SUPPLEMENT: ADAPTIVE ESTIMATION IN STRUCTURED FACTOR MODELS WITH APPLICATIONS TO OVERLAPPING CLUSTERING

By Xin Bing and Florentina Bunea and Yang Ning and Marten Wegkamp

Cornell University

APPENDIX A: APPENDIX

A.1. Proofs of the results from Section 2. We begin by stating and proving two lemmata that are crucial for the main results of this section. All results are proved under the condition that model 1.1 and (i) - (iii) hold.

LEMMA 1. *For any $a \in [K]$ and $i \in I_a$, we have*

(a) $|\Sigma_{ij}| = C_{aa}$ *for all* $j \in I_a$,
(b) $|\Sigma_{ij}| < C_{aa}$ *for all* $j \notin I_a$.

PROOF. For given $i \in [p]$, we define the set $s(i) := \{1 \le a \le K : A_{ia} \neq 0\}$. For any $i \in I_a$ and $j \neq i$, we have

$$\begin{aligned} |\Sigma_{ij}| &= \Big|\sum_{a\in s(i)} A_{ia}\Big(\sum_{b\in s(j)} A_{jb}C_{ab}\Big)\Big| \\ &= \Big|\sum_{b\in s(j)} A_{jb}C_{ab}\Big| \text{ from the definition of } I_a \\ &\le \sum_{b\in s(j)} |A_{jb}| \cdot \max_{b\in s(i)} |C_{ab}| \\ &\le C_{aa} \text{ using conditions (i) and (iii).} \end{aligned}$$

Furthermore, using conditions (i) and (iii), we observe that we have equality in the above display for $j \in I_a$, and strict inequality for $j \notin I_a$, which proves the lemma. □

LEMMA 2. *We have*

(a) $S_i \cap I \neq \varnothing$, *for any* $i \in [p]$,
(b) $S_i \cup \{i\} = I_a$ *and* $M_i = C_{aa}$, *for any* $i \in I_a$ *and* $a \in [K]$,

where M_i and S_i are defined in (2.2) and (2.3), respectively.

PROOF. Lemma 1 implies that, for any $i \in I_a$, $M_i = C_{aa}$ and $S_i = I_a \backslash \{i\}$, which proves part (b).
From the result of part (b), it remains to show $S_i \cap I \neq \varnothing$ for any $i \notin I$. Let $i \notin I$ be fixed. We have

$$\begin{aligned} M_i &= \max_{j \neq i} |\Sigma_{ij}| = \max_{j \neq i} \left| \sum_{b \in s(j)} A_{jb} \left(\sum_{a \in s(i)} A_{ia} C_{ab} \right) \right| \\ &\leq \max_{j \neq i} \max_{b \in s(j)} \left| \sum_{a \in s(i)} A_{ia} C_{ab} \right| = \max_{j \neq i} \left| \sum_{a \in s(i)} A_{ia} C_{ab^*} \right| \end{aligned} \tag{A.1}$$

for some $b^* \in [K]$. A direct computation yields $|\Sigma_{ij}| = |\sum_{a \in s(i)} A_{ia} C_{ab^*}|$ for any $j \in I_{b^*}$, that is, the maximum M_i of $|\Sigma_{ij}|$ is achieved at all $j \in I_{b^*}$. Since $I_{b^*} \neq \varnothing$ by condition (ii), this completes the proof of claim (a). □

Proof of Theorem 1. We have all the necessary ingredients to proceed with the proof of the main result of this section.

*Proof of (**a**).* We first show the sufficiency part. Consider any $i \in [p]$ with $M_i = M_j$ for all $j \in S_i$. Part (a) of Lemma 2 states that there exists a $j \in I_a \cap S_i$ for some $a \in [K]$. For this $j \in I_a$, we have $M_j = C_{aa}$ from part (b) of Lemma 2. Invoking our premise $M_j = M_i$ as $j \in S_i$, we conclude that $M_i = C_{aa}$, that is, $\max_{k \neq i} |\Sigma_{ik}| = C_{aa}$. By Lemma 1, the maximum is achieved for any pair $i, k \in I_a$. However, if $i \notin I_a$, we have that $|\Sigma_{ik}| < C_{aa}$ for all $k \neq i$. Hence $i \in I_a$ and this concludes the proof of the sufficiency part.

It remains to prove the necessity part. Let $i \in I_a$ for some $a \in [K]$ and $j \in S_i$. Lemma 2 implies that $j \in I_a$ and $M_i = C_{aa}$. Since $j \in S_i$, we have $|\Sigma_{ij}| = M_i = C_{aa}$, while $j \in I_a$ yields $|\Sigma_{jk}| \leq C_{aa}$ for all $k \neq j$, and $|\Sigma_{jk}| = C_{aa}$ for $k \in I_a$, as a result of Lemma 1. Hence, $M_j = \max_{k \neq j} |\Sigma_{jk}| = C_{aa} = M_i$ for any $j \in S_i$, which proves our claim.

*Proof of (**b**).* We start with the following constructive approach. Let $N = [p]$ be the set of all variable indices and $O = \varnothing$. Let M_i and S_i be defined in (2.2) and (2.3), respectively.

(1) Choose $i \in N$ and calculate S_i and M_i.
   (a) If $M_i = M_j$, for all $j \in S_i$, set $I^{(i)} := S_i \cup \{i\}$, $O = O \cup \{i\}$ and $N = N \setminus I^{(i)}$.

(b) Otherwise, replace N by $N \setminus \{i\}$.

(2) Repeat step (1) until $N = \varnothing$.

We show that $\{I^{(i)} : \ i \in O\} = \mathcal{I}$. Let $i \in O$ be arbitrary fixed. By (**a**), we have $i \in I$. Thus, there exists $a \in [K]$ such that $i \in I_a$. By Lemma 2, $i \in I_a$ implies $I_a = S_i \cup \{i\} = I^{(i)}$. On the other hand, let $a \in [K]$ be arbitrary fixed. By condition (ii), there exists at least one $j \in I_a$. Once again, by part (b) of Lemma 2, if $j \in I_a$, then $S_j \cup \{j\} = I_a$, that is, $I^{(j)} = I_a$. □

Proof of Theorem 2. Theorem 1 shows that Σ uniquely defines I and its partition $\mathcal{I}$, up to permutation of labels. Given I and its partition $\mathcal{I} = \{I_1, \ldots, I_K\}$, for any $i \in I$, there exists a unique $1 \leq a \leq K$ such that $i \in I_a$. Then we set $|A_{i\cdot}| = e_a$, the canonical basis vector in $\mathbb{R}^K$ that contains 1 in position a and is zero otherwise. Thus, the $|I| \times K$ matrix A_I with rows $A_{i\cdot}$ is uniquely defined up to multiplication with a signed permutation matrix P.

We show below that A_J is also identifiable up to a signed permutation matrix. We begin by observing that, for each $i \in I_k$, for some $k \in [K]$, and any $j \in J$, Model 1.1 implies

$$\Sigma_{ij} = \sum_{a \in s(i)} \sum_{b \in s(j)} A_{ia} A_{jb} C_{kb} = A_{ik} \sum_{b \in s(j)} A_{jb} C_{kb}$$

and since $A_{ik}^2 = 1$, we obtain

$$A_{ik} \Sigma_{ij} = C_{k\cdot}^T A_{j\cdot}$$

and, after averaging over all $i \in I_k$,

$$C_{k\cdot}^T A_{j\cdot} = \frac{1}{|I_k|} \sum_{i \in I_k} A_{ik} \Sigma_{ij}.$$

Repeating this for every $k \in [K]$, we obtain the formula

$$CA_{j\cdot} \quad = \quad \left(\frac{1}{|I_1|} \sum_{i \in I_1} A_{i1} \Sigma_{ij}, \ldots, \frac{1}{|I_K|} \sum_{i \in I_K} A_{iK} \Sigma_{ij} \right)^T := \theta^j.$$

The covariance matrix C can be uniquely constructed from Σ via

$$C_{aa} \quad = \quad \frac{1}{|I_a|(|I_a| - 1)} \sum_{i,j \in I_a, i \neq j} |\Sigma_{ij}|$$

for any $a \in [K]$, and

$$C_{ab} \quad = \quad \frac{1}{|I_a||I_b|} \sum_{i \in I_a, j \in I_b} A_{ia} A_{jb} \Sigma_{ij}$$

for $a, b \in [K]$ with $a \neq b$. Notice that $\min_{a \in [K]} |I_a| \geq 2$, which is part of our model requirement (ii), is needed for the construction of C_{aa}. Since the covariance matrix C is assumed to be positive definite, $A_{j\cdot} = C^{-1}\theta^j$, for each $j \in J$, which shows that A_J can be determined uniquely from Σ up to a signed permutation. Therefore, A_J is identifiable which concludes the proof. □

A.2. Proofs of the results from Section 4.1 . The proof of Theorem 3 will repeatedly use Lemma 3, stated and proved below. Let

$$\widehat{M}_i := \max_{j \in [p] \setminus \{i\}} |\widehat{\Sigma}_{ij}|. \tag{A.2}$$

LEMMA 3. *Under the conditions in Theorem 3, for any $i \in I_a$ with some $a \in [K]$, the following inequalities hold on the event $\mathcal{E}$:*

$$\left| |\widehat{\Sigma}_{ij}| - |\widehat{\Sigma}_{ik}| \right| \leq 2\delta, \qquad \text{for all } j, k \in I_a \setminus \{i\} \text{ and } j \neq k; \tag{A.3}$$

$$|\widehat{\Sigma}_{ij}| - |\widehat{\Sigma}_{ik}| > 2\delta, \qquad \text{for all } j \in I_a \setminus \{i\},\ k \notin (I_a \cup J_1^a); \tag{A.4}$$

$$|\widehat{\Sigma}_{ij}| - |\widehat{\Sigma}_{ik}| < 2\delta, \qquad \text{for all } j \in J_1^a \text{ and } k \in I_a \setminus \{i\}. \tag{A.5}$$

For any $i \in J_1^a$, we have

$$\widehat{M}_i - |\widehat{\Sigma}_{ij}| \leq 2\delta, \qquad \text{for any } j \in I_a. \tag{A.6}$$

PROOF OF LEMMA 3. For the entire proof, we work on the event $\mathcal{E}$ defined in (4.1). To prove (A.3), we observe that, for any $i, j, k \in I_a$, $\Sigma_{ij} = \Sigma_{ik} = C_{aa}$ by Lemma 1, whence

$$\left| |\widehat{\Sigma}_{ij}| - |\widehat{\Sigma}_{ik}| \right| \quad \leq \quad \left| |\Sigma_{ij}| - |\Sigma_{ik}| \right| + 2\delta = 2\delta.$$

To prove (A.4), we first observe that, for any $j \in I_a$, $|\Sigma_{ij}| = C_{aa}$ by Lemma 1, whence

$$|\widehat{\Sigma}_{ij}| \overset{\mathcal{E}}{\geq} C_{aa} - \delta. \tag{A.7}$$

Next, we notice that, for any $\ell \in [p]$,

$$|\Sigma_{i\ell}| \;=\; \left|\sum_{b=1}^{K} A_{\ell b}C_{ab}\right| = \left|A_{\ell a}C_{aa} + \sum_{b\neq a} A_{\ell b}C_{ab}\right|$$

$$\overset{(iii)}{\leq} \; |A_{\ell a}|C_{aa} + (1-|A_{\ell a}|)(C_{aa}-\nu) \;=\; C_{aa} - (1-|A_{\ell a}|)\nu. \tag{A.8}$$

For any $j \in I_a$ and $k \in [p]\backslash(I_a \cup J_1^a)$, the definition of J_1 implies $|A_{ka}| \leq 4\delta/\nu$, hence

$$|\widehat{\Sigma}_{ik}| \overset{\mathcal{E}}{\leq} |\Sigma_{ik}| + \delta \overset{(A.8)}{\leq} C_{aa} - (1-|A_{ka}|)\nu + \delta \leq C_{aa} - \nu + 5\delta,$$

so that

$$|\widehat{\Sigma}_{ij}| - |\widehat{\Sigma}_{ik}| \overset{\mathcal{E}}{\geq} |\Sigma_{ij}| - \delta - |\widehat{\Sigma}_{ik}| \geq |\Sigma_{ij}| - C_{aa} + \nu - 6\delta > 2\delta,$$

by using $\nu > 8\delta \cdot (\|C\|_\infty/\nu) \geq 8\delta$. To prove (A.5), observe that, for any $j \in J_1^a$ and $k \in I_a \setminus \{i\}$,

$$|\widehat{\Sigma}_{ij}| \overset{(A.8)}{\leq} C_{aa} - (1-|A_{ja}|)\nu + \delta < C_{aa} + \delta = |\Sigma_{ik}| + \delta \overset{\mathcal{E}}{\leq} |\widehat{\Sigma}_{ik}| + 2\delta.$$

