \Lambda_\theta(\cdot) | \theta \in \mathbb{R} \setminus \{0\} \}.
\]
At first, recall that \( \mathcal{F}_1' \) is a VC class if and only if the between graph set
\[
\mathcal{C} := \{ C_\theta | \theta \in \mathbb{R} \setminus \{0\} \}
\]
with \( C_\theta := \{(x,t) \in \mathbb{R}^+ \times \mathbb{R} | 0 \leq t \leq \Lambda_\theta(x) \) or \( \Lambda_\theta(x) \leq t \leq 0 \} \) is a VC class (cf. [9], page 152). We now consider the dual class (cf. [1]) of \( \mathcal{C} \) given by
\[
\mathcal{D} := \{ D(x,t) | (x,t) \in \mathbb{R}^+ \times \mathbb{R} \}
\]
with
\[
D(x,t) := \{ \theta \in \mathbb{R} \setminus \{0\} | (x,t) \in C_\theta \}
\]
\[
= \{ \theta \in \mathbb{R} \setminus \{0\} | 0 \leq t \leq \dot{\Lambda}_\theta(x) \) or \( \dot{\Lambda}_\theta(x) \leq t \leq 0 \} .
\]
For the derivative of \( \Lambda_\theta(x) \), we have
\[
\dot{\Lambda}_\theta(x) = \frac{1}{\theta^2} \left( (\theta \log(x) - 1) x^\theta + 1 \right) \geq 0
\]
\[
\Leftrightarrow \quad (\theta \log(x) - 1) x^\theta \geq -1
\]
\[
\Leftrightarrow \quad \log(x^\theta) \geq \frac{x^\theta - 1}{x^\theta}
\]
which is true for all \( x \) and \( \theta \). Since \( \Lambda_\theta(x) \) is continuous and monotone increasing in \( \theta \) the set \( D(x,t) \) is the union of at most two intervals in \( \mathbb{R} \setminus \{0\} \) and therefore \( \mathcal{D} \) is a VC class which by Proposition 2.12 in [1] implies that \( \mathcal{C} \) is a VC class. By the same argument as above, we have to prove that
\[
\mathcal{D}' = \{ D'(x,t) | (x,t) \in \mathbb{R}^+ \times \mathbb{R} \}
\]
is a VC class with
\[
D'(x,t) := \{ \theta \in \mathbb{R} \setminus \{0\} | 0 \leq t \leq \dot{\Lambda}_\theta(x) \}
\]
since \( \dot{\Lambda}_\theta(x) \geq 0 \). The second derivative with respect to \( \theta \) is given by
\[
\ddot{\Lambda}_\theta(x) = \frac{1}{\theta^3} \left( \left( \log(x^\theta) - 1 \right)^2 x^\theta + x^\theta - 2 \right).
\]
The case \( x = 1 \) directly implies \( \dot{\Lambda}_\theta(x) = 0 \). We substitute \( z = x^\theta \) in \( f(x^\theta) \) and notice that

\[
f'(z) = (\log(z) - 1)^2 + 2(\log(z) - 1) + 1 = (\log(z))^2 \geq 0.
\]

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

\[
\begin{align*}
 x > 1, \ \theta > 0 \\
 0 < x < 1, \ \theta < 0 \\
 x > 1, \ \theta < 0 \\
 0 < x < 1, \ \theta > 0
\end{align*}
\]

\( \Rightarrow \)

\[
\begin{align*}
 x^\theta > 1 \\
 0 < x^\theta < 1
\end{align*}
\]

and the coefficient \( 1/\theta^3 \) imply

\[
\dot{\Lambda}_\theta(x) = \begin{cases} 
\geq 0 & \text{for } x \geq 1 \\
< 0 & \text{for } x < 1.
\end{cases}
\]

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

\[
\tilde{F}_2 = \{ \Phi_\theta(\cdot) | \theta \in \mathbb{R} \setminus \{0,2\} \},
\]

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

\[
\tilde{C} := \{ \tilde{C}_\theta | \theta \in \mathbb{R} \setminus \{0,2\} \}
\]

with \( \tilde{C}_\theta := \{ (x,t) \in \mathbb{R} \times \mathbb{R} | 0 \leq t \leq \Phi_\theta(x) \text{ or } \Phi_\theta(x) \leq t \leq 0 \} \). Since \( \Phi_\theta(x) \geq 0 \) for \( x \geq 0 \) and \( \Phi_\theta(x) < 0 \) for \( x < 0 \), we have

\[
\tilde{C}_\theta := \{ (x,t) \in \mathbb{R} \times \mathbb{R} | 0 \leq t \leq \Phi_\theta(x) \text{ or } \Phi_\theta(x) \leq t \leq 0 \}
= \{ (x,t) \in \mathbb{R}^+_0 \times \mathbb{R} | 0 \leq t \leq \Phi_\theta(x) \} \cup \{ (x,t) \in \mathbb{R}^- \times \mathbb{R} | \Phi_\theta(x) \leq t \leq 0 \}
= \{ (x,t) \in \mathbb{R}^+_0 \times \mathbb{R} | 0 \leq t \leq \Lambda_\theta(x+1) \}
\quad \cup \{ (x,t) \in \mathbb{R}^- \times \mathbb{R} | -\Lambda_{2-\theta}(1-x) \leq t \leq 0 \}
= \tilde{C}_{\theta,1} \cup \tilde{C}_{\theta,2}.
\]

The sets

\[
\tilde{C}_1 := \{ \tilde{C}_{\theta,1} | \theta \in \mathbb{R} \setminus \{0,2\} \} \text{ and } \tilde{C}_2 := \{ \tilde{C}_{\theta,2} | \theta \in \mathbb{R} \setminus \{0,2\} \}
\]
are VC classes as shown above. By Lemma 2.6.17 (iii) from [9],
\[ \hat{C}_1 \cup \hat{C}_2 = \{ \hat{C}_{\theta,1} \cup \hat{C}_{\theta,2} | \hat{C}_{\theta,1} \in \hat{C}_1, \hat{C}_{\theta,2} \in \hat{C}_2 \} \]
is a VC class which contains \( \hat{C} \). The proof for the class of derivatives can be shown analogously. ■

APPENDIX B: UNIFORM CONVERGENCE RATES FOR THE LASSO

In this chapter, we consider the high-dimensional transformation model introduced in Section 2. For every \( \theta \in \Theta \), we assume a linear model
\[ \Lambda_\theta(Y) = X^T \beta_\theta + \varepsilon_\theta \]
with \( E[\varepsilon_\theta | X] = 0 \). We provide results for uniform estimation rates in \( \theta \) of the lasso estimator
\[ \hat{\beta}_\theta := \arg \max_\beta E_n[(\Lambda_\theta(Y) - X^T \beta)^2] + \frac{\lambda}{n}||\Psi_\theta \beta||_1, \]
where the penalty term is given by \( \lambda = c\sqrt{n} \Phi^{-1}(1 - \gamma/(2p)) \) with suitable constants \( c > 1 \) and \( \gamma \in [1/n, 1 \log(n)] \). The penalty loadings \( \Psi_\theta \) are estimated as in the Algorithm 6.1 of [3]. To provide our results in Theorem 5, we impose the following assumptions.

Assumptions B1-B7.

The following assumptions hold uniformly in \( n \geq n_0 \) and \( P \in P_n \):

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

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

B3 It holds \( \sup_{j=1,\ldots,p_n} |X_j| \leq C < \infty \) a.s. for a constant \( C \) independent of \( n \).

B4 Uniformly in \( \theta \), the conditional variance of the error term is bounded
\[ 0 < c \leq \inf_{\theta \in \Theta} E[\varepsilon_\theta^2 | X] \leq \sup_{\theta \in \Theta} E[\varepsilon_\theta^2 | X] \leq C < \infty. \]

B5 The transformations are measurable and the class of transformations \( \mathcal{F}_\Lambda := \{ \Lambda_\theta(\cdot) | \theta \in \Theta \} \) has VC index \( C_\Lambda \) and an envelope \( F_\Lambda \) with \( E[F_\Lambda(Y)^2] < \infty \).

B6 The transformations are differentiable with respect to \( \theta \) and it holds \( \sup_{\theta \in \Theta} E[(\hat{\Lambda}_\theta(Y))^2] \leq C < \infty \).
B7 With probability $1 - o(1)$, the empirical minimum and maximum sparse eigenvalues are bounded from zero and above, namely

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

Theorem 5. Under Assumptions B1-B7 above, uniformly for all $P \in \mathcal{P}_n$ with probability $1 - o(1)$, it holds:

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

Proof. We verify the Assumption 6.1 from Belloni et al. (2017) [3]. Due to Assumptions B1 and B2, the Condition 6.1(i) is satisfied. Needless to say, the Assumption 6.1(ii) holds for a compact $\Theta \subset \mathbb{R}$. Remark that Assumptions B4 and B5 imply the Condition 6.1 (iii). Due to Assumption B3, the conditions in 6.1(iv)(a) are satisfied and we can omit the $X$ in the technical conditions in 6.1(iv)(b). The eigenvalue Condition 6.1(iv)(c) is the same as in B7. Therefore, we have to show with probability $1 - o(1)$:

1. $\sup_{\theta \in \Theta} (|E_n - E|_{\mathcal{P}_n}^2 \vee |E_n - E|_{\mathcal{P}_n}^2)^{1/2} = O(\delta_n)$
2. $n^{1/2} \sup_{|\theta - \theta'| \leq 1/n} |E_n [\varepsilon_\theta - \varepsilon_{\theta'}]| = O(\delta_n)$ and
3. $\log(p \vee n)^{1/2} \sup_{|\theta - \theta'| \leq 1/n} E_n [(\varepsilon_\theta - \varepsilon_{\theta'})^2]^{1/2} = O(\delta_n)$.

Since $\mathcal{F}_\lambda$ is a VC class of functions with VC index $C_\lambda$, we have by Theorem 2.6.7 in [9]

$$\log N(\varepsilon\|F\|_{Q,2}, \mathcal{F}_\lambda, L_2(Q)) \leq C'_\lambda \log(C''_\lambda/\varepsilon),$$

