Supplementary Material to "SOFAR: Large-Scale Association Network Learning"

Yoshimasa Uematsu[1], Yingying Fan[1], Kun Chen[2], Jinchi Lv[1] and Wei Lin[3]

University of Southern California[1], University of Connecticut[2] and Peking University[3]

This Supplementary Material contains the proofs of Theorems 1–3 and additional technical details.

A Proofs of main results

To ease the technical presentation, we introduce some necessary notation. Recall that $\mathbf{A}^* = \mathbf{U}^*\mathbf{D}^*$, $\mathbf{B}^* = \mathbf{V}^*\mathbf{D}^*$, $\mathbf{A} = \mathbf{U}\mathbf{D}$, and $\mathbf{B} = \mathbf{V}\mathbf{D}$. Denote by $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{C}} - \mathbf{C}^*$, $\widehat{\mathbf{\Delta}}^d = \widehat{\mathbf{D}} - \mathbf{D}^*$, $\widehat{\mathbf{\Delta}}^a = \widehat{\mathbf{A}} - \mathbf{A}^*$, and $\widehat{\mathbf{\Delta}}^b = \widehat{\mathbf{B}} - \mathbf{B}^*$ the different estimation errors, and $\mathrm{FS}(\widehat{\mathbf{M}}) = |\{(i,j) : \mathrm{sgn}(\widehat{m}_{ij}) \neq \mathrm{sgn}(m^*_{ij})\}|$ the total number of falsely discovered signs of an estimator $\widehat{\mathbf{M}} = (\widehat{m}_{ij})$ for matrix $\mathbf{M}^* = (m^*_{ij})$. For $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_m) \in \mathbb{R}^{m\times m}$, we define $\mathbf{D}^-$ as a diagonal matrix with $\mathrm{rank}(\mathbf{D}^-) = \mathrm{rank}(\mathbf{D})$ and jth diagonal entry $d_j^- = d_j^{-1}1\{d_j > 0\}$, and define $\mathbf{D}^{*-}$ based on $\mathbf{D}^*$ similarly. For any matrices $\mathbf{M}_1$ and $\mathbf{M}_2$, denote by $\langle \mathbf{M}_1, \mathbf{M}_2\rangle = \mathrm{tr}(\mathbf{M}_1^T\mathbf{M}_2)$. Hereafter we use c to denote a generic positive constant whose value may vary from line to line.

A.1 Proof of Theorem 1

We prove the bounds in (11)–(13) separately. Recall that $s = \|\mathbf{C}^*\|_0$ and define a space

$$\mathcal{C}_0 = \{\mathbf{M} \in \mathbb{R}^{p\times q} : m_{ij} = 0 \text{ for } (i,j) \neq S\},$$

where S stands for the support of $\mathbf{C}^*$. We also denote by $\mathcal{C}_0^\perp$ the orthogonal complement of $\mathcal{C}_0$.

Part 1: Proof of bound (11). The proof is composed of two steps. We first derive the *deterministic* error bound (11) under the assumption that

$$\|n^{-1}\mathbf{X}^T\mathbf{E}\|_\infty \leq \lambda_0/2 \tag{A.1}$$

holds almost surely in the first step and then verify that condition (A.1) holds with high probability in the second step.

Step 1. Since the objective function is convex, the global optimality of $\widetilde{\mathbf{C}}$ implies

$$(2n)^{-1}\|\mathbf{Y} - \mathbf{X}\widetilde{\mathbf{C}}\|_F^2 + \lambda_0\|\widetilde{\mathbf{C}}\|_1 \leq (2n)^{-1}\|\mathbf{Y} - \mathbf{X}\mathbf{C}^*\|_F^2 + \lambda_0\|\mathbf{C}^*\|_1.$$

Then letting $\widetilde{\mathbf{\Delta}} \equiv \widetilde{\mathbf{C}} - \mathbf{C}^*$, we see that

$$(2n)^{-1}\|\mathbf{X}\widetilde{\mathbf{\Delta}}\|_F^2 \leq \langle n^{-1}\mathbf{X}^T\mathbf{E}, \widetilde{\mathbf{\Delta}}\rangle + \lambda_0(\|\mathbf{C}^*\|_1 - \|\widetilde{\mathbf{\Delta}} + \mathbf{C}^*\|_1). \tag{A.2}$$

By Hölder's inequality and the assumed condition (A.1), it holds that

$$\langle n^{-1}\mathbf{X}^T\mathbf{E}, \widetilde{\mathbf{\Delta}}\rangle \leq \|n^{-1}\mathbf{X}^T\mathbf{E}\|_\infty\|\widetilde{\mathbf{\Delta}}\|_1 \leq 2^{-1}\lambda_0\|\widetilde{\mathbf{\Delta}}\|_1. \tag{A.3}$$

By the triangle inequality, we have

$$\lambda_0(\|\mathbf{C}^*\|_1 - \|\widetilde{\mathbf{\Delta}} + \mathbf{C}^*\|_1) \leq \lambda_0\|\widetilde{\mathbf{\Delta}}\|_1. \tag{A.4}$$

Therefore, (A.2) together with Lemma 4 in Section B.2 and (A.3)–(A.4) entails that

$$2c_2\|\widetilde{\mathbf{\Delta}}\|_F^2 \leq 2n^{-1}\|\mathbf{X}\widetilde{\mathbf{\Delta}}\|_F^2 \leq 6\lambda_0\|\widetilde{\mathbf{\Delta}}\|_1. \tag{A.5}$$

Meanwhile, since $n^{-1}\|\mathbf{X}\widetilde{\mathbf{\Delta}}\|_F^2$ is nonnegative (A.2) is also bounded from below as

$$0 \leq \langle n^{-1}\mathbf{X}^T\mathbf{E}, \widetilde{\mathbf{\Delta}}\rangle + \lambda_0(\|\mathbf{C}^*\|_1 - \|\widetilde{\mathbf{\Delta}} + \mathbf{C}^*\|_1). \tag{A.6}$$

Note that $\mathbf{C}^*_{\mathcal{C}_0^\perp} = \mathbf{0}$ in our model. Hence it follows from the triangle inequality and decomposability of the nuclear norm that

$$\begin{aligned}
\lambda_0(\|\mathbf{C}^*\|_1 - \|\widetilde{\mathbf{\Delta}} + \mathbf{C}^*\|_1) &= \lambda_0(\|\mathbf{C}^*_{\mathcal{C}_0} + \mathbf{C}^*_{\mathcal{C}_0^\perp}\|_1 - \|\widetilde{\mathbf{\Delta}}_{\mathcal{C}_0} + \widetilde{\mathbf{\Delta}}_{\mathcal{C}_0^\perp} + \mathbf{C}^*_{\mathcal{C}_0} + \mathbf{C}^*_{\mathcal{C}_0^\perp}\|_1) \\
&\leq \lambda_0(\|\mathbf{C}^*_{\mathcal{C}_0}\|_1 + \|\mathbf{C}^*_{\mathcal{C}_0^\perp}\|_1 - \|\mathbf{C}^*_{\mathcal{C}_0} + \widetilde{\mathbf{\Delta}}_{\mathcal{C}_0^\perp}\|_1 + \|\mathbf{C}^*_{\mathcal{C}_0^\perp} + \widetilde{\mathbf{\Delta}}_{\mathcal{C}_0}\|_1) \\
&= \lambda_0(\|\widetilde{\mathbf{\Delta}}_{\mathcal{C}_0}\|_1 - \|\widetilde{\mathbf{\Delta}}_{\mathcal{C}_0^\perp}\|_1).
\end{aligned} \tag{A.7}$$

Thus by (A.3) and (A.7), we can bound (A.6) from above as

$$\begin{aligned}
0 &\leq 2^{-1}\lambda_0\|\widetilde{\mathbf{\Delta}}\|_1 + \lambda_0(\|\widetilde{\mathbf{\Delta}}_{\mathcal{C}_0}\|_1 - \|\widetilde{\mathbf{\Delta}}_{\mathcal{C}_0^\perp}\|_1) \\
&\leq 2^{-1}\lambda_0(\|\widetilde{\mathbf{\Delta}}_{\mathcal{C}_0}\|_1 + \|\widetilde{\mathbf{\Delta}}_{\mathcal{C}_0^\perp}\|_1) + \lambda_0(\|\widetilde{\mathbf{\Delta}}_{\mathcal{C}_0}\|_1 - \|\widetilde{\mathbf{\Delta}}_{\mathcal{C}_0^\perp}\|_1) \\
&= 2^{-1}\lambda_0(3\|\widetilde{\mathbf{\Delta}}_{\mathcal{C}_0}\|_1 - \|\widetilde{\mathbf{\Delta}}_{\mathcal{C}_0^\perp}\|_1),
\end{aligned}$$

which can be equivalently rewritten as

$$\lambda_0\|\widetilde{\mathbf{\Delta}}_{\mathcal{C}_0^\perp}\|_1 \leq 3\lambda_0\|\widetilde{\mathbf{\Delta}}_{\mathcal{C}_0}\|_1. \tag{A.8}$$