So far, we have proved (A.3) - (A.5) and it remains to show (A.6). For any $i \in J_1^a$, we have, for some $c \in [K]$,

$$\begin{aligned}\widehat{M}_i &\overset{\mathcal{E}}{\leq} \max_{k\in[p]\backslash i} |\Sigma_{ik}| + \delta \overset{(A.1)}{=} \left|\sum_{b=1}^{K} A_{ib}C_{bc}\right| + \delta \\ &\overset{(*)}{\leq} \left|\sum_{b=1}^{K} A_{ib}C_{ba}\right| + \delta = |\Sigma_{ij}| + \delta \overset{\mathcal{E}}{\leq} |\widehat{\Sigma}_{ij}| + 2\delta.\end{aligned}$$

It remains to show that inequality $(*)$ holds, for any $c \neq a$. On the one hand, we have

$$\left|\sum_{b=1}^{K} A_{ib}C_{bc}\right| \leq |A_{ia}||C_{ac}| + (1-|A_{ia}|)C_{cc} \overset{(iii)}{\leq} |A_{ia}|(C_{aa}-\nu) + (1-|A_{ia}|)C_{cc},$$

while on the other hand, we find

$$\left|\sum_{b=1}^{K} A_{ib}C_{ab}\right| \overset{(iii)}{\geq} |A_{ia}||C_{aa}| - (1-|A_{ia}|)(C_{aa}-\nu).$$

Combining the preceding two display yields

$$\left|\sum_{b=1}^{K} A_{ib}C_{ab}\right| - \left|\sum_{b=1}^{K} A_{ib}C_{bc}\right| \geq \nu - (1-|A_{ia}|)(C_{aa}+C_{cc}).$$

The term on the right is positive, since condition (4.7) guarantees that

$$\nu > \frac{4\delta}{\nu}(C_{aa}+C_{cc}) \geq (1-|A_{ia}|)(C_{aa}+C_{cc}),$$

where the last inequality is due to the definition of J_1. This concludes the proof. □

Lemma 3 remains valid under the conditions of Remark 3 in which case $J_1 = \varnothing$ and we only need $\nu > 4\delta$ to prove (A.4).

Proof of Theorem 3. We work on the event $\mathcal{E}$ throughout the proof. Without loss of generality, we assume that the label permutation π is the identity. We start by pointing out that the following three claims are sufficient to prove (a) - (c). Let $\widehat{I}^{(i)}$ be defined in step 4 of Algorithm 1.

(1) For any $i \in J \setminus J_1$, we have $Pure(i) = False$.
(2) For any $i \in I_a$ and $a \in [K]$, we have $Pure(i) = True$, $I_a \subseteq \widehat{I}^{(i)}$ and $\widehat{I}^{(i)} \setminus I_a \subseteq J_1^a$.
(3) For any $i \in J_1^a$ and $a \in [K]$, we have $I_a \subseteq \widehat{I}^{(i)}$.

If we can prove these claims, then (1) implies that none of variables in $J \setminus J_1$ will be selected in any set of $\widehat{\mathcal{I}}$ via $i \in J \setminus J_1$. (2) implies that for any $a \in [K]$, there exists $\widehat{I}_a$ such that $I_a \subseteq \widehat{I}_a$ and $\widehat{I}_a \setminus I_a \subseteq J_1^a$. Moreover, this together with MERGE in Algorithm 1 prevents $\widehat{I}_a$ from selecting any variable from $[p]\setminus(I_a \cup J_1^a)$. Finally, (3) guarantees that none of pure variables will be excluded by any $i \in J_1$ in the MERGE step. Thus, $\widehat{K} = K$ and $\widehat{\mathcal{I}} = \{\widehat{I}_1, \ldots, \widehat{I}_K\}$ is the desired partition. Therefore, in the following we proceed to prove (1) - (3).

To prove (1), let $i \in J \setminus J_1$ be fixed. We first prove that $Pure(i) = False$ when $\widehat{I}^{(i)} \cap I \neq \varnothing$. It suffices to show that, there exists $j \in \widehat{I}^{(i)}$ such that the following does not hold

$$\widehat{M}_j - |\widehat{\Sigma}_{ij}| \leq 2\delta. \tag{A.9}$$

Let $\widehat{I}^{(i)} \cap I \neq \varnothing$, so there exists $j \in I_b \cap \widehat{I}^{(i)}$ for some $b \in [K]$. For such j, we have $|\Sigma_{ij}| = |\sum_{a=1}^{K} A_{ia}C_{ab}|$ and

(A.10)

$$|\widehat{\Sigma}_{ij}| \overset{\mathcal{E}}{\leq} \left|\sum_{a=1}^{K} A_{ia}C_{ab}\right| + \delta \overset{(iii)}{\leq} |A_{ib}|C_{bb} + (1-|A_{ib}|)(C_{bb}-\nu) + \delta < C_{bb} - 3\delta,$$

using the definition of J_1 to justify the last inequality. On the other hand, since $j \in I_b$, part (b) of Lemma 2 implies

$$\widehat{M}_j = \max_{k\in[p]\setminus\{i\}} |\widehat{\Sigma}_{jk}| \overset{\mathcal{E}}{\geq} \max_{k\in[p]\setminus\{i\}} |\Sigma_{jk}| - \delta = C_{bb} - \delta. \tag{A.11}$$

Combining (A.10) with (A.11) gives $\widehat{M}_j - |\widehat{\Sigma}_{ij}| > 2\delta$. This shows that for any $i \in J \setminus J_1$, if $\widehat{I}^{(i)} \cap I \neq \varnothing$, then $Pure(i) = False$. Therefore, to complete the proof of (1), we show $\widehat{I}^{(i)} \cap I = \varnothing$ is impossible when $i \in J \setminus J_1$ under our assumptions. If $\widehat{I}^{(i)} \cap I = \varnothing$, then there exists some $j \in J \cap \widehat{I}^{(i)}$ and

$$|\Sigma_{ij}| = \left|\sum_{b=1}^{K}\sum_{a=1}^{K} A_{ia}A_{jb}C_{ab}\right| \leq \max_{1\leq b\leq K}\left|\sum_{a=1}^{K} A_{ia}C_{ab}\right| = \left|\sum_{a=1}^{K} A_{ia}C_{ab^*}\right| = |\Sigma_{ik}|$$

for some $b^* \in [K]$ and any $k \in I_{b^*}$ (the set I_{b^*} is non-empty by condition (ii)). Therefore,

$$|\widehat{\Sigma}_{ij}| - |\widehat{\Sigma}_{ik}| \overset{\mathcal{E}}{\leq} |\Sigma_{ij}| - |\Sigma_{ik}| + 2\delta \leq 2\delta$$

However, since $\widehat{I}^{(i)} \cap I = \varnothing$ and $k \in I_{b^*}$, we know $k \notin \widehat{I}^{(i)}$, which implies

$$|\widehat{\Sigma}_{ij}| - |\widehat{\Sigma}_{ik}| > 2\delta,$$

from Step 4 of Algorithm 1. The last two displays contradict each other, and we conclude that, for any $i \in J \setminus J_1$, $\widehat{I}^{(i)} \cap I \neq \varnothing$.

To prove (2), let $i \in I_a$ be arbitrarily fixed with some $a \in [K]$. We first show that $Pure(i) = True$. From steps 7 - 8 of Algorithm 1, it suffices to show that, for any $j \in \widehat{I}^{(i)}$, (A.9) holds. From (A.4) in Lemma 3, given Step 4 of Algorithm 1, we know that, for any $j \in \widehat{I}^{(i)}$, $j \in I_a \cup J_1^a$. Thus, we write $\widehat{I}^{(i)} = (\widehat{I}^{(i)} \cap I_a) \cup (\widehat{I}^{(i)} \cap J_1^a)$. For any $j \in \widehat{I}^{(i)} \cap I_a$, by the same reasoning, $\widehat{M}_j$ is achieved by some element in either I_a or J_1^a. For both cases, since $i, j \in I_a$ and $i \neq j$, (A.3) and (A.5) in Lemma 3 guarantee that (A.9) holds. On the other hand, for any $j \in \widehat{I}^{(i)} \cap J_1^a$, (A.6) in Lemma 3 implies that (A.9) still holds. Thus, we have shown that, for any $i \in I_a$, $Pure(i) = True$. To show $I_a \subseteq \widehat{I}^{(i)}$, let any $j \in I_a \setminus \{i\}$ and observe that $\widehat{M}_i$ can only be achieved by indices in $I_a \cup J_1^a$. In both cases, (A.3) and (A.5) imply $j \in \widehat{I}^{(i)}$. Thus, $I_a \subseteq \widehat{I}^{(i)}$. Finally, $\widehat{I}^{(i)} \setminus I_a \subseteq J_1^a$ follows immediately from (A.4).

We conclude the proof by noting that (3) immediately follows from (A.6). □

A.3. Proofs of the results from Section 4.2. We divide the proof of Theorem 4 into three steps:

Step 1. We show that there exists a signed permutation $\widehat{P}$ such that the columns of $\widehat{A}_I$ aligns with those of A_I in terms of label and sign, as detailed in Lemma 4;

Step 2. We write $\bar{A} = A\widehat{P}$, and prove first the error bounds for $\widehat{A}_{\widehat{I}} - \bar{A}_{\widehat{I}}$;

Step 3. We prove the error bounds for $\widehat{A} - \bar{A} = \widehat{A} - A\widehat{P}$, with the same $\widehat{P}$, which further implies that $\widehat{P}$ aligns the columns of $\widehat{A}$ and A.

LEMMA 4. *Under conditions of Theorem 4, there exists a signed permutation matrix Q such that $\bar{A} = AQ$ satisfies that* $\operatorname{sign}(\bar{A}_{ia}) = \operatorname{sign}(\widehat{A}_{ia})$ *for any $i \in \widehat{I}_a$ with each $a \in [K]$.*

PROOF OF LEMMA 4. Theorem 3 guarantees $\widehat{K} = K$, $I \subseteq \widehat{I} \subseteq I \cup J_1$ and $I_{\pi(a)} \subseteq \widehat{I}_a \subseteq I_{\pi(a)} \cup J_1^{\pi(a)}$, with high probability, for any $a \in [K]$ and some label permutation π. Let us write $Q = Q_1 Q_2$, with the unsigned permutation matrix Q_1 which relabels the columns of A_I according to those of $\widehat{A}_I$, and with $Q_2 = \operatorname{diag}(q_1, \ldots, q_K)$ with $q_a \in \{+1, -1\}$ for each $a \in [K]$.

Denoting $\breve{A} = AQ_1$, we proceed to show that, for each $a \in [K]$, $\operatorname{sign}(\widehat{A}_{ia}) = \operatorname{sign}(\breve{A}_{ia}) \cdot q_a$ holds for any $i \in \widehat{I}_a$, in which case each q_a can be uniquely constructed. Since $\widehat{I}_a \subseteq I_{\pi(a)} \cup J_1^{\pi(a)}$, it suffices to prove that, for any $a \in [K]$,

(A.12)
$$\frac{\operatorname{sign}(\widehat{A}_{ia})}{\operatorname{sign}(\breve{A}_{ia})} = \frac{\operatorname{sign}(\widehat{A}_{ja})}{\operatorname{sign}(\breve{A}_{ja})}, \qquad \text{for any } i, j \in I_{\pi(a)} \text{ or } i, j \in J_1^{\pi(a)} \text{ with } i \neq j.$$

From the definition of A_I and the way we construct $\widehat{A}_I$, for any $i, j \in I_{\pi(a)}$ or $i, j \in J_1^{\pi(a)}$, we consider the following two cases:

If $\operatorname{sign}(A_{i\pi(a)}) = \operatorname{sign}(A_{j\pi(a)})$, this implies $\operatorname{sign}(\breve{A}_{ia}) = \operatorname{sign}(\breve{A}_{ja})$. To show $\widehat{A}_{ia} = \widehat{A}_{ja}$, from (3.1), we need to show $i, j \in \widehat{I}_a^1$ or $i, j \in \widehat{I}_a^2$ which is equivalent to show $\widehat{\Sigma}_{ij} > 0$. For any $i, j \in I_{\pi(a)}$ or $i, j \in J_1^{\pi(a)}$ with $i \neq j$, display (4.5) gives $|A_{k\pi(a)}| \geq 1 - 4\delta/\nu$ and $\sum_{b \neq \pi(a)} |A_{kb}| \leq 4\delta/\nu$, for $k = i, j$.

