Supplementary Appendix for “Generalized Liquid Association Analysis for Multimodal Data Integration”

Lexin Li[†], Jing Zeng[‡], and Xin Zhang[‡]

[†]*University of California at Berkeley and* [‡]*Florida State University*

S1 Additional Simulations with Different Functions

We present some additional numerical results, by considering different function forms for $f_k(\cdot)$, $k = 1, 2$, in $\boldsymbol{f}(\boldsymbol{\Gamma}_3^\top \boldsymbol{Z}) = \mathrm{diag}\{f_1(\boldsymbol{\Gamma}_3^\top \boldsymbol{Z}), f_2(\boldsymbol{\Gamma}_3^\top \boldsymbol{Z})\} \in \mathbb{R}^{2\times 2}$. Particularly, we consider a family of tanh-type functions,

$$f(a; \rho, \xi) = \rho \left(\frac{2}{1 + e^{-2\xi a}} - 1 \right), \quad 0 < \rho \leq 1, \;\; \xi > 0.$$

The slope of this function at $a = 0$ reflects the strength of the change of correlation at point $Z = 0$ for any pair (X_i, Y_j), $i = 1, \ldots, p_1$, $j = 1, \ldots, p_2$. This slope is controlled by both ρ and ξ. When ρ is fixed, the function $f(a; \rho, \xi)$ at $a = 0$ becomes steeper as ξ increases. Figure S1 shows two such functions, where $\rho = 0.95$, and $\xi = 1$ and 5, respectively. When $\xi \to \infty$, $f(a; \rho, \xi)$ is a step function. When $\rho = 1$ and $\xi = 1$, $f(a; \rho, \xi)$ is a standard tanh function.

We adopt the simulation setup as in Scenario 1 in Section 5 of the paper, but with the new function $f(a; \rho, \xi)$. That is, we increase the dimensions of $\boldsymbol{X}$ and $\boldsymbol{Y}$ as $p_1 = p_2 = \{100, 200, 300,$

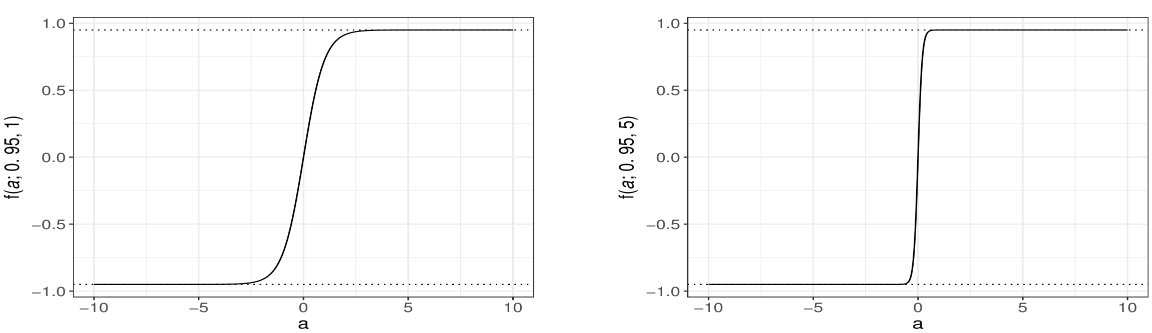

Figure S1: The tanh-type functions: $f(a; 0.95, 1)$ on the left panel, and $f(a; 0.95, 5)$ on the right.