We are now ready to derive the error bound. For a generic positive constant c, (A.5) is bounded from

above by the decomposability of the ℓ_1-norm and (A.8) as

$$c\|\widetilde{\mathbf{\Delta}}\|_F^2 \leq \lambda_0\|\widetilde{\mathbf{\Delta}}\|_1 = \lambda_0\|\widetilde{\mathbf{\Delta}}_{\mathcal{C}_0}\|_1 + \lambda_0\|\widetilde{\mathbf{\Delta}}_{\mathcal{C}_0^\perp}\|_1 \leq 4\lambda_0\|\widetilde{\mathbf{\Delta}}_{\mathcal{C}_0}\|_1. \tag{A.9}$$

Using the subspace compatibility conditions (see the proof of Theorem 1 of [53]), we can show that

$$\|\widetilde{\mathbf{\Delta}}_{\mathcal{C}_0}\|_1 \leq s^{1/2}\|\widetilde{\mathbf{\Delta}}_{\mathcal{C}_0}\|_F \leq s^{1/2}\|\widetilde{\mathbf{\Delta}}\|_F.$$

Therefore, with c changed appropriately (A.9) can be further bounded as

$$\|\widetilde{\mathbf{\Delta}}\|_F^2 \leq cs^{1/2}\lambda_0\|\widetilde{\mathbf{\Delta}}\|_F.$$

This consequently yields the desired error bound

$$\|\widetilde{\mathbf{\Delta}}\|_F \leq cs^{1/2}\lambda_0,$$

which completes the first step of the proof.

Step 2. Let $\mathbf{x}_i$ and $\mathbf{e}_j$ denote the ith and jth columns of $\mathbf{X} \in \mathbb{R}^{n\times p}$ and $\mathbf{E} \in \mathbb{R}^{n\times q}$, respectively. Since $\|\mathbf{X}^T\mathbf{E}\|_\infty = \max_{1\leq i\leq p}\max_{1\leq j\leq q}|\mathbf{x}_i^T\mathbf{e}_j|$, using Bonferroni's inequality and the Gaussianity of $\mathbf{e}_j$ we deduce

$$\begin{aligned} P\left(n^{-1}\|\mathbf{X}^T\mathbf{E}\|_\infty \geq \lambda_0\right) &\leq \sum_{i=1}^{p}\sum_{j=1}^{q} P\left(n^{-1}|\mathbf{x}_i^T\mathbf{e}_j| \geq \lambda_0\right) \\ &\leq 2\sum_{i=1}^{p}\sum_{j=1}^{q}\exp\left(-\frac{n^2\lambda_0^2}{2\mathbb{E}|\mathbf{x}_i^T\mathbf{e}_j|^2}\right). \end{aligned} \tag{A.10}$$

Since $\mathbf{e}_j$ is distributed as $N\left(0, \sigma_j^2\mathbf{I}_n\right)$, it holds that

$$\mathbb{E}|\mathbf{x}_i^T\mathbf{e}_j|^2 = \sigma_j^2\mathbf{x}_i^T\mathbf{x}_i \leq \sigma_{\max}^2 n. \tag{A.11}$$

By the assumption $\lambda_0^2 = c_0^2\sigma_{\max}^2 n^{-1}\log(pq)$ and (A.10)–(A.11), the upper bound on the probability in (A.10) can be further bounded from above by

$$2pq\exp\left\{-(c_0^2/2)\log(pq)\right\} = 2(pq)^{1-c_0^2/2},$$

which concludes the proof for bound (11).

Part 2: Proofs of bounds (12) and (13). Both inequalities (12) and (13) are direct consequences of Lemma 3 in Section B.1 and bound (11). This completes the proof of Theorem 1.

A.2 Proof of Theorem 2

Recall that we solve SOFAR in a local neighborhood $\mathcal{P}_n$ of the initial solution $\widetilde{\mathbf{C}}$. It follows that $\|\widehat{\mathbf{\Delta}}\|_F \le \|\widehat{\mathbf{C}} - \widetilde{\mathbf{C}}\|_F + \|\widetilde{\mathbf{C}} - \mathbf{C}^*\|_F \le 3R_n \le cs^{1/2}\lambda_{\max}$, where $\mathcal{P}_n$ is defined in (14), R_n is as in Theorem 1, and c is some generic positive constant. Thus by Lemma 3, we have

$$\|\widehat{\mathbf{\Delta}}^a\|_F + \|\widehat{\mathbf{\Delta}}^b\|_F + \|\widehat{\mathbf{\Delta}}^d\|_F \le c\eta_n\|\widehat{\mathbf{\Delta}}\|_F \tag{A.12}$$

$$\le cs^{1/2}\lambda_{\max}\eta_n, \tag{A.13}$$

where $\eta_n = 1 + \delta^{-1/2}\big(\sum_{j=1}^r (d_1^*/d_j^*)^2\big)^{1/2}$. Note that under Conditions 1 and 2, Lemma 4 and Lemma 1 in Section A.3 entail that

$$\|\widehat{\mathbf{\Delta}}\|_F^2 \le cn^{-1}\|\mathbf{X}\widehat{\mathbf{\Delta}}\|_F^2 \le c\lambda_{\max}\Big(\|\widehat{\mathbf{\Delta}}^d\|_1 + \|\widehat{\mathbf{\Delta}}^a\|_1 + \|\widehat{\mathbf{\Delta}}^b\|_1\Big). \tag{A.14}$$

Furthermore, it follows from the Cauchy–Schwarz inequality and (A.12) that

$$\begin{aligned}
&\|\widehat{\mathbf{\Delta}}^a\|_1 + \|\widehat{\mathbf{\Delta}}^d\|_1 + \|\widehat{\mathbf{\Delta}}^b\|_1 \\
&\le \max\{\|\widehat{\mathbf{\Delta}}^d\|_0, \|\widehat{\mathbf{\Delta}}^a\|_0, \|\widehat{\mathbf{\Delta}}^b\|_0\}^{1/2}\Big(\|\widehat{\mathbf{\Delta}}^a\|_F + \|\widehat{\mathbf{\Delta}}^d\|_F + \|\widehat{\mathbf{\Delta}}^b\|_F\Big) \\
&\le c\eta_n\{\|\widehat{\mathbf{\Delta}}^d\|_0 + \|\widehat{\mathbf{\Delta}}^a\|_0 + \|\widehat{\mathbf{\Delta}}^b\|_0\}^{1/2}\|\widehat{\mathbf{\Delta}}\|_F.
\end{aligned} \tag{A.15}$$

Combining (A.15) and (A.14) leads to

$$\|\widehat{\mathbf{\Delta}}\|_F \le c\lambda_{\max}\eta_n\{\|\widehat{\mathbf{\Delta}}^d\|_0 + \|\widehat{\mathbf{\Delta}}^a\|_0 + \|\widehat{\mathbf{\Delta}}^b\|_0\}^{1/2}. \tag{A.16}$$

We next provide an upper bound for $\|\widehat{\mathbf{\Delta}}^d\|_0 + \|\widehat{\mathbf{\Delta}}^a\|_0 + \|\widehat{\mathbf{\Delta}}^b\|_0$. Since $(\widehat{\mathbf{D}}, \widehat{\mathbf{A}}, \widehat{\mathbf{B}})$ and $(\mathbf{D}^*, \mathbf{A}^*, \mathbf{B}^*)$ are elements in $\mathcal{D} \times \mathcal{A} \times \mathcal{B}$ by Condition 1, we have

$$\mathrm{FS}(\widehat{\mathbf{D}})^{1/2}\tau \le \|\widehat{\mathbf{\Delta}}^d\|_F, \quad \mathrm{FS}(\widehat{\mathbf{A}})^{1/2}\tau \le \|\widehat{\mathbf{\Delta}}^a\|_F, \quad \text{and} \quad \mathrm{FS}(\widehat{\mathbf{B}})^{1/2}\tau \le \|\widehat{\mathbf{\Delta}}^b\|_F. \tag{A.17}$$