Thus, using $\text{sign}(A_{i\pi(a)}) = \text{sign}(A_{j\pi(a)})$, we have

$$\begin{aligned}
\Sigma_{ij} &= A_{i\pi(a)}A_{j\pi(a)}C_{\pi(a)\pi(a)} + A_{i\pi(a)}\sum_{c\neq a}A_{jc}C_{\pi(a)c} + A_{j\pi(a)}\sum_{b\neq a}A_{ib}C_{\pi(a)b} \\
&\quad + \sum_{b,c\neq\pi(a)}A_{ib}A_{jc}C_{bc} \\
&\overset{(iii)}{\geq} A_{i\pi(a)}A_{j\pi(a)}C_{\pi(a)\pi(a)} - |A_{i\pi(a)}|(1-|A_{j\pi(a)}|)(C_{\pi(a)\pi(a)}-\nu) \\
&\quad - |A_{j\pi(a)}|(1-|A_{i\pi(a)}|)(C_{\pi(a)\pi(a)}-\nu) - \sum_{b,c\neq\pi(a)}A_{ib}A_{jc}C_{bc} \\
&\geq \left(1-\frac{4\delta}{\nu}\right)^2 C_{\pi(a)\pi(a)} - \frac{8\delta}{\nu}\cdot\left(1-\frac{4\delta}{\nu}\right)C_{\pi(a)\pi(a)} - \frac{16\delta^2}{\nu^2}C_{b^*b^*} + 8\delta \\
&\geq \left(1+\frac{48\delta^2}{\nu^2}-\frac{16\delta}{\nu}\right)C_{\pi(a)\pi(a)} - \frac{16\delta^2}{\nu^2}C_{b^*b^*} + 8\delta,
\end{aligned}$$

for some $b^* \neq \pi(a)$. Since (4.7) implies $8\delta C_{b^*b^*} < \nu^2$ and $\nu > 8\delta$, on the event $\mathcal{E}$, we have $\widehat{\Sigma}_{ij} \geq \Sigma_{ij} - \delta > 3\delta > 0$.

If $\text{sign}(A_{i\pi(a)}) \neq \text{sign}(A_{j\pi(a)})$, this gives $\text{sign}(\breve{A}_{ia}) \neq \text{sign}(\breve{A}_{ja})$. Similarly, to show $\widehat{A}_{ia} \neq \widehat{A}_{ja}$, we prove $\widehat{\Sigma}_{ij} < 0$. Using the same arguments yields

$$\widehat{\Sigma}_{kl} \overset{\mathcal{E}}{\leq} \Sigma_{kl} + \delta < -3\delta < 0.$$

Therefore, given $\widehat{\mathcal{I}} = \{\widehat{I}_a\}_{a\in[K]}$, we can construct the signed permutation $\widehat{P} = Q$ which alligns the columns of A_I with those of $\widehat{A}_I$. □

For ease of notation and without loss of generality, we make the blanket assumption that the signed permutation $\widehat{P}$ is the identity so that $\bar{A} = A$ for the remainder of the proof. We note that the signed permutation $\widehat{P}$ will be the same when estimating each row $A_{j\cdot}$ for $j \in J$.

Proof of step 2: From the construction of $\widehat{A}_{\widehat{I}}$ and parts (a) - (c) in Theorem 3, we can write, for each $a \in [K]$, $\widehat{I}_a = I_a \cup L_a$ with $L_a := \widehat{I}_a \cap J_1^a$. For any $i \in \widehat{I}_a$, the definitions of I and J_1^a imply $|A_{ia}| \geq 1 - 4\delta/\nu$. Since Lemma 4 guarantees that $\text{sign}(A_{ia}) = \text{sign}(\widehat{A}_{ia})$, we have

$$\|\widehat{A}_{\widehat{I}} - A_{\widehat{I}}\|_\infty = \max_{i\in\widehat{I}}\|\widehat{A}_{i\cdot} - A_{i\cdot}\|_\infty \leq \frac{4}{\nu}\delta.$$

Let $s_i = \|A_{i\cdot}\|_0$ for $i \in [p]$. Then, for any $i \in \widehat{I}$, we have

$$\left\|\widehat{A}_{j\cdot} - \bar{A}_{j\cdot}\right\|_q \leq \frac{4}{\nu}s_i^{1/q}\delta, \qquad 1 \leq q \leq \infty.$$

□

For **Step 3** of the proof of Theorem 4, we will make use of the results of Lemmas 5 and 6, stated here first and proved at the end of this section, in order to preserve the flow of the presentation.

LEMMA 5. *Under the conditions of Theorem 4, on the event $\mathcal{E}$, we have*

$$\|\widehat{C} - C\|_\infty \le 2\delta', \qquad \max_{j\in\widehat{J}} \|\widehat{\theta}^j - \theta^j\|_\infty \le \delta', \tag{A.13}$$

where δ' is given in (4.8).

LEMMA 6. *Under the conditions of Theorem 4, on the event $\mathcal{E}$, we have $\beta_a^j = 0$ implies $\widehat{\beta}_a^j = 0$, for any $j \in \widehat{J}$ and $a \in [\widehat{K}]$.*

Proof of Step 3. For each $j \in \widehat{J}$, recall that $\beta^j = C^{-1}\theta^j = \Omega\theta^j$ since C is invertible. Also recall that $\bar{\beta}^j = \widehat{\Omega}\widehat{\theta}^j$. We first show $\|\bar{\beta}^j - \beta^j\|_\infty \le 5\|\Omega\|_{\infty,1}\delta'$. For notational convenience, we remove all the super indices. From Lemma 5, the following event

$$\mathcal{E}' = \left\{ \|\widehat{C} - C\|_\infty \le 2\delta',\ \max_{j\in\widehat{J}} \|\widehat{\theta}^j - \theta^j\|_\infty \le \delta' \right\},$$

is implied by the event $\mathcal{E} = \mathcal{E}_\delta$. On the event $\mathcal{E}'$, the true $\Omega := C^{-1}$ satisfies the constraint since

$$\|\Omega\widehat{C} - I\|_\infty \ = \ \|\Omega(\widehat{C} - C)\|_\infty \ \le \ \|\widehat{C} - C\|_\infty \|\Omega\|_{\infty,1} \ \le \ 2\delta'\|\Omega\|_{\infty,1}.$$

Then the pair $(\|\Omega\|_{\infty,1}, \Omega)$ of (t, Ω) is feasible. Consequently, the optimality and feasibility of $(\widehat{t}, \widehat{\Omega})$ imply

$$\|\widehat{\Omega}\|_{\infty,1} \le \widehat{t} \le \|\Omega\|_{\infty,1}, \quad \|\widehat{\Omega}\widehat{C} - I\|_\infty \le 2\delta'\widehat{t} \le 2\delta'\|\Omega\|_{\infty,1}. \tag{A.14}$$

Then, on the event $\mathcal{E}'$, we obtain

$$\begin{aligned}
\|\bar{\beta} - \beta\|_\infty &= \|\widehat{\Omega}\widehat{\theta} - \widehat{\Omega}\theta + \widehat{\Omega}\theta - \beta\|_\infty \\
&\le \|\widehat{\Omega}\|_{\infty,1}\|\widehat{\theta} - \theta\|_\infty + \|\widehat{\Omega}\theta - \beta\|_\infty \\
&\le \delta'\|\widehat{\Omega}\|_{\infty,1} + \|\widehat{\Omega}C\beta - \beta\|_\infty \\
&\le \delta'\|\Omega\|_{\infty,1} + \|\widehat{\Omega}C - I\|_\infty\|\beta\|_1 \\
&\le \delta'\|\Omega\|_{\infty,1} + \|\widehat{\Omega}\widehat{C} - I\|_\infty + \|\widehat{\Omega}\widehat{C} - \widehat{\Omega}C\|_\infty \quad (\text{since } \|\beta\|_1 \le 1) \\
&\le 3\delta'\|\Omega\|_{\infty,1} + \|\widehat{\Omega}\|_{\infty,1}\|\widehat{C} - C\|_\infty \\
&\le 5\delta'\|\Omega\|_{\infty,1}.
\end{aligned}$$

The feasibility of $\widehat{\beta}^j$ implies that $\|\widehat{\beta}^j - \bar{\beta}^j\|_\infty \leq \mu$. By the triangle inequality, we obtain

$$\|\widehat{\beta}^j - \beta^j\|_\infty \;\leq\; \|\widehat{\beta}^j - \bar{\beta}^j\|_\infty + \|\bar{\beta}^j - \beta^j\|_\infty \;\leq\; 2\mu,$$

since $\mu = 5\delta'\|\Omega\|_{\infty,1}$. Then following from Lemma 6 and using $\widehat{K} = K$ on the event $\mathcal{E}$ gives

$$\|\widehat{A}_{j\cdot} - A_{j\cdot}\|_q = \left(\sum_{a=1}^{K} |\widehat{\beta}_a^j - \beta_a^j|^q\right)^{1/q} = \left(\sum_{a\in s_j} |\widehat{\beta}_a^j - \beta_a^j|^q\right)^{1/q} \leq\; 2s_j^{1/q}\mu,$$

for any $1 \leq q \leq \infty$. This completes the proof of the last step and of Theorem 4. □

To conclude this section we give below the proofs of the intermediary results used in the proof.

Proof of Lemma 5. On the event $\mathcal{E}$, we showed that $\widehat{K} = K$. Then, from the definition of $\widehat{C}_{aa}$, we have

$$\begin{aligned}\max_{1\leq a\leq K} |\widehat{C}_{aa} - C_{aa}| \;&\leq\; \max_{1\leq a\leq K} \frac{1}{|\widehat{I}_a|(|\widehat{I}_a|-1)} \sum_{i,j\in\widehat{I}_a, i\neq j} \big||\widehat{\Sigma}_{ij}| - C_{aa}\big| \\ &\overset{\mathcal{E}}{\leq}\; \delta + \frac{1}{|\widehat{I}_a|(|\widehat{I}_a|-1)} \sum_{i,j\in\widehat{I}_a, i\neq j} \big||\Sigma_{ij}| - C_{aa}\big|.\end{aligned}$$

Theorem 3 states that, on the event $\mathcal{E}$, $\widehat{I}_a = I_a \cup L_a$ where $L_a = \widehat{I}_a \cap J_1^a$, for any $a \in [K]$. Therefore, we consider the following three cases:
(1) For any $i, j \in I_a$ and $i \neq j$, Lemma 1 implies $\big||\Sigma_{ij}| - C_{aa}\big| = 0$.
(2) For any $i \in I_a$ and $j \in L_a$, the definition of J_1^a gives

$$\big||\Sigma_{ij}| - C_{aa}\big| \leq \big(1 - |A_{ja}|\big)(2C_{aa} - \nu) \leq \frac{8\delta}{\nu}\|C\|_\infty - 4\delta.$$

(3) For any $i, j \in L_a$ and $i \neq j$, since $i, j \in J_1^a$, we know $|A_{ka}| \geq 1 - 4\delta/\nu$

and $\sum_{b\neq a}|A_{kb}| \leq 4\delta/\nu$, for $k = i, j$. Thus,

$$\begin{aligned}
\big||\Sigma_{ij}| - C_{aa}\big| &\leq \big(1 - |A_{ia}||A_{ja}|\big)C_{aa} + |A_{ia}|\sum_{c\neq a}|A_{jc}||C_{ac}| + |A_{ja}|\sum_{b\neq a}|A_{ib}||C_{ab}| \\
&\quad + \sum_{b,c\neq a}|A_{ib}A_{jc}|C_{bc} \\
&\leq \big(1 - |A_{ia}||A_{ja}|\big)C_{aa} + |A_{ia}|(1-|A_{ja}|)(C_{aa}-\nu) \\
&\quad + |A_{ja}|(1-|A_{ib}|)(C_{aa}-\nu) + (1-|A_{ib}|)(1-|A_{jc}|)C_{b^*b^*} \\
&\leq \left[1 - \left(1 - \frac{4\delta}{\nu}\right)^2\right] C_{aa} + \frac{8\delta}{\nu}(C_{aa}-\nu) + \frac{16\delta^2}{\nu^2}C_{b^*b^*} \\
&\leq \frac{16\delta}{\nu}\|C\|_\infty - 8\delta, \qquad \text{(by (4.7))}.
\end{aligned}$$

for some $b^* \neq a$, where we use the definition of J_1 in the third inequality. Therefore, by combining cases (1) - (3), we have

$$\begin{aligned}
\max_{1\leq a\leq K}|\widehat{C}_{aa} - C_{aa}| &\leq \delta + \frac{|I_a||L_a| + |L_a|(|L_a|-1)}{|\widehat{I}_a|(|\widehat{I}_a|-1)} \cdot \left(\frac{16\delta\|C\|_\infty}{\nu} - 8\delta\right) \\
&\leq \left(\frac{16}{\nu}\|C\|_\infty - 7\right)\delta.
\end{aligned}$$

where the last inequality comes from that $|L_a| + |I_a| = |\widehat{I}_a|$. For the off-diagonal entries, since $\mathrm{sign}(\widehat{A}_{ia}) = \mathrm{sign}(A_{ia})$, for any $i \in \widehat{I}$ and $a \in [K]$, we have

$$\max_{1\leq a,b\leq K, a\neq b}|\widehat{C}_{ab} - C_{ab}| \leq \delta + \frac{1}{|\widehat{I}_a||\widehat{I}_b|}\sum_{i\in\widehat{I}_a, j\in\widehat{I}_b}\big||\Sigma_{ij}| - |C_{ab}|\big|,$$

we consider the following three cases:
(1) For any $i \in I_a, j \in I_b$, we have $|\Sigma_{ij}| - |C_{ab}| = 0$.
(2) For any $i \in I_a$, $j \in J_1^b$, we have

$$\big||\Sigma_{ij}| - |C_{ab}|\big| \leq (1 - |A_{jb}|)|C_{ab}| + \sum_{c\neq b}|A_{jc}||C_{ac}| \leq \frac{8\delta}{\nu}\|C\|_\infty - 4\delta.$$

