Supplement to "Bounded support in linear random coefficient models: Identification and variable selection"

Philipp Hermann and Hajo Holzmann*
Department of Mathematics and Computer Science
Philipps-Universität Marburg
{herm, holzmann}@mathematik.uni-marburg.de

February 26, 2024

Abstract

As supplementary material for the paper 'Bounded support in linear random coefficient models: Identification and variable selection' we provide the proofs of the results in Section 3.1, a primal-dual witness condition of the adaptive LASSO as well as an analysis for estimating the mean vector with diverging number of parameters.

6 Proofs for Section 3.1

Proof of Proposition 3.1

Proof of Proposition 3.1. From Theorem 2.4, under the assumptions of the proposition the matrix

$$S = \left[\mathrm{v}\Big(\big(1, \mathbf{W_1}^\top\big)^\top \Big), \ldots, \mathrm{v}\Big(\big(1, \mathbf{W}_{\mathbf{p(p+1)/2}}^\top\big)^\top \Big) \right]^\top$$

is of full rank with positive probability. Therefore, the random positive semi-definite matrix

$$\frac{1}{n} \left(\mathbb{X}_n^\sigma\right)^\top \mathbb{X}_n^\sigma = \frac{1}{n} \sum_{i=1}^n \mathrm{v}\Big(\big(1, \mathbf{W_i}^\top\big)^\top \Big) \, \mathrm{v}\Big(\big(1, \mathbf{W_i}^\top\big)^\top \Big)^\top$$

for $n \geq p(p+1)/2$ is positive definite with positive probability. Hence its expected value, which equals C^σ, is positive definite. □

Proof of Theorem 3.2

Turning to the proof of Theorem 3.2, recall the decomposition

$$\varepsilon_n^\sigma = \delta_n + \zeta_n + \xi_n \tag{6.1}$$

*Corresponding author. Prof. Dr. Hajo Holzmann, Department of Mathematics and Computer Science, Philipps-Universität Marburg, Hans-Meerweinstr., 35043 Marburg, Germany

with

$$\begin{aligned}\delta_n &\coloneqq \Big(\mathrm{v}\big(\mathbf{X_1}\big)^\top \mathrm{vec}\big(D_1 - \Sigma^*\big), \ldots, \mathrm{v}\big(\mathbf{X_n}\big)^\top \mathrm{vec}\big(D_n - \Sigma^*\big)\Big)^\top, \\ \zeta_n &\coloneqq \Big(\mathrm{v}\big(\mathbf{X_1}\big)^\top \mathrm{vec}\big(E_n\big), \ldots, \mathrm{v}\big(\mathbf{X_n}\big)^\top \mathrm{vec}\big(E_n\big)\Big)^\top, \\ \xi_n &\coloneqq \Big(\mathrm{v}\big(\mathbf{X_1}\big)^\top \mathrm{vec}\big(F_{n,1}\big), \ldots, \mathrm{v}\big(\mathbf{X_n}\big)^\top \mathrm{vec}\big(F_{n,n}\big)\Big)^\top.\end{aligned} \tag{6.2}$$

of the error term (3.3).

Lemma 6.1. *Under the conditions of Theorem 3.2, we have that*

$$\frac{1}{\sqrt{n}} \big(\mathbb{X}_n^\sigma\big)^\top \big(\zeta_n + \xi_n\big) = \mathrm{o}_{\mathbb{P}}(1).$$

The proofs of the previous as well as the following lemma are deferred to the end of this section.

Lemma 6.2. *Set* $Z_n^{\sigma,1} = \frac{1}{\sqrt{n}} \big(\mathbb{X}_n^\sigma\big)^\top \delta_n$, *then*

$$\mathbb{E}\big[Z_n^{\sigma,1} \,\big|\, \mathbb{X}_n^\sigma\big] = \mathbf{0}_{p(p+1)/2} \quad \textit{and} \quad \mathbb{C}\mathrm{ov}\big(Z_n^{\sigma,1} \,\big|\, \mathbb{X}_n^\sigma\big) = \frac{1}{n} \big(\mathbb{X}_n^\sigma\big)^\top \Omega_n^\sigma \, \mathbb{X}_n^\sigma,$$

where Ω_n^σ *is a diagonal matrix with entries* $\mathrm{v}(\mathbf{X_1})^\top \Psi^* \, \mathrm{v}(\mathbf{X_1}), \ldots, \mathrm{v}(\mathbf{X_n})^\top \Psi^* \, \mathrm{v}(\mathbf{X_n})$. *In particular,* $\mathbb{C}\mathrm{ov}(Z_n^{\sigma,1}) = \mathrm{B}^\sigma$ *and* $Z_n^{\sigma,1} = \mathcal{O}_{\mathbb{P}}(1)$.

Proof of Theorem 3.2. We shall use the primal-dual witness characterization of the adaptive LASSO in Lemma 7.1 in this supplement, Section 7, to prove the sign-consistency (3.7), and the Lindeberg-Feller central limit theorem for random vectors, see van der Vaart (1998, Proposition 2.27), to prove the asymptotic normality (3.8). For more details see also the proof of Theorem 3.5. By Lemmas 6.1 and 6.2, setting

$$\mathrm{P}_{\mathbb{X}^\sigma_{\mathrm{n,S_\sigma}}} = \mathrm{I}_n - \mathbb{X}^\sigma_{n,S_\sigma} \Big(\big(\mathbb{X}^\sigma_{n,S_\sigma}\big)^\top \mathbb{X}^\sigma_{n,S_\sigma}\Big)^{-1} \big(\mathbb{X}^\sigma_{n,S_\sigma}\big)^\top,$$

we have that

$$\frac{1}{\sqrt{n}} \big(\mathbb{X}^\sigma_{n,S_\sigma^c}\big)^\top \mathrm{P}_{\mathbb{X}^\sigma_{\mathrm{n,S_\sigma}}} \varepsilon_n^\sigma = \mathcal{O}_{\mathbb{P}}(1).$$