By the definition of $\mathrm{FS}(\widehat{\mathbf{A}})$, it holds that $\|\widehat{\mathbf{\Delta}}^a\|_0 \le s_a + \mathrm{FS}(\widehat{\mathbf{A}})$. Similar inequalities hold for $\|\widehat{\mathbf{\Delta}}^b\|_0$ and $\|\widehat{\mathbf{\Delta}}^d\|_0$. Therefore, it follows from (A.17) and (A.12) that

$$\begin{aligned}
\|\widehat{\mathbf{\Delta}}^d\|_0 + \|\widehat{\mathbf{\Delta}}^a\|_0 + \|\widehat{\mathbf{\Delta}}^b\|_0 &\le r + s_a + s_b + \mathrm{FS}(\widehat{\mathbf{D}}) + \mathrm{FS}(\widehat{\mathbf{A}}) + \mathrm{FS}(\widehat{\mathbf{B}}) \\
&\le r + s_a + s_b + \tau^{-2}\Big(\|\widehat{\mathbf{\Delta}}^a\|_F + \|\widehat{\mathbf{\Delta}}^b\|_F + \|\widehat{\mathbf{\Delta}}^d\|_F\Big)^2 \\
&\le r + s_a + s_b + c(\eta_n/\tau)^2\|\widehat{\mathbf{\Delta}}\|_F^2.
\end{aligned} \tag{A.18}$$

Plugging (A.18) into (A.16) yields

$$\|\widehat{\mathbf{\Delta}}\|_F \le c\lambda_{\max}\eta_n\Big(r + s_a + s_b + c(\eta_n/\tau)^2\|\widehat{\mathbf{\Delta}}\|_F^2\Big)^{1/2}.$$

Thus solving for $\|\widehat{\mathbf{\Delta}}\|_F$ gives

$$\|\widehat{\mathbf{\Delta}}\|_F \leq \frac{c(r+s_a+s_b)^{1/2}\lambda_{\max}\eta_n}{\{1-c\lambda_{\max}^2(\eta_n^2/\tau)^2\}^{1/2}}, \tag{A.19}$$

which together with Theorem 1 results in the first inequality in Theorem 2.

Plugging (A.19) into (A.12), we deduce

$$\|\widehat{\mathbf{\Delta}}^a\|_F + \|\widehat{\mathbf{\Delta}}^b\|_F + \|\widehat{\mathbf{\Delta}}^d\|_F \leq \frac{c(r+s_a+s_b)^{1/2}\lambda_{\max}\eta_n^2}{\{1-c\lambda_{\max}^2(\eta_n^2/\tau)^2\}^{1/2}},$$

which along with (A.13) entails the second inequality in Theorem 2. Note that plugging (A.19) into (A.18) and combining terms yield

$$\begin{aligned}\|\widehat{\mathbf{\Delta}}^d\|_0 + \|\widehat{\mathbf{\Delta}}^a\|_0 + \|\widehat{\mathbf{\Delta}}^b\|_0 &\leq (r+s_a+s_b)\left[1+\frac{c\lambda_{\max}^2(\eta_n^2/\tau)^2}{1-c\lambda_{\max}^2(\eta_n^2/\tau)^2}\right] \\ &= (r+s_a+s_b)[1+o(1)],\end{aligned}$$

which gives the third inequality in Theorem 2.

We now plug the above inequality and (A.19) into (A.15). Then it holds that

$$\|\widehat{\mathbf{\Delta}}^a\|_1 + \|\widehat{\mathbf{\Delta}}^d\|_1 + \|\widehat{\mathbf{\Delta}}^b\|_1 \leq \frac{c(r+s_a+s_b)\lambda_{\max}\eta_n^2}{1-c\lambda_{\max}^2(\eta_n^2/\tau)^2}, \tag{A.20}$$

which yileds the fourth inequality in Theorem 2. Finally, it follows from Lemma 1 and (A.20) that

$$n^{-1}\|\mathbf{X}\widehat{\mathbf{\Delta}}\|_F^2 \leq \frac{c(r+s_a+s_b)\lambda_{\max}^2\eta_n^2}{1-c\lambda_{\max}^2(\eta_n^2/\tau)^2},$$

which establishes the fifth inequality in the theorem and concludes the proof of Theorem 2.

A.3 Lemma 1 and its proof

Lemma 1. *Under the conditions of Theorem 2, with at least probability as specified in (15) we have*

$$n^{-1}\|\mathbf{X}\widehat{\mathbf{\Delta}}\|_F^2 \leq c\lambda_{\max}\left(\|\widehat{\mathbf{\Delta}}^d\|_1 + \|\widehat{\mathbf{\Delta}}^a\|_1 + \|\widehat{\mathbf{\Delta}}^b\|_1\right),$$

where c is some positive constant.

Proof of Lemma 1. Denote by $\mathcal{E}_2$ the event on which inequalities (A.25)–(A.27) hold. Then by Lemma 2 in Section A.4, we see that event $\mathcal{E}_2$ holds with probability bound as specified in (15). We will prove Lemma 1 by conditioning on event $\mathcal{E}_2$. Since the SOAR estimator is the minimizer in the

neighborhood $\mathcal{P}_n$ defined in (14), it holds that

$$
\begin{aligned}
&(2n)^{-1}\|\mathbf{Y}-\mathbf{X}\widehat{\mathbf{U}}\widehat{\mathbf{D}}\widehat{\mathbf{V}}^T\|_F^2+\lambda_d\|\widehat{\mathbf{D}}\|_1+\lambda_a\rho_a(\widehat{\mathbf{A}})+\lambda_b\rho_b(\widehat{\mathbf{B}})\\
&\leq (2n)^{-1}\|\mathbf{Y}-\mathbf{X}\mathbf{U}^*\mathbf{D}^*(\mathbf{V}^*)^T\|_F^2+\lambda_d\|\mathbf{D}^*\|_1+\lambda_a\rho_a(\mathbf{A}*)+\lambda_b\rho_b(\mathbf{B}^*).
\end{aligned}
$$

Let $\widehat{\mathbf{\Delta}}=\widehat{\mathbf{C}}-\mathbf{C}^*$. Rearranging terms in the above inequality leads to

$$
\begin{aligned}
(2n)^{-1}\|\mathbf{X}\widehat{\mathbf{\Delta}}\|_F^2 &\leq \langle n^{-1}\mathbf{X}^T\mathbf{E},\widehat{\mathbf{\Delta}}\rangle\\
&+\lambda_d\left(\|\mathbf{D}^*\|_1-\|\widehat{\mathbf{D}}\|_1\right)+\lambda_a\left(\rho_a(\mathbf{A}^*)-\rho_a(\widehat{\mathbf{A}})\right)+\lambda_b\left(\rho_b(\mathbf{B}^*)-\rho_b(\widehat{\mathbf{B}})\right).
\end{aligned}
\tag{A.21}
$$

By the definition of $\mathbf{D}^-$, the estimation error can be decomposed as

$$
\begin{aligned}
\widehat{\mathbf{\Delta}}&\equiv\widehat{\mathbf{U}}\widehat{\mathbf{D}}\widehat{\mathbf{V}}^T-\mathbf{U}^*\mathbf{D}^*\mathbf{V}^{*T}=\widehat{\mathbf{A}}\widehat{\mathbf{D}}^-\widehat{\mathbf{B}}^T-\mathbf{A}^*\mathbf{D}^{*-}\mathbf{B}^{*T}\\
&=\widehat{\mathbf{\Delta}}^a(\widehat{\mathbf{B}}\widehat{\mathbf{D}}^-)^T-\mathbf{U}^*\widehat{\mathbf{\Delta}}^d(\widehat{\mathbf{B}}\widehat{\mathbf{D}}^-)^T+\mathbf{U}^*(\widehat{\mathbf{\Delta}}^b)^T.
\end{aligned}
$$