(3) For any $i \in J_1^a$, $j \in J_1^b$, we obtain

$$\Sigma_{ij} = A_{ia}A_{jb}C_{ab} + A_{ia}\sum_{d\neq b}A_{jd}C_{ad} + \sum_{c\neq a}A_{ic}\sum_{d\in s(j)}A_{jd}C_{cd}.$$

Thus,

$$\begin{aligned}\big||\Sigma_{ij}|-|C_{ab}|\big| &\le (1-|A_{ia}||A_{jb}|)|C_{ab}| + |A_{ia}|(1-|A_{jb}|)\|C\|_\infty + (1-|A_{ia}|)\|C\|_\infty \\ &\le \left(\frac{8\delta}{\nu}-\frac{16\delta^2}{\nu^2}\right)(C_{aa}-\nu) + \left(\frac{8\delta}{\nu}-\frac{16\delta^2}{\nu^2}\right)\|C\|_\infty \\ &\le \frac{16\delta}{\nu}\|C\|_\infty - 8\delta. \qquad (\text{by } \nu < \|C\|_\infty)\end{aligned}$$

Therefore, combining the three cases gives

$$\begin{aligned}\max_{1\le a,b\le K, a\neq b}|\widehat{C}_{ab}-C_{ab}| &\le \delta + \frac{|I_a||L_b|+|L_a||I_b|+2|L_a||L_b|}{2|\widehat{I}_a||\widehat{I}_b|}\cdot\left(\frac{16\delta\|C\|_\infty}{\nu}-8\delta\right) \\ &\le \left(\frac{16}{\nu}\|C\|_\infty - 7\right)\delta.\end{aligned}$$

Combining the diagonal and off-diagonal cases yields

$$\|\widehat{C}-C\|_\infty \;\le\; \left(\frac{16}{\nu}\|C\|_\infty - 7\right)\delta \;\le\; 2\delta'$$

We now proceed to bound $\max_{j\in\widehat{J}}\|\widehat{\theta}^j-\theta^j\|_\infty$. From $\mathrm{sign}(A_{ia})=\mathrm{sign}(\widehat{A}_{ia})$ for any $i\in\widehat{I}$, we obtain

$$\max_{j\in\widehat{J}}\|\widehat{\theta}^j-\theta^j\|_\infty \;\;\le\;\; \delta + \max_{a\in[K],j\in\widehat{J}}\frac{1}{|\widehat{I}_a|}\sum_{i\in\widehat{I}_a}\Big|\Sigma_{ij}-\sum_{b\in s(j)}A_{jb}C_{ab}\Big|.$$

Since for any $i\in I_a$ and any $j\in J$, $\Sigma_{ij}=\sum_{b\in s(j)}A_{jb}C_{ab}$, we focus on the case when $i\in L_a$. For any $i\in L_a$ and $j\in J$, (A.1) yields

$$\sum_{b\in s(j)}A_{jb}C_{ab}-\Sigma_{ij} \;\;=\;\; (1-A_{ia})\sum_{b\in s(j)}A_{jb}C_{ab} - \sum_{c\neq a}A_{ic}\sum_{b\in s(j)}A_{jb}C_{bc},$$

which, by the definition of J_1, implies

$$\begin{aligned}&\left|\sum_{b\in s(j)}A_{jb}C_{ab}-\Sigma_{ij}\right| \\ &\quad\le (1-|A_{ia}|)|A_{jd}||C_{ad}| + (1-|A_{ia}|)|A_{jd'}||C_{ad'}| \quad (\text{for some } d,d'\in[K]) \\ &\quad\le \frac{4\delta}{\nu}(2\|C\|_\infty-\nu) \;\le\; \frac{8\delta}{\nu}\|C\|_\infty - 4\delta.\end{aligned}$$

Since we have $\widehat{J}\subseteq J$, we have

$$\max_{j\in\widehat{J}}\|\widehat{\theta}^j-\theta^j\|_\infty \le \delta + \max_a\frac{|L_a|}{|\widehat{I}_a|}\cdot\left(\frac{8}{\nu}\|C\|_\infty-4\right)\delta \le \left(\frac{8}{\nu}\|C\|_\infty-3\right)\delta = \delta',$$

which concludes the proof of Lemma 5. □

Proof of Lemma 6. Let $j \in \widehat{J}$ be arbitrarily fixed and $\widehat{\beta}^j$ be the optimal solution of (3.11) with $\mu = 5\|\Omega\|_{\infty,1}\delta'$. For simplicity, we remove the super indices. Starting with the following Karush-Kuhn-Tucker condition:

$$\operatorname{sign}(\widehat{\beta}_a) + \lambda_a \operatorname{sign}(\widehat{\beta}_a - \bar{\beta}_a) = 0, \tag{A.15}$$

subject to

$$\lambda_a(|\widehat{\beta}_a - \bar{\beta}_a| - \mu) = 0, \quad \lambda_a \geq 0, \quad \text{for } a = \{1, \ldots, K\}, \tag{A.16}$$

we obtain

$$0 = \operatorname{sign}(\widehat{\beta}_a)\left(\widehat{\beta}_a - \bar{\beta}_a\right) + \lambda_a\left|\widehat{\beta}_a - \bar{\beta}_a\right| \overset{(A.16)}{=} \operatorname{sign}(\widehat{\beta}_a)\left(\widehat{\beta}_a - \bar{\beta}_a\right) + \lambda_a\mu, \tag{A.17}$$

by multiplying both sides of (A.15) by $\widehat{\beta}_a - \bar{\beta}_a$. In what follows we prove that if $\beta_a = 0$, for some a, then $\widehat{\beta}_a = 0$. Since this is true when $\lambda_a = 0$ from (A.15), we only consider when $\lambda_a \neq 0$. Note this implies $|\widehat{\beta}_a - \bar{\beta}_a| = \mu$ from (A.16). If we assume $\widehat{\beta}_a > 0$, then (A.17) gives

$$\bar{\beta}_a - \widehat{\beta}_a = \lambda_a\mu.$$

Since $|\widehat{\beta}_a - \bar{\beta}_a| = \mu$, we further obtain $\lambda_a = 1$ and

$$\bar{\beta}_a = \mu + \widehat{\beta}_a > \mu. \tag{A.18}$$

Recall that $\|\beta - \bar{\beta}\|_\infty \leq \mu$. This implies $\bar{\beta}_a \leq \mu + |\beta_a| = \mu$, which contradicts (A.18), so $\widehat{\beta}_a$ cannot be strictly positive. Similarly, $\widehat{\beta}_a < 0$ cannot hold based on similar arguments. Thus, $\widehat{\beta}_a = 0$ from which we conclude $\operatorname{supp}(\widehat{\beta}^j) \subseteq \operatorname{supp}(\beta^j)$ for any $j \in \widehat{J}$. □

Proof of Theorem 5. Estimation of the submatrix A_I is as in Step 2 of the proof of Theorem 4. We denote by $\widehat{\beta}^j_D$, $j \in \widehat{J}$, the minimizer of (3.13) under the constraint (3.14). First, we observe that the true β^j satisfies the constraint (3.14) on the event $\mathcal{E}$. Indeed,

$$\begin{aligned}
\|\widehat{C}\beta^j - \widehat{\theta}^j\|_\infty &\leq \|\widehat{C}\beta^j - C\beta^j\|_\infty + \|C\beta^j - \widehat{\theta}^j\|_\infty \\
&\leq \|\widehat{C} - C\|_\infty\|\beta^j\|_1 + \|\theta^j - \widehat{\theta}^j\|_\infty \\
&\leq \|\widehat{C} - C\|_\infty + \|\theta^j - \widehat{\theta}^j\|_\infty \\
&\leq 3\delta' = \lambda',
\end{aligned}$$

by Lemma 5. Second, this implies, on the event $\mathcal{E}$, that $\|\widehat{\beta}_D^j\|_1 \leq \|\beta^j\|_1$ and $\widehat{\beta}_D^j - \beta^j$ is in the cone $\mathcal{C}_S$ with $S = \mathrm{supp}(\beta^j)$ by a standard argument. Finally, by the definition of the ℓ_q-sensitivity of C and the feasibility of $\widehat{\beta}_D^j$, we get for $\Delta = \widehat{\beta}_D^j - \beta^j$

$$\begin{aligned}
&\|\Delta\|_q \kappa_q(C,s) \\
&\quad \leq \|C\Delta\|_\infty \\
&\quad \leq \|C\widehat{\beta}_D^j - \widehat{\theta}^j\|_\infty + \|\widehat{\theta}^j - \theta^j\|_\infty \quad (\text{since } \theta^j = C\beta^j) \\
&\quad \leq \|\widehat{C}\widehat{\beta}_D^j - \widehat{\theta}^j\|_\infty + \|\widehat{C} - C\|_\infty \|\widehat{\beta}_D^j\|_1 + \|\widehat{\theta}^j - \theta^j\|_\infty \\
&\quad \leq \|\widehat{C}\widehat{\beta}_D^j - \widehat{\theta}^j\|_\infty + \|\widehat{C} - C\|_\infty + \|\widehat{\theta}^j - \theta^j\|_\infty \quad (\text{since } \|\widehat{\beta}_D^j\|_1 \leq 1) \\
&\quad \leq 2\lambda'
\end{aligned}$$

and the conclusion (4.11) follows. It remains to prove the second inequality (4.12). First, we observe that $\|v\|_q \leq \|v\|_\infty (2s)^{1/q}$ for all $v \in \mathcal{C}_S$ and $s = |S|$ by the following computation:

$$\begin{aligned}
\|v\|_q^q &\leq \|v\|_1 \|v\|_\infty^{q-1} \\
&\leq 2\|v_S\|_1 \|v\|_\infty^{q-1} \quad (\text{ since } v \in \mathcal{C}_S) \\
&\leq 2s\|v\|_\infty^q.
\end{aligned}$$

This implies that $\kappa_q(C,s) \geq (2s)^{-1/q}\kappa_\infty(C,s)$, and clearly $[\kappa_\infty(C,s)]^{-1} \leq \|C^{-1}\|_{\infty,1}$ for all $s \leq K$, with equality for $s = K$. Now (4.12) follows from (4.11). □

Proof of Theorem 6. Without loss of generality, we assume that $\lambda_1(C) < \infty$, since otherwise the lower bound is trivially zero.
First we construct a set of "hypotheses" of A. Let

$$\mathcal{M} := \{v \in \{0,1\}^K : d_H(0,v) = s\}$$

where $d_H(\cdot)$ denotes the Hamming distance between two binary vectors. Following Lemma A.3 in Rigollet and Tsybakov (2011) when $s \leq 4K/5$, there exists $\mathcal{M}' \subset \mathcal{M}$ such that, for any $w^{(i)} \neq w^{(j)} \in M'$,

$$d_H\left(w^{(i)}, w^{(j)}\right) > s/16, \tag{A.19}$$

and

$$\log|\mathcal{M}'| \geq c_0 s \log(K/s), \tag{A.20}$$

for some constant $c_0 > 0$. We let $w^{(0)} = (0, \ldots, 0) \in \mathbb{R}^K$. Then, we choose

$$A^{(j)} = \begin{bmatrix} B \\ \eta \left(w^{(j)}\right)^T \end{bmatrix} \in \mathbb{R}^{p \times K}, \quad \text{for each } j = 0, 1, \ldots, |\mathcal{M}'|, \tag{A.21}$$

where

$$B = \begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_K \end{bmatrix} \in \mathbb{R}^{(p-1)\times K}, \quad B_k = \begin{bmatrix} e_k^T \\ e_k^T \\ \vdots \\ e_k^T \end{bmatrix} \in \mathbb{R}^{|I_k| \times K}, \quad \text{for } k \in [K], \tag{A.22}$$

and

$$\eta = \sqrt{\frac{c_0 \sigma^2}{8\lambda_1(C)}} \sqrt{\frac{\log(K/s)}{n}}. \tag{A.23}$$

We use e_k to denote the canonical basis of K dimensional space and $\mathbf{0}$ to denote the zero vector. Note that, for each B_k, the only non-zero values are at the kth column. By specifying as above, we choose $\sum_{k=1}^K |I_k| = p - 1$ and consider the $A^{(j)}$ with only one non-pure row. It is easy to verify that $A^{(j)} \in \mathcal{A}_s$ for each $j = 0, 1, \ldots, |\mathcal{M}'|$ under (4.13).