In addition, the requirements $\sqrt{n}\,\lambda_n^\sigma \to 0$ and $\sqrt{n}\,\big(\widehat{\sigma}_n^{\,\mathrm{init}} - \sigma^*\big) = \mathcal{O}_{\mathbb{P}}(1)$ in Theorem 3.2 lead to

$$0 \leq \frac{\sqrt{n}\,\lambda_n^\sigma}{\big|\widehat{\sigma}_{n,k}^{\,\mathrm{init}}\big|} \leq \frac{\sqrt{n}\,\lambda_n^\sigma}{\Big|\big|\sigma_k^*\big| - \big|\widehat{\sigma}_{n,k}^{\,\mathrm{init}} - \sigma_k^*\big|\Big|} \xrightarrow{\mathbb{P}} 0 \tag{6.3}$$

for all $k \in S_\sigma$ since $|\sigma_k^*| > 0$ for these k. This implies

$$\begin{aligned}\sqrt{n} \Bigg[\big(\mathbb{X}^\sigma_{n,S_\sigma^c}\big)^\top \mathbb{X}^\sigma_{n,S_\sigma} \Big(\big(\mathbb{X}^\sigma_{n,S_\sigma}\big)^\top \mathbb{X}^\sigma_{n,S_\sigma}\Big)^{-1} \Bigg(\lambda_n^\sigma \left(\frac{1}{\big|\widehat{\sigma}^{\,\mathrm{init}}_{n,S_\sigma}\big|} \odot \mathrm{sign}\big(\sigma^*_{S_\sigma}\big)\right)\Bigg) + \frac{1}{n} \big(\mathbb{X}^\sigma_{n,S_\sigma^c}\big)^\top \mathrm{P}_{\mathbb{X}^\sigma_{\mathrm{n,S_\sigma}}} \varepsilon_n^\sigma\Bigg] \\ = \mathcal{O}_{\mathbb{P}}(1)\,\mathrm{o}_{\mathbb{P}}(1) + \mathcal{O}_{\mathbb{P}}(1) = \mathcal{O}_{\mathbb{P}}(1).\end{aligned} \tag{6.4}$$

Moreover, $\sqrt{n}\,(\widehat{\sigma}_n^{\,\mathrm{init}} - \sigma^*) = \mathcal{O}_{\mathbb{P}}(1)$ implies also $\sqrt{n}\,\widehat{\sigma}_{n,k}^{\,\mathrm{init}} = \mathcal{O}_{\mathbb{P}}(1)$ for all $k \in S_\sigma^c$ since $\sigma_k^* = 0$ for these k. Thus, by the second requirement $n\,\lambda_n^\sigma \to \infty$ on the regularization parameter it follows that

$$\frac{\sqrt{n}\,\lambda_n^\sigma}{\bigl|\widehat{\sigma}_{n,k}^{\,\mathrm{init}}\bigr|} = \frac{n\,\lambda_n^\sigma}{\sqrt{n}\,\bigl|\widehat{\sigma}_{n,k}^{\,\mathrm{init}}\bigr|} \xrightarrow{\mathbb{P}} \infty$$

for all $k \in S_\sigma^c$. Together with (6.4) this implies the first condition (7.1) of Lemma 7.1 with high probability for a sufficient large number n of observations. Furthermore, let

$$\widetilde{\sigma}_{n,S_\sigma} = \sigma_{S_\sigma}^* + \left(\frac{1}{n}\left(\mathbb{X}_{n,S_\sigma}^\sigma\right)^\top \mathbb{X}_{n,S_\sigma}^\sigma\right)^{-1}\left(\frac{1}{n}\left(\mathbb{X}_{n,S_\sigma}^\sigma\right)^\top \varepsilon_n^\sigma - \lambda_n^\sigma\left(\frac{1}{|\widehat{\sigma}_{n,S_\sigma}^{\,\mathrm{init}}|} \odot \operatorname{sign}\bigl(\sigma_{S_\sigma}^*\bigr)\right)\right).$$

Then we obtain

$$\sqrt{n}\,\bigl(\widetilde{\sigma}_{n,S_\sigma} - \sigma_{S_\sigma}^*\bigr) = \left(\frac{1}{n}\left(\mathbb{X}_{n,S_\sigma}^\sigma\right)^\top \mathbb{X}_{n,S_\sigma}^\sigma\right)^{-1} \frac{1}{\sqrt{n}}\left(\mathbb{X}_{n,S_\sigma}^\sigma\right)^\top \varepsilon_n^\sigma + \mathrm{o}_{\mathbb{P}}(1)$$

by (6.3). Moreover, with Lemmas 6.1 and 6.2 it follows that

$$\begin{aligned}\sqrt{n}\,\bigl(\widetilde{\sigma}_{n,S_\sigma} - \sigma_{S_\sigma}^*\bigr) &= \left(\frac{1}{n}\left(\mathbb{X}_{n,S_\sigma}^\sigma\right)^\top \mathbb{X}_{n,S_\sigma}^\sigma\right)^{-1} \frac{1}{\sqrt{n}}\left(\mathbb{X}_{n,S_\sigma}^\sigma\right)^\top \delta_n + \mathrm{o}_{\mathbb{P}}(1) \\ &= \mathcal{O}_{\mathbb{P}}(1) + \mathrm{o}_{\mathbb{P}}(1) = \mathcal{O}_{\mathbb{P}}(1)\,,\end{aligned} \tag{6.5}$$

which leads to $\widetilde{\sigma}_{n,S_\sigma} - \sigma_{S_\sigma}^* = \mathrm{o}_{\mathbb{P}}(1)$. Therefore the second condition, $\operatorname{sign}\bigl(\widetilde{\sigma}_{n,S_\sigma}\bigr) = \operatorname{sign}\bigl(\sigma_{S_\sigma}^*\bigr)$, of Lemma 7.1 is also satisfied with high probability for large n. Sign-consistency of the adaptive LASSO and $\widehat{\sigma}_{n,S_\sigma}^{\mathrm{AL}} = \widetilde{\sigma}_{n,S_\sigma}$ is the consequence.

Note that for the asymptotic normality (3.8) of the rescaled estimation error only the first term in (6.5) is crucial. Hence we consider the random vectors

$$Z_n^{\sigma,1} = \frac{1}{\sqrt{n}}\left(\mathbb{X}_n^\sigma\right)^\top \delta_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n \bigl(e_i^\top \delta_n\bigr)\,\mathrm{v}(\mathbf{X_i}) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \Bigl(\mathrm{v}(\mathbf{X_i})^\top \mathrm{vec}\bigl(D_i - \Sigma^*\bigr)\Bigr)\,\mathrm{v}(\mathbf{X_i})\,,$$

where $D_i = \bigl(\mathbf{A_i} - \mu^*\bigr)\bigl(\mathbf{A_i} - \mu^*\bigr)^\top$ and δ_n is defined in (6.2). Now we want to apply the Lindeberg-Feller central limit theorem for the array

$$Q_{n,i} = \frac{1}{\sqrt{n}}\Bigl(\mathrm{v}(\mathbf{X_i})^\top \mathrm{vec}\bigl(D_i - \Sigma^*\bigr)\Bigr)\,\mathrm{v}(\mathbf{X_i})\,, \qquad i = 1, \ldots, n\,,$$