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

$$\mathcal{F}'_\lambda := \{ E[\lambda_\theta(\cdot)|X] | \theta \in \Theta \}$$

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

$$\mathcal{E}'_\lambda := \{ (\lambda_\theta(\cdot) - E[\lambda_\theta(\cdot)|X])^2 | \theta \in \Theta \}. $$
with envelope \((F_\Lambda + F'_\Lambda)^2\). By Lemma L.2 in the supplement to [3], we have
\[(B.2)\]
where the supremum on the left-hand side is taken over all finitely discrete probability measures \(Q'\) with
\[\|F'_\Lambda\|^2_{Q',2} := E_{Q'}[(E[F_\Lambda(Y)|X])^2] \equiv E_{Q'}[(E[F_\Lambda|X])^2] < \infty.\]
Since \(E^2 \subset (F_\Lambda - F'_\Lambda)^2\), it follows by Lemma L.1 in the supplement to [3] for any \(\hat{Q}\) with \(E_{\hat{Q}}[(F_\Lambda + F'_\Lambda)^4] < \infty\) and \(0 < \epsilon \leq 1\)
\[
\begin{align*}
\log N(\epsilon\|F_\Lambda + F'_\Lambda\|_{Q,2}, E^2, L_2(\hat{Q})) & \leq 2 \log N\left(\frac{\epsilon}{2}\|F_\Lambda + F'_\Lambda\|_{Q,2}, F_\Lambda - F'_\Lambda, L_2(\hat{Q})\right) \\
& \leq 2 \log N\left(\frac{\epsilon}{4}\|F_\Lambda\|_{Q,2}, F_\Lambda, L_2(\hat{Q})\right) + 2 \log N\left(\frac{\epsilon}{4}\|F'_\Lambda\|_{Q,2}, F'_\Lambda, L_2(\hat{Q})\right) \\
& \leq 2 \sup_{Q'} \log N\left(\frac{\epsilon}{4}\|F'_\Lambda\|_{Q',2}, F'_\Lambda, L_2(Q')\right) \\
& \quad + 2 \sup_{Q'} \log N\left(\frac{\epsilon}{4}\|F'_\Lambda\|_{Q',2}, F'_\Lambda, L_2(Q')\right) \\
& \leq 4 \sup_{Q'} \log N\left(\frac{\epsilon^2}{256}\|F_\Lambda\|_{Q,2}, F_\Lambda, L_2(Q)\right),
\end{align*}
\]
where we used \((B.2)\) in the last step. We conclude
\[
\begin{align*}
\log N(\epsilon\|F_\Lambda + F'_\Lambda\|_{Q,2}, E^2, L_2(\hat{Q})) & \leq 4C'_\Lambda \log(256C'_\Lambda/\epsilon^2) \\
& = 16C'_\Lambda \log(16\sqrt{C'_\Lambda}/\epsilon)
\end{align*}
\]
by \((B.1)\). Under \(B5\) for all \(r \in \{1, 2, 3\}\), it holds
\[
E[F^{2r}] = E[(E[F_\Lambda|X])^{2r}] \leq E\left[E\left[(F_\Lambda)^{2r}|X\right]\right] = E[F^{2r}] < \infty,
\]
which implies
\[
\begin{align*}
E[(F_\Lambda + F'_\Lambda)^4] = E[F^{4}] + E[F^{4}] + 6 E[F^{2}F^{2}] \\
\quad \leq E[F^{4}] \leq \sqrt{E[F^{4}]E[F^{4}]} \\
\quad + 4 E[F^{2}F^{2}] + 4 E[F^{2}F^{2}] \\
\quad \leq \sqrt{E[F^{2}]E[F^{2}]} \leq \sqrt{E[F^{2}]E[F^{2}]}
\end{align*}
\]
Remark that
\[
\mathbb{E} \left[ \sup_{\theta \in \Theta} \varepsilon_\theta^2 \right] \leq \mathbb{E} \left[ (F_\Lambda + F'_\Lambda)^2 \right] \leq C < \infty.
\]
We have
\[
\sqrt{n} \sup_{\theta \in \Theta} |(\mathbb{E}_n - \mathbb{E})\varepsilon_\theta^2| = \sup_{g \in \mathcal{E}_\Lambda^2} |G_n(g)|.
\]
For every \(\sigma_C^2\) with \(\sup_{g \in \mathcal{E}_\Lambda^2} \mathbb{E}[g^2] \leq \sigma_C^2 \leq \mathbb{E}[(F_\Lambda + F'_\Lambda)^4] := G_1 < \infty\) and universal constants \(K\) and \(K_2\) with probability not less than \(1 - \Omega(1)\), it holds
\[
\sup_{g \in \mathcal{E}_\Lambda^2} |G_n(g)| 
\leq 2K \left( S\sigma_C^2 \log(AG_1^{1/2}/\sigma_C) \right)^{1/2} + SG_1^{1/2} \log(AG_1^{1/2}/\sigma_C) 
+ K_2(\sigma_C \log(n)^{1/2} + G_1^{1/2} \log(n)) 
= O(\log(n))
\]
by Lemma 1 in [4] with \(q = 2, t = \log(n), A = 16\sqrt{C_\Lambda} \) and \(S = 16C'\).
Therefore, it follows with probability \(1 - \Omega(1)\)
\[
\sup_{\theta \in \Theta} |(\mathbb{E}_n - \mathbb{E})\varepsilon_\theta^2| = O \left( \frac{\log(n)}{\sqrt{n}} \right).
\]
Analogously, it can be shown with probability \(1 - \Omega(1)\)
\[
\sup_{\theta \in \Theta} |(\mathbb{E}_n - \mathbb{E})\Lambda_\theta(Y)^2| = O \left( \frac{\log(n)}{\sqrt{n}} \right).
\]
(1) follows by Assumption B2.

Further, we have
\[
\sup_{|\theta - \theta'| \leq 1/n} |\mathbb{E}_n [\varepsilon_\theta - \varepsilon_\theta']| = \sup_{|\theta - \theta'| \leq 1/n} \frac{1}{\sqrt{n}} |G_n(\varepsilon_\theta - \varepsilon_\theta')| 
\]
Define \(\mathcal{E}_\Lambda' := \{\varepsilon_\theta - \varepsilon_{\theta'} | \theta, \theta' \in \Theta\}\) and \(\mathcal{E}_\Lambda := \{\varepsilon_\theta = (\Lambda_\theta(\cdot) - \mathbb{E}[\Lambda_\theta(\cdot)|X]) | \theta \in \Theta\}\). Using the same argument as above, we obtain
\[
\log N(\varepsilon\|2(F_\Lambda + F'_\Lambda)\|_{\tilde{Q},2}, \mathcal{E}_\Lambda', L_2(\tilde{Q}))
\]
\[ \leq \log N \left( \frac{\varepsilon}{2} \|2F_\Lambda\|_{Q,2}, \mathcal{F}_\Lambda, L_2(\tilde{Q}) \right) + \log N \left( \frac{\varepsilon}{2} \|2F'_\Lambda\|_{\bar{Q},2}, \mathcal{F}'_\Lambda, L_2(\bar{Q}) \right) \]

\[ \leq \sup_Q \log N \left( \varepsilon \|F_\Lambda\|_{Q,2}, \mathcal{F}_\Lambda, L_2(Q) \right) + \sup_Q \log N \left( \varepsilon \|F'_\Lambda\|_{Q',2}, \mathcal{F}'_\Lambda, L_2(Q') \right) \]

\[ \leq 2 \sup_Q \log N \left( \left( \frac{\varepsilon}{4} \right)^2 \|F_\Lambda\|_{Q,2}, \mathcal{F}_\Lambda, L_2(Q) \right) \]

\[ \leq 4C'_\Lambda \log(4\sqrt{C'_\Lambda}/\varepsilon). \]

Since

\[ \mathcal{E}'_\Lambda := \{ \varepsilon_\theta - \varepsilon_{\theta'} \mid \theta, \theta' \in \Theta, |\theta - \theta'| \leq 1/n \} \subset \mathcal{E}'_\Lambda, \]

we can use Lemma 1 in [4] again, since we obtain the same envelope and bound for the entropy as for \( \mathcal{E}'_\Lambda \).

We achieve for every \( \sigma_n^2 \) with \( \sup_{g \in E'_\Lambda} \mathbb{E}[g^2] \leq \sigma_n^2 \leq \mathbb{E}[4(F_\Lambda + F'_\Lambda)^2] := G_2 \) and universal constants \( K \) and \( K_2 \) with probability at least \( 1 - (1/\log(n)) \)

\[ \sup_{g \in E'_\Lambda} |G_n(g)| \]

\[ \leq 2K \left[ \left( S\sigma_n^2 \log(AG_2^{1/2}/\sigma_n) \right)^{1/2} + n^{-\frac{1}{4}} S^2 \mathbb{E}((F_\Lambda + F'_\Lambda)^4)^{1/4} \log(A\sqrt{G_2^{1/2}/\sigma_n}) \right] \]

\[ + K_2(\sigma_n \log(n)^{1/2} + n^{-\frac{1}{4}} 2\mathbb{E}((F_\Lambda + F'_\Lambda)^4)^{1/4} \log(n)) \]

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

We have

\[ \sup_{|\theta - \theta'| \leq \frac{1}{n}} \mathbb{E}[(\varepsilon_\theta - \varepsilon_{\theta'})^2] \]

\[ = \sup_{|\theta - \theta'| \leq \frac{1}{n}} \mathbb{E} \left[ \left( \Lambda_\theta(Y) - \mathbb{E}[\Lambda_\theta(Y)|X] - \Lambda_{\theta'}(Y) + \mathbb{E}[\Lambda_{\theta'}(Y)|X] \right)^2 \right] \]

\[ = \sup_{|\theta - \theta'| \leq \frac{1}{n}} \mathbb{E} \left[ \left( \left( \Lambda_\theta(Y) - \Lambda_{\theta'}(Y) \right) - \left( \mathbb{E}[\Lambda_\theta(Y)|X] - \mathbb{E}[\Lambda_{\theta'}(Y)|X] \right) \right)^2 \right] \]

\[ = \sup_{|\theta - \theta'| \leq \frac{1}{n}} \left( \mathbb{E} \left[ (\Lambda_\theta(Y) - \Lambda_{\theta'}(Y))^2 \right] + \mathbb{E} \left[ (\Lambda_\theta(Y) - \Lambda_{\theta'}(Y))^2 | X \right] \right) \]

\[ \leq \mathbb{E} \left[ (\Lambda_\theta(Y) - \Lambda_{\theta'}(Y))^2 | X \right] \]

\[ - 2 \mathbb{E} \left[ (\Lambda_\theta(Y) - \Lambda_{\theta'}(Y)) (\Lambda_\theta(Y) - \Lambda_{\theta'}(Y)) | X \right] \]

\[ \leq \sup_{|\theta - \theta'| \leq \frac{1}{n}} 2 \mathbb{E} \left[ (\Lambda_\theta(Y) - \Lambda_{\theta'}(Y))^2 \right] \]
\[
\begin{align*}
\sup_{|\theta - \theta'| \leq 1/n} 2\mathbb{E} \left[ (\theta - \theta')^2 (\dot{A}_\theta(Y))^2 \right] \\
\leq \frac{2}{n^2} \sup_{\theta \in \Theta} \mathbb{E} \left[ (\dot{A}_\theta(Y))^2 \right] = O(n^{-2}).
\end{align*}
\]

Therefore, we can choose \( \sigma_n^2 = O(n^{-2}) \) and obtain with probability \( 1 - o(1) \)
\[
n^{1/2} \sup_{|\theta - \theta'| \leq 1/n} |\mathbb{E}_n [\varepsilon_{\theta} - \varepsilon_{\theta'}]| = \sup_{|\theta - \theta'| \leq 1/n} |G_n(\varepsilon_{\theta} - \varepsilon_{\theta'})| \\
= \sup_{g \in E_n} |G_n(g)| \\
= O \left( \frac{\log(n)}{n^{1/4}} \right) = O(\delta_n).
\]

For (3), we can use the same arguments as above and we remark
\[
\sup_{|\theta - \theta'| \leq 1/n} \mathbb{E}_n \left[ (\varepsilon_{\theta} - \varepsilon_{\theta'})^2 \right] \leq \sup_{|\theta - \theta'| \leq 1/n} \mathbb{E} \left[ (\varepsilon_{\theta} - \varepsilon_{\theta'})^2 \right] \\
+ \sup_{|\theta - \theta'| \leq 1/n} \left( \mathbb{E}_n \left[ (\varepsilon_{\theta} - \varepsilon_{\theta'})^2 \right] - \mathbb{E} \left[ (\varepsilon_{\theta} - \varepsilon_{\theta'})^2 \right] \right) \\
\leq \sup_{g \in E_n^{2'}} \frac{1}{\sqrt{n}} G_n(g) + O(n^{-2})
\]

with \( E_n^{2'} := \{(\varepsilon_{\theta} - \varepsilon_{\theta'})^2 | \theta, \theta' \in \Theta \} \). The entropy of this class is bounded by
\[
\log N(\varepsilon \| 4(F_\Lambda + F'_\Lambda)^2 \|_{Q,2}, E_n^{2'}, L_2(\hat{Q})) \\
\leq 2 \log N \left( \frac{\varepsilon}{2} \| 4(F_\Lambda + F'_\Lambda) \|_{Q,2}, E_n^{2'}, L_2(\hat{Q}) \right) \\
\leq 2 \log N \left( \frac{\varepsilon}{4} \| 4F_\Lambda \|_{Q,2}, F_{\Lambda}, L_2(\hat{Q}) \right) + 2 \log N \left( \frac{\varepsilon}{4} \| 4F'_\Lambda \|_{Q,2}, F'_{\Lambda}, L_2(\hat{Q}) \right) \\
\leq 2 \sup_{Q} \log N (\varepsilon \| F_\Lambda \|_{Q,2}, F_{\Lambda}, L_2(Q)) + 2 \sup_{Q'} \log N (\varepsilon \| F'_\Lambda \|_{Q',2}, F'_{\Lambda}, L_2(Q')) \\
\leq 4 \sup_{Q} \log N \left( \frac{\varepsilon}{4} \| F_\Lambda \|_{Q,2}, F_{\Lambda}, L_2(Q) \right) \\
\leq 8C' \log \left( \frac{4C'' \varepsilon}{\delta} \right).
\]