We denote by $\mathcal{KL}(\mathbb{P}, \mathbb{Q})$ the Kullback-Leibler divergence between two probability distributions $\mathbb{P}$ and $\mathbb{Q}$. Since we particularize into one choice of C, we write $\mathbb{P}_A := \mathbb{P}_{A,C}$ for simplicity. In order to apply Theorem 2.5 in Tsybakov (2009) to prove (4.14), for fixed $\alpha \in (0, 1/8)$, we need to check the following three conditions:

(a) $\mathcal{KL}(\mathbb{P}_{A^{(i)}}, \mathbb{P}_{A^{(0)}}) \le \alpha \log |\mathcal{M}'|$, for each $i = 1, \ldots, |\mathcal{M}'|$.

(b) For any $0 \le i < j \le |\mathcal{M}'|$, with some constant $c' > 0$,

$$L_q\left(A^{(i)}, A^{(j)}\right) \ge c' s^{1/q} \sqrt{\frac{\log(K/s)}{n}}.$$

(c) $L_q(\,\cdot\,)$ satisfies the triangle inequality.

To show (a), since $X \sim N(0, ACA^T + \sigma^2 \boldsymbol{I}_p)$, invoking Lemma 7 gives

$$\mathcal{KL}\left(\mathbb{P}_{A^{(i)}}, \mathbb{P}_{A^{(0)}}\right) \le \lambda_1(C) \frac{n\eta^2 s}{2\sigma^2} \le \frac{1}{16} \log |M'|, \quad \forall\, i = 1, \ldots, |M'|, \tag{A.24}$$

by using (A.20) and (A.23).

To prove (b), for any $i = 1, \ldots, |\mathcal{M}'|$, observe that

$$L_q\left(A^{(i)}, A^{(0)}\right) = \eta \|w^{(i)}\|_q = s^{1/q}\eta$$

and, for any $i \neq j$ different from 0,

$$L_q\left(A^{(i)}, A^{(j)}\right) = \eta \|w^{(i)} - w^{(j)}\|_q \geq (s/16)^{1/q}\eta \geq (s^{1/q}\eta)/16,$$

by using (A.19). Combining these two and using the expression of η yield

$$L_q\left(A^{(i)}, A^{(j)}\right) \geq c' s^{1/q} \sqrt{\frac{\sigma^2}{\lambda_1(C)}} \sqrt{\frac{\log(K/s)}{n}}, \tag{A.25}$$

for $0 \leq i < j \leq |\mathcal{M}'|$.

Finally, we verify (c) by showing that $L_q(\cdot)$ satisfies the triangle inequality. Consider $(A, \widetilde{A}, \widehat{A})$ and observe that

$$\begin{aligned}
L_q(A, \widetilde{A}) &= \min_{P \in \mathcal{H}_K} \|AP - \widetilde{A}\|_{\infty,q} \\
&= \min_{P,Q \in \mathcal{H}_K} \|AP - \widetilde{A}Q\|_{\infty,q} \\
&\leq \min_{P,Q \in \mathcal{H}_K} \left(\|AP - \widehat{A}\|_{\infty,q} + \|\widehat{A} - \widetilde{A}Q\|_{\infty,q}\right) \\
&= \min_{P \in \mathcal{H}_K} \|AP - \widehat{A}\|_{\infty,q} + \min_{Q \in \mathcal{H}_K} \|\widehat{A} - \widetilde{A}Q\|_{\infty,q} \\
&= L_q(A, \widehat{A}) + L_q(\widetilde{A}, \widehat{A}).
\end{aligned}$$

Therefore, we conclude the proof of (4.14) by invoking the Theorem 2.5 in Tsybakov (2009). □

LEMMA 7. *Assume model (1.1) and $X \sim N_p(\mathbf{0}, ACA^T + \sigma^2 \boldsymbol{I}_p)$. Let $A^{(0)}$ and $A^{(i)}$ be constructed as (A.21) and (A.22), for any $1 \leq i \leq M'$ with M' satisfying (A.20). Let $\mathbb{P}_{A^{(0)}}$ and $\mathbb{P}_{A^{(i)}}$ be the probability densities of X parametrized by $A^{(0)}$ and $A^{(i)}$, respectively. Then we have*

$$\mathcal{KL}\left(\mathbb{P}_{A^{(i)}}, \mathbb{P}_{A^{(0)}}\right) \leq \lambda_1(C) \frac{n\eta^2 s}{2\sigma^2}. \tag{A.26}$$

PROOF OF LEMMA 7. From the property of Kullback-Leibler divergence, we only need to verify the case when $n = 1$. We consider arbitrary $A^{(i)}$ constructed as (A.21) and (A.22) for some $0 \leq i \leq M'$. For notational

simplicity, we write $A = A^{(i)} = (B^T, \xi)^T$ where $\xi = \eta w^{(i)} \in \mathbb{R}^K$. For this $A \in \mathbb{R}^{p\times K}$, from (A.21) and (A.22), we observe that

$$\Sigma = ACA^T + \Gamma = \begin{bmatrix} BCB^T + \sigma^2 \boldsymbol{I}_{p-1} & BC\xi \\ \xi^T CB^T & \xi^T C\xi + \sigma^2 \end{bmatrix} := \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

Similarly, for any $\widetilde{A} \neq A$ constructed in the same way, we have

$$\widetilde{\Sigma} = \widetilde{A}C\widetilde{A}^T + \sigma^2 \boldsymbol{I}_{p-1} = \begin{bmatrix} BCB^T + \Gamma_B & BC\widetilde{\xi} \\ \widetilde{\xi}^T CB^T & \widetilde{\xi}^T C\widetilde{\xi} + \sigma^2 \end{bmatrix} := \begin{bmatrix} \Sigma_{11} & \widetilde{\Sigma}_{12} \\ \widetilde{\Sigma}_{21} & \widetilde{\Sigma}_{22} \end{bmatrix}$$

Recall that the Kullback-Leibler divergence between two p-dimensional multivariate normal distributions $\mathcal{N}_0 := N_p(\mathbf{0}, \Sigma)$ and $\mathcal{N}_1 := N_p(\mathbf{0}, \widetilde{\Sigma})$ is given by

$$\mathcal{KL}(\mathbb{P}_{\widetilde{A}}, \mathbb{P}_A) = \frac{1}{2}\left[\mathrm{tr}\left(\Sigma^{-1}\widetilde{\Sigma}\right) - p + \log\left(\frac{\det\Sigma}{\det\widetilde{\Sigma}}\right)\right]. \tag{A.27}$$

By using the formula of the inverse of a block matrix

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{bmatrix},$$

for square matrices A and D and non-singular matrices A and $D - CA^{-1}B$, we have

$$\Sigma^{-1} = \begin{bmatrix} \Sigma_{11}^{-1} + \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22\cdot1}^{-1}\Sigma_{21}\Sigma_{11}^{-1} & -\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22\cdot1}^{-1} \\ -\Sigma_{22\cdot1}^{-1}\Sigma_{21}\Sigma_{11}^{-1} & \Sigma_{22\cdot1}^{-1} \end{bmatrix} := \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix},$$

with $\Sigma_{22\cdot1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$. This gives

$$\mathrm{tr}\left(\Sigma^{-1}\widetilde{\Sigma}\right) = \underbrace{\mathrm{tr}\left(\Omega_{11}\Sigma_{11} + \Omega_{12}(BC\widetilde{\xi})^T\right)}_{T_1} + \underbrace{\Omega_{21}BC\widetilde{\xi} + \Omega_{22}(\widetilde{\xi}^T C\widetilde{\xi} + \sigma^2)}_{T_2}.$$

We first calculate T_1 by observing that

$$\begin{aligned} T_1 &= \mathrm{tr}\left(\boldsymbol{I}_{p-1} + \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22\cdot1}^{-1}\Sigma_{21} - \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22\cdot1}^{-1}\widetilde{\xi}^T CB^T\right) \\ &= p - 1 + \mathrm{tr}\left(\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22\cdot1}^{-1}\Delta^T CB^T\right) \qquad \left(\Sigma_{21} = \xi^T CB^T\right) \\ &= p - 1 + \Sigma_{22\cdot1}^{-1}\Delta^T CB^T\Sigma_{11}^{-1}BC\xi \end{aligned} \tag{A.28}$$

where $\Delta := \xi - \widetilde{\xi} \in \mathbb{R}^K$. On the other hand, we have

$$T_2 = \Sigma_{22\cdot1}^{-1}\left(\widetilde{\xi}^T C\widetilde{\xi} + \sigma^2 - \xi^T CB^T\Sigma_{11}^{-1}BC\widetilde{\xi}\right). \tag{A.29}$$

Since our specification of $A = A^{(0)}$ and $\widetilde{A} = A^{(i)}$ in (A.21) and (A.22) gives $\xi = \mathbf{0}$ and $\widetilde{\xi} = \eta w^{(i)}$, it implies $\|\widetilde{\xi}\|^2 = \eta^2 s$ and

$$\begin{aligned} \Sigma_{22\cdot 1} &= \xi^T C \xi + \sigma^2 - \xi^T C B^T \Sigma_{11}^{-1} B C \xi = \sigma^2, \\ \widetilde{\Sigma}_{22\cdot 1} &= \sigma^2 + \widetilde{\xi}^T \left(C - C B^T \Sigma_{11}^{-1} B C \right) \widetilde{\xi}. \end{aligned} \tag{A.30}$$

Hence combining (A.28) with (A.29) yields

$$\operatorname{tr}\left(\Sigma^{-1}\widetilde{\Sigma}\right) = p + \frac{\widetilde{\xi}^T C \widetilde{\xi}}{\sigma^2} \leq p + \frac{\eta^2 s}{\sigma^2}\lambda_1(C). \tag{A.31}$$

To calculate the determinant of Σ and $\widetilde{\Sigma}$, recall that the inverse formula of a block matrix is

$$\det \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \det(A)\det(D - CA^{-1}B)$$

for any invertible matrix A. We thus obtain

$$\det \Sigma = \det \Sigma_{11} \cdot \Sigma_{22\cdot 1}, \quad \det \widetilde{\Sigma} = \det \Sigma_{11} \cdot \widetilde{\Sigma}_{22\cdot 1},$$

from which, the display (A.30) further gives

$$\log\left(\frac{\det \Sigma}{\det \widetilde{\Sigma}}\right) = \log \Sigma_{22\cdot 1} - \log \widetilde{\Sigma}_{22\cdot 1} = \log \sigma^2 - \log\left(\sigma^2 + \xi^T M \xi\right)$$

with $M := C - CB^T\Sigma_{11}^{-1}BC$. It is easy to see that M is positive definite. Indeed, since

$$\left\| C^{1/2} B^T \Sigma_{11}^{-1} B C^{1/2} \right\|_{op} = \left\| (BCB^T + \sigma^2 \boldsymbol{I}_{p-1})^{-1} BCB^T \right\|_{op} < 1,$$

$\lambda_{\min}(M) > 0$ follows from

$$M = C^{1/2}\left(\boldsymbol{I}_p - C^{1/2} B^T \Sigma_{11}^{-1} B C^{1/2}\right) C^{1/2}$$

and an application of Weyl's inequality. This implies

$$\log\left(\frac{\det \Sigma}{\det \widetilde{\Sigma}}\right) < 0. \tag{A.32}$$

Finally, plugging (A.31) and (A.32) into (A.27) concludes the proof of Lemma 7. □

A.4. Proofs for the results from Section 4.3. We first prove the three statements of Theorem 7, then present the proofs of Remark 5. Without loss of generality, we assume that the signed permutation P is identity.

Proof of Theorem 7. We first give the proof for part (a). Then, for ease of the presentation, we prove part (c) first and then part (b).

Proof of part (a). Recall that Lemma 6 immediately implies $\mathrm{supp}(\widehat{A}_{\widehat{J}}) \subseteq \mathrm{supp}(A_{\widehat{J}})$. In addition, Theorem 3 yields $\widehat{I}_a \subseteq I_a \cup J_1^a$, for any $a \in \widehat{K}$. From the way we construct $\widehat{A}_{\widehat{I}}$, we have $\mathrm{supp}(\widehat{A}_{\widehat{I}}) \subseteq \mathrm{supp}(A_{\widehat{I}})$. Therefore, we have proved $\mathrm{supp}(\widehat{A}) \subseteq \mathrm{supp}(A)$.