of random vectors. These are independent and identically distributed in each row (for fixed n) since $(\mathbf{X_1}^\top, \mathbf{A_1}^\top)^\top, \ldots, (\mathbf{X_n}^\top, \mathbf{A_n}^\top)^\top$ are independent and identically distributed. Furthermore, they are centered,

$$\mathbb{E}\bigl[Q_{n,i}\bigr] = \frac{1}{\sqrt{n}}\,\mathbb{E}\Bigl[\mathbb{E}\Bigl[\mathrm{v}(\mathbf{X_i})^\top \mathrm{vec}\bigl(D_i - \Sigma^*\bigr)\,\Big|\,\mathbb{X}_n^\sigma\Bigr]\,\mathrm{v}(\mathbf{X_i})\Bigr] = \frac{1}{\sqrt{n}}\,\mathbb{E}\bigl[0 \cdot \mathrm{v}(\mathbf{X_i})\bigr] = \mathbf{0}_{p(p+1)/2}\,,$$

and for the sum of the covariance matrices

$$\sum_{i=1}^n \mathbb{C}\mathrm{ov}\bigl(Q_{n,i}\bigr) = \mathbb{C}\mathrm{ov}\left(\sum_{i=1}^n Q_{n,i}\right) = \mathbb{C}\mathrm{ov}\bigl(Z_n^{\sigma,1}\bigr)$$

we get by Lemma 6.2

$$\sum_{i=1}^{n} \mathbb{C}\mathrm{ov}\big(Q_{n,i}\big) = \mathrm{B}^{\sigma}\,.$$

Moreover, we obtain for arbitrary $\delta > 0$ the equation

$$\sum_{i=1}^{n} \mathbb{E}\Big[\|Q_{n,i}\|_2^2\ \mathbb{1}\big\{\|Q_{n,i}\|_2 > \delta\big\}\Big] = \mathbb{E}\Big[\mathrm{v}\big(\mathbf{X}\big)^{\top}\mathrm{vec}\big(D-\Sigma^*\big)\,\mathrm{v}\big(\mathbf{X}\big)^{\top}\mathrm{vec}\big(D-\Sigma^*\big)\,\mathrm{v}\big(\mathbf{X}\big)^{\top}\mathrm{v}\big(\mathbf{X}\big) \cdot \mathbb{1}\big\{\mathrm{v}(\mathbf{X})^{\top}\mathrm{vec}(D-\Sigma^*)\,\mathrm{v}(\mathbf{X})^{\top}\mathrm{vec}(D-\Sigma^*)\,\mathrm{v}(\mathbf{X})^{\top}\mathrm{v}(\mathbf{X}) > \delta^2 n\big\}\Big]\,.$$

The expected mean $\mathbb{E}\big[\mathrm{v}(\mathbf{X})^{\top}\mathrm{vec}(D-\Sigma^*)\,\mathrm{v}(\mathbf{X})^{\top}\mathrm{vec}(D-\Sigma^*)\,\mathrm{v}(\mathbf{X})^{\top}\mathrm{v}(\mathbf{X})\big]$ exists because of Assumption 1 and the Cauchy Schwarz inequality. Thus we get

$$\lim_{n\to\infty}\sum_{i=1}^{n} \mathbb{E}\Big[\|Q_{n,i}\|_2^2\ \mathbb{1}\big\{\|Q_{n,i}\|_2 > \delta\big\}\Big] = 0$$

by Lebesgue's dominated convergence theorem, which coincides with Lindeberg's condition, see van der Vaart (1998, Proposition 2.27). Hence the mentioned proposition implies the weak convergence

$$Z_n^{\sigma,1} = \frac{1}{\sqrt{n}}\big(\mathbb{X}_n^{\sigma}\big)^{\top}\delta_n = \sum_{i=1}^{n} Q_{n,i} \ \xrightarrow{d}\ Q \sim \mathcal{N}_{p(p+1)/2}\Big(\mathbf{0}_{p(p+1)/2}, \mathrm{B}^{\sigma}\Big)\,,$$

respectively

$$\frac{1}{\sqrt{n}}\big(\mathbb{X}_{n,S_\sigma}^{\sigma}\big)^{\top}\delta_n \ \xrightarrow{d}\ Q_{S_\sigma} \sim \mathcal{N}_{s_\sigma}\big(\mathbf{0}_{s_\sigma}, \mathrm{B}_{S_\sigma S_\sigma}^{\sigma}\big)\,.$$

So all in all a multivariate version of Slutsky's theorem, see for example van der Vaart (1998, Theorem 2.7, Lemma 2.8), together with equation (6.5) leads to

$$\sqrt{n}\,\big(\widehat{\sigma}_{n,S_\sigma}^{\mathrm{AL}} - \sigma_{S_\sigma}^{*}\big) \ \xrightarrow{d}\ \big(\mathrm{C}_{\mathrm{S}_\sigma\mathrm{S}_\sigma}^{\sigma}\big)^{-1} Q_{S_\sigma}\,.$$

In addition, it follows that

$$\big(\mathrm{C}_{\mathrm{S}_\sigma\mathrm{S}_\sigma}^{\sigma}\big)^{-1} Q_{S_\sigma} \sim \mathcal{N}_{s_\sigma}\Big(\mathbf{0}_{s_\sigma}, \big(\mathrm{C}_{\mathrm{S}_\sigma\mathrm{S}_\sigma}^{\sigma}\big)^{-1}\mathrm{B}_{S_\sigma S_\sigma}^{\sigma}\big(\mathrm{C}_{\mathrm{S}_\sigma\mathrm{S}_\sigma}^{\sigma}\big)^{-1}\Big)$$

by the symmetry of $\mathrm{C}_{\mathrm{S}_\sigma\mathrm{S}_\sigma}^{\sigma}$ and the properties of the multivariate normal distribution, and hence the asymptotic normality (3.8). □

Proof of Lemma 6.1. We prove Lemma 6.1 in two steps. First we show that

$$\frac{1}{\sqrt{n}}\big(\mathbb{X}_n^{\sigma}\big)^{\top}\zeta_n = \mathrm{o}_{\mathbb{P}}\,(1)\,. \tag{6.6}$$

We obtain

$$\frac{1}{\sqrt{n}}\left(\mathbb{X}_n^{\sigma}\right)^{\top}\zeta_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(e_i^{\top}\zeta_n\right)\mathrm{v}\left(\mathbf{X_i}\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\mathrm{v}\left(\mathbf{X_i}\right)^{\top}\mathrm{vec}\left(E_n\right)\Big)\mathrm{v}\left(\mathbf{X_i}\right)$$
$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Bigg(\sum_{q=1}^{p(p+1)/2}\mathrm{v}\left(\mathbf{X_i}\right)_q\mathrm{vec}\left(E_n\right)_q\Bigg)\mathrm{v}\left(\mathbf{X_i}\right)$$
$$= \sum_{q=1}^{p(p+1)/2}\sqrt{n}\,\mathrm{vec}\left(E_n\right)_q\Bigg(\frac{1}{n}\sum_{i=1}^{n}\mathrm{v}\left(\mathbf{X_i}\right)_q\mathrm{v}\left(\mathbf{X_i}\right)\Bigg), \tag{6.7}$$

where

$$E_n = \left(\mu^* - \widehat{\mu}_n\right)\left(\mu^* - \widehat{\mu}_n\right)^{\top}.$$