For every \( \sigma_n^2 \) with \( \sup_{g \in E_n^{2'}} \mathbb{E}[g^2] \leq \sigma_n^2 \leq \mathbb{E}[16(F_\Lambda + F'_\Lambda)^4] := G_3 < \infty \) and universal constants \( K \) and \( K_2 \) with probability not less than \( 1 - (1/\log(n)) \),
it holds
\[
\sup_{g \in \mathcal{E}'} |G_n(g)| \leq 2K \left[ \left( S\sigma_C^2 \log(AG_3^{1/2}/\sigma_C) \right)^{1/2} + S G_3^{1/2} \log(AG_3^{1/2}/\sigma_C) \right] + K_2(\sigma_C \log(n))^{1/2} + G_3^{1/2} \log(n)) = O(\log(n))
\]
by Lemma 1 in [4] with \( q = 2, t = \log(n), A = 4\sqrt{C_A'}, S = 8C_A' \).

We conclude
\[
\sup_{|\theta - \theta'| \leq 1/n} \mathbb{E}_n \left[ (\varepsilon_{\theta} - \varepsilon_{\theta'})^2 \right] = O\left( \frac{\log n}{\sqrt{n}} \right)
\]
and therefore
\[
\log(p \lor n)^{1/2} \sup_{|\theta - \theta'| \leq 1/n} \mathbb{E}_n \left[ (\varepsilon_{\theta} - \varepsilon_{\theta'})^2 \right]^{1/2} = O(\delta_n)
\]
since \( n^{-1/3} \log(p \lor n) \leq \delta_n \).

APPENDIX C: INFERENCE IN GENERAL Z-ESTIMATION PROBLEMS

In this chapter, we consider a general Z-problem, where target parameter \( \theta_0 \) obeys the moment condition
\[
\mathbb{E} \left[ \psi((Y, X), \theta_0, h_0(\theta_0, X)) \right] = 0.
\]

In this setting, the unknown nuisance function
\[
h_0(\theta, X) = (h_{0,1}(\theta, X), \ldots, h_{0,m}(\theta, X)) \in \mathcal{H}
\]
may depend on the target parameter \( \theta \). The central theorem is a statement about the asymptotic distribution of an estimate which solves
\[
(C.1) \quad \left| \mathbb{E}_n \left[ \psi((Y, X), \hat{\theta}, h_0(\hat{\theta}, X)) \right] \right| = \inf_{\theta \in \Theta} \left| \mathbb{E}_n \left[ \psi((Y, X), \theta, h_0(\theta, X)) \right] \right| + \epsilon_n,
\]
where \( \epsilon_n = o(n^{-1/2}) \) is the numerical tolerance. We need a more general form of the conditions in Section 3.
Assumptions C1-C7.

The following assumptions hold uniformly in \( n \geq n_0 \) and \( P \in \mathcal{P}_n \):

**C1** The true parameter \( \theta_0 \) obeys the moment condition

\[
\mathbb{E} \left[ \psi((Y, X), \theta_0, h_0) \right] = 0.
\]

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

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

**C4** For all \( \tilde{h} \in \tilde{\mathcal{H}} \), the score \( \psi \) obeys the Neyman orthogonality property

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

**C5** For all \( \theta \in \Theta \), the class of functions

\[
\Psi(\theta) = \left\{ (y, x) \mapsto \psi((y, x), \theta, \tilde{h}(\theta, x), \tilde{h} \in \tilde{\mathcal{H}}(\theta)) \right\}
\]

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

\[
\mathbb{E}\left[ |\tilde{\psi}(Y, X)|^q \right] \leq C.
\]

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

\[
\sup_{Q} \log N(\varepsilon ||\tilde{\psi}||_{Q,2}, \Psi(\theta), L_2(Q)) \leq C_1 \log \left( \frac{C_2 (p \lor n)}{\varepsilon} \right)
\]

with \( C_1 \) and \( C_2 \) being independent from \( \theta \).

**C6** (i) For a positive sequence \( \rho_n \downarrow 0 \) with

\[
n^{-1/2} \left( s^{1/2} \log (p \lor n)^{1/2} + n^{-1/2} + s^{1/2} \log (p \lor n) \right) = O(\rho_n),
\]

we have

\[
\sup_{\theta \in \Theta, \tilde{h} \in \tilde{\mathcal{H}}(\theta)} \left| \mathbb{E}[\psi((Y, X)), \theta, h_0(\theta, X)] - \mathbb{E}[\psi((Y, X)), \theta, \tilde{h}(\theta, X)] \right| \leq C \rho_n.
\]
(ii) We define
\[
\sup E \left[ \psi((Y, X), \theta, \tilde{h}(\theta, X)) - \psi((Y, X), \theta_0, h_0(\theta_0, X)) \right]^2 \equiv r_n,
\]
where the supremum is taken over all \( \theta \) with \( |\theta - \theta_0| \leq C \rho_n \) and \( \tilde{h} \in \tilde{H} \), meaning
\[
\sup \equiv \sup_{\theta:|\theta - \theta_0| \leq C \rho_n, \tilde{h} \in \tilde{H}(\theta)}.
\]
and it holds \( R_n := r_n s^2 \log \left( \frac{(n \vee n)}{r_n} \right)^{\frac{1}{2}} + n^{-\frac{1}{2}} + \frac{1}{q} s \log \left( \frac{(n \vee n)}{r_n} \right) = o(1) \) with \( q \) as in Assumption C5.

(iii) It holds
\[
R'_n := \sup \left| \partial^2_{\psi} E \left[ \psi((Y, X), \theta_0 + r(\theta - \theta_0), h_0 + r(\tilde{h} - h_0)) \right] \right|
\]
= \( o(n^{-1/2}) \),
where \( \sup \equiv \sup_{r \in (0, 1), \theta:|\theta - \theta_0| \leq C \rho_n, \tilde{h} \in \tilde{H}(\theta)} \).

**C7** For \( h \in \tilde{H} \), the function
\[
\theta \mapsto E \left[ \psi((Y, X), \theta, h(\theta, X)) \right]
\]
is differentiable in a neighborhood of \( \theta_0 \) and, for all \( \theta \in \Theta \), the identification relation
\[
2 |E[\psi((Y, X)), \theta, h_0(\theta, X)]| \geq |\Gamma(\theta - \theta_0)| \wedge c_0
\]
is satisfied with \( \Gamma := \partial_\theta E \left[ \psi((Y, X), \theta_0, h_0(\theta_0, X)) \right] > c_1 \).
Since the nuisance functions depend on the target parameter, the conditions ensure that they can be estimated uniformly over all \( \theta \) with a sufficiently fast rate.

**Theorem 6.** Under the Assumptions C1-C7, an estimator \( \hat{\theta} \) of the form in (C.1) obeys
\[
\sqrt{n}(\hat{\theta} - \theta_0) = -\sqrt{n} \Gamma^{-1} E_n \left[ \psi((Y, X), \theta_0, h_0(\theta_0, X)) \right] + O_p(\tilde{R}_n) \xrightarrow{D} \mathcal{N}(0, \Sigma),
\]
where
\[ \tilde{R}_n := R_n + \sqrt{n}R'_n = o(1) \]
and
\[ \Sigma := \mathbb{E} \left[ \Gamma^{-2} \psi^2((Y, X), \theta_0, h_0(\theta_0, X)) \right] \]
with \( \Gamma = \partial_{\theta} \mathbb{E} \left[ \psi((Y, X), \theta_0, h_0(\theta_0, X)) \right] \).

**Remark C.1.**
The setting and the theorem is almost identical to Assumption 3.4 and Theorem 3.3 in Chernozhukov et al. (2017) [5]. Their theorem holds for dependent nuisance functions, but the entropy condition may be hard to verify in some settings:

Suppose the unknown nuisance function \( h_0 \) is a linear function of \( X \), where the coefficients \( \beta_0(\theta) (\|\beta_0(\theta)\|_0 \leq s \) for all \( \theta \) are dependent on the target parameter. If \( h_0(\theta, X) = X\beta_0(\theta) \) is estimated by the lasso estimator, the uniform covering entropy of
\[ \mathcal{F}_h := \left\{ \psi(\cdot, \theta, h(\theta, \cdot)), \theta \in \Theta \right\} \]
may not fulfill the desired condition. This is because the uniform covering entropy of the class
\[ \mathcal{H} := \left\{ h(\theta, \cdot) : \mathcal{X} \to \mathbb{R} | h(\theta, X) = \beta(\theta)X, \|\beta(\theta)\|_0 \leq s, \theta \in \Theta \right\} \]
can not be bounded by representing the class as the union over sets with a bounded VC index (see for example Belloni et al. (2014) [4]) since the indices which differ from zero may vary for each \( \theta \). In their example, the estimation of the average treatment effect, this problem does not occur because the estimated nuisance functions do not depend on the target parameter. To bypass this, we rely on a slightly different set of entropy conditions which enables us to restrict the entropy of the classes uniformly over all \( \theta \in \Theta \).

**Proof.** We are using similar arguments as in proof of Theorem 2 from Belloni, Chernozhukov and Kato (2014) [4]. We prove the theorem under an arbitrary sequence \( P = P_n \in \mathcal{P}_n \). Therefore, the dependence of \( P \) on \( n \) can be suppressed. Let \( \rho_n \) be an positive sequence converging to zero.

**Step 1.**
Let \( \hat{\theta} \) be an arbitrary estimator fulfilling \( |\hat{\theta} - \theta_0| \leq C\rho_n \) with probability \( 1 - o(1) \). We aim to prove that with probability \( 1 - o(1) \)
\[ \mathbb{E}_n \left[ \psi((Y, X), \hat{\theta}, \hat{h}(\hat{\theta}, X)) \right] \]
By Assumption C1, we can expand the term
\[
\begin{align*}
\mathbb{E}_n \left[ \psi((Y, X), \hat{\theta}, \hat{h}(\hat{\theta}, X)) \right] \\
&= \mathbb{E}_n \left[ \psi((Y, X), \hat{\theta}, \hat{h}(\hat{\theta}, X)) \right] + \mathbb{E} \left[ \psi((Y, X), \theta_0, h_0(\theta_0, X)) \right] \\
&+ \mathbb{E}_n \left[ \psi((Y, X), \hat{\theta}, \hat{h}(\hat{\theta}, X)) \right] - \mathbb{E}_n \left[ \psi((Y, X), \theta_0, h_0(\theta_0, X)) \right] \\
&+ \mathbb{E} \left[ \psi((Y, X), \hat{\theta}, \hat{h}(\hat{\theta}, X)) \right] - \mathbb{E} \left[ \psi((Y, X), \hat{\theta}, \hat{h}(\hat{\theta}, X)) \right] \\
&= \mathbb{E}_n \left[ \psi((Y, X), \theta_0, h_0(\theta_0, X)) \right] + \mathbb{E} \left[ \psi((Y, X), \theta_0, h_0(\theta_0, X)) \right] \\
&+ \mathbb{E}_n \left[ \psi((Y, X), \theta_0, h_0(\theta_0, X)) \right] - \mathbb{E} \left[ \psi((Y, X), \theta_0, h_0(\theta_0, X)) \right] \\
&- \left( \mathbb{E}_n \left[ \psi((Y, X), \theta_0, h_0(\theta_0, X)) \right] - \mathbb{E} \left[ \psi((Y, X), \theta_0, h_0(\theta_0, X)) \right] \right) \\
&= I + II + III - IV.
\end{align*}
\]
with
\[ \Psi'(\theta) = \left\{ (y, x) \mapsto \psi((y, x), \theta, \tilde{h}(\theta, x)) - \psi((y, x), \theta_0, h_0(\theta_0, x)), \tilde{h} \in \tilde{\mathcal{H}}(\theta) \right\} \]
and envelope \(2\tilde{\psi}\). Here, we used Assumption C5 and that with probability \(1 - o(1)\) we have \(\tilde{h}(\theta, X), h_0(\theta, X) \in \tilde{\mathcal{H}}(\theta)\) for all \(\theta \in \Theta\) by Assumption C3.