On the other hand, for any $(j,a) \in \mathrm{supp}(A_{J_2})$, we know $|\beta_a^j| > 2\mu$. This and the fact that $\|\widehat{\beta}^j - \beta^j\|_\infty \le 2\mu$, immediately gives

$$|\widehat{\beta}_a^j| \ge |\beta_a^j| - \|\widehat{\beta}^j - \beta^j\|_\infty > 0,$$

which implies $\mathrm{supp}(A_{J_2}) \subseteq \mathrm{supp}(\widehat{A}_{J_2})$.

To show $\mathrm{sign}(\widehat{A}_{\widehat{S}}) = \mathrm{sign}(A_{\widehat{S}})$, since Lemma 4 guarantees $\mathrm{sign}(\widehat{A}_{ia}) = \mathrm{sign}(A_{ia})$ for any $(i,a) \in \widehat{S}$ and $i \in \widehat{I}$, we focus on any fixed $(j,a) \in \widehat{S}$ and $j \in \widehat{J}$. First, we consider the case $\widehat{A}_{ja} = \widehat{\beta}_a^j > 0$. Removing super indices, if $\widehat{\beta}_a > 0$, (A.18) gives $\bar{\beta}_a > \mu$. Thus, $\beta_a \ge \bar{\beta}_a - \|\beta - \bar{\beta}\|_\infty > 0$ by recalling $\|\beta - \bar{\beta}\|_\infty \le \mu$. So far, we have shown that, for any $\widehat{A}_{ja} > 0$, $(j,a) \in \widehat{S}$ and $j \in \widehat{J}$, we have $A_{ja} > 0$. Since the same argument holds for any $\widehat{A}_{ja} < 0$, the proof of $\mathrm{sign}(\widehat{A}_{\widehat{S}}) = \mathrm{sign}(A_{\widehat{S}})$ is completed.

Proof of part (c). Recall that, for any $i \in [p]$ and $a \in [K]$,

$$i \in G_a \iff A_{ia} \neq 0, \qquad i \in \widehat{G}_a \iff \widehat{A}_{ia} \neq 0.$$

We start our proof by rewriting the equivalent expression of TFPP and TFNP:

$$\mathrm{TFPP} = \frac{\sum_{i\in[p],a\in[K]} 1\{A_{ia} = 0, \widehat{A}_{ia} \neq 0\}}{\sum_{i\in[p],a\in[K]} 1\{A_{ia} = 0\}},$$

$$\mathrm{TFNP} = \frac{\sum_{i\in[p],a\in[K]} 1\{A_{ia} \neq 0, \widehat{A}_{ia} = 0\}}{\sum_{i\in[p],a\in[K]} 1\{A_{ia} \neq 0\}}.$$

We first show TFPP $= 0$. From the result of part (a), we know $\mathrm{supp}(\widehat{A}) \subseteq \mathrm{supp}(A)$. Thus,

$$\sum_{i\in[p],a\in[K]} 1\{A_{ia} = 0, \widehat{A}_{ia} \neq 0\} = 0,$$

which implies TFPP $= 0$.

In order to prove the result of TFNP, observe

$$\sum_{i\in[p],a\in[K]} 1\{A_{ia}\neq 0\} = |I| + \sum_{i\in J} s_i. \tag{A.33}$$

with $s_i = \|A_{i\cdot}\|_0$ for each $j \in J$. For given $\widehat{I}$, we partition $[p] = I \cup J_1 \cup J_2 \cup J_3 = I \cup (L_1 \cup L_2) \cup J_2 \cup J_3$ with $L_1 = \widehat{I} \cap J_1$ and $L_2 = J_1 \setminus L_1$. Let us consider the set $I \cup L_1$ first. Theorem 3 implies $I \cup L_1 = \widehat{I}$ and $\widehat{I}_a \setminus I_a \subseteq J_1^a$. From the way we construct $\widehat{A}_{\widehat{I}}$, we have

$$\sum_{i\in I\cup L_1,a\in[K]} 1\{A_{ia}\neq 0, \widehat{A}_{ia}=0\} = \sum_{i\in L_1,a\in[K]} 1\{A_{ia}\neq 0, \widehat{A}_{ia}=0\}.$$

Since the definition of J_1 implies that, for any $j \in J_1^a$ and $a \in [K]$, $|A_{ja}| \geq 1 - 4\delta/\nu$ and $|A_{jb}| \leq 4\delta/\nu$, for any $b \neq a$, this implies

$$\sum_{a\in[K]} 1\{A_{ia}\neq 0, \widehat{A}_{ia}=0\} = \sum_{b\neq a} 1\{A_{ib}\neq 0\} = t_i,$$

for any $i \in J_1^a \cap L_1$ and $a \in [K]$. Thus, we have

$$\sum_{i\in I\cup L_1,a\in[K]} 1\{A_{ia}\neq 0, \widehat{A}_{ia}=0\} = \sum_{i\in L_1} t_i. \tag{A.34}$$

Next we consider the set L_2. On the event $\mathcal{E}$, for any $i \in J_1^a \cap L_2$, we have

$$|\widehat{A}_{ia}| \geq |A_{ia}| - \|\widehat{A} - A\|_\infty \geq 1 - \frac{4\delta}{\nu} - 2\mu > 0.$$

Thus, $\widehat{A}_{ia} \neq 0$, which implies

$$\sum_{i\in L_2,a\in[K]} 1\{A_{ia}\neq 0, \widehat{A}_{ia}=0\} \leq \sum_{i\in L_2} t_i. \tag{A.35}$$

Then we consider the set J_2. Part (a) gives $\mathrm{supp}(A_{J_2}) = \mathrm{supp}(\widehat{A}_{J_2})$ which yields

$$\sum_{i\in J_2,a\in[K]} 1\{A_{ia}\neq 0, \widehat{A}_{ia}=0\} = 0. \tag{A.36}$$

Finally, we consider the set J_3. By examining the proof of Part (a), it is easy to verify that $\widehat{A}_{ja} \neq 0$ if $|A_{ja}| \geq (2\mu) \vee (4\delta/\nu)$, for any $j \in J_3$ and $a \in [K]$. Thus,

$$\sum_{i\in J_3,a\in[K]} 1\{A_{ia}\neq 0, \widehat{A}_{ia}=0\} \leq \sum_{i\in J_3} t_i. \tag{A.37}$$

At last, combining (A.33) - (A.37) gives

$$\mathrm{TFNP} = \frac{\sum_{i\in[p],a\in[K]} 1\{A_{ia}\neq 0, \widehat{A}_{ia}=0\}}{\sum_{i\in[p],a\in[K]} 1\{A_{ia}\neq 0\}} \leq \frac{\sum_{j\in J_1\cup J_3} t_j}{|I|+\sum_{j\in J} s_j}.$$

Proof of part (b). Similarly, we can express $\mathrm{GFPP}(\widehat{G}_a)$ and $\mathrm{GFNP}(\widehat{G}_a)$ by the following:

$$\begin{aligned}\mathrm{GFPP}(\widehat{G}_a) &= \frac{\sum_{i\in[p]} 1\{A_{ia}=0, \widehat{A}_{ia}\neq 0\}}{\sum_{i\in[p]} 1\{A_{ia}=0\}},\\ \mathrm{GFNP}(\widehat{G}_a) &= \frac{\sum_{i\in[p]} 1\{A_{ia}\neq 0, \widehat{A}_{ia}=0\}}{\sum_{i\in[p]} 1\{A_{ia}\neq 0\}}.\end{aligned}$$

For any given $a\in[\widehat{K}]$, $\mathrm{GFPP}(\widehat{G}_a)=0$ follows immediately by noting that

$$0 = \mathrm{TFPP} \geq \frac{|(G_a)^c\cap \widehat{G}_a|}{\sum_{b=1}^K |(G_b)^c|} = \frac{|(G_a)^c|}{\sum_{b=1}^K |(G_b)^c|}\mathrm{GFPP}(\widehat{G}_a),$$

with the convention $\mathrm{GFPP}(\widehat{G}_a)=0$ if $(G_a)^c=0$. To show the expression of $\mathrm{GFNP}(\widehat{G}_a)$, by the definition of I and Theorem 3, we obtain

$$\sum_{i\in I} 1\{A_{ia}\neq 0, \widehat{A}_{ia}=0\}=0, \qquad \sum_{i\in I} 1\{A_{ia}\neq 0\} = |I_a|.$$

The latter immediately implies

$$\sum_{i\in[p]} 1\{A_{ia}\neq 0\} = |I_a| + \sum_{i\in J} s_i^a$$

In addition, following the same arguments in the proof of part (b), we have

$$\sum_{i\in J_2} 1\{A_{ia}\neq 0, \widehat{A}_{ia}=0\}=0, \quad \sum_{i\in J_1\cup J_3} 1\{A_{ia}\neq 0, \widehat{A}_{ia}=0\} = \sum_{i\in J_1\cup J_3\setminus J_1^a} t_i^a.$$

Combining these two concludes the proof. □

Proofs of Remark 5. We briefly verify the first claim. It suffices to verify $\|A_{j\cdot}\|_1\leq 1$ which is equivalent with $\|\widehat{\beta}^j\|_1\leq 1$, for any $j\in\widehat{J}$. Recall (3.11), since β^j is feasible, the optimality of $\widehat{\beta}^j$ immediately gives $\|\widehat{\beta}^j\|_1\leq 1$. □

To verify the expression of TFNP in the second claim, we assume $t_j = t$ and $s_j = s$, for $j \in J$, and $|J_1| + |J_3| = \alpha(|I| + |J_3|)$. Note that $|I| + |J_1| + |J_2| + |J_3| = p$ implies $|J_1| + |J_3| = \alpha p/(1+\alpha)$. We therefore obtain

$$\begin{aligned} \text{TFNP} &\leq \frac{t(|J_1| + |J_3|)}{s(|J_1| + |J_3|) + s|J_2| + |I|} = \frac{t}{s + \frac{s|J_2|+|I|}{\alpha} \cdot \frac{1+\alpha}{p}} \\ &= t \Big/ \left(s + \frac{1}{\alpha} \cdot \frac{|I| + s|J_2|}{|I| + |J_2|} \right) \qquad \big(\text{using } (1+\alpha)(|I| + |J_2|) = p\big), \end{aligned}$$

as desired. □

We verify the third claim. On the event $\mathcal{E}$, when $J_2 = J$, Remark 3 yields $\widehat{I} = I$, $\widehat{\mathcal{I}} = \mathcal{I}$ and $\widehat{J} = J$. After careful examination of the proof of Lemma 5, we derive that $\|\widehat{C} - C\|_\infty \leq \delta$ and $\max_{j \in J} \|\widehat{\theta}^j - \theta\|_\infty \leq \delta$, on the event $\mathcal{E}$. Therefore, choosing $\lambda = \delta$ and $\mu = 3\|\Omega\|_{\infty,1}\delta$ proves the claim, following the proof of Theorems 4 and 7 step by step. □

Finally, we verify the fourth claim on the hard-threshold estimator $\widetilde{\beta}^j$ for any $j \in J$. For simplicity, we remove the super indices. Recall that, $\widetilde{\beta}$ is defined coordinate-wisely by $\bar{\beta}_a 1\{|\bar{\beta}_a| > \mu\}$ with $\mu = 5\|\Omega\|_{\infty,1}\delta'$.

First, we show $\|\widetilde{\beta} - \beta\|_\infty \leq 2\mu$. For any $a \in [K]$ such that $|\bar{\beta}_a| \leq \mu$, we have

$$|\widetilde{\beta}_a - \beta_a| = |\beta_a| \leq \|\bar{\beta} - \beta\|_\infty + |\bar{\beta}_a| \leq 2\mu,$$

while the same bound is obtained above for the case $|\bar{\beta}_a| > \mu$. This proves $\|\widetilde{A} - A\|_\infty \leq 2\mu$ where $\widetilde{A}$ combines $\widehat{A}_{\widehat{I}}$ and $\widetilde{\beta}^j$ for each $j \in \widehat{J}$. To prove the same rate in Theorem 4 for $\widetilde{A}$, it suffices to show that Lemma 6 still holds for $\widetilde{\beta}^j$. Recall that, on the event $\mathcal{E}$, we have $\|\bar{\beta} - \beta\|_\infty \leq \mu$. For any $\beta_a = 0$, we thus have $|\bar{\beta}_a| \leq \|\bar{\beta} - \beta\|_\infty \leq \mu$, which implies $\widetilde{\beta}_a = 0$. This concludes the proof of Theorem 4 for $\widetilde{A}$.