The above decomposition together with Hölder's inequality entails that the following inequality

$$
\begin{aligned}
&\langle n^{-1}\mathbf{X}^T\mathbf{E},\widehat{\mathbf{\Delta}}\rangle\\
&=\langle n^{-1}\mathbf{X}^T\mathbf{E}\widehat{\mathbf{B}}\widehat{\mathbf{D}}^-,\widehat{\mathbf{\Delta}}^a\rangle-\langle n^{-1}\mathbf{U}^{*T}\mathbf{X}^T\mathbf{E}\widehat{\mathbf{B}}\widehat{\mathbf{D}}^-,\widehat{\mathbf{\Delta}}^d\rangle+\langle n^{-1}\mathbf{U}^{*T}\mathbf{X}^T\mathbf{E},\widehat{\mathbf{\Delta}}^{bT}\rangle\\
&\leq\|n^{-1}\mathbf{X}^T\mathbf{E}\widehat{\mathbf{B}}\widehat{\mathbf{D}}^-\|_\infty\|\widehat{\mathbf{\Delta}}^a\|_1+\|n^{-1}\mathbf{U}^{*T}\mathbf{X}^T\mathbf{E}\widehat{\mathbf{B}}\widehat{\mathbf{D}}^-\|_\infty\|\widehat{\mathbf{\Delta}}^d\|_1+\|n^{-1}\mathbf{U}^{*T}\mathbf{X}^T\mathbf{E}\|_\infty\|\widehat{\mathbf{\Delta}}^b\|_1\\
&\leq\lambda_a\|\widehat{\mathbf{\Delta}}^a\|_1+\lambda_d\|\widehat{\mathbf{\Delta}}^d\|_1+\lambda_b\|\widehat{\mathbf{\Delta}}^b\|_1
\end{aligned}
\tag{A.22}
$$

holds on event $\mathcal{E}_2$.

By the triangle inequality for the ℓ_1-norm and Condition 4, we deduce

$$
\begin{aligned}
&\lambda_d\left(\|\mathbf{D}^*\|_1-\|\widehat{\mathbf{D}}\|_1\right)+\lambda_a\left(\rho_a(\mathbf{A}^*)-\rho_a(\widehat{\mathbf{A}})\right)+\lambda_b\left(\rho_b(\mathbf{B}^*)-\rho_b(\widehat{\mathbf{B}})\right)\\
&\leq\lambda_d\|\widehat{\mathbf{\Delta}}^d\|_1+\lambda_a\|\widehat{\mathbf{\Delta}}^a\|_1+\lambda_b\|\widehat{\mathbf{\Delta}}^b\|_1.
\end{aligned}
\tag{A.23}
$$

Thus plugging (A.22) and (A.23) into (A.21) yields

$$
\begin{aligned}
(cn)^{-1}\|\mathbf{X}\widehat{\mathbf{\Delta}}\|_F^2&\leq\lambda_d\|\widehat{\mathbf{\Delta}}^d\|_1+\lambda_a\|\widehat{\mathbf{\Delta}}^a\|_1+\lambda_b\|\widehat{\mathbf{\Delta}}^b\|_1\\
&\leq\lambda_{\max}\left(\|\widehat{\mathbf{\Delta}}^d\|_1+\|\widehat{\mathbf{\Delta}}^a\|_1+\|\widehat{\mathbf{\Delta}}^b\|_1\right)
\end{aligned}
\tag{A.24}
$$

with $\lambda_{\max}=\max(\lambda_d,\lambda_a,\lambda_b)$, which completes the proof of Lemma 1.

A.4 Lemma 2 and its proof

Lemma 2. *Under the conditions of Theorem 2, with at least probability as specified in (15) the following inequalities hold*

$$\sup_{(\mathbf{B},\mathbf{D})\in\mathcal{P}_n} \|n^{-1}\mathbf{U}^{*T}\mathbf{X}^T\mathbf{E}\mathbf{B}\mathbf{D}^-\|_\infty \le \lambda_d, \tag{A.25}$$

$$\sup_{(\mathbf{B},\mathbf{D})\in\mathcal{P}_n} \|n^{-1}\mathbf{X}^T\mathbf{E}\mathbf{B}\mathbf{D}^-\|_\infty \le \lambda_a, \tag{A.26}$$

$$\sup_{(\mathbf{B},\mathbf{D})\in\mathcal{P}_n} \|n^{-1}\mathbf{U}^{*T}\mathbf{X}^T\mathbf{E}\|_\infty \le \lambda_b. \tag{A.27}$$

Proof of Lemma 2. Recall that $\widetilde{\mathcal{P}}_n = \{\mathbf{C} : \|\mathbf{C}-\widetilde{\mathbf{C}}\|_F \le 2R_n\}$, where $\widetilde{\mathbf{C}}$ is the initial Lasso estimator and $R_n = c(n^{-1}s\log(pq))^{1/2}$ is as defined in Theorem 1. It follows from Theorem 1 that the true regression coefficient matrix $\mathbf{C}^*$ falls in the neighborhood $\widetilde{\mathcal{P}}_n$ with probability at least $1-2(pq)^{1-c_0^2/2}$, where $c_0 > \sqrt{2}$ is some constant given in Theorem 1. Note that the neighborhood $\widetilde{\mathcal{P}}_n$ shrinks asymptotically as $n\to\infty$ since $R_n^2 = O(n^{\alpha+\beta/2+\gamma-1})$ and $\alpha+\beta/2+\gamma < \alpha+\beta+\gamma < 1$ holds under our assumptions. In order to deal with the nonconvexity of the objective function, we exploit the framework of convexity-assisted nonconvex optimization (CANO) and solve the SOFAR optimization problem in the shrinking local region $\mathcal{P}_n = \widetilde{\mathcal{P}}_n \cap (\mathcal{C}\times\mathcal{D}\times\mathcal{A}\times\mathcal{B})$ as defined in (14).

Observe that for any $\mathbf{C}\in\widetilde{\mathcal{P}}_n$, by the triangle inequality it holds that

$$\|\mathbf{C}-\mathbf{C}^*\|_F \le \|\mathbf{C}-\widetilde{\mathbf{C}}\|_F + \|\widetilde{\mathbf{C}}-\mathbf{C}^*\|_F \le 3R_n;$$

that is, with probability at least $1-2(pq)^{1-c_0^2/2}$, $\widetilde{\mathcal{P}}_n \subset \{\mathbf{C} : \|\mathbf{C}-\mathbf{C}^*\|_F \le 3R_n\}$. Further, by Lemma 3 we have $\{\mathbf{C} : \|\mathbf{C}-\mathbf{C}^*\|_F \le 3R_n\} \subset \mathcal{E}_1$, where

$$\begin{aligned}\mathcal{E}_1 = \{\mathbf{C}\equiv\mathbf{A}\mathbf{D}^-\mathbf{B} : \|\mathbf{D}-\mathbf{D}^*\|_F \le 3R_n,\\ \|\mathbf{A}-\mathbf{A}^*\|_F + \|\mathbf{B}-\mathbf{B}^*\|_F \le 3c\eta_n R_n\}\end{aligned} \tag{A.28}$$

with $c > 0$ some constant. Combining the above results yields that with probability at least $1-2(pq)^{1-c_0^2/2}$, $\mathcal{P}_n \subset \widetilde{\mathcal{P}}_n \subset \mathcal{E}_1$, which entails

$$P\Big(\mathcal{P}_n \not\subset \mathcal{E}_1\Big) \le 2(pq)^{1-c_0^2/2}. \tag{A.29}$$

We next establish that (A.25)–(A.27) hold with asymptotic probability one. Note that it follows from

the definition of conditional probability and (A.29) that

$$
\begin{aligned}
&P\Big(\sup_{\mathbf{C}\in\mathcal{P}_n}\|n^{-1}\mathbf{U}^*\mathbf{X}^T\mathbf{E}\mathbf{B}\mathbf{D}^-\|_\infty>\lambda_d\Big)\\
&\quad\le P\Big(\sup_{\mathbf{C}\in\mathcal{P}_n}\|n^{-1}\mathbf{U}^*\mathbf{X}^T\mathbf{E}\mathbf{B}\mathbf{D}^-\|_\infty>\lambda_d\Big|\mathcal{P}_n\subset\mathcal{E}_1\Big)+P\Big(\mathcal{P}_n\not\subset\mathcal{E}_1\Big)\\
&\quad\le P\Big(\sup_{\mathbf{C}\in\mathcal{E}_1}\|n^{-1}\mathbf{U}^*\mathbf{X}^T\mathbf{E}\mathbf{B}\mathbf{D}^-\|_\infty>\lambda_d)+2(pq)^{1-c_0^2/2}.
\end{aligned}
$$

Thus to prove (A.25), we only need to show that

$$
\sup_{\mathbf{C}\in\mathcal{E}_1}\|n^{-1}\mathbf{U}^*\mathbf{X}^T\mathbf{E}\mathbf{B}\mathbf{D}^-\|_\infty\le\lambda_d \tag{A.30}
$$

holds with asymptotic probability one. Similarly, to show (A.26) and (A.27) we only need to prove that

$$
\sup_{\mathbf{C}\in\mathcal{E}_1}\|n^{-1}\mathbf{X}^T\mathbf{E}\mathbf{B}\mathbf{D}^-\|_\infty\le\lambda_a, \tag{A.31}
$$

$$
\sup_{\mathbf{C}\in\mathcal{E}_1}\|n^{-1}\mathbf{U}^{*T}\mathbf{X}^T\mathbf{E}\|_\infty\le\lambda_b \tag{A.32}
$$

hold with asymptotic probability one. We next proceed to prove (A.30)–(A.32) hold with asymptotic probability one.