Recall that
\[ \sup_{\theta} \log N(\varepsilon ||2\tilde{\psi}||_{Q,2}, \Psi'(\theta), L_2(Q)) \leq C_1 s \log \left( \frac{C_2 (p \vee n)}{\varepsilon} \right) \]
for constants \(C_1\) and \(C_2\) being independent from \(\theta\). We want to apply Lemma 1 from Belloni, Chernozhukov and Kato (2014) [4]. By Assumption C6, we have
\[ \sup_{\theta : |\theta - \theta_0| \leq C \rho_n, f \in \Psi'(\theta)} \mathbb{E} \left[ f^2((Y, X)) \right] = \sup_{\theta : |\theta - \theta_0| \leq C \rho_n, \tilde{h} \in \tilde{\mathcal{H}}(\theta)} \mathbb{E} \left[ \left( \psi((Y, X), \theta, \tilde{h}(\theta, X)) - \psi((Y, X), \theta_0, h_0(\theta_0, X)) \right)^2 \right] =: r_n^2 \]
with \(r_n s^{\frac{1}{2}} \log \left( \frac{p \vee n}{\rho_n} \right)^{\frac{1}{2}} + n^{-\frac{1}{2} + \frac{1}{q}} s \log \left( \frac{p \vee n}{\rho_n} \right) = o(1)\). Choosing \(\sigma_n^2 = r_n^2\) and \(\max_{q \in \{2, 4\}} \mathbb{E}[(\tilde{\psi}(Y, X))^q] \leq C\), the first inequality of Lemma 1 in [4] implies
\[ \mathbb{E} \left[ \sup_{f \in \Psi'(\theta)} |G_n(f)| \right] \leq K \left( C_1 s \sigma_n^2 \log \left( \frac{C_2 (p \vee n) C_1^2}{\sigma_n} \right) \right)^{\frac{1}{2}} + n^{-\frac{1}{2} + \frac{1}{q}} C_1 s C_1^{\frac{1}{2}} \log \left( \frac{C_2 (p \vee n) C_1^2}{\sigma_n} \right) \]
\[ \leq K' \left( \sigma_n \left( s \log \left( \frac{p \vee n}{\sigma_n} \right) \right)^{\frac{1}{2}} + n^{-\frac{1}{2} + \frac{1}{q}} s \log \left( \frac{p \vee n}{\sigma_n} \right) \right). \]
Applying the second part of Lemma 1 in [4] with \(t = \log(n)\), we obtain
\[ n^{\frac{1}{2}} |III - IV| \leq \sup_{\theta : |\theta - \theta_0| \leq C \rho_n} \left( \sup_{f \in \Psi'(\theta)} |G_n(f)| \right) \]
\[ \leq \sup_{\theta : |\theta - \theta_0| \leq C \rho_n} \left( 2 \mathbb{E} \left[ \sup_{f \in \Psi'(\theta)} |G_n(f)| \right] \right) \]
\[ + K_q \left( \sigma_n \log(n)^{\frac{1}{2}} + n^{-\frac{1}{2} + \frac{1}{q}} C_1^{\frac{1}{2}} \log(n) \right) \].
\[
\leq K'_q \left( \sigma_n \left( s \log \left( \frac{p \vee n}{\sigma_n} \right) \right) \right)^{1/2} + n^{-1/2} \frac{1}{q} s \log \left( \frac{p \vee n}{\sigma_n} \right)
= o(1).
\]

Now, we expand the term II. Let \( \tilde{h} \in \tilde{H} \) and \( \tilde{\theta} \in \Theta \). By Taylor expansion of the function \( r \mapsto \mathbb{E} \left[ \psi((Y, X), \theta_0 + r(\tilde{\theta} - \theta_0), h_0 + r(\tilde{h} - h_0)) \right] \), we have by Assumption C2
\[
\mathbb{E} \left[ \psi((Y, X), \tilde{\theta}, \tilde{h}) \right] = \mathbb{E} \left[ \psi((Y, X), \theta_0, h_0) \right] + \partial_r \left\{ \mathbb{E} \left[ \psi((Y, X), \theta_0 + r(\tilde{\theta} - \theta_0), h_0 + r(\tilde{h} - h_0)) \right] \right\}_{r=0} + \frac{1}{2} \partial_r^2 \left\{ \mathbb{E} \left[ \psi((Y, X), \theta_0 + r(\tilde{\theta} - \theta_0), h_0 + r(\tilde{h} - h_0)) \right] \right\}_{r=\bar{r}}
\]
for some \( \bar{r} \in (0, 1) \). Due to the orthogonality condition in C4, we have
\[
\partial_r \left\{ \mathbb{E} \left[ \psi((Y, X), \theta_0 + r(\tilde{\theta} - \theta_0), h_0 + r(\tilde{h} - h_0)) \right] \right\}_{r=0} = D_0[\tilde{h} - h_0]
\]
\[
= \partial_r \left\{ \mathbb{E} \left[ \psi((Y, X), \theta_0 + r(\tilde{\theta} - \theta_0), h_0 + r(\tilde{h} - h_0)) \right] \right\}_{r=0} - \mathbb{E} \left[ \psi((Y, X), \theta_0, h_0 + r(\tilde{h} - h_0)) \right]
\]
\[
= \partial_r \left\{ r(\tilde{\theta} - \theta_0) \partial_{\theta} \mathbb{E} \left[ \psi((Y, X), \theta, h_0 + r(\tilde{h} - h_0)) \right] \right\}_{\theta \in [\theta_0, \theta_0 + (\tilde{\theta} - \theta_0)]}_{r=0} = (\tilde{\theta} - \theta_0) \partial_{\theta} \mathbb{E} \left[ \psi((Y, X), \theta_0, h_0) \right].
\]

By Assumption C6, we have
\[
\left| \partial_r^2 \left\{ \mathbb{E} \left[ \psi((Y, X), \theta_0 + r(\tilde{\theta} - \theta_0), h_0 + r(\tilde{h} - h_0)) \right] \right\} \right|_{r=\bar{r}} = o(n^{-1/2})
\]
and therefore
\[
\mathbb{E} \left[ \psi((Y, X), \tilde{\theta}, \tilde{h}) \right] = \Gamma(\tilde{\theta} - \theta_0) + o(n^{-1/2}).
\]
In total, we obtain
\[ E_n [\psi((Y, X), \hat{\theta}, \hat{h}(<\hat{\theta}, X))] = E_n [\psi((Y, X), \theta_0, h_0(\theta_0, X))] + \Gamma(\hat{\theta} - \theta_0) + o(n^{-\frac{1}{2}}) \]
with probability 1 − o(1).

**Step 2.**
We want to prove that with probability 1 − o(1)
\[ \inf_{\theta \in \Theta} |E_n [\psi((Y, X), \theta, \hat{h}(<\theta, X))]| = o(n^{-\frac{1}{2}}). \]
Define
\[ \theta^* := \theta_0 - \Gamma^{-1} E_n [\psi((Y, X), \theta_0, h_0(\theta_0, X))]. \]
By the central limit theorem, it holds
\[ |\theta^* - \theta_0| = |\Gamma^{-1} E_n [\psi((Y, X), \theta_0, h_0(\theta_0, X))]| \leq C \rho_n. \]
Using Step 1, we obtain with probability 1 − o(1)
\[ \inf_{\theta \in \Theta} |E_n [\psi((Y, X), \theta, \hat{h}(<\theta, X))]| \leq |E_n [\psi((Y, X), \theta^*, \hat{h}(<\theta^*, X))]| = o(n^{-\frac{1}{2}}) \]
by inserting the definition of \( \theta^* \).

**Step 3.**
We aim to show that the estimated \( \hat{\theta} \) converges towards \( \theta_0 \), meaning with probability 1 − o(1)
\[ |\hat{\theta} - \theta_0| \leq C \rho_n. \]
By definition of \( \hat{\theta} \) and Step 2, we have
\[ |E_n [\psi((Y, X), \hat{\theta}, \hat{h}(<\hat{\theta}, X))]| = o(n^{-\frac{1}{2}}). \]
Since \( \hat{h}(\theta) \in \hat{\mathcal{H}}(\theta) \) with probability 1 − o(1) for all \( \theta \in \Theta \), we have
\[ \sup_{\theta \in \Theta} |E_n [\psi((Y, X), \theta, \hat{h}(\theta, X))] - E [\psi((Y, X), \theta, \hat{h}(\theta, X))]| \leq \sup_{\theta \in \Theta} \left( n^{-\frac{1}{2}} \sup_{g \in \mathcal{V}(\theta)} |G_n(g)| \right) \]
\[
= O \left( n^{-1/2} \left( s^{1/2} \log(p \vee n)^{1/2} + n^{-1/2 + 1/s} s \log(p \vee n) \right) \right),
\]
where we used Lemma 1 in [4] and \( \mathbb{E}\left[ (\bar{\psi}(Y, X))^2 \right] \leq C \) as in Step 1. Combining this with the triangle inequality, we obtain
\[
\left| \mathbb{E}\left[ \psi((Y, X), \hat{\theta}, h_0(\hat{\theta}, X)) \right] \right|
\leq \sup_{\theta \in \Theta, \hat{h} \in \tilde{H}(\theta)} \left| \mathbb{E}_n\left[ \psi((Y, X), \theta, \hat{h}(\theta, X)) \right] - \mathbb{E}\left[ \psi((Y, X), \theta, \hat{h}(\theta, X)) \right] \right|
+ \sup_{\theta \in \Theta, \hat{h}(\theta) \in \tilde{H}(\theta)} \left| \mathbb{E}_n\left[ \psi((Y, X), \hat{\theta}, \hat{h}(\theta, X)) \right] \right|
\leq C \rho_n
\]
by Assumption C6. Hence, it follows by Assumption C7
\[
|\Gamma(\hat{\theta} - \theta_0)| \wedge c_0 \leq 2 \mathbb{E}[\psi((Y, X)), \hat{\theta}, h_0(\hat{\theta}, X)] \leq C \rho_n
\]
with probability \( 1 - o(1) \) and dividing by \( \Gamma > c_1 \) gives the claim of this step.