To show part (a) of Theorem 7, let $\widehat{S}$ denote the support of $\widetilde{A}$ and we write $(i,a) \in \widehat{S}$ if $|\widetilde{A}_{ia}| \neq 0$. Let $(i,a) \in \widehat{S}$ be arbitrary fixed and consider the following two cases:

- If $i \in \widehat{I}$, from Theorem 3 and the way we construct $\widehat{A}_{\widehat{I}}$, we have $|\widetilde{A}_{ia}| = 1$. Thus, $|A_{ia}| \geq |\widetilde{A}_{ia}| - \|\widetilde{A} - A\|_\infty \geq 1 - 2\mu > 0$.
- If $i \in \widehat{J}$, then $|\widetilde{A}_{ia}| = |\widetilde{\beta}^i_a| = |\bar{\beta}^i_a| > \mu$. Therefore, $|A_{ia}| = |\beta^i_a| \geq |\bar{\beta}^i_a| - \|\bar{\beta}^i - \beta^i\|_\infty > 0$.

Thus, we have proved that $\text{supp}(\widetilde{A}) \subseteq \text{supp}(A)$. To show $\text{supp}(A_{J_2}) \subseteq \text{supp}(\widetilde{A})$, for any $(i,a) \in \text{supp}(A_{J_2})$, by the definition of J_2, $|A_{ia}| > 2\mu$. Thus, $|\widetilde{A}_{ia}| \geq |A_{ia}| - \|\widetilde{A} - A\|_\infty \geq 0$. Therefore, $(i,a) \in \text{supp}(\widetilde{A})$.

To show $\operatorname{sign}(\widetilde{A}_{\widehat{S}}) = \operatorname{sign}(A_{\widehat{S}})$, since Lemma 4 guarantees $\operatorname{sign}(\widehat{A}_{ia}) = \operatorname{sign}(A_{ia})$ for any $(i,a) \in \widehat{S}$ and $i \in \widehat{I}$, we focus on each $(i,a) \in \widehat{S}$ and $i \in \widehat{J}$. Assuming $\widetilde{A}_{ia} = \widetilde{\beta}^i_a > 0$, we know $\widetilde{\beta}^i_a = \bar{\beta}^i_a > \mu$. Since $A_{ia} = \beta^i_a \geq \bar{\beta}^i_a - \|\bar{\beta}^i - \beta^i\|_\infty > 0$, we have proved that $A_{ia} > 0$ for any $\widetilde{A}_{ia} > 0$ with $(i,a) \in \widehat{S}$ and $i \in \widehat{J}$. Since the same argument holds for any $\widetilde{A}_{ia} < 0$, we conclude the proof of $\operatorname{sign}(\widetilde{A}_{\widehat{S}}) = \operatorname{sign}(A_{\widehat{S}})$.

The same conclusion in part (b) and (c) of Theorem 7 holds for GFPP, GFNP, TFPP and TFNP based on the hard-threshold estimator $\widetilde{A}$, as it shares the same property in part (a). □

APPENDIX B: CROSS-VALIDATION ILLUSTRATION

We consider a simple case, when C is diagonal and the signed permutation matrix P is I, to illustrate our cross-validation method.

Example 1. Let $C = \operatorname{diag}(\tau,\tau,\tau)$, $\mathcal{I} = \big\{\{1,2\},\{3,4\},\{5,6\}\big\}$ and

$$A = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & -1 \\ 0.4 & 0.6 & 0 \\ -0.5 & 0 & 0.4 \end{bmatrix}, \quad A_I C A_I^T = \begin{bmatrix} * & \tau & 0 & 0 & 0 & 0 \\ \tau & * & 0 & 0 & 0 & 0 \\ 0 & 0 & * & \tau & 0 & 0 \\ 0 & 0 & \tau & * & 0 & 0 \\ 0 & 0 & 0 & 0 & * & \tau \\ 0 & 0 & 0 & 0 & \tau & * \end{bmatrix},$$

where we use $*$ to reflect the fact that our algorithm ignores the diagonal elements. For the true I and $\mathcal{I}$, we have $\widehat{A}_I = A_I$,

$$\begin{aligned} \left\|\widehat{\Sigma}^{(1)}_{II} - A_I \widehat{C} A_I^T\right\|_{\text{F-off}} &\leq \left\|\widehat{\Sigma}^{(1)}_{II} - \Sigma_{II}\right\|_{\text{F-off}} + \left\|A_I \widehat{C} A_I^T - \Sigma_{II}\right\|_{\text{F-off}} \\ &\leq \left\|\widehat{\Sigma}^{(1)}_{II} - \Sigma_{II}\right\|_{\text{F-off}} + \sqrt{|I|(|I|-1)} \cdot \|\widehat{C} - C\|_\infty. \end{aligned}$$

For

$$\epsilon = \left(\max_{i \neq j} \left|\widehat{\Sigma}^{(1)}_{ij} - \Sigma_{ij}\right|\right) \vee \left(\max_{i \neq j} \left|\widehat{\Sigma}^{(2)}_{ij} - \Sigma_{ij}\right|\right),$$

we obtain

$$CV(\mathcal{I}) = \frac{1}{\sqrt{|I|\big(|I|-1\big)}} \left\|\widehat{\Sigma}^{(1)}_{II} - A_I \widehat{C} A_I^T\right\|_{\text{F-off}} \leq 2\epsilon.$$

Suppose that $\widehat{\mathcal{I}} = \big\{\{1,2\},\{3,5\},\{4,6\}\big\}$, so $\widehat{I} = I$, yet $\widehat{\mathcal{I}} \neq \mathcal{I}$, we would have

$$\widehat{A}_{\widehat{I}}\widehat{C}\widehat{A}_{\widehat{I}}^T = \begin{bmatrix} * & \widehat{\tau}_1 & 0 & 0 & 0 & 0 \\ \widehat{\tau}_1 & * & 0 & 0 & 0 & 0 \\ 0 & 0 & * & 0 & \widehat{\tau}_2 & 0 \\ 0 & 0 & 0 & * & 0 & \widehat{\tau}_3 \\ 0 & 0 & \widehat{\tau}_2 & 0 & * & 0 \\ 0 & 0 & 0 & \widehat{\tau}_3 & 0 & * \end{bmatrix},$$

$$\widehat{A}_{\widehat{I}}\widehat{C}\widehat{A}_{\widehat{I}}^T - \Sigma_{\widehat{I}\widehat{I}} = \begin{bmatrix} * & \Delta\tau_1 & 0 & 0 & 0 & 0 \\ \Delta\tau_1 & * & 0 & 0 & 0 & 0 \\ 0 & 0 & * & \boldsymbol{-\tau} & \boldsymbol{\widehat{\tau}_2} & 0 \\ 0 & 0 & \boldsymbol{-\tau} & * & 0 & \boldsymbol{\widehat{\tau}_3} \\ 0 & 0 & \boldsymbol{\widehat{\tau}_2} & 0 & * & \boldsymbol{-\tau} \\ 0 & 0 & 0 & \boldsymbol{\widehat{\tau}_3} & \boldsymbol{-\tau} & * \end{bmatrix}.$$

Here $\Delta\tau_a = \widehat{\tau}_a - \tau_a$, using estimates $\widehat{\tau}_a$ defined in lieu of $\widehat{C}_{aa}$ from (3.6) for each $a \in [\widehat{K}]$. Thus, the cross-validation criterion in (5.1) would satisfy

$$CV(\widehat{\mathcal{I}}) \geq \frac{\big\|\widehat{A}_{\widehat{I}}\widehat{C}\widehat{A}_{\widehat{I}}^T - \Sigma_{\widehat{I}\widehat{I}}\big\|_{\text{F-off}} - \big\|\widehat{\Sigma}^{(1)}_{\widehat{I}\widehat{I}} - \Sigma_{\widehat{I}\widehat{I}}\big\|_{\text{F-off}}}{\sqrt{|\widehat{I}|(|\widehat{I}|-1)}} \geq \sqrt{\frac{4\tau^2 + 2\widehat{\tau}_2^2 + 2\widehat{\tau}_3^2}{|\widehat{I}|(|\widehat{I}|-1)}} - 2\epsilon.$$

From noting that $|\widehat{\tau}_a - \tau| \leq \epsilon$, for $a = 2, 3$, it gives

$$CV(\widehat{\mathcal{I}}) \geq \sqrt{\frac{4\tau^2 - 4\tau\epsilon + 2\epsilon^2}{15}} - 2\epsilon > 2\epsilon \geq CV(\mathcal{I}),$$

for $\tau \geq 9\epsilon$. We conclude in this example, with $\widehat{I} = I$, incorrectly specifying $\mathcal{I}$ will induce a large loss. It is easily verified that this is also the case when $\widehat{I} = I$ but $\widehat{K} \neq K$ and $\widehat{\mathcal{I}} \neq \mathcal{I}$.

On the other hand, suppose we mistakenly included some non-pure variable in $\widehat{I}$. For instance, suppose we found $\widehat{\mathcal{I}} = \big\{\{1,2\},\{3,4\},\{5,6,7\}\big\}$. Then we would have

$$\Sigma_{\widehat{I}'\widehat{I}'} = \begin{bmatrix} * & \tau & 0 & 0 & 0 & 0 & 0.4\tau \\ \tau & * & 0 & 0 & 0 & 0 & -0.4\tau \\ 0 & 0 & * & \tau & 0 & 0 & 0.6\tau \\ 0 & 0 & \tau & * & 0 & 0 & 0.6\tau \\ 0 & 0 & 0 & 0 & * & \tau & 0 \\ 0 & 0 & 0 & 0 & \tau & * & 0 \\ 0.4\tau & -0.4\tau & 0.6\tau & 0.6\tau & 0 & 0 & * \end{bmatrix},$$

and

$$\widehat{A}_{\widehat{I}'}\widehat{C}\widehat{A}_{\widehat{I}'}^T = \begin{bmatrix} * & \widehat{\tau}_1 & 0 & 0 & 0 & 0 & 0 \\ \widehat{\tau}_1 & * & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & * & \widehat{\tau}_2 & 0 & 0 & 0 \\ 0 & 0 & \widehat{\tau}_2 & * & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & * & \widehat{\tau}_3 & \widehat{\tau}_3 \\ 0 & 0 & 0 & 0 & \widehat{\tau}_3 & * & \widehat{\tau}_3 \\ 0 & 0 & 0 & 0 & \widehat{\tau}_3 & \widehat{\tau}_3 & * \end{bmatrix}.$$

We thus have

$$\widehat{A}_{\widehat{I}'}\widehat{C}\widehat{A}_{\widehat{I}'}^T - \Sigma_{\widehat{I}'\widehat{I}'} = \begin{bmatrix} * & \Delta\tau_1 & 0 & 0 & 0 & 0 & \boldsymbol{-0.4\tau} \\ \Delta\tau_1 & * & 0 & 0 & 0 & 0 & \boldsymbol{0.4\tau} \\ 0 & 0 & * & \Delta\tau_2 & 0 & 0 & \boldsymbol{-0.6\tau} \\ 0 & 0 & \Delta\tau_2 & * & 0 & 0 & \boldsymbol{-0.6\tau} \\ 0 & 0 & 0 & 0 & * & \Delta\tau_3 & \boldsymbol{\widehat{\tau}_3} \\ 0 & 0 & 0 & 0 & \Delta\tau_3 & * & \boldsymbol{\widehat{\tau}_3} \\ \boldsymbol{-0.4\tau} & \boldsymbol{0.4\tau} & \boldsymbol{-0.6\tau} & \boldsymbol{-0.6\tau} & \boldsymbol{\widehat{\tau}_3} & \boldsymbol{\widehat{\tau}_3} & * \end{bmatrix}$$

and, by similar arguments, for $\tau \geq 12\epsilon$, we find

$$CV(\widehat{\mathcal{I}}') \geq \sqrt{\frac{4\widehat{\tau}_3^2 + 4 \times 0.36\tau^2 + 4 \times 0.16\tau^2}{42}} - 2\epsilon > 2\epsilon.$$

Thus, the cross-validation loss in this example will be large even if only one non-pure variable is mistakenly classified as pure variable. In rare cases, the cross-validation criterion might miss a very small subset of I but this can be rectified in our later estimation of A_J.