Denote by $\mathbf{x}_i$ and $\mathbf{e}_j$ the ith and jth columns of $\mathbf{X}\in\mathbb{R}^{n\times p}$ and $\mathbf{E}\in\mathbb{R}^{n\times q}$, respectively. Let $\mathbf{x}_i^*$ and $\mathbf{e}_j^*$ be the ith and jth columns of $\mathbf{X}^*\equiv\mathbf{X}\mathbf{U}^*\in\mathbb{R}^{n\times q}$ and $\mathbf{E}^*\equiv\mathbf{E}\mathbf{V}^*\in\mathbb{R}^{n\times q}$, respectively. It is seen that the last $q-r$ columns of $\mathbf{X}^*$ and $\mathbf{E}^*$ are all zero. First, we show that (A.30) holds with significant probability. The decomposition

$$
\mathbf{B}\mathbf{D}^-=\mathbf{V}^*+\mathbf{\Delta}^b\mathbf{D}^-+\mathbf{V}^*\mathbf{D}^*\mathbf{\Delta}^{d-}
$$

and the triangle inequality lead to

$$
\|n^{-1}\mathbf{X}^{*T}\mathbf{E}\mathbf{B}\mathbf{D}^-\|_\infty\le\|n^{-1}\mathbf{X}^{*T}\mathbf{E}^*\|_\infty+\|n^{-1}\mathbf{X}^{*T}\mathbf{E}\mathbf{\Delta}^b\mathbf{D}^-\|_\infty+\|n^{-1}\mathbf{X}^{*T}\mathbf{E}^*\mathbf{D}^*\mathbf{\Delta}^{d-}\|_\infty,
$$

where $\mathbf{\Delta}^{d-}=\mathbf{D}^--\mathbf{D}^{*-}=\mathrm{diag}\{d_j^{-1}-(d_j^*)^{-1}\}$. Thus it holds that

$$
\begin{aligned}
&P\left(\sup_{\mathbf{C}\in\mathcal{E}_1}\|n^{-1}\mathbf{X}^{*T}\mathbf{E}\mathbf{B}\mathbf{D}^-\|_\infty\ge\lambda_d\right)\le P\left(\|n^{-1}\mathbf{X}^{*T}\mathbf{E}^*\|_\infty\ge\lambda_d/3\right)\\
&+P\left(\sup_{\mathbf{C}\in\mathcal{E}_1}\|n^{-1}\mathbf{X}^{*T}\mathbf{E}\mathbf{\Delta}^b\mathbf{D}^-\|_\infty\ge\lambda_d/3\right)+P\left(\sup_{\mathbf{C}\in\mathcal{E}_1}\|n^{-1}\mathbf{X}^{*T}\mathbf{E}^*\mathbf{D}^*\mathbf{\Delta}^{d-})\|_\infty\ge\lambda_d/3\right).
\end{aligned}
\tag{A.33}
$$

Let us consider the first term on the right hand side of (A.33). Since $\mathbf{E}\sim N(\mathbf{0},\mathbf{I}_n\otimes\mathbf{\Sigma})$ by Condition 3, the jth column vector of $\mathbf{E}^*$, $\mathbf{e}_j^*=\mathbf{E}\mathbf{v}_j^*$ with $\mathbf{v}_j^*$ the jth column vector of $\mathbf{V}^*$, is distributed as

$N\left(0, \mathbf{v}_j^{*T}\mathbf{\Sigma}\mathbf{v}_j^* I_n\right)$. Furthermore, note that $\|\mathbf{X}^{*T}\mathbf{E}^*\|_\infty = \max_{1\le i\le q}\max_{1\le j\le q}|\mathbf{x}_i^{*T}\mathbf{e}_j^*|$ and

$$\mathbb{E}|\mathbf{x}_i^{*T}\mathbf{e}_j^*|^2 = \mathbf{v}_j^{*T}\mathbf{\Sigma}\mathbf{v}_j^*\mathbf{x}_i^{*T}\mathbf{x}_i^* \le \alpha_{\max}\mathbf{u}_i^{*T}\mathbf{X}^T\mathbf{X}\mathbf{u}_i^* \le \alpha_{\max}c_3 n \le cn, \tag{A.34}$$

where $\alpha_{\max}$ denotes the maximum eigenvalue of $\mathbf{\Sigma}$ and the second inequality follows from Condition 2 and the fact that $\mathbf{u}_i^* = \mathbf{0}$ for $i = r+1, \cdots, q$. Therefore, it follows from Bonferroni's inequality, the Gaussianity of $\mathbf{e}_j^*$, and (A.34) that for $\lambda_d^2 = c_1^2 n^{-1}\log(pr)$,

$$\begin{aligned} P\left(n^{-1}\|\mathbf{X}^{*T}\mathbf{E}^*\|_\infty \ge \lambda_d/3\right) &\le \sum_{i=1}^r\sum_{j=1}^r P\left(n^{-1}|\mathbf{x}_i^{*T}\mathbf{e}_j^*| \ge \lambda_d/3\right) \\ &\le 2\sum_{i=1}^r\sum_{j=1}^r \exp\left(-\frac{n^2\lambda_d^2/9}{2\mathbb{E}|\mathbf{x}_i^{*T}\mathbf{e}_j^*|^2}\right) \\ &\le 2r^2\exp\left(-\frac{n^2c_1^2n^{-1}\log(pr)}{18cn}\right) \\ &= 2r^2(pr)^{-c_1^2/c}. \end{aligned} \tag{A.35}$$

We now consider the second term on the right hand side of (A.33). Some algebra gives

$$\begin{aligned} \|n^{-1}\mathbf{X}^{*T}\mathbf{E}\mathbf{\Delta}^b\mathbf{D}^-\|_\infty &= \|n^{-1}(\mathbf{I}_q\otimes\mathbf{X}^{*T}\mathbf{E})\operatorname{vec}(\mathbf{\Delta}^b\mathbf{D}^-)\|_\infty \\ &\le \max_{1\le i\le r}\sum_{j=1}^q |n^{-1}\mathbf{x}_i^{*T}\mathbf{e}_j|\,\|\operatorname{vec}(\mathbf{\Delta}^b\mathbf{D}^-)\|_\infty \\ &\le q\max_{1\le i\le r}\max_{1\le j\le q}|n^{-1}\mathbf{x}_i^{*T}\mathbf{e}_j|\,\|(\mathbf{D}^-\otimes\mathbf{I}_q)\operatorname{vec}(\mathbf{\Delta}^b)\|_\infty \\ &\le q\|\mathbf{D}^-\|_\infty\max_{1\le i\le r}\max_{1\le j\le q}|n^{-1}\mathbf{x}_i^{*T}\mathbf{e}_j|\,\|\operatorname{vec}(\mathbf{\Delta}^b)\|_\infty. \end{aligned}$$