Step 4.
Because of Step 3, we are able to use Step 1 for the estimated parameter and obtain with probability \( 1 - o(1) \)
\[
\mathbb{E}_n\left[ \psi((Y, X), \hat{\theta}, \hat{h}(\theta, X)) \right] = \mathbb{E}_n\left[ \psi((Y, X), \theta_0, h_0(\theta_0, X)) \right]
+ \Gamma(\hat{\theta} - \theta) + \mathcal{O}\left(n^{-1/2}\right).
\]
By Step 2, we have
\[
\Gamma(\hat{\theta} - \theta)
= \mathbb{E}_n\left[ \psi((Y, X), \hat{\theta}, \hat{h}(\theta, X)) \right] - \mathbb{E}_n\left[ \psi((Y, X), \theta_0, h_0(\theta_0, X)) \right] + \mathcal{O}\left(n^{-1/2}\right)
= o\left(n^{-1/2}\right)
\]
\[
= - \left( \mathbb{E}_n\left[ \psi((Y, X), \theta_0, h_0(\theta_0, X)) \right] - \mathbb{E}\left[ \psi((Y, X), \theta_0, h_0(\theta_0, X)) \right] \right)
+ \mathcal{O}\left(n^{-1/2}\right)
\]
Using the central limit theorem, we conclude with probability \( 1 - o(1) \)
\[
n^{1/2}(\hat{\theta} - \theta)
\]
\[ \begin{align*}
&= -\Gamma^{-1} n^{\frac{1}{2}} \left( \mathbb{E}_n \left[ \psi((Y, X), \theta_0, h_0(\theta_0, X)) \right] - \mathbb{E} \left[ \psi((Y, X), \theta_0, h_0(\theta_0, X)) \right] \right) \\
&\quad + o(1)
\end{align*} \]

with

\[ \Sigma := \text{Var} \left( \Gamma^{-1} \psi((Y, X), \theta_0, h_0(\theta_0, X)) \right) = \mathbb{E} \left[ \Gamma^{-2} \psi^2((Y, X), \theta_0, h_0(\theta_0, X)) \right]. \]

APPENDIX D: SIMULATION RESULTS

![Graph showing coverage for increasing number of regressors](image)

**Fig 3.** Coverage for increasing number of regressors
| n   | p   | s  | SNR        | Estimator  | Acceptance rate | MAE | rel. MSE |
|-----|-----|----|------------|------------|-----------------|-----|---------|
| 100 | 20  | 5  | 1.0        | -0.00055821| 0.928           | 0.021| 1.806   |
| 100 | 20  | 5  | 3.0        | -0.00109542| 0.952           | 0.015| 3.328   |
| 100 | 20  | 10 | 1.0        | -0.00029139| 0.938           | 0.010| 1.704   |
| 100 | 20  | 10 | 3.0        | 0.00073964 | 0.950           | 0.011| 3.536   |
| 100 | 50  | 5  | 1.0        | -0.0010804 | 0.940           | 0.021| 1.912   |
| 100 | 50  | 5  | 3.0        | -0.00149765| 0.966           | 0.020| 1.516   |
| 100 | 50  | 10 | 1.0        | 0.00011707 | 0.948           | 0.016| 1.832   |
| 100 | 50  | 10 | 3.0        | -0.0014085 | 0.970           | 0.017| 2.505   |
| 100 | 50  | 20 | 1.0        | -0.0003969 | 0.956           | 0.010| 1.868   |
| 100 | 50  | 20 | 3.0        | 0.00015181 | 0.926           | 0.013| 3.759   |
| 100 | 100 | 5  | 1.0        | 0.0004696  | 0.946           | 0.020| 1.694   |
| 100 | 100 | 5  | 3.0        | 0.0013731  | 0.972           | 0.020| 1.605   |
| 100 | 100 | 10 | 1.0        | -0.0005059 | 0.938           | 0.015| 2.020   |
| 100 | 100 | 10 | 3.0        | 0.00051767 | 0.966           | 0.018| 3.738   |
| 100 | 100 | 20 | 1.0        | -0.0016135 | 0.960           | 0.012| 1.779   |
| 100 | 100 | 20 | 3.0        | 0.00055394 | 0.944           | 0.013| 3.526   |
| 200 | 20  | 5  | 1.0        | 0.00031100 | 0.938           | 0.014| 1.520   |
| 200 | 20  | 5  | 3.0        | -0.00059165| 0.942           | 0.013| 0.941   |
| 200 | 20  | 10 | 1.0        | -0.0027602 | 0.974           | 0.020| 1.835   |
| 200 | 20  | 10 | 3.0        | 0.00062493 | 0.962           | 0.009| 1.888   |
| 200 | 20  | 20 | 1.0        | -0.0010337 | 0.934           | 0.009| 1.744   |
| 200 | 20  | 20 | 3.0        | 0.00030134 | 0.936           | 0.009| 3.389   |
| 200 | 50  | 5  | 1.0        | -0.0007235 | 0.924           | 0.010| 1.326   |
| 200 | 50  | 5  | 3.0        | 0.0002937  | 0.934           | 0.010| 0.961   |
| 200 | 50  | 10 | 1.0        | 0.00037448 | 0.928           | 0.010| 1.795   |
| 200 | 50  | 10 | 3.0        | 0.0001872  | 0.986           | 0.010| 1.580   |
| 200 | 50  | 20 | 1.0        | -0.0000121 | 0.936           | 0.007| 1.859   |
| 200 | 50  | 20 | 3.0        | 0.00010199 | 0.930           | 0.009| 3.708   |
| 200 | 100 | 5  | 1.0        | -0.0009874 | 0.918           | 0.015| 1.429   |
| 200 | 100 | 5  | 3.0        | -0.00147473| 0.942           | 0.012| 1.089   |
| 200 | 100 | 10 | 1.0        | -0.00073476| 0.918           | 0.010| 1.996   |
| 200 | 100 | 10 | 3.0        | 0.00025756 | 0.964           | 0.020| 2.657   |
| 200 | 100 | 20 | 1.0        | 0.00029025 | 0.924           | 0.007| 1.881   |
| 200 | 100 | 20 | 3.0        | 0.00147962 | 0.950           | 0.008| 3.608   |
| 200 | 200 | 5  | 1.0        | -0.0009955 | 0.952           | 0.033| 1.525   |
| 200 | 200 | 5  | 3.0        | 0.00039661 | 0.936           | 0.019| 0.967   |
| 200 | 200 | 10 | 1.0        | -0.00037153| 0.920           | 0.010| 1.890   |
| 200 | 200 | 10 | 3.0        | 0.00018176 | 0.958           | 0.011| 2.988   |
| 200 | 200 | 20 | 1.0        | -0.00025439| 0.942           | 0.007| 1.774   |
| 200 | 200 | 20 | 3.0        | 0.00016864 | 0.918           | 0.009| 3.497   |
| 200 | 500 | 5  | 1.0        | -0.0012946 | 0.942           | 0.013| 1.690   |
| 200 | 500 | 5  | 3.0        | 0.0003737  | 0.946           | 0.016| 1.084   |
| 200 | 500 | 10 | 1.0        | -0.0003134 | 0.942           | 0.010| 1.930   |
| 200 | 500 | 10 | 3.0        | -0.0001818 | 0.958           | 0.021| 2.693   |
| 200 | 500 | 20 | 1.0        | -0.0004502 | 0.930           | 0.007| 2.117   |
| 200 | 500 | 20 | 3.0        | -0.0006815 | 0.946           | 0.009| 4.106   |

Table 1: Simulation for $\Sigma^{(A)} = I_p$
| n    | p    | s    | SNR | Estimator | Acceptance rate | MAE | rel. MSE |
|------|------|------|-----|-----------|-----------------|-----|---------|
| 100  | 20   | 5    | 1.0 | 0.00035548 | 0.932           | 0.0161 | 1.2439  |
| 100  | 20   | 5    | 3.0 | -0.00097029 | 0.956           | 0.0147 | 1.0356  |
| 100  | 20   | 10   | 1.0 | -0.00046551 | 0.960           | 0.0098 | 1.6340  |
| 100  | 20   | 10   | 3.0 | 0.00092449  | 0.962           | 0.0107 | 1.5360  |
| 100  | 20   | 20   | 1.0 | 0.00017477  | 0.904           | 0.0080 | 1.7442  |
| 100  | 20   | 20   | 3.0 | 0.00025592  | 0.940           | 0.0094 | 2.7994  |
| 100  | 50   | 5    | 1.0 | -0.00025392 | 0.942           | 0.0154 | 1.3759  |
| 100  | 50   | 5    | 3.0 | 0.00093743  | 0.960           | 0.0152 | 1.0898  |
| 100  | 50   | 10   | 1.0 | -0.00032109 | 0.936           | 0.0111 | 1.7082  |
| 100  | 50   | 10   | 3.0 | 0.00047115  | 0.966           | 0.0112 | 1.7237  |
| 100  | 50   | 20   | 1.0 | -0.0002715  | 0.934           | 0.0078 | 1.8586  |
| 100  | 50   | 20   | 3.0 | 0.00018370  | 0.952           | 0.0093 | 3.2917  |
| 100  | 100  | 5    | 1.0 | -0.00095235 | 0.938           | 0.0147 | 1.4640  |
| 100  | 100  | 5    | 3.0 | -0.00019675 | 0.940           | 0.0159 | 1.2573  |
| 100  | 100  | 10   | 1.0 | 0.00029316  | 0.960           | 0.0080 | 1.7442  |
| 100  | 100  | 10   | 3.0 | 0.00013246  | 0.962           | 0.0075 | 1.8949  |
| 100  | 100  | 20   | 1.0 | 0.00021921  | 0.960           | 0.0093 | 3.3579  |
| 100  | 200  | 5    | 1.0 | 0.00009516  | 0.942           | 0.0108 | 1.9001  |
| 100  | 200  | 5    | 3.0 | 0.000031272 | 0.982           | 0.0142 | 1.1441  |
| 100  | 200  | 10   | 1.0 | 0.00056643  | 0.956           | 0.0111 | 1.8399  |
| 100  | 200  | 10   | 3.0 | 0.000082101 | 0.974           | 0.0125 | 2.2858  |
| 100  | 200  | 20   | 1.0 | -0.00038676 | 0.968           | 0.0073 | 1.7771  |
| 100  | 200  | 20   | 3.0 | 0.00019953  | 0.966           | 0.0088 | 3.3579  |