By the assumption on $\widehat{\mu}_n$ we get $e_k^{\top}E_n\,e_l = \left(\widehat{\mu}_{n,k} - \mu_k^*\right)\left(\widehat{\mu}_{n,l} - \mu_l^*\right) = \mathcal{O}_{\mathbb{P}}\left(1/n\right)$ for $k,l \in \{1\ldots,p\}$, and hence also

$$\sqrt{n}\,\mathrm{vec}\left(E_n\right)_q = \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right) \tag{6.8}$$

for all $q \in \{1,\ldots,p(p+1)/2\}$. Furthermore, the random vectors $Q_i^q = \mathrm{v}\left(\mathbf{X_i}\right)_q\mathrm{v}\left(\mathbf{X_i}\right)$ are independent and identically distributed with

$$\mathbb{E}\Big[\left\|Q_i^q\right\|_2\Big] \le \mathbb{E}\Big[\left\|Q_i^q\right\|_1\Big] = \mathbb{E}\Bigg[\Big\|\mathrm{v}\left(\mathbf{X_i}\right)_q\mathrm{v}\left(\mathbf{X_i}\right)\Big\|_1\Bigg] = \sum_{r=1}^{p(p+1)/2}\mathbb{E}\Bigg[\Big|\mathrm{v}\left(\mathbf{X_i}\right)_r\mathrm{v}\left(\mathbf{X_i}\right)_q\Big|\Bigg] < \infty,$$

so that by the law of large numbers

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{v}\left(\mathbf{X_i}\right)_q\mathrm{v}\left(\mathbf{X_i}\right) = \mathcal{O}_{\mathbb{P}}\left(1\right) \tag{6.9}$$

for all $q \in \{1,\ldots,p(p+1)/2\}$ follows. In summary, (6.7), (6.8) and (6.9) lead to (6.6).

In the second step, consider

$$\frac{1}{\sqrt{n}}\left(\mathbb{X}_n^{\sigma}\right)^{\top}\xi_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Bigg(\sum_{q=1}^{p(p+1)/2}\mathrm{v}\left(\mathbf{X_i}\right)_q\mathrm{vec}\left(F_{n,i}\right)_q\Bigg)\mathrm{v}\left(\mathbf{X_i}\right),$$

where $F_{n,i} = \left(\mathbf{A_i} - \mu^*\right)\left(\mu^* - \widehat{\mu}_n\right)^{\top} + \left(\mu^* - \widehat{\mu}_n\right)\left(\mathbf{A_i} - \mu^*\right)^{\top}$. Then we obtain analogously

$$\frac{1}{\sqrt{n}}\left(\mathbb{X}_n^{\sigma}\right)^{\top}\xi_n = \sum_{k,l=1}^{p}\sqrt{n}\left(\widehat{\mu}_{n,k} - \mu_k^*\right)\Bigg(-\frac{2}{n}\sum_{i=1}^{n}X_{i,k}X_{i,l}\left(A_{i,l} - \mu_l^*\right)\mathrm{v}\left(\mathbf{X_i}\right)\Bigg)$$
$$= \mathcal{O}_{\mathbb{P}}\left(1\right)\mathrm{o}_{\mathbb{P}}\left(1\right) = \mathrm{o}_{\mathbb{P}}\left(1\right),$$

since

$$\mathbb{E}\Big[X_{1,k}X_{1,l}\left(A_{1,l} - \mu_l^*\right)\mathrm{v}\left(\mathbf{X_1}\right)\Big] = 0$$

by the independence of $\mathbf{X_1}$ and $\mathbf{A_1}$. □

Proof of Lemma 6.2. We obtain by simple calculation $\mathbb{E}\big[\delta_n \,\big|\, \mathbb{X}_n^\sigma\big] = \mathbf{0}_n$ and $\mathbb{C}\mathrm{ov}\big(\delta_n \,\big|\, \mathbb{X}_n^\sigma\big) = \Omega_n^\sigma$, hence

$$\mathbb{E}\big[Z_n^{\sigma,1} \,\big|\, \mathbb{X}_n^\sigma\big] = \frac{1}{\sqrt{n}} \left(\mathbb{X}_n^\sigma\right)^\top \mathbb{E}\big[\delta_n \,\big|\, \mathbb{X}_n^\sigma\big] = \mathbf{0}_{p(p+1)/2}$$

and

$$\mathbb{C}\mathrm{ov}\big(Z_n^{\sigma,1} \,\big|\, \mathbb{X}_n^\sigma\big) = \frac{1}{n} \left(\mathbb{X}_n^\sigma\right)^\top \mathbb{C}\mathrm{ov}\big(\delta_n \,\big|\, \mathbb{X}_n^\sigma\big)\, \mathbb{X}_n^\sigma = \frac{1}{n} \left(\mathbb{X}_n^\sigma\right)^\top \Omega_n^\sigma\, \mathbb{X}_n^\sigma \,.$$

For random variables Q_1, Q_2 and Q_3 the law of total covariance implies the decomposition

$$\mathbb{C}\mathrm{ov}\big(Q_1, Q_2\big) = \mathbb{E}\Big[\mathbb{C}\mathrm{ov}\big(Q_1, Q_2 \,\big|\, Q_3\big)\Big] + \mathbb{C}\mathrm{ov}\Big(\mathbb{E}\big[Q_1 \,\big|\, Q_3\big], \mathbb{E}\big[Q_2 \,\big|\, Q_3\big]\Big) \,.$$

This can be extended to random vectors and covariance matrices and hence we obtain

$$\begin{aligned} \mathbb{C}\mathrm{ov}\big(Z_n^{\sigma,1}\big) &= \mathbb{E}\Big[\mathbb{C}\mathrm{ov}\big(Z_n^{\sigma,1} \,\big|\, \mathbb{X}_n^\sigma\big)\Big] + \mathbb{C}\mathrm{ov}\Big(\mathbb{E}\big[Z_n^{\sigma,1} \,\big|\, \mathbb{X}_n^\sigma\big]\Big) \\ &= \mathbb{E}\left[\frac{1}{n} \left(\mathbb{X}_n^\sigma\right)^\top \Omega_n^\sigma\, \mathbb{X}_n^\sigma\right] = \mathrm{B}^\sigma \,. \end{aligned}$$