Since we solve SOFAR in the local neighborhood $\mathcal{P}_n$ defined in (14), by Condition 1 we have $\|\mathbf{D}^-\|_\infty \le \tau^{-1}$ for any $\mathbf{C} \equiv \mathbf{A}\mathbf{D}^-\mathbf{B} \in \mathcal{P}_n$. Thus by (A.28), the second term in the upper bound of (A.33) can be bounded as

$$\begin{aligned} \sup_{\mathbf{C}\in\mathcal{E}_1}\|n^{-1}\mathbf{X}^{*T}\mathbf{E}\mathbf{\Delta}^b\mathbf{D}^-\|_\infty &\le (q/\tau)\max_{1\le i\le r}\max_{1\le j\le q}|n^{-1}\mathbf{x}_i^{*T}\mathbf{e}_j|\sup_{\mathcal{E}_1}\|\operatorname{vec}(\mathbf{\Delta}^b)\|_\infty \\ &\le (q/\tau)\max_{1\le i\le r}\max_{1\le j\le q}|n^{-1}\mathbf{x}_i^{*T}\mathbf{e}_j|\sup_{\mathcal{E}_1}\|\mathbf{\Delta}^b\|_F \\ &\le 3c(q/\tau)\eta_n R_n\max_{1\le i\le r}\max_{1\le j\le q}|n^{-1}\mathbf{x}_i^{*T}\mathbf{e}_j|. \end{aligned} \tag{A.36}$$

Similarly to (A.34), we can show that

$$\mathbb{E}|\mathbf{x}_i^{*T}\mathbf{e}_j|^2 \le \sigma_j^2 c_3 n \le \sigma_{\max}^2 c_3 n \le cn. \tag{A.37}$$

Therefore, in view of (A.36), (A.37), $R_n^2 = O(sn^{-1}\log(pq))$, and $p \ge q$, the same inequality as (A.35)

results in

$$
\begin{aligned}
&P\left(\sup_{\mathcal{E}_1}\|n^{-1}\mathbf{X}^{*T}\mathbf{E}\boldsymbol{\Delta}^{b}\mathbf{D}^{-}\|_\infty \geq \lambda_d/3\right)\\
&= P\left(3c(q/\tau)\eta_n R_n \max_{1\leq i\leq r}\max_{1\leq j\leq q}|n^{-1}\mathbf{x}_i^{*T}\mathbf{e}_j| \geq \lambda_d/3\right)\\
&\leq 2\sum_{i=1}^{r}\sum_{j=1}^{q}\exp\left(-\frac{n^2\lambda_d^2}{81c(q/\tau)^2\eta_n^2 R_n^2\mathbb{E}|\mathbf{x}_i^{*T}\mathbf{e}_j|^2}\right)\\
&= 2qr\exp\left(-\frac{c_1^2 n}{c(q/\tau)^2\eta_n^2 s}\right),
\end{aligned}
\tag{A.38}
$$

where c is some positive constant.

It remains to investigate the third term on the right hand side of (A.33). Since $\mathbf{D}^*\boldsymbol{\Delta}^{d-}$ is a diagonal matrix whose (k,k)th entry is given by $(d_k^* - d_k)/d_k$ with $\mathrm{rank}(\mathbf{D}^*\boldsymbol{\Delta}^{d-}) \leq r$, the last $q-r$ columns of both $\mathbf{X}^*$ and $\mathbf{E}^*$ are zero, and $\mathbf{D}\in\mathcal{D}$, we have

$$
\begin{aligned}
\sup_{\mathbf{C}\in\mathcal{E}_1}\|n^{-1}\mathbf{X}^{*T}\mathbf{E}^*\mathbf{D}^*\boldsymbol{\Delta}^{d-}\|_\infty &\leq \max_{1\leq i\leq r}\max_{1\leq j\leq r}|n^{-1}\mathbf{x}_i^{*T}\mathbf{e}_j^*|\sup_{\mathcal{E}}\|\mathbf{D}^*\boldsymbol{\Delta}^{d-}\|_\infty\\
&\leq \tau^{-1}\max_{1\leq i\leq r}\max_{1\leq j\leq r}|n^{-1}\mathbf{x}_i^{*T}\mathbf{e}_j^*|\max_{1\leq k\leq r}|d_k^* - d_k|\\
&\leq \tau^{-1}\max_{1\leq i\leq r}\max_{1\leq j\leq r}|n^{-1}\mathbf{x}_i^{*T}\mathbf{e}_j^*|\|\boldsymbol{\Delta}^{d}\|_F\\
&\leq 3(R_n/\tau)\max_{1\leq i\leq r}\max_{1\leq j\leq r}|n^{-1}\mathbf{x}_i^{*T}\mathbf{e}_j^*|.
\end{aligned}
\tag{A.39}
$$

Then by (A.34) and (A.39), the same inequality yields

$$
\begin{aligned}
P\left(\sup\|n^{-1}\mathbf{X}^{*T}\mathbf{E}^*\mathbf{D}^*\boldsymbol{\Delta}^{d-}\|_\infty \geq \lambda_d/3\right) &\leq P\left(3(R_n/\tau)\max_{1\leq i\leq r}\max_{1\leq j\leq r}|n^{-1}\mathbf{x}_i^{*T}\mathbf{e}_j^*| \geq \lambda_d/3\right)\\
&\leq 2\sum_{i=1}^{r}\sum_{j=1}^{r}\exp\left(-\frac{n^2\lambda_d^2}{81c(R_n/\tau)^2\mathbb{E}|\mathbf{x}_i^{*T}\mathbf{e}_j^*|^2}\right)\\
&\leq 2r^2\exp\left(-\frac{c_1^2 n^2 n^{-1}\log(pr)}{csn^{-1}\log(pq)\tau^{-2}n}\right)\\
&\leq 2r^2\exp\left(-\frac{c_1^2\tau^2 n}{cs}\right).
\end{aligned}
\tag{A.40}
$$

Therefore, combining (A.35), (A.38), and (A.40) with (A.33) gives the probability bound

$$
\begin{aligned}
&P\left(\sup_{\mathbf{C}\in\mathcal{E}_1}\|n^{-1}\mathbf{X}^{*T}\mathbf{E}\mathbf{B}\mathbf{D}^{-}\|_\infty \geq \lambda_d\right)\\
&\leq 2r^2(pr)^{-c_1^2/c} + 2rq\exp\left(-\frac{c_1^2 n}{c(q/\tau)^2\eta_n^2 s}\right) + 2r^2\exp\left(-\frac{c_1^2\tau^2 n}{cs}\right).
\end{aligned}
\tag{A.41}
$$

We next prove that (A.31) holds with high probability. The arguments are similar to those for proving (A.30) except that $\mathbf{X}^*$ is replaced with $\mathbf{X}$ in the proof of (A.25). More specifically, note that we have

the following decomposition of probability bound

$$P\left(\sup_{\mathcal{E}_1}\|n^{-1}\mathbf{X}^T\mathbf{E}\mathbf{B}\mathbf{D}^-\|_\infty \geq \lambda_a\right) \leq P\left(\|n^{-1}\mathbf{X}^T\mathbf{E}^*\|_\infty \geq \lambda_a/3\right) \tag{A.42}$$
$$+ P\left(\sup_{\mathcal{E}_1}\|n^{-1}\mathbf{X}^T\mathbf{E}\boldsymbol{\Delta}^b\mathbf{D}^-\|_\infty \geq \lambda_a/3\right) + P\left(\sup_{\mathcal{E}_1}\|n^{-1}\mathbf{X}^T\mathbf{E}^*\mathbf{D}^*\boldsymbol{\Delta}^{d-})\|_\infty \geq \lambda_a/3\right).$$

Thus, it suffices to bound the probabilities on the right hand side of (A.42). Let us consider the first term. Observe that

$$\mathbb{E}|\mathbf{x}_i^T\mathbf{e}_j^*|^2 \leq \alpha_{\max}\mathbf{x}_i^T\mathbf{x}_i = \alpha_{\max}n \leq cn,$$

where c is some positive constant. Thus, setting $\lambda_a^2 = c_1^2 n^{-1}\log(pr)$ and noting that $\mathbf{E}^*$ has only r nonzero columns lead to the bound

$$\begin{aligned} P\left(n^{-1}\|\mathbf{X}^T\mathbf{E}^*\|_\infty \geq \lambda_a/3\right) &\leq 2\sum_{i=1}^{p}\sum_{j=1}^{r}\exp\left(-\frac{n^2\lambda_a^2}{8\mathbb{E}|\mathbf{x}_i^T\mathbf{e}_j^*|^2}\right) \\ &\leq 2pr\exp\left(-\frac{c_1^2 n^2 n^{-1}\log(pr)}{cn}\right) \\ &\leq 2(pr)^{1-c_1^2/c}. \end{aligned} \tag{A.43}$$