Table 2: Simulation for $\Sigma^{(X)} = \Sigma_1^{(X)}$
| n   | p   | s   | SNR | Estimator  | Acceptance rate | MAE  | rel. MSE |
|-----|-----|-----|-----|------------|-----------------|------|----------|
| 100 | 20  | 5   | 1.0 | -0.00029259 | 0.952           | 0.0203 | 1.7828   |
| 100 | 20  | 5   | 3.0 | -0.00018328 | 0.968           | 0.0203 | 1.4251   |
| 100 | 20  | 10  | 1.0 | 0.00014132  | 0.928           | 0.0157 | 1.8490   |
| 100 | 20  | 10  | 3.0 | 0.000109876 | 0.956           | 0.0186 | 3.7599   |
| 100 | 20  | 20  | 1.0 | 0.000159371 | 0.938           | 0.0111 | 1.8879   |
| 100 | 20  | 20  | 3.0 | 0.00015822  | 0.948           | 0.0126 | 3.7501   |
| 100 | 50  | 5   | 1.0 | -0.00160281 | 0.938           | 0.0226 | 1.6711   |
| 100 | 50  | 5   | 3.0 | -0.00065212 | 0.984           | 0.0216 | 1.8807   |
| 100 | 50  | 10  | 1.0 | 0.0014570   | 0.966           | 0.0145 | 1.9117   |
| 100 | 50  | 10  | 3.0 | 0.00130942  | 0.964           | 0.0196 | 3.7486   |
| 100 | 50  | 20  | 1.0 | 0.00159371  | 0.938           | 0.0111 | 1.8879   |
| 100 | 50  | 20  | 3.0 | 0.00153853  | 0.972           | 0.0133 | 3.5190   |
| 100 | 100 | 5   | 1.0 | -0.0014132  | 0.968           | 0.0157 | 1.9941   |
| 100 | 100 | 5   | 3.0 | -0.00018987 | 0.956           | 0.0186 | 3.7599   |
| 100 | 100 | 10  | 1.0 | -0.00019387 | 0.938           | 0.0111 | 1.8879   |
| 100 | 100 | 10  | 3.0 | 0.000155053 | 0.972           | 0.0133 | 3.5190   |
| 200 | 20  | 5   | 1.0 | -0.00066470 | 0.956           | 0.0122 | 1.1421   |
| 200 | 20  | 5   | 3.0 | 0.00007279  | 0.942           | 0.0124 | 0.9439   |
| 200 | 20  | 10  | 1.0 | 0.00023762  | 0.924           | 0.0100 | 1.7593   |
| 200 | 20  | 10  | 3.0 | 0.00056502  | 0.958           | 0.0101 | 1.2329   |
| 200 | 20  | 20  | 1.0 | 0.00062338  | 0.924           | 0.0073 | 1.7263   |
| 200 | 20  | 20  | 3.0 | 0.0019359   | 0.956           | 0.0084 | 3.2262   |
| 200 | 50  | 5   | 1.0 | 0.00131883  | 0.946           | 0.0141 | 1.3133   |
| 200 | 50  | 5   | 3.0 | 0.00028543  | 0.946           | 0.0135 | 0.9588   |
| 200 | 50  | 10  | 1.0 | -0.00064094 | 0.958           | 0.0107 | 1.8158   |
| 200 | 50  | 10  | 3.0 | -0.00046929 | 0.952           | 0.0104 | 1.5506   |
| 200 | 50  | 20  | 1.0 | -0.00047503 | 0.948           | 0.0072 | 1.8612   |
| 200 | 50  | 20  | 3.0 | 0.00028000  | 0.932           | 0.0091 | 3.7182   |
| 200 | 100 | 5   | 1.0 | -0.00080455 | 0.938           | 0.0138 | 1.3802   |
| 200 | 100 | 5   | 3.0 | 0.00006808  | 0.954           | 0.0129 | 1.0877   |
| 200 | 100 | 10  | 1.0 | -0.00077886 | 0.928           | 0.0106 | 1.9660   |
| 200 | 100 | 10  | 3.0 | 0.00072632  | 0.968           | 0.0102 | 2.9484   |
| 200 | 100 | 20  | 1.0 | -0.0003086  | 0.944           | 0.0073 | 1.8808   |
| 200 | 100 | 20  | 3.0 | -0.00028688 | 0.950           | 0.0085 | 3.6061   |
| 200 | 200 | 5   | 1.0 | -0.00010218 | 0.912           | 0.0142 | 1.5353   |
| 200 | 200 | 5   | 3.0 | 0.000138252 | 0.946           | 0.0136 | 0.9609   |
| 200 | 200 | 10  | 1.0 | 0.00004042  | 0.932           | 0.0101 | 1.8956   |
| 200 | 200 | 10  | 3.0 | -0.00043170 | 0.962           | 0.0109 | 2.2782   |
| 200 | 200 | 20  | 1.0 | -0.00028175 | 0.932           | 0.0075 | 1.7815   |
| 200 | 200 | 20  | 3.0 | 0.00030455  | 0.946           | 0.0085 | 3.5073   |
| 200 | 500 | 5   | 1.0 | -0.00045556 | 0.930           | 0.0143 | 1.6798   |
| 200 | 500 | 5   | 3.0 | 0.00023905  | 0.940           | 0.0139 | 1.0843   |
| 200 | 500 | 10  | 1.0 | 0.00016880  | 0.932           | 0.0110 | 1.9023   |
| 200 | 500 | 10  | 3.0 | -0.00135486 | 0.960           | 0.0118 | 2.7441   |
| 200 | 500 | 20  | 1.0 | -0.0005315  | 0.936           | 0.0078 | 2.0996   |
| 200 | 500 | 20  | 3.0 | -0.0009986  | 0.932           | 0.0093 | 4.0653   |

Table 3: Simulation for $\Sigma^{(X)} = \sum_{i=2}^{X}$
| n  | p  | s  | SNR   | Estimator | Acceptance rate | MAE   | rel. MSE | rel. MSE |
|----|----|----|-------|-----------|----------------|-------|----------|----------|
| 100| 20 | 5  | 1.0   | 0.99990985 | 0.956        | 0.0531| 1.7955   |
| 100| 20 | 5  | 3.0   | 0.99701129 | 0.956        | 0.0460| 1.2454   |
| 100| 20 | 10 | 1.0   | 0.99732572 | 0.946        | 0.0440| 1.8621   |
| 100| 20 | 10 | 3.0   | 1.00201985 | 0.954        | 0.0492| 3.2990   |
| 100| 20 | 20 | 1.0   | 1.00463863 | 0.932        | 0.0402| 1.7770   |
| 100| 20 | 20 | 3.0   | 0.9917208 | 0.940        | 0.0411| 3.5338   |
| 100| 50 | 5  | 1.0   | 1.00133561 | 0.942        | 0.0532| 1.9071   |
| 100| 50 | 5  | 3.0   | 0.9966997 | 0.982        | 0.0478| 1.4648   |
| 100| 50 | 10 | 1.0   | 1.0054266 | 0.940        | 0.0450| 1.8360   |
| 100| 50 | 10 | 3.0   | 1.0017462 | 0.960        | 0.0469| 3.4738   |
| 100| 50 | 20 | 1.0   | 0.9837549 | 0.964        | 0.0371| 1.8729   |
| 100| 50 | 20 | 3.0   | 0.9961815 | 0.938        | 0.0426| 3.7640   |
| 100| 100| 5  | 1.0   | 1.0027212 | 0.956        | 0.0533| 1.6976   |
| 100| 100| 5  | 3.0   | 0.9972378 | 0.954        | 0.0509| 1.6225   |
| 100| 100| 10 | 1.0   | 1.0029172 | 0.944        | 0.0461| 2.0220   |
| 100| 100| 10 | 3.0   | 1.0000260 | 0.942        | 0.0499| 3.8505   |
| 100| 100| 20 | 1.0   | 1.0004159 | 0.952        | 0.0385| 1.8904   |
| 100| 100| 20 | 3.0   | 0.9983754 | 0.964        | 0.0371| 1.8729   |
| 200| 20 | 5  | 1.0   | 1.0001244 | 0.938        | 0.0361| 1.1689   |
| 200| 20 | 5  | 3.0   | 1.0005385 | 0.936        | 0.0313| 0.9433   |
| 200| 20 | 10 | 1.0   | 1.0043023 | 0.944        | 0.0306| 1.7735   |
| 200| 20 | 10 | 3.0   | 0.9997127 | 0.960        | 0.0265| 1.2521   |
| 200| 20 | 20 | 1.0   | 1.0021089 | 0.952        | 0.0494| 3.7504   |
| 200| 20 | 20 | 3.0   | 0.9976642 | 0.936        | 0.0307| 3.4005   |
| 200| 50 | 5  | 1.0   | 0.9971819 | 0.954        | 0.0328| 1.3104   |
| 200| 50 | 5  | 3.0   | 1.0018026 | 0.942        | 0.0301| 0.9625   |
| 200| 50 | 10 | 1.0   | 0.9996908 | 0.974        | 0.0481| 1.8679   |
| 200| 50 | 10 | 3.0   | 1.0022502 | 0.962        | 0.0460| 1.9035   |
| 200| 50 | 20 | 1.0   | 0.9971724 | 0.960        | 0.0265| 1.7735   |
| 200| 50 | 20 | 3.0   | 1.0004432 | 0.952        | 0.0494| 3.7504   |
| 200| 100| 5  | 1.0   | 0.9985371 | 0.960        | 0.0365| 1.3969   |
| 200| 100| 5  | 3.0   | 0.9962946 | 0.952        | 0.0284| 1.0922   |
| 200| 100| 10 | 1.0   | 1.0049785 | 0.946        | 0.0326| 1.9988   |
| 200| 100| 10 | 3.0   | 0.9996408 | 0.960        | 0.0286| 1.9835   |
| 200| 100| 20 | 1.0   | 0.9985907 | 0.938        | 0.0273| 1.8807   |
| 200| 100| 20 | 3.0   | 0.9962509 | 0.952        | 0.0294| 3.6156   |
| 200| 200| 5  | 1.0   | 1.0024386 | 0.952        | 0.0363| 1.5089   |
| 200| 200| 5  | 3.0   | 1.0004254 | 0.946        | 0.0318| 0.9710   |
| 200| 200| 10 | 1.0   | 1.0015006 | 0.956        | 0.0312| 1.8897   |
| 200| 200| 10 | 3.0   | 1.0005871 | 0.972        | 0.0295| 2.2308   |
| 200| 200| 20 | 1.0   | 1.0037436 | 0.948        | 0.0208| 2.7388   |
| 200| 200| 20 | 3.0   | 1.0007742 | 0.954        | 0.0293| 3.4977   |
| 200| 500| 5  | 1.0   | 1.0033787 | 0.934        | 0.0394| 1.7058   |
| 200| 500| 5  | 3.0   | 1.0000420 | 0.946        | 0.0319| 1.0865   |
| 200| 500| 10 | 1.0   | 0.9996577 | 0.948        | 0.0303| 1.9287   |
| 200| 500| 10 | 3.0   | 0.9986465 | 0.962        | 0.0327| 2.8139   |
| 200| 500| 20 | 1.0   | 1.0018244 | 0.950        | 0.0271| 2.1191   |
| 200| 500| 20 | 3.0   | 1.0009260 | 0.948        | 0.0297| 4.1020   |

Table 4: Simulation for $\Sigma^{(A)} = I_p$
| n   | p   | s   | SNR | Estimator | Acceptance rate | MAE | rel. MSE |
|-----|-----|-----|-----|-----------|-----------------|-----|---------|
| 100 | 20  | 5   | 1.0 | 0.99619347 | 0.924           | 0.0463 | 1.2505  |
| 100 | 20  | 5   | 3.0 | 0.99833394 | 0.960           | 0.0374 | 1.0330  |
| 100 | 20  | 10  | 1.0 | 0.99814415 | 0.946           | 0.0403 | 1.6382  |
| 100 | 20  | 10  | 3.0 | 0.99861606 | 0.980           | 0.0346 | 1.5484  |
| 100 | 20  | 20  | 1.0 | 1.00332660 | 0.946           | 0.0321 | 1.7477  |
| 100 | 50  | 5   | 1.0 | 1.00242808 | 0.952           | 0.0345 | 2.8320  |
| 100 | 50  | 5   | 3.0 | 0.99862564 | 0.950           | 0.0374 | 1.0930  |
| 100 | 50  | 10  | 1.0 | 0.99914415 | 0.976           | 0.0365 | 2.1079  |
| 100 | 50  | 10  | 3.0 | 1.00405369 | 0.970           | 0.0361 | 1.7522  |
| 100 | 50  | 20  | 1.0 | 1.00049886 | 0.942           | 0.0333 | 1.8622  |
| 100 | 50  | 20  | 3.0 | 1.00275828 | 0.954           | 0.0354 | 3.3531  |
| 100 | 100 | 5   | 1.0 | 0.99964792 | 0.958           | 0.0436 | 1.4512  |
| 100 | 100 | 5   | 3.0 | 1.00018733 | 0.970           | 0.0378 | 1.2533  |
| 100 | 100 | 10  | 1.0 | 0.99887052 | 0.970           | 0.0365 | 1.9048  |
| 100 | 100 | 10  | 3.0 | 0.99914381 | 0.976           | 0.0365 | 2.1079  |
| 100 | 100 | 20  | 1.0 | 1.00325414 | 0.940           | 0.0341 | 1.8941  |
| 100 | 100 | 20  | 3.0 | 1.00049886 | 0.942           | 0.0333 | 1.8622  |