Boundedness in probability follows since by the law of large numbers,

$$\frac{1}{n} \left(\mathbb{X}_n^\sigma\right)^\top \Omega_n^\sigma\, \mathbb{X}_n^\sigma = \frac{1}{n} \sum_{i=1}^n \Big(\mathrm{v}\big(\mathbf{X_i}\big)^\top \Psi^*\, \mathrm{v}\big(\mathbf{X_i}\big)\Big)\, \mathrm{v}\big(\mathbf{X_i}\big)\, \mathrm{v}\big(\mathbf{X_i}\big)^\top \to \mathrm{B}^\sigma \,.$$

□

7 The adaptive LASSO

We look for a fixed number $n \in \mathbb{N}$ of observations at the ordinary linear regression model

$$\mathbb{Y}_n = \mathbb{X}_n\, \beta^* + \varepsilon_n \,,$$

where $\mathbb{Y}_n \in \mathbb{R}^n$ is the vector of the response variables, $\mathbb{X}_n \in \mathbb{R}^{n \times p}$ the deterministic design matrix, $\beta^* \in \mathbb{R}^p$ the unknown coefficient vector and $\varepsilon_n \in \mathbb{R}^n$ represents additive noise. Moreover, we allow the coefficients β^* to be sparse, in other words it is $s \leq p$ for

$$S = \mathrm{supp}\big(\beta^*\big) = \Big\{k \in \{1, \ldots, p\} \,\Big|\, \beta_k^* \neq 0\Big\} \,, \qquad s = |S| \,.$$

In addition, let $S^c = \{1, \ldots, p\} \setminus S$ be the relative complement of S. Because of the sparsity of the coefficients the linear regression model can also be expressed by

$$\mathbb{Y}_n = \mathbb{X}_{n,S}\, \beta_S^* + \varepsilon_n \,.$$

Consider the adaptive LASSO estimator with regularization parameter $\lambda_n > 0$, given by

$$\widehat{\beta}_n^{\mathrm{AL}} \in \rho_{n,\lambda_n}^{\mathrm{AL}} := \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \left(\frac{1}{n} \left\|\mathbb{Y}_n - \mathbb{X}_n\, \beta\right\|_2^2 + 2\lambda_n \sum_{k=1}^p \frac{|\beta_k|}{\big|\widehat{\beta}_{n,k}^{\mathrm{init}}\big|}\right) ,$$

where $\widehat{\beta}_n^{\mathrm{init}} \in \mathbb{R}^p$ is an initial estimator of β^*. If $\widehat{\beta}_{n,k}^{\mathrm{init}} = 0$, we require $\beta_k = 0$ in the above definition.

Lemma 7.1 (Primal-dual witness characterization of the adaptive LASSO)**.** *Assume* $s \leq n$ *and* $\operatorname{rank}(\mathbb{X}_{n,S}) = s$*. If*

$$\left| \mathbb{X}_{n,S^c}^\top \mathbb{X}_{n,S} \Big(\mathbb{X}_{n,S}^\top \mathbb{X}_{n,S} \Big)^{-1} \lambda_n \left(\frac{1}{\big|\widehat{\beta}_{n,S}^{\text{init}}\big|} \odot \operatorname{sign}\big(\beta_S^*\big) \right) + \frac{1}{n} \mathbb{X}_{n,S^c}^\top \mathrm{P}_{\mathbb{X}_{\mathrm{n,S}}^\perp} \varepsilon_n \right| < \frac{\lambda_n}{\big|\widehat{\beta}_{n,S^c}^{\text{init}}\big|} \tag{7.1}$$

with

$$\mathrm{P}_{\mathbb{X}_{\mathrm{n,S}}^\perp} \coloneqq \mathrm{I}_n - \mathbb{X}_{n,S} \Big(\mathbb{X}_{n,S}^\top \mathbb{X}_{n,S} \Big)^{-1} \mathbb{X}_{n,S}^\top$$

holds, and

$$\widetilde{\beta}_{n,S} = \beta_S^* + \left(\frac{1}{n} \mathbb{X}_{n,S}^\top \mathbb{X}_{n,S} \right)^{-1} \left(\frac{1}{n} \mathbb{X}_{n,S}^\top \varepsilon_n - \lambda_n \left(\frac{1}{\big|\widehat{\beta}_{n,S}^{\text{init}}\big|} \odot \operatorname{sign}\big(\beta_S^*\big) \right) \right)$$

satisfies $\operatorname{sign}\big(\widetilde{\beta}_{n,S}\big) = \operatorname{sign}\big(\beta_S^*\big)$*, then the unique adaptive LASSO solution* $\rho_{n,\lambda_n}^{\mathrm{AL}} = \big\{\widehat{\beta}_n^{\mathrm{AL}}\big\}$ *satisfies*

$$\operatorname{sign}\big(\widehat{\beta}_n^{\mathrm{AL}}\big) = \operatorname{sign}\big(\beta^*\big)\,,\ \widehat{\beta}_{n,S}^{\mathrm{AL}} = \widetilde{\beta}_{n,S} \text{ and } \widehat{\beta}_{n,S^c}^{\mathrm{AL}} = \mathbf{0}_{|S^c|}\,.$$

Proof. Cf. Lemma 12.1 in Zhou et al. (2009) with $\vec{w} = \big(1/|\widehat{\beta}_{n,1}^{\text{init}}|, \ldots, 1/|\widehat{\beta}_{n,p}^{\text{init}}|\big)^\top \in \mathbb{R}^p$. □

8 Estimating the means with diverging number p of parameters

The model is given in vector-matrix form by

$$\mathbb{Y}_n^\mu = \mathbb{X}_n^\mu\, \mu^* + \varepsilon_n^\mu\,,$$

where

$$\mathbb{Y}_n^\mu \coloneqq \big(Y_1, \ldots, Y_n\big)^\top, \quad \mathbb{X}_n^\mu \coloneqq \big[\mathbf{X_1}, \ldots, \mathbf{X_n}\big]^\top, \quad \varepsilon_n^\mu \coloneqq \Big(\mathbf{X_1}^\top\big(\mathbf{A_1} - \mu^*\big), \ldots, \mathbf{X_n}^\top\big(\mathbf{A_n} - \mu^*\big)\Big)^\top.$$