We next consider the second probability bound on the right hand side of (A.42). Since

$$\mathbb{E}|\mathbf{x}_i^T\mathbf{e}_j|^2 \leq \sigma_{\max}^2\mathbf{x}_i^T\mathbf{x}_i = \sigma_{\max}^2 n \leq cn,$$

by replacing $\max_{1\leq i\leq r}$ in (A.36) and (A.38) with $\max_{1\leq i\leq p}$ we deduce

$$P\left(\sup_{\mathbf{C}\in\mathcal{E}_1}\|n^{-1}\mathbf{X}^T\mathbf{E}\boldsymbol{\Delta}^b\mathbf{D}^-\|_\infty \geq \lambda_a/3\right) \leq 2pr\exp\left(-\frac{c_1^2 n}{c(q/\tau)^2\eta_n^2 s}\right). \tag{A.44}$$

It remains to study the third probability bound on the right hand side of (A.42). Similarly, replacing $\max_{1\leq i\leq r}$ in (A.39) and (A.40) with $\max_{1\leq i\leq p}$ yields

$$P\left(\sup_{\mathbf{C}\in\mathcal{E}_1}\|n^{-1}\mathbf{X}^T\mathbf{E}^*\mathbf{D}^*\boldsymbol{\Delta}^{d-}\|_\infty \geq \lambda_a/3\right) \leq 2pr\exp\left(-\frac{c_1^2\tau^2 n}{cs}\right). \tag{A.45}$$

Thus combining (A.43)–(A.45), we can bound (A.42) as

$$\begin{aligned} &P\left(\sup_{\mathbf{C}\in\mathcal{E}_1}\|n^{-1}\mathbf{X}^T\mathbf{E}\mathbf{B}\mathbf{D}^-\|_\infty \geq \lambda_a\right) \\ &\leq 2(pr)^{1-c_1^2/c} + 2pr\exp\left(-\frac{c_1^2 n}{c(q/\tau)^2\eta_n^2 s}\right) + 2pr\exp\left(-\frac{c_1^2\tau^2 n}{cs}\right). \end{aligned} \tag{A.46}$$

Finally, we show that condition (A.32) holds with large probability. Choosing $\lambda_b^2 = c_1^2 n^{-1}\log(pr)$

results in

$$P\left(n^{-1}\|\mathbf{X}^{*T}\mathbf{E}\|_\infty \geq \lambda_b\right) \leq 2\sum_{i=1}^{r}\sum_{j=1}^{q}\exp\left(-\frac{n^2\lambda_b^2}{2\mathbb{E}|\mathbf{x}_i^{*T}\mathbf{e}_j|^2}\right)$$
$$\leq 2qr\exp\left(-\frac{c_1^2 n^2 n^{-1}\log(pr)}{cn}\right)$$
$$\leq 2qr(pr)^{-c_1^2/c}. \tag{A.47}$$

Consequently, for the given set of regularization parameters $(\lambda_d, \lambda_a, \lambda_b)$ it follows from (A.41), (A.46), and (A.47) that conditions (A.30)–(A.32) hold simultaneously with probability at least

$$1 - \left\{2(pr)^{1-c_1^2/c} + 2pr\exp\left(-\frac{c_1^2 n}{c(q/\tau)^2\eta_n^2 s}\right)\right\},$$

where we have used the facts of $c_1^2 > c$ and $p \geq q \geq 1$. Moreover, to check that the probability bound converges to one, since $c_1^2 > c$ it is sufficient to show that

$$2pr\exp\left(-\frac{c_1^2 n}{c(q/\tau)^2\eta_n^2 s}\right)$$

converges to zero. This follows immediately from the assumptions of $\log p = O(n^\alpha)$, $q = O(n^{\beta/2})$, $s = O(n^\gamma)$, and $\eta_n/\tau = o(n^{(1-\alpha-\beta-\gamma)/2})$, which concludes the proof of Lemma 2.

A.5 Proof of Theorem 3

Recall that the theoretical results for the SOFAR estimator established in the paper hold simultaneously over the set of all local minimizers in a neighborhood of the initial Lasso estimator. Thus we aim to establish the convergence of the SOFAR algorithm when supplied the initial Lasso estimator. Note that the equivalent form of the SOFAR problem (21) with the slack variables $\mathbf{A}$ and $\mathbf{B}$ can be solved using the augmented Lagrangian form with sufficiently large penalty parameter $\mu > 0$. From now on, we fix parameter μ and the set of Lagrangian multipliers $\mathbf{\Gamma}$, and thus work with the objective function $L_\mu(\mathbf{\Theta}, \mathbf{\Omega}; \mathbf{\Gamma})$.

By the nature of the block coordinate descent algorithm applied to $(\mathbf{U}, \mathbf{V}, \mathbf{D}, \mathbf{A}, \mathbf{B})$, the sequence $(L_\mu(\cdot))$ of values of the objective function $L_\mu(\mathbf{\Theta}, \mathbf{\Omega}; \mathbf{\Gamma})$ is decreasing. Clearly the function $L_\mu(\mathbf{\Theta}, \mathbf{\Omega}; \mathbf{\Gamma})$ is bounded from below. Thus the sequence $(L_\mu(\cdot))$ converges. Since the rank parameter m is fixed in the SOFAR algorithm, we assume for simplicity that the diagonal matrix $\mathbf{D}^k$ of singular values has all the diagonal entries bounded away from zero, since otherwise we can solve the SOFAR problem with a smaller rank m.

By assumption, we have

$$\sum_{k=1}^{\infty}[\Delta L_\mu(\mathbf{U}^k)]^{1/2} < \infty,\ \sum_{k=1}^{\infty}[\Delta L_\mu(\mathbf{V}^k)]^{1/2} < \infty,\ \text{ and }\ \sum_{k=1}^{\infty}[\Delta L_\mu(\mathbf{D}^k)]^{1/2} < \infty,$$

where $\Delta L_\mu(\cdot)$ stands for the decrease in $L_\mu(\cdot)$ by a block update. Note that the $\mathbf{U}$-space with constraint $\mathbf{U}^T\mathbf{U} = \mathbf{I}_m$ is a Stiefel manifold which is compact and smooth; see, e.g., [47] for a brief review of the geometry of Stiefel manifold. Since the $\mathbf{D}$-sequence is always positive definite by assumption, the objective function along the $\mathbf{U}$-block with all the other four blocks fixed is convex and has positive curvature bounded away from zero along any direction in the $\mathbf{U}$-space. By definition, $\mathbf{U}^k$ is the minimizer of such a restricted objective function, which entails that the gradient of this function at $\mathbf{U}^k$ on the Stiefel manifold vanishes. Thus it follows easily from the mean value theorem and the fact of positive curvature that $\Delta L_\mu(\mathbf{U}^k)$ is bounded from below by some positive constant δ times $d_g^2(\mathbf{U}^k, \mathbf{U}^{k-1})$, where $d_g(\cdot,\cdot)$ denotes the distance function on the Stiefel manifold. Then it holds that

$$\sum_{k=1}^{\infty} d_g(\mathbf{U}^k, \mathbf{U}^{k-1}) \leq \delta^{-1/2} \sum_{k=1}^{\infty} [\Delta L_\mu(\mathbf{U}^k)]^{1/2} < \infty,$$

which along with the triangle inequality entails that $(\mathbf{U}^k)$ is a Cauchy sequence on the Stiefel manifold. Therefore, the sequence $(\mathbf{U}^k)$ converges to a limit point $\mathbf{U}_*$ on the Stiefel manifold which is a local solution along the $\mathbf{U}$-block. Similarly, we can show that the sequence $(\mathbf{V}^k)$ also converges to a limit point $\mathbf{V}_*$ on the Stiefel manifold that is a local solution along the $\mathbf{V}$-block.

Recall that the diagonal matrix $\mathbf{D}^k$ of singular values is assumed to have all the diagonal entries bounded away from zero. Since we have shown that the sequences $(\mathbf{U}^k)$ and $(\mathbf{V}^k)$ converge to limit points $\mathbf{U}_*$ and $\mathbf{V}_*$ on the Stiefel manifolds, respectively, it follows from the fact that both $\mathbf{U}_*$ and $\mathbf{V}_*$ have full column rank m that as k becomes large, the objective function along the $\mathbf{D}$-block with all the other four blocks fixed is convex and has positive curvature bounded away from zero. Thus an application of similar arguments as above yields that the sequence $(\mathbf{D}^k)$ also converges to a limit point $\mathbf{D}_*$.