Table 5: Simulation for $\Sigma^{(X)} = \Sigma^{(X)}_1$
| n    | p    | s    | SNR | Estimator | Acceptance rate | MAE rel. MSE |
|------|------|------|-----|-----------|-----------------|--------------|
| 100  | 20   | 5    | 1.0 | 1.00367279 | 0.940           | 0.0537       | 1.7783      |
| 100  | 20   | 5    | 3.0 | 0.9994812  | 0.970           | 0.0440       | 1.4556      |
| 100  | 20   | 10   | 1.0 | 1.00152401 | 0.928           | 0.0453       | 1.8476      |
| 100  | 20   | 10   | 3.0 | 0.99976291 | 0.942           | 0.0477       | 3.5271      |
| 100  | 20   | 20   | 1.0 | 0.99963396 | 0.952           | 0.0381       | 1.8688      |
| 100  | 20   | 20   | 3.0 | 1.00001843 | 0.946           | 0.0409       | 3.7607      |
| 100  | 50   | 5    | 1.0 | 0.99656704 | 0.948           | 0.0540       | 1.6708      |
| 100  | 50   | 5    | 3.0 | 0.99950611 | 0.970           | 0.0533       | 1.9694      |
| 100  | 50   | 10   | 1.0 | 1.00102807 | 0.938           | 0.0321       | 1.7606      |
| 100  | 50   | 10   | 3.0 | 0.99999890 | 0.954           | 0.0275       | 1.8236      |
| 100  | 50   | 20   | 1.0 | 1.00100783 | 0.940           | 0.0264       | 1.7254      |
| 100  | 50   | 20   | 3.0 | 0.99932297 | 0.964           | 0.0274       | 3.1870      |
| 200  | 20   | 5    | 1.0 | 1.00265436 | 0.954           | 0.0326       | 1.1445      |
| 200  | 20   | 5    | 3.0 | 1.00058134 | 0.954           | 0.0301       | 1.9432      |
| 200  | 20   | 10   | 1.0 | 1.00125043 | 0.938           | 0.0321       | 1.7606      |
| 200  | 20   | 10   | 3.0 | 0.99998890 | 0.960           | 0.0257       | 1.8236      |
| 200  | 20   | 20   | 1.0 | 1.00160783 | 0.940           | 0.0264       | 1.7254      |
| 200  | 20   | 20   | 3.0 | 0.99932297 | 0.964           | 0.0274       | 3.1870      |
| 200  | 50   | 5    | 1.0 | 1.00079590 | 0.954           | 0.0341       | 1.3372      |
| 200  | 50   | 5    | 3.0 | 0.99869385 | 0.926           | 0.0324       | 0.9649      |
| 200  | 50   | 10   | 1.0 | 1.00102807 | 0.954           | 0.0310       | 1.8102      |
| 200  | 50   | 10   | 3.0 | 1.00030919 | 0.954           | 0.0283       | 1.6297      |
| 200  | 50   | 20   | 1.0 | 1.00097676 | 0.952           | 0.0270       | 1.8610      |
| 200  | 50   | 20   | 3.0 | 0.99814638 | 0.944           | 0.0300       | 3.0498      |
| 200  | 100  | 5    | 1.0 | 0.99798485 | 0.930           | 0.0373       | 1.3546      |
| 200  | 100  | 5    | 3.0 | 1.00224359 | 0.950           | 0.0294       | 1.0932      |
| 200  | 100  | 10   | 1.0 | 0.99968150 | 0.940           | 0.0318       | 1.9679      |
| 200  | 100  | 10   | 3.0 | 0.99912236 | 0.974           | 0.0286       | 2.5742      |
| 200  | 100  | 20   | 1.0 | 1.00001755 | 0.954           | 0.0258       | 1.8803      |
| 200  | 100  | 20   | 3.0 | 0.99963970 | 0.936           | 0.0296       | 3.6140      |
| 200  | 200  | 5    | 1.0 | 0.99632996 | 0.950           | 0.0345       | 1.4789      |
| 200  | 200  | 5    | 3.0 | 0.99969194 | 0.954           | 0.0303       | 0.9649      |
| 200  | 200  | 10   | 1.0 | 1.00113110 | 0.932           | 0.0320       | 1.8961      |
| 200  | 200  | 10   | 3.0 | 1.00093688 | 0.968           | 0.0294       | 2.2247      |
| 200  | 200  | 20   | 1.0 | 1.00027202 | 0.950           | 0.0275       | 1.8049      |
| 200  | 200  | 30   | 3.0 | 1.00006203 | 0.956           | 0.0287       | 3.5083      |
| 200  | 500  | 5    | 1.0 | 1.00762721 | 0.952           | 0.0384       | 1.7089      |
| 200  | 500  | 5    | 3.0 | 1.00006196 | 0.966           | 0.0298       | 1.0830      |
| 200  | 500  | 10   | 1.0 | 0.99913096 | 0.942           | 0.0318       | 1.9029      |
| 200  | 500  | 10   | 3.0 | 1.00015333 | 0.956           | 0.0327       | 2.7849      |
| 200  | 500  | 20   | 1.0 | 1.00062420 | 0.942           | 0.0267       | 2.0994      |
| 200  | 500  | 20   | 3.0 | 1.00125883 | 0.932           | 0.0300       | 4.0673      |

Table 6: Simulation for $\Sigma^{(X)} = \Sigma_{2}^{(X)}$
**APPENDIX E: APPLICATION: FIGURES AND TABLES**

**Fig 4.** Empirical wage distribution from the US survey data

**Fig 5.** Comparison of the Q-Q plots
| Variable                  | Type        | Baseline Category                      |
|--------------------------|-------------|----------------------------------------|
| Female                   | binary      |                                        |
| Marital status           | six categories | never married, single                   |
| Race                     | four categories | White                                  |
| English language skills   | five categories | speaks only English                    |
| Hispanic                 | binary      |                                        |
| Veteran Status           | binary      |                                        |
| Industry                 | 14 categories | wholesale trade                         |
| Occupation               | 26 categories | management, science, arts               |
| Region (US census)       | nine categories | New England division                   |
| Experience (years)       | continuous  |                                        |
| Experience squared       | continuous  |                                        |
| Years of Education       | continuous  |                                        |
| Family Size              | continuous  |                                        |
| Number of own young children | continuous                  |
| Field of degree          | 37 categories | administration, teaching                |

**Table 7**  
*List of Regressors*

**Fig 6.** Transformation function for $\theta = 0$ (dashed) and $\theta = \hat{\theta}$ (solid)
APPENDIX F: ADDITIONAL SIMULATIONS

F.1. Approximately Sparse Setting. In the approximately sparse setting the coefficients are set to

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

The other parameters are chosen as in simulations in the main text (Section 4), but to restrict the calculation time we focus on the correlation structure $\Sigma_{1}(X)$. The results for Box-Cox-Transformations ($\theta_0 = 0$) are presented in Table 9 and the results for Yeo-Johnson Power Transformations ($\theta_0 = 1$) in Table 10. We remark that the case $p = 20$ and $s = 20$ is not contained in both tables since these settings coincide with the exactly sparse setting. The results are similar to the exactly sparse setting and the acceptance rate is close to the nominal level.
| n  | p  | s  | SNR | Estimator   | Acceptance rate | MAE    | rel. MSE |
|----|----|----|-----|-------------|-----------------|--------|---------|
| 100| 20 | 5  | 1.0 | -0.00072435| 0.946           | 0.0147 | 1.2323  |
| 100| 20 | 5  | 3.0 | 0.00065075 | 0.944           | 0.0139 | 1.0325  |
| 100| 20 | 10 | 1.0 | -0.00035129| 0.950           | 0.0108 | 1.6230  |
| 100| 20 | 10 | 3.0 | -0.00068912| 0.949           | 0.0114 | 1.4992  |
| 100| 50 | 5  | 1.0 | 0.00039227 | 0.970           | 0.0136 | 1.1084  |
| 100| 50 | 5  | 3.0 | 0.00037004 | 0.952           | 0.0103 | 1.7093  |
| 100| 50 | 10 | 1.0 | 0.00040427 | 0.952           | 0.0111 | 1.7451  |
| 100| 50 | 10 | 3.0 | 0.00024774 | 0.948           | 0.0071 | 1.8708  |
| 100| 50 | 20 | 1.0 | 0.00032668 | 0.946           | 0.0092 | 3.3351  |
| 100| 100| 5  | 1.0 | 0.00115031 | 0.958           | 0.0143 | 1.4935  |
| 100| 100| 5  | 3.0 | -0.00002014| 0.976           | 0.0147 | 1.1812  |
| 100| 100| 10 | 1.0 | -0.00066524| 0.952           | 0.0105 | 1.8505  |
| 100| 100| 10 | 3.0 | -0.00072896| 0.966           | 0.0119 | 2.0950  |
| 100| 100| 20 | 1.0 | -0.00033613| 0.936           | 0.0079 | 1.8906  |
| 100| 100| 20 | 3.0 | -0.00003507| 0.962           | 0.0091 | 3.4120  |
| 100| 200| 5  | 1.0 | -0.0067739 | 0.976           | 0.0151 | 1.4985  |
| 100| 200| 5  | 3.0 | -0.0000964 | 0.966           | 0.0151 | 1.1812  |
| 100| 200| 10 | 1.0 | -0.00071120| 0.952           | 0.0105 | 1.8505  |
| 100| 200| 10 | 3.0 | -0.0014934 | 0.980           | 0.0130 | 2.2642  |
| 100| 200| 20 | 1.0 | -0.00103713| 0.946           | 0.0080 | 1.7758  |
| 100| 200| 20 | 3.0 | -0.00008740| 0.962           | 0.0094 | 3.3879  |
| 200| 20 | 5  | 1.0 | -0.00104238| 0.924           | 0.0095 | 0.9621  |
| 200| 20 | 5  | 3.0 | 0.00119542 | 0.928           | 0.0098 | 0.9285  |
| 200| 20 | 10 | 1.0 | 0.00034137 | 0.942           | 0.0067 | 1.1898  |
| 200| 20 | 10 | 3.0 | -0.00036637| 0.932           | 0.0068 | 1.0075  |
| 200| 50 | 5  | 1.0 | 0.00029033 | 0.938           | 0.0100 | 1.0325  |
| 200| 50 | 5  | 3.0 | 0.00128785 | 0.946           | 0.0095 | 0.9765  |
| 200| 50 | 10 | 1.0 | 0.00027922 | 0.948           | 0.0069 | 1.2909  |
| 200| 50 | 10 | 3.0 | 0.00015796 | 0.950           | 0.0067 | 1.0655  |
| 200| 50 | 20 | 1.0 | 0.00014660 | 0.932           | 0.0053 | 1.7986  |
| 200| 50 | 20 | 3.0 | 0.00027307 | 0.948           | 0.0054 | 1.8437  |
| 200| 100| 5  | 1.0 | 0.00339317 | 0.944           | 0.0087 | 1.1958  |
| 200| 100| 5  | 3.0 | -0.0000793 | 0.946           | 0.0090 | 1.1458  |
| 200| 100| 10 | 1.0 | 0.00027762 | 0.946           | 0.0069 | 1.4877  |
| 200| 100| 10 | 3.0 | 0.00127725 | 0.946           | 0.0070 | 1.2023  |
| 200| 100| 20 | 1.0 | -0.00028952| 0.944           | 0.0051 | 1.8415  |
| 200| 100| 20 | 3.0 | 0.00049090 | 0.954           | 0.0060 | 2.0796  |
| 200| 200| 5  | 1.0 | 0.00045282 | 0.926           | 0.0101 | 1.0675  |
| 200| 200| 5  | 3.0 | 0.00005082 | 0.946           | 0.0096 | 0.9836  |
| 200| 200| 10 | 1.0 | 0.00026888 | 0.944           | 0.0076 | 1.4918  |
| 200| 200| 10 | 3.0 | 0.00052876 | 0.954           | 0.0070 | 1.1149  |
| 200| 200| 20 | 1.0 | 0.00010103 | 0.940           | 0.0051 | 1.7536  |
| 200| 200| 20 | 3.0 | -0.00068245| 0.966           | 0.0057 | 2.2181  |
| 200| 500| 5  | 1.0 | 0.00058465 | 0.906           | 0.0104 | 1.2234  |
| 200| 500| 5  | 3.0 | -0.00013389| 0.926           | 0.0097 | 1.1230  |
| 200| 500| 10 | 1.0 | 0.00012013 | 0.942           | 0.0075 | 1.6718  |
| 200| 500| 10 | 3.0 | 0.00021128 | 0.944           | 0.0070 | 1.2930  |
| 200| 500| 20 | 1.0 | 0.00019671 | 0.944           | 0.0052 | 2.0953  |
| 200| 500| 20 | 3.0 | -0.00008077| 0.974           | 0.0060 | 3.0412  |