APPENDIX C: ADDITIONAL SIMULATION RESULTS

C.1. Related work on the estimation of A. As we explained in Section 4.4, the existing procedures for estimating A in (1.1) are developed for models satisfying identifiability conditions different than our (i)-(iii). Specifically, Bai and Li (2012) propose to first optimize, via EM, a quasi-likelihood objective under the identifiability conditions (a) $C = \boldsymbol{I}_K$ and (b) $A^T\Gamma^{-1}A$ is diagonal. The major advantage of this setting is that the computationally demanding EM algorithm only needs to determine A and Γ as $C = \boldsymbol{I}_K$ is given. The EM algorithm, however, is only guaranteed to find stationary point $\widehat{B}$ with the property that $\widehat{B}^T\widehat{B}$ is diagonal. In the context of this problem, as the authors note, the EM algorithm requires a delicate initialization and is computationally demanding, even if only one

of K, n and p is moderately large. Next, the authors propose to link this estimator with an estimator of a model no longer satisfying (a) and (b) as identifiability conditions, but satisfying instead (1) C is an arbitrary positive definite matrix; (2) There exists a *known* set S of K pure variables, with only one pure variable per latent factor allowed. No further sparsity conditions on A are imposed. To estimate A under (2), they suggest to solve for A and C the equation $ACA^T = \widehat{B}\widehat{B}^T$. This yields the estimator $\widetilde{A} = \widehat{B}\widehat{B}_S^{-1}$ of A. However, when K is relatively large, $\widehat{B}_S$ may not be invertible, and the estimator may not exist. Finally, although $\widetilde{A}_S = \widehat{B}_S\widehat{B}_S^{-1} = \boldsymbol{I}_K$, the submatrix $\widetilde{A}_{S^c}$ is not sparse in general. One possibility is to threshold $\widetilde{A}$, but it is unclear how to choose the correct threshold level, for the following reason. Although the authors establish the asymptotic limit of the MLE of A under (1) and (2), the estimator of A explained above is not guaranteed to be the MLE in this model: if it exists, it is a transformation of a stationary point that estimates parameters under the model specifications (a) and (b), different from (1) and (2). The immediate practical implication is that the variation of $\widehat{A}$ around A under (1) and (2) is not known, which makes the thresholding level of $\widehat{A}$ difficult to assess. For all these reasons, we cannot compare numerically our estimation procedure with the procedure proposed in Bai and Li (2012), even in the (unrealistic) case when the pure variable set is known.

C.2. LOVE for non-overlapping cluster estimation. In applications, one may not have prior information on whether the clusters may overlap or not. Thus, one would prefer a clustering method that works well in both overlapping and non-overlapping scenarios. In the previous section, we have demonstrated that LOVE outperforms the existing clustering methods if data are generated from a model that yields variable clusters with overlaps. In this section, we study the numerical performance of the proposed method under non-overlapping data generating schemes.

To generate data with non-overlapping clusters, we set the number of variables in each cluster to be 20. We generate the diagonal elements of C from the uniform distribution in $[1, 2]$ and use the same method as in Section 5.2 to generate the off-diagonal elements. The variance σ_j^2 of the error E_j is generated from the uniform distribution in $[3, 4]$. In Table 1, we compare the sensitivity and specificity of the proposed method with the CORD estimator (Bunea, Giraud and Luo, 2016a) under non-overlapping scenarios, where the sensitivity and specificity are defined in (C.1). The CORD estimator can be viewed as a benchmark method for variable clustering without overlaps and is shown to outperform K-means and hierarchical clustering, via an

extensive numerical study presented in Bunea, Giraud and Luo (2016a). For this reason, we only focus on the comparison between LOVE and CORD. From Table 1, we see that for small p (i.e., $p = 100$) the performance of LOVE is only slightly worse than CORD. As p increases, the specificity of LOVE and that of CORD remain close to 1, but LOVE yields in fact higher sensitivity than CORD when $n = 300$. This confirms that the performance of the proposed method is comparable to the benchmark method under non-overlapping scenarios. Of course, LOVE is much more flexible as it can detect possible overlaps.

TABLE 1

Sensitivity (SN) and specificity (SP) of the proposed method (LOVE) and CORD under non-overlapping scenarios. Numbers in parentheses are the simulation standard errors.

p	$n = 300$				$n = 500$			
	LOVE		CORD		LOVE		CORD	
	SN	SP	SN	SP	SN	SP	SN	SP
100	0.87	0.90	0.92	0.98	0.93	0.97	0.98	1.00
	(0.09)	(0.10)	(0.05)	(0.02)	(0.05)	(0.03)	(0.02)	(0.01)
500	0.86	0.98	0.82	0.98	0.87	0.99	0.94	1.00
	(0.05)	(0.01)	(0.03)	(0.01)	(0.04)	(0.00)	(0.02)	(0.01)
1000	0.84	0.97	0.78	0.97	0.87	1.00	0.90	1.00
	(0.05)	(0.02)	(0.03)	(0.01)	(0.04)	(0.00)	(0.02)	(0.01)

C.3. Comparison with other overlapping clustering algorithms. We adopt the same data generating procedure except that we set $K = 10$ and the negative entries of A are replaced by their absolute values, since existing overlapping clustering algorithms typically return an estimator of A with positive entries. We compare the proposed method with the following overlapping clustering algorithms: fuzzy K-means, and fuzzy K-medoids (Krishnapuram et al., 2001), the latter being more robust to noise and outliers. We describe the methods briefly in what follows. Both of them aim to estimate a degree of membership matrix $M \in \mathbb{R}^{p\times K}$ by minimizing the average within-cluster L_2 or L_1 distances (Bezdek, 2013). Specifically, denote $\widetilde{X}_j = (X_{1j}, ..., X_{nj})$, and $\widetilde{X} = \{\widetilde{X}_1, ..., \widetilde{X}_p\}$. Let $W = \{w_1, ..., w_K\}$, where $w_k \in \mathbb{R}^n$, be a subset of $\widetilde{X}$ with K elements. The fuzzy algorithms aim to find the set W such that $J(W)$ defined as

$$J(W) = \sum_{j=1}^{p}\sum_{k=1}^{K} M_{jk} r(\widetilde{X}_j, w_k),$$

is minimized. Here, $M_{jk} > 0$ can be interpreted as the degree of membership matrix which is a known function of $r(\widetilde{X}_j, w_k)$. Some commonly used expressions of M_{jk} are shown by Krishnapuram et al. (2001). In addition, $r(\widetilde{X}_j, w_k)$ is a measure of dissimilarity between $\widetilde{X}_j$ and w_k. For instance, if $r(x, \theta) = \|x - \theta\|_2^2$, this corresponds to the fuzzy K-means. Similarly, the fuzzy K-medoids is given by $r(x, \theta) = \|x - \theta\|_1$. Since searching over all possible subsets of $\widetilde{X}$ is computationally infeasible, an approximate algorithm for minimizing $J(W)$ is proposed by Krishnapuram et al. (2001), we refer to their original paper for further details.

Their degree of membership matrix M plays the same role as our allocation matrix A, but is typically non-sparse. In order to construct overlapping clusters based on M one needs to specify a cut-off value v and assign variable j to cluster k if $M_{jk} > v$. Moreover, the number of clusters K is a required input of the algorithm. In the simulations presented in this section we set $K = 10$ for these two methods, which have been implemented by the functions `FKM` and `FKM.med` in R.

We compare their performance with our proposed method LOVE. We emphasize that our method does not require the specification of K and that the tuning parameters are chosen in a data adaptive fashion, as explained in the previous sections. We follow the pairwise approach of Wiwie, Baumbach and Röttger (2015) for this comparison. Recall that $\mathcal{G} = (G_1, ..., G_K)$ denotes the true overlapping clusters. For notational simplicity, we use $\widehat{\mathcal{G}} = (\widehat{G}_1, ..., \widehat{G}_{\widehat{K}})$ to denote clusters computed from an algorithm. Since LOVE estimates the number of clusters, we allow $\widehat{K}$ to be different from K. For any pair $1 \le j < k \le p$, define

$$
\begin{aligned}
TP_{jk} &= \mathbf{1}\left\{\text{if } j,k \in G_a \text{ and } j,k \in \widehat{G}_b \text{ for some } 1 \le a \le K \text{ and } 1 \le b \le \widehat{K}\right\},\\
TN_{jk} &= \mathbf{1}\left\{\text{if } j,k \notin G_a \text{ and } j,k \notin \widehat{G}_b \text{ for any } 1 \le a \le K \text{ and } 1 \le b \le \widehat{K}\right\},\\
FP_{jk} &= \mathbf{1}\left\{\text{if } j,k \notin G_a \text{ for any } 1 \le a \le K \text{ and } j,k \in \widehat{G}_b \text{ for some } 1 \le b \le \widehat{K}\right\},\\
FN_{jk} &= \mathbf{1}\left\{\text{if } j,k \in G_a \text{ for some } 1 \le a \le K \text{ and } j,k \notin \widehat{G}_b \text{ for any } 1 \le b \le \widehat{K}\right\}.
\end{aligned}
$$

and we define

$$
\begin{aligned}
TP &= \sum_{1\le j<k\le p} TP_{jk},\ TN = \sum_{1\le j<k\le p} TN_{jk},\\
FP &= \sum_{1\le j<k\le p} FP_{jk},\ FN = \sum_{1\le j<k\le p} FN_{jk}.
\end{aligned}
$$

We use sensitivity (SN) and specificity (SP) to evaluate the performance of

different methods, where

$$SP = \frac{TN}{TN + FP}, \text{ and } SN = \frac{TP}{TP + FN}. \tag{C.1}$$

Recall that for the fuzzy methods, variable j belongs to cluster k if the estimated membership matrix M_{jk} is beyond a cut-off v, i.e., $M_{jk} > v$. We search for the optimal cut-off v in a grid $\{0.01, 0.1, ..., 0.3\}$ such that $SP+SN$ is maximized. The corresponding sensitivity and specificity for LOVE, fuzzy K-means (F-Kmeans) and fuzzy K-medoids (F-Kmed) are shown in Figure 1. To save space, we only present the results for $p = 500$ since the other scenarios illustrate the same patterns. The following findings are observed. First, the F-Kmeans is superior to F-Kmed in most scenarios in terms of both sensitivity and specificity. Second, LOVE clearly outperforms these two existing methods and its specificity and sensitivity are very close to 1, which implies that our method leads to very few false positives and false negatives. The conclusions hold with n from 300 to 1000. Moreover, we reiterate that the true value $K = 10$ is used as input in the competing methods, whereas it is estimated from the data in LOVE. This illustrates the net advantage of the proposed method over the existing overlapping clustering methods, for data generated from Model (1.1).

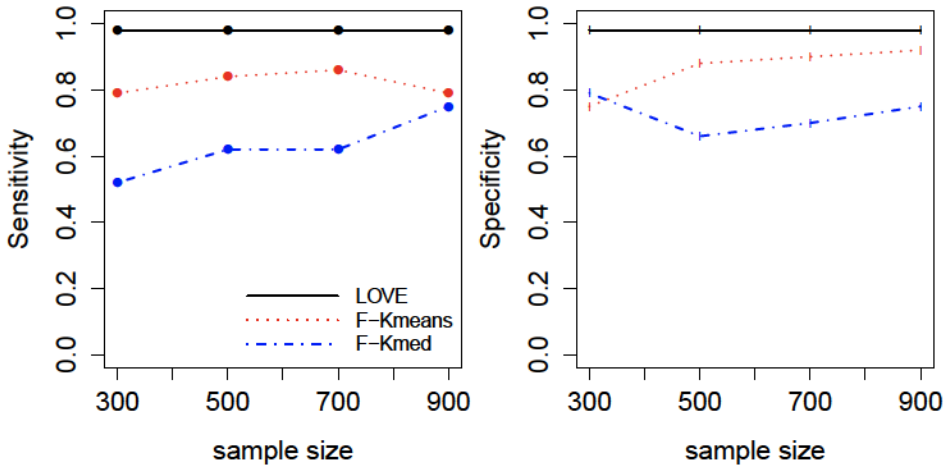

Fig 1: Plot of specificity and sensitivity for LOVE, fuzzy K-means (F-Kmeans), and fuzzy K-medoids (F-Kmed) when $p = 500$.

REFERENCES

Bai, J. and Li, K. (2012). Statistical analysis of factor models of high dimension. *Ann. Statist.* **40** 436–465.

Bezdek, J. C. (2013). *Pattern recognition with fuzzy objective function algorithms*. Springer Science & Business Media.

Bunea, F., Giraud, C. and Luo, X. (2016a). Minimax Optimal Variable Clustering in G-models via Cord. *arXiv preprint arXiv:1508.01939*.

Krishnapuram, R., Joshi, A., Nasraoui, O. and Yi, L. (2001). Low-complexity fuzzy relational clustering algorithms for web mining. *IEEE transactions on Fuzzy Systems* **9** 595–607.

Rigollet, P. and Tsybakov, A. (2011). Exponential Screening and optimal rates of sparse estimation. *Ann. Statist.* **39** 731–771.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer, New York.

Wiwie, C., Baumbach, J. and Röttger, R. (2015). Comparing the performance of biomedical clustering methods. *Nature methods* **12** 1033–1038.

X. Bing
Department of Statistics and Data Science
Cornell University
Ithaca, New York 14853-3801
USA
E-mail: xb43@cornell.edu

F. Bunea
Department of Statistics and Data Science
Cornell University
Ithaca, New York 14853-3801
USA
E-mail: fb238@cornell.edu

Y. Ning
Department of Statistics and Data Science
Cornell University
Ithaca, New York 14853-3801
USA
E-mail: yn265@cornell.edu

M. Wegkamp
Department of Mathematics &
Department of Statistics and Data Science
Cornell University
Ithaca, New York 14853-3801
USA
E-mail: mhw73@cornell.edu