Then the adaptive LASSO estimator with regularization parameter $\lambda_n^\mu > 0$ is given by

$$\widehat{\mu}_n^{\mathrm{AL}} \in \rho_{\mu,n,\lambda_n^\mu}^{\mathrm{AL}} \coloneqq \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \left(\frac{1}{n} \left\| \mathbb{Y}_n^\mu - \mathbb{X}_n^\mu\, \beta \right\|_2^2 + 2\lambda_n^\mu \sum_{k=1}^p \frac{|\beta_k|}{\big|\widehat{\mu}_{n,k}^{\text{init}}\big|} \right), \tag{8.1}$$

where $\widehat{\mu}_n^{\text{init}} \in \mathbb{R}^p$ is an initial estimator of μ^*. Note that if $\widehat{\mu}_{n,k}^{\text{init}} = 0$, we require again $\beta_k = 0$. Further, let

$$\mathrm{C}^\mu \coloneqq \mathbb{E}\big[\mathbf{X}\,\mathbf{X}^\top\big]\,, \qquad\qquad \mathrm{B}^\mu \coloneqq \mathbb{E}\Big[\big(\mathbf{X}^\top \Sigma^*\,\mathbf{X}\big)\,\mathbf{X}\,\mathbf{X}^\top\Big]\,,$$

and we denote by

$$S_\mu \coloneqq \operatorname{supp}(\mu^*) = \Big\{k \in \{1, \ldots, p\} \,\Big|\, \mu_k^* \neq 0\Big\}\,, \qquad s_\mu \coloneqq |S_\mu|\,,$$

the support of the mean vector μ^*. $S_\mu^c \coloneqq \{1, \ldots, p\} \setminus S_\mu$ is again the corresponding relative complement.

Assumption 3 (Growing p)**.** We assume that $(\mathbf{X_i}^\top, \mathbf{A_i}^\top)^\top$, $i = 1, \dots, n$, are identically distributed, and that

(A9) the random coefficients $\mathbf{A}$ have finite second moments,

(A10) the covariate vector $\mathbf{X}$ is sub-Gaussian,

(A11) $c_{\mathrm{C}^\mu,\mathrm{l}} \le \lambda_{\min}\big(\mathrm{C}^\mu\big) \le \lambda_{\max}\big(\mathrm{C}^\mu\big) \le c_{\mathrm{C}^\mu,\mathrm{u}}$ for some positive constants $0 < c_{\mathrm{C}^\mu,\mathrm{l}} \le c_{\mathrm{C}^\mu,\mathrm{u}} < \infty$, where $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the minimal and maximal eigenvalues of a symmetric matrix A,

(A12) $\lambda_{\max}\big(\mathrm{B}^\mu\big) \le c_{\mathrm{B}^\mu,\mathrm{u}}$ for some positive constant $c_{\mathrm{B}^\mu,\mathrm{u}} > 0$,

(A13) $\lim_{n\to\infty} p/n = 0$.

Theorem 8.1 (Variable selection for growing p)**.** *Let Assumption 3 be satisfied, and assume that for the initial estimator $\widehat{\mu}_n^{\mathrm{init}}$ in the adaptive LASSO $\widehat{\mu}_n^{\mathrm{AL}}$ in (8.1) we have $\sqrt{n/p}\,\big\|\widehat{\mu}_n^{\mathrm{init}} - \mu^*\big\|_2 = \mathcal{O}_{\mathbb{P}}(1)$. Moreover, if the regularization parameter is chosen as $\lambda_n^\mu \to 0$,*

$$\sqrt{s_\mu\, n}\,\lambda_n^\mu \,/(\mu^*_{\min}\sqrt{p}) \to 0\,, \quad \sqrt{p}/(\mu^*_{\min}\sqrt{n}) \to 0\,, \quad n\,\lambda_n^\mu/p \to \infty$$

with $\mu^*_{\min} := \min_{k\in S_\mu} |\mu_k^*|$, then it follows that $\widehat{\mu}_n^{\mathrm{AL}}$ is sign-consistent,

$$\mathbb{P}\Big(\mathrm{sign}\big(\widehat{\mu}_n^{\mathrm{AL}}\big) = \mathrm{sign}\big(\mu^*\big)\Big) \to 1\,. \tag{8.2}$$

For the proof of Theorem 8.1 we need the following auxiliary lemma.

Lemma 8.2. *Set $Z_n^\mu = \frac{1}{n}\left(\mathbb{X}_n^\mu\right)^\top \varepsilon_n^\mu$, then $\|Z_n^\mu\|_2 = \mathcal{O}_{\mathbb{P}}\left(\sqrt{p/n}\right)$.*

Proof of Lemma 8.2. It is

$$\begin{aligned}
\mathbb{E}\Big[\|Z_n^\mu\|_2^2\Big] &= \frac{1}{n^2}\,\mathbb{E}\Big[(\varepsilon_n^\mu)^\top \mathbb{X}_n^\mu\,(\mathbb{X}_n^\mu)^\top \varepsilon_n^\mu\Big] = \frac{1}{n^2}\,\mathbb{E}\Big[\mathrm{trace}\big((\mathbb{X}_n^\mu)^\top \varepsilon_n^\mu\,(\varepsilon_n^\mu)^\top \mathbb{X}_n^\mu\big)\Big] \\
&= \frac{1}{n^2}\,\mathbb{E}\Big[\mathrm{trace}\big((\mathbb{X}_n^\mu)^\top \mathbb{E}\big[\varepsilon_n^\mu\,(\varepsilon_n^\mu)^\top \,\big|\, \mathbb{X}_n^\mu\big]\,\mathbb{X}_n^\mu\big)\Big] \\
&= \frac{1}{n}\,\mathrm{trace}\left(\mathbb{E}\left[\frac{1}{n}\,(\mathbb{X}_n^\mu)^\top \Omega_n^\mu\,\mathbb{X}_n^\mu\right]\right),
\end{aligned}$$

where $\Omega_n^\mu = \mathbb{C}\mathrm{ov}\big(\varepsilon_n^\mu \,\big|\, \mathbb{X}_n^\mu\big)$ is a diagonal matrix with entries $\mathbf{X_1}^\top \Sigma^*\,\mathbf{X_1}^\top, \dots, \mathbf{X_n}^\top \Sigma^*\,\mathbf{X_n}^\top$. It is obvious that

$$\mathbb{E}\left[\frac{1}{n}\,(\mathbb{X}_n^\mu)^\top \Omega_n^\mu\,\mathbb{X}_n^\mu\right] = \mathrm{B}^\mu\,,$$

and hence we obtain by Assumption (A12) the estimate

$$\mathbb{E}\Big[\|Z_n^\mu\|_2^2\Big] = \frac{\mathrm{trace}\big(\mathrm{B}^\mu\big)}{n} \le \frac{\lambda_{\max}\big(\mathrm{B}^\mu\big)\,p}{n} \le \frac{c_{\mathrm{B}^\mu,\mathrm{u}}\,p}{n}\,.$$