Table 9: Simulation for Box-Cox-Transformations
### Table 10: Simulation for Yeo-Johnson Power Transformations

| n  | p  | s  | SNR  | Estimator  | Acceptance rate | MAE rel. MSE |
|----|----|----|------|------------|-----------------|--------------|
| 100| 20 | 5  | 1.0  | 1.00056735 | 0.962           | 0.0418       | 1.2250       |
| 100| 20 | 5  | 3.0  | 0.99968158 | 0.932           | 0.0410       | 1.0245       |
| 100| 20 | 10 | 1.0  | 0.9996806  | 0.942           | 0.0382       | 1.6179       |
| 100| 20 | 10 | 3.0  | 0.99983138 | 0.960           | 0.0373       | 1.5512       |
| 100| 50 | 5  | 1.0  | 1.00428870 | 0.950           | 0.0421       | 1.3767       |
| 100| 50 | 5  | 3.0  | 0.99942787 | 0.950           | 0.0391       | 1.7085       |
| 100| 50 | 10 | 1.0  | 0.99916016 | 0.966           | 0.0357       | 1.7955       |
| 100| 50 | 10 | 3.0  | 0.99842764 | 0.946           | 0.0321       | 1.8719       |
| 100| 50 | 20 | 1.0  | 0.99724029 | 0.966           | 0.0355       | 3.4082       |
| 100| 50 | 20 | 3.0  | 0.99724029 | 0.966           | 0.0355       | 3.4082       |
| 100| 100| 5  | 1.0  | 1.0045974  | 0.978           | 0.0424       | 1.5090       |
| 100| 100| 5  | 3.0  | 0.99536763 | 0.968           | 0.0383       | 1.1875       |
| 100| 100| 10 | 1.0  | 0.9877481  | 0.972           | 0.0377       | 1.8432       |
| 100| 100| 10 | 3.0  | 1.00074576 | 0.962           | 0.0398       | 2.1258       |
| 100| 200| 5  | 1.0  | 1.00045974 | 0.978           | 0.0424       | 1.5090       |
| 100| 200| 5  | 3.0  | 0.99536763 | 0.968           | 0.0383       | 1.1875       |
| 200| 20 | 5  | 1.0  | 0.99985307 | 0.928           | 0.0292       | 0.9671       |
| 200| 20 | 5  | 3.0  | 0.99996887 | 0.962           | 0.0255       | 0.9295       |
| 200| 20 | 10 | 1.0  | 1.00054556 | 0.924           | 0.0259       | 1.1881       |
| 200| 20 | 10 | 3.0  | 0.99888749 | 0.942           | 0.0223       | 1.0106       |
| 200| 50 | 5  | 1.0  | 0.99122336 | 0.916           | 0.0312       | 1.0292       |
| 200| 50 | 5  | 3.0  | 0.99876047 | 0.944           | 0.0258       | 0.9765       |
| 200| 50 | 10 | 1.0  | 0.99938511 | 0.946           | 0.0251       | 1.2924       |
| 200| 50 | 10 | 3.0  | 0.99885312 | 0.938           | 0.0237       | 1.0708       |
| 200| 20 | 10 | 1.0  | 0.99941435 | 0.936           | 0.0224       | 1.7849       |
| 200| 20 | 10 | 3.0  | 0.99886322 | 0.960           | 0.0224       | 1.8152       |
| 200| 100| 5  | 1.0  | 1.00073382 | 0.932           | 0.0301       | 1.1934       |
| 200| 100| 5  | 3.0  | 0.99614218 | 0.926           | 0.0268       | 1.1418       |
| 200| 100| 10 | 1.0  | 1.00113064 | 0.936           | 0.0271       | 1.5229       |
| 200| 100| 10 | 3.0  | 1.00177678 | 0.956           | 0.0214       | 1.1959       |
| 200| 200| 5  | 1.0  | 1.00129852 | 0.938           | 0.0288       | 1.0590       |
| 200| 200| 5  | 3.0  | 0.99850564 | 0.940           | 0.0275       | 0.9827       |
| 200| 200| 10 | 1.0  | 0.99588019 | 0.938           | 0.0262       | 1.5113       |
| 200| 200| 10 | 3.0  | 1.00066971 | 0.954           | 0.0233       | 1.1361       |
| 200| 200| 20 | 1.0  | 0.99697192 | 0.940           | 0.0239       | 1.7520       |
| 200| 200| 20 | 3.0  | 1.00029713 | 0.962           | 0.0216       | 2.2585       |
| 200| 500| 5  | 1.0  | 1.00007598 | 0.962           | 0.0279       | 1.2183       |
| 200| 500| 5  | 3.0  | 0.99809707 | 0.936           | 0.0276       | 1.1211       |
| 200| 500| 10 | 1.0  | 1.00142777 | 0.936           | 0.0272       | 1.6619       |
| 200| 500| 10 | 3.0  | 0.99917894 | 0.962           | 0.0226       | 1.2654       |
| 200| 500| 20 | 1.0  | 0.99981870 | 0.940           | 0.0242       | 2.1018       |
| 200| 500| 20 | 3.0  | 0.99695433 | 0.964           | 0.0231       | 2.9583       |
F.2. Inference on Regression Coefficients. In the following example, let $\mathcal{F}_\Lambda$ be the class of Yeo-Johnson Power Transformations. We generate the data as in Section 4. The data generating process is given by

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

where the coefficients are set to

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

with $s = 5$ and $X \sim \mathcal{N}(0, I_p)$. We set the transformation parameter $\theta_0 = 1.5$ and generate $n = 100$ observations (with $p = 200$ and SNR= 1).

First, we use our approach to estimate the transformation parameter $\theta_0$. As illustrated in the Figure 7, the transformation reduces the skewness of the error distribution.

Fig 7. Estimated errors on an independent testing sample of size $n_{test} = 1000$ (with identity function).

For the class of Yeo-Johnson Power Transformations, $\theta = 1$ corresponds to the “naive” estimate ($\Lambda_1(\cdot)$ is the identity function with corresponding estimator $\hat{\beta}_1$). In a second stage, we employ the double machine learning estimator from [5] with outcome $\Lambda_{\theta}(Y)$ to obtain estimates for the coefficient $\beta_{\theta,j}$ and the corresponding 95% confidence intervals for $\theta \in \{\theta_0, \hat{\theta}, 1\}$ and $j = \{1, 6\}$. We choose $j = 6$ since by construction $\beta_{\theta,6} = 0$ for all choices of $\theta$, due to $X_6$ being independent from $Y$. The calculations are performed using rlassoEffects from the R package hdm, version 0.3.1 by [6] which can be downloaded from CRAN. To make all estimators comparable, each outcome $\Lambda_{\theta}(Y)$ is standardized before the double machine learning estimate is applied. We repeat this procedure 1000 times and save the estimators $\hat{\beta}_{\theta,j}$ and corresponding confidence intervals. Finally, we compare the variance of
the estimators over all repetitions and the average width of the estimated confidence intervals.

| Transformation parameter $\theta$ | 1       | $\hat{\theta}$ | $\theta_0$ |
|-----------------------------------|---------|----------------|------------|
| Variance of $\hat{\beta}_{\theta,1}$ ($\cdot 10^3$) | 7.380114 | 6.433512 | 6.448069 |
| Variance of $\hat{\beta}_{\theta,6}$ ($\cdot 10^3$) | 9.001942 | 6.851171 | 6.806154 |
| Average width of CI for $\beta_{\theta,1}$ | 0.3645432 | 0.3262797 | 0.3269129 |
| Average width of CI for $\beta_{\theta,6}$ | 0.3747205 | 0.3250437 | 0.3252659 |
| Coverage for $\beta_{\theta,1}$ | -       | 0.933    | 0.934    |
| Coverage for $\beta_{\theta,6}$ | 0.950   | 0.951    | 0.952    |

The variance of the regression coefficient $\hat{\beta}_{\theta,6}$ is reduced by 23.9% with respect to the "naive" model for the estimated transformation parameter and the average width of confidence interval reduced by 13.3% percent. In this setting, the coverage can not be calculated for $\beta_{1,1}$ since we do not know the true coefficient (for $\theta = \hat{\theta}$, we calculated the coverage assuming $\beta_{\theta_0,1}$ is the true coefficient). Overall, the transformation reduces the variance of the estimated regression coefficients as the tails of the errors are behaving more nicely due to the normality.

**F.3. Non-Normal Errors.** In this chapter, we check the performance of our proposed method under non-normal errors. The same simulation is run as in Section 4 with $n = 100$ observations, but we simulate errors according to a $t$-distribution with $df$ degrees of freedom

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

We focus on the correlation structure $\Sigma_1^{(X)}$ and the Box-Cox-Transformations ($\theta_0 = 0$). We set $s = 20$ and vary the degrees of freedom.
Figure 8 displays the effect of non-normal errors on the coverage. If the deviation from the normal distribution is high (low number of degrees of freedom) the coverage differs largely from the nominal level of 95%. With an increasing number of regressors the coverage gradually approaches the nominal level.
REFERENCES

[1] Assouad, P. (1983). Densité et dimension. In Annales de l’Institut Fourier 33 233–282.
[2] Belloni, A., Chernozhukov, V., Chetverikov, D. and Wei, Y. (2018). Uniformly valid post-regularization confidence regions for many functional parameters in z-estimation framework. Annals of statistics 46 3643.
[3] Belloni, A., Chernozhukov, V., Fernández-Val, I. and Hansen, C. (2017). Program Evaluation and Causal Inference With High-Dimensional Data. Econometrica 85 233–298.
[4] Belloni, A., Chernozhukov, V. and Kato, K. (2014). Uniform post selection inference for LAD regression and other Z-estimation problems Technical Report, cemmap working paper, Centre for Microdata Methods and Practice.
[5] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2017). Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal.
[6] Chernozhukov, V., Hansen, C. and Spindler, M. (2016). hdm: High-Dimensional Metrics. R Journal 8 185-199.
[7] Neumeyer, N., Noh, H. and Van Keilegom, I. (2016). Heteroscedastic semiparametric transformation models: estimation and testing for validity. Statistica Sinica 26 925–954.
[8] Quiroz, A. J., Nakamura, M. and Pérez, F. J. (1996). Estimation of a multivariate Box-Cox transformation to elliptical symmetry via the empirical characteristic function. Annals of the Institute of Statistical Mathematics 48 687–709.
[9] Van der Vaart, A. and Wellner, J. (1996). Weak convergence and empirical processes.