With the established convergence results of the sequences $(\mathbf{U}^k)$, $(\mathbf{V}^k)$, and $(\mathbf{D}^k)$, the convergence of the sequences $(\mathbf{A}^k)$ and $(\mathbf{B}^k)$ follows easily from the convergence property of the block coordinate descent algorithm applied to separable convex problems [61], by noting that the objective function with $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{D}$ replaced by their limit points is jointly convex in $\mathbf{A}$ and $\mathbf{B}$ since the penalty functions $\rho_a(\cdot)$ and $\rho_b(\cdot)$ are assumed to be convex. This completes the proof of Theorem 3.

B Additional technical details

B.1 Lemma 3 and its proof

Lemma 3. *Under Condition 5, we have for any matrix* $\mathbf{C} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ *and* $\mathbf{C}^* = \mathbf{U}^*\mathbf{D}^*\mathbf{V}^{*T}$ *with* $\|\mathbf{C} - \mathbf{C}^*\|_2 \leq d_1^*$ *that*

$$\|\mathbf{D} - \mathbf{D}^*\|_F \leq \|\mathbf{C} - \mathbf{C}^*\|_F,$$
$$\|\mathbf{A} - \mathbf{A}^*\|_F + \|\mathbf{B} - \mathbf{B}^*\|_F \leq c\eta_n \|\mathbf{C} - \mathbf{C}^*\|_F,$$

where $\eta_n = 1 + \delta^{-1/2}\big(\sum_{j=1}^{r}(d_1^*/d_j^*)^2\big)^{1/2}$ *and* $c > 0$ *is some constant.*

Proof of Lemma 3. It is well known that the inequality

$$\|\mathbf{D} - \mathbf{D}^*\|_F \le \|\mathbf{C} - \mathbf{C}^*\|_F$$

holds; see, for example, [51]. It remains to show the second desired inequality. Recall that $\mathbf{A}^* = \mathbf{U}^*\mathbf{D}^*$. By the decomposition

$$\mathbf{C} - \mathbf{C}^* = (\mathbf{A} - \mathbf{A}^*)\mathbf{V}^T + \mathbf{A}^*(\mathbf{V} - \mathbf{V}^*)^T$$

and the unitary property of the Frobenius norm, we have

$$\|\mathbf{A} - \mathbf{A}^*\|_F \le \|\mathbf{C} - \mathbf{C}^*\|_F + \|\mathbf{D}^*(\mathbf{V} - \mathbf{V}^*)^T\|_F. \tag{A.48}$$

Let us examine the second term on the right hand side of (A.48). To do so, we apply Theorem 3 of [65] to $\mathbf{V} - \mathbf{V}^*$ columnwise to avoid the identifiability issue. When $r = 1$ or 2, it holds that

$$\|\mathbf{v}_1 - \mathbf{v}_1^*\|_2 \le \frac{cd_1^*\|\mathbf{C} - \mathbf{C}^*\|_F}{\delta^{1/2}(d_1^*)^2}, \quad \|\mathbf{v}_r - \mathbf{v}_r^*\|_2 \le \frac{cd_1^*\|\mathbf{C} - \mathbf{C}^*\|_F}{\delta^{1/2}(d_r^*)^2}. \tag{A.49}$$

When $r \ge 3$, in addition to (A.49) we have for $j = 2, \ldots, r-1$,

$$\|\mathbf{v}_j - \mathbf{v}_j^*\|_2 \le \frac{c(2d_1^* + \|\mathbf{C} - \mathbf{C}^*\|_2)\|\mathbf{C} - \mathbf{C}^*\|_F}{\min(d_{j-1}^{*2} - d_j^{*2}, d_j^{*2} - d_{j+1}^{*2})},$$

where $c > 0$ is some constant. Since Condition 5 gives $d_{j-1}^{*2} - d_j^{*2} \ge \delta^{1/2}(d_{j-1}^*)^2 \ge \delta^{1/2}(d_j^*)^2$, it follows from the assumption $\|\mathbf{C} - \mathbf{C}^*\|_2 \le d_1^*$ that the above inequality can be further bounded as

$$\|\mathbf{v}_j - \mathbf{v}_j^*\|_2 \le \frac{c(2d_1^* + \|\mathbf{C} - \mathbf{C}^*\|_2)\|\mathbf{C} - \mathbf{C}^*\|_F}{\min(d_{j-1}^{*2} - d_j^{*2}, d_j^{*2} - d_{j+1}^{*2})} \le \frac{cd_1^*\|\mathbf{C} - \mathbf{C}^*\|_F}{\delta^{1/2}(d_j^*)^2}.$$

Thus these inequalities entail that

$$\|\mathbf{D}^*(\mathbf{V} - \mathbf{V}^*)^T\|_F^2 = \sum_{j=1}^{r} d_j^{*2}\|\mathbf{v}_j - \mathbf{v}_j^*\|_2^2 \le (c/\delta)\|\mathbf{C} - \mathbf{C}^*\|_F^2 \sum_{j=1}^{r}(d_1^*/d_j^*)^2. \tag{A.50}$$

Consequently, combining (A.48) and (A.50) leads to the bound

$$\|\mathbf{A} - \mathbf{A}^*\|_F \le \|\mathbf{C} - \mathbf{C}^*\|_F + (c/\delta^{1/2})\|\mathbf{C} - \mathbf{C}^*\|_F \left\{\sum_{j=1}^{r}(d_1^*/d_j^*)^2\right\}^{1/2}.$$

On the other hand, the bound for $\|\mathbf{B} - \mathbf{B}^*\|_F$ can be obtained by the decomposition $\mathbf{C} - \mathbf{C}^* = \mathbf{U}(\mathbf{B} - \mathbf{B}^*)^T + (\mathbf{U} - \mathbf{U}^*)\mathbf{B}^{*T}$ and similar arguments. Therefore, adding both bounds together and enlarging

the positive constant c conclude the proof of Lemma 3.

B.2 Lemma 4 and its proof

Lemma 4. *Under Conditions 1 and 2, it holds for any* $\mathbf{C} \in \mathcal{C}$ *that*

$$n^{-1}\|\mathbf{X}(\mathbf{C} - \mathbf{C}^*)\|_F^2 \geq c_2\|\mathbf{C} - \mathbf{C}^*\|_F^2.$$

Proof of Lemma 4. Denote by $\mathbf{\Delta} = \mathbf{C} - \mathbf{C}^*$, $\mathbf{W} = \mathbf{I}_q \otimes \mathbf{X}$, and $\boldsymbol{\delta} = \mathrm{vec}(\mathbf{\Delta})$, where $\mathbf{I}_q$ is the $q \times q$ identity matrix. It follows from the triangle inequality and Condition 1 that

$$\begin{aligned}\|\boldsymbol{\delta}\|_0 &= \|\,\mathrm{vec}(\mathbf{C}) - \mathrm{vec}(\mathbf{C}^*)\|_0 \leq \|\,\mathrm{vec}(\mathbf{C})\|_0 + \|\,\mathrm{vec}(\mathbf{C}^*)\|_0 \\ &< \kappa_{c_2}/2 + \kappa_{c_2}/2 = \kappa_{c_2}.\end{aligned}$$

Note that the singular values of $\mathbf{W}$ are the same as those of the original design matrix $\mathbf{X}$ with the multiplicity of each singular value multiplied by q. This entails that the robust spark of $\mathbf{W}$ is equal to that of $\mathbf{X}$, which is κ_{c_2} for a given positive constant c_2. Thus by the definition of the robust spark, we obtain

$$n^{-1}\|\mathbf{X}\mathbf{\Delta}\|_F^2 = n^{-1}\|\mathbf{W}\boldsymbol{\delta}\|_2^2 = n^{-1}\|\mathbf{W}_{\mathrm{supp}(\boldsymbol{\delta})}\boldsymbol{\delta}_{\mathrm{supp}(\boldsymbol{\delta})}\|_2^2 \geq c_2\|\boldsymbol{\delta}\|_2^2 = c_2\|\mathbf{\Delta}\|_F^2,$$

where the subscript $\mathrm{supp}(\boldsymbol{\delta})$ denotes the restriction of the matrix to the corresponding columns or that of the vector to the corresponding components. This completes the proof of Lemma 4.