Markov's inequality implies the assertion. □

Proof of Theorem 8.1. We shall use the primal-dual witness characterization of the adaptive LASSO in Lemma 7.1 in Section 7 to prove the sign-consistency (8.2). We obtain by Assumption (A10) and Wainwright (2019, Theorem 6.5) that

$$\left\| \frac{1}{n} (\mathbb{X}_n^\mu)^\top \mathbb{X}_n^\mu - \mathrm{C}^\mu \right\|_{\mathrm{M},2} = \mathcal{O}_{\mathbb{P}} \left(\sqrt{p/n} \right),$$

which implies together with the Assumptions (A11) and (A13) the invertibility of the Gram matrix for large n, and hence by Loh and Wainwright (2017, Lemma 11) we get also

$$\left\| \left(\frac{1}{n} (\mathbb{X}_n^\mu)^\top \mathbb{X}_n^\mu \right)^{-1} - \left(\mathrm{C}^\mu \right)^{-1} \right\|_{\mathrm{M},2} = \mathcal{O}_{\mathbb{P}} \left(\sqrt{p/n} \right).$$

Furthermore, basic properties of the ℓ_2 operator norm lead to

$$\left\| \left(\mathbb{X}_{n,S_\mu^c}^\mu \right)^\top \mathbb{X}_{n,S_\mu}^\mu \left(\left(\mathbb{X}_{n,S_\mu}^\mu \right)^\top \mathbb{X}_{n,S_\mu}^\mu \right)^{-1} - \mathrm{C}_{S_\mu^c S_\mu}^\mu \left(\mathrm{C}_{S_\mu S_\mu}^\mu \right)^{-1} \right\|_{\mathrm{M},2} = \mathcal{O}_{\mathbb{P}} \left(\sqrt{p/n} \right).$$

In particular, this implies

$$\left\| \left(\frac{1}{n} (\mathbb{X}_n^\mu)^\top \mathbb{X}_n^\mu \right)^{-1} \right\|_{\mathrm{M},2} = \mathcal{O}_{\mathbb{P}} (1), \qquad \left\| \left(\mathbb{X}_{n,S_\mu^c}^\mu \right)^\top \mathbb{X}_{n,S_\mu}^\mu \left(\left(\mathbb{X}_{n,S_\mu}^\mu \right)^\top \mathbb{X}_{n,S_\mu}^\mu \right)^{-1} \right\|_{\mathrm{M},2} = \mathcal{O}_{\mathbb{P}} (1). \tag{8.3}$$

Moreover, let $\widehat{\mu}_{n,\min}^{\mathrm{init}} := \min_{k \in S_\mu} |\widehat{\mu}_{n,k}^{\mathrm{init}}|$, then

$$\left| \frac{\widehat{\mu}_{n,\min}^{\mathrm{init}} - \mu_{\min}^*}{\mu_{\min}^*} \right| \leq \frac{1}{\mu_{\min}^*} \left\| \widehat{\mu}_n^{\mathrm{init}} - \mu^* \right\|_2 = \mathcal{O}_{\mathbb{P}} \left(\frac{\sqrt{p}}{\mu_{\min}^* \sqrt{n}} \right) = \mathrm{o}_{\mathbb{P}} (1)$$

since $\sqrt{n/p} \left\| \widehat{\mu}_n^{\mathrm{init}} - \mu^* \right\|_2 = \mathcal{O}_{\mathbb{P}} (1)$ and $\sqrt{p}/(\mu_{\min}^* \sqrt{n}) \to 0$. This implies

$$\left(1 + \frac{\widehat{\mu}_{n,\min}^{\mathrm{init}} - \mu_{\min}^*}{\mu_{\min}^*} \right)^{-1} = \mathcal{O}_{\mathbb{P}} (1),$$

and hence we obtain

$$\begin{aligned}
\sqrt{\frac{n}{p}} \left\| \lambda_n^\mu \left(\frac{1}{|\widehat{\mu}_{n,S_\mu}^{\mathrm{init}}|} \odot \operatorname{sign}\left(\mu_{S_\mu}^* \right) \right) \right\|_2 &\leq \frac{\sqrt{n}\, \lambda_n^\mu}{\sqrt{p}} \left\| \frac{1}{|\widehat{\mu}_{n,S_\mu}^{\mathrm{init}}|} \right\|_2 \leq \frac{\sqrt{s_\mu\, n}\, \lambda_n^\mu}{\sqrt{p}} \left\| \frac{1}{|\widehat{\mu}_{n,S_\mu}^{\mathrm{init}}|} \right\|_\infty \\
&= \frac{\sqrt{s_\mu\, n}\, \lambda_n^\mu}{\sqrt{p}} \left(\widehat{\mu}_{n,\min}^{\mathrm{init}} \right)^{-1} \\
&= \frac{\sqrt{s_\mu\, n}\, \lambda_n^\mu}{\sqrt{p}} \left(\mu_{\min}^* \right)^{-1} \left(1 + \frac{\widehat{\mu}_{n,\min}^{\mathrm{init}} - \mu_{\min}^*}{\mu_{\min}^*} \right)^{-1} \\
&= \frac{\sqrt{s_\mu\, n}\, \lambda_n^\mu}{\mu_{\min}^* \sqrt{p}}\, \mathcal{O}_{\mathbb{P}} (1) \\
&= \mathrm{o}_{\mathbb{P}} (1)
\end{aligned} \tag{8.4}$$

since $\sqrt{s_\mu\, n}\, \lambda_n^\mu / (\mu^*_{\min}\sqrt{p}) \to 0$ by assumption. It follows that

$$\begin{aligned}
&\sqrt{\frac{n}{p}} \left\| \left(\mathbb{X}^\mu_{n,S^c_\mu}\right)^\top \mathbb{X}^\mu_{n,S_\mu} \left(\left(\mathbb{X}^\mu_{n,S_\mu}\right)^\top \mathbb{X}^\mu_{n,S_\mu} \right)^{-1} \left(\lambda_n^\mu \left(\frac{1}{|\widehat{\mu}^{\text{init}}_{n,S_\mu}|} \odot \operatorname{sign}\left(\mu^*_{S_\mu}\right) \right) \right) + \frac{1}{n} \left(\mathbb{X}^\mu_{n,S^c_\mu}\right)^\top \mathrm{P}_{\mathbb{X}^\mu_{\mathrm{n},\mathrm{S}_\mu}} \varepsilon^\mu_n \right\|_2 \\
&\quad \le \left\| \left(\mathbb{X}^\mu_{n,S^c_\mu}\right)^\top \mathbb{X}^\mu_{n,S_\mu} \left(\left(\mathbb{X}^\mu_{n,S_\mu}\right)^\top \mathbb{X}^\mu_{n,S_\mu} \right)^{-1} \right\|_{\mathrm{M},2} \sqrt{\frac{n}{p}} \left\| \lambda_n^\mu \left(\frac{1}{|\widehat{\mu}^{\text{init}}_{n,S_\mu}|} \odot \operatorname{sign}\left(\mu^*_{S_\mu}\right) \right) \right\|_2 \\
&\qquad + \sqrt{\frac{n}{p}} \left\| \frac{1}{n} \left(\mathbb{X}^\mu_{n,S^c_\mu}\right)^\top \varepsilon^\mu_n \right\|_2 + \left\| \left(\mathbb{X}^\mu_{n,S^c_\mu}\right)^\top \mathbb{X}^\mu_{n,S_\mu} \left(\left(\mathbb{X}^\mu_{n,S_\mu}\right)^\top \mathbb{X}^\mu_{n,S_\mu} \right)^{-1} \right\|_{\mathrm{M},2} \sqrt{\frac{n}{p}} \left\| \frac{1}{n} \left(\mathbb{X}^\mu_{n,S_\mu}\right)^\top \varepsilon^\mu_n \right\|_2 \\
&\quad = \mathcal{O}_{\mathbb{P}}(1)\, \mathrm{o}_{\mathbb{P}}(1) + \mathcal{O}_{\mathbb{P}}(1) + \mathcal{O}_{\mathbb{P}}(1) \\
&\quad = \mathcal{O}_{\mathbb{P}}(1) \qquad (8.5)
\end{aligned}$$

by Lemma 8.2 and (8.3), where

$$\mathrm{P}_{\mathbb{X}^\mu_{\mathrm{n},\mathrm{S}_\mu}} = \mathrm{I}_n - \mathbb{X}^\mu_{n,S_\mu} \left(\left(\mathbb{X}^\mu_{n,S_\mu}\right)^\top \mathbb{X}^\mu_{n,S_\mu} \right)^{-1} \left(\mathbb{X}^\mu_{n,S_\mu}\right)^\top .$$

Furthermore, it is

$$\frac{|\widehat{\mu}^{\text{init}}_{n,k}|}{\lambda^\mu_n} \le \frac{\left\|\widehat{\mu}^{\text{init}}_{n,S^c_\mu}\right\|_2}{\lambda^\mu_n} = \frac{\left\|\widehat{\mu}^{\text{init}}_{n,S^c_\mu} - \mu^*_{S^c_\mu}\right\|_2}{\lambda^\mu_n} \le \frac{\left\|\widehat{\mu}^{\text{init}}_{n} - \mu^*\right\|_2}{\lambda^\mu_n} = \frac{\sqrt{n/p}\left\|\widehat{\mu}^{\text{init}}_{n} - \mu^*\right\|_2}{\sqrt{n/p}\,\lambda^\mu_n}$$

for all $k \in S^c$. The condition $\sqrt{n/p}\left\|\widehat{\mu}^{\text{init}}_n - \mu^*\right\|_2 = \mathcal{O}_{\mathbb{P}}(1)$ together with $n\,\lambda^\mu_n/p \to \infty$ implies the convergence

$$\frac{|\widehat{\mu}^{\text{init}}_{n,k}|}{\sqrt{n/p}\,\lambda^\mu_n} = \frac{1}{n\,\lambda^\mu_n/p}\,\mathcal{O}_{\mathbb{P}}(1) = \mathrm{o}_{\mathbb{P}}(1).$$

Hence it follows by (8.5) that the first condition (7.1) of Lemma 7.1 is satisfied with high probability for a sufficient large sample size n. Furthermore, let

$$\widetilde{\mu}_{n,S_\mu} = \mu^*_{S_\mu} + \left(\frac{1}{n} \left(\mathbb{X}^\mu_{n,S_\mu}\right)^\top \mathbb{X}^\mu_{n,S_\mu} \right)^{-1} \left(\frac{1}{n} \left(\mathbb{X}^\mu_{n,S_\mu}\right)^\top \varepsilon^\mu_n - \lambda^\mu_n \left(\frac{1}{|\widehat{\mu}^{\text{init}}_{n,S_\mu}|} \odot \operatorname{sign}\left(\mu^*_{S_\mu}\right) \right) \right).$$

Then we obtain

$$\begin{aligned}
\sqrt{\frac{n}{p}} \left\|\widetilde{\mu}_{n,S_\mu} - \mu^*_{S_\mu}\right\|_2 &\le \left\| \left(\frac{1}{n} \left(\mathbb{X}^\mu_{n,S_\mu}\right)^\top \mathbb{X}^\mu_{n,S_\mu} \right)^{-1} \right\|_{\mathrm{M},2} \left(\sqrt{\frac{n}{p}} \left\| \frac{1}{n} \left(\mathbb{X}^\mu_{n,S_\mu}\right)^\top \varepsilon^\mu_n \right\|_2 \right. \\
&\qquad \left. + \sqrt{\frac{n}{p}} \left\| \lambda^\mu_n \left(\frac{1}{|\widehat{\mu}^{\text{init}}_{n,S_\mu}|} \odot \operatorname{sign}\left(\mu^*_{S_\mu}\right) \right) \right\|_2 \right) \\
&= \mathcal{O}_{\mathbb{P}}(1)\big(\mathcal{O}_{\mathbb{P}}(1) + \mathrm{o}_{\mathbb{P}}(1)\big) = \mathcal{O}_{\mathbb{P}}(1)
\end{aligned}$$

by (8.3), (8.4) and Lemma 8.2. In particular, this implies

$$\left\|\widetilde{\mu}_{n,S_\mu} - \mu^*_{S_\mu}\right\|_2 = \mathcal{O}_{\mathbb{P}}\left(\sqrt{p/n}\right) = \mathrm{o}_{\mathbb{P}}(1)$$

by Assumption (A13), and hence the second condition, $\operatorname{sign}\big(\widetilde{\mu}_{n,S_\mu}\big) = \operatorname{sign}\big(\mu^*_{S_\mu}\big)$, of Lemma 7.1 is also satisfied with high probability for large sample sizes n. Sign-consistency of the adaptive LASSO and $\widehat{\mu}^{\mathrm{AL}}_{n,S_\mu} = \widetilde{\mu}_{n,S_\mu}$ is the consequence. □

References

Loh, P.-L. and M. J. Wainwright (2017). Support recovery without incoherence: A case for nonconvex regularization. *The Annals of Statistics 45*(6), 2455–2482.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, Volume 48. Cambridge University Press.

Zhou, S., S. van de Geer, and P. Bühlmann (2009). Adaptive lasso for high dimensional regression and Gaussian graphical modeling. *arXiv preprint arXiv:0903.2515*.