p	Method	TPR-1 (%)	FPR-1 (%)	TPR-2 (%)	FPR-2 (%)	D
100	GLAA	1.000 (0.000)	0.000 (0.000)	1.000 (0.000)	0.000 (0.000)	0.131 (0.004)
	ULA	0.932 (0.010)	0.004 (0.001)	0.94 (0.009)	0.003 (0.000)	0.822 (0.002)
	PMD	0.734 (0.034)	0.682 (0.030)	0.708 (0.032)	0.680 (0.030)	0.976 (0.002)
	SCCA	0.498 (0.036)	0.520 (0.033)	0.52 (0.034)	0.521 (0.033)	0.986 (0.002)
200	GLAA	0.996 (0.003)	0.001 (0.001)	0.996 (0.003)	0.001 (0.001)	0.148 (0.009)
	ULA	0.668 (0.020)	0.009 (0.001)	0.704 (0.020)	0.008 (0.001)	0.929 (0.003)
	PMD	0.790 (0.032)	0.743 (0.029)	0.788 (0.031)	0.742 (0.029)	0.987 (0.001)
	SCCA	0.622 (0.032)	0.612 (0.028)	0.630 (0.033)	0.610 (0.028)	0.993 (0.001)
300	GLAA	0.952 (0.021)	0.002 (0.001)	0.954 (0.020)	0.003 (0.001)	0.181 (0.02)
	ULA	0.498 (0.020)	0.009 (0.000)	0.478 (0.022)	0.009 (0.000)	0.973 (0.002)
	PMD	0.762 (0.033)	0.723 (0.029)	0.706 (0.035)	0.721 (0.029)	0.993 (0.001)
	SCCA	0.640 (0.026)	0.653 (0.018)	0.618 (0.025)	0.657 (0.017)	0.997 (0.000)
400	GLAA	0.738 (0.042)	0.022 (0.006)	0.742 (0.042)	0.022 (0.006)	0.387 (0.039)
	ULA	0.328 (0.022)	0.009 (0.000)	0.336 (0.020)	0.008 (0.000)	0.989 (0.001)
	PMD	0.756 (0.031)	0.731 (0.027)	0.78 (0.029)	0.735 (0.027)	0.994 (0.001)
	SCCA	0.598 (0.026)	0.597 (0.012)	0.614 (0.024)	0.590 (0.012)	0.997 (0.000)
500	GLAA	0.682 (0.042)	0.035 (0.008)	0.664 (0.043)	0.033 (0.008)	0.471 (0.041)
	ULA	0.186 (0.019)	0.008 (0.000)	0.234 (0.018)	0.008 (0.000)	0.993 (0.001)
	PMD	0.778 (0.031)	0.757 (0.026)	0.820 (0.029)	0.754 (0.027)	0.995 (0.000)
	SCCA	0.522 (0.022)	0.525 (0.008)	0.532 (0.023)	0.522 (0.008)	0.998 (0.000)

Table S1: Simulation results for the f_k functions with $\xi = 1$. The reported are the average TPR and FPR for the variable selection accuracy, and D for the subspace estimation accuracy, with the standard errors in the parenthesis. The results are over 100 replications.

$400, 500\}$, while fixing the dimension of $\boldsymbol{Z}$ at $p_3 = 1$. We set $r_1 = r_2 = 2$, $s_1 = s_2 = 5$, and $n = 500$. We consider two choices of $f(a; \rho, \xi)$: $f_1(a) = f(a; 0.95, 1)$, $f_2(a) = f(a; 0.85, 1)$; and $f_1(a) = f(a; 0.95, 5)$, $f_2(a) = f(a; 0.85, 5)$. Tables S1 and S2 report the accuracy of variable selection and subspace estimation. We observe similar qualitative patterns as in Table 1 of the paper that our proposed GLAA method clearly outperforms all the alternative methods.

S2 Algorithmic Convergence

We study the algorithmic convergence of our Algorithm 1 developed in Section 3 of the paper. We also carry out a simulation study to examine empirically both algorithmic and statistical convergence as the number of iterations increases.

Our main strategy is to first utilize the result from Theorem 3 that our method can consistently

p	Method	TPR-1 (%)	FPR-1 (%)	TPR-2 (%)	FPR-2 (%)	D
100	GLAA	1.000 (0.000)	0.000 (0.000)	1.000 (0.000)	0.000 (0.000)	0.098 (0.003)
	ULA	0.996 (0.003)	0.000 (0.000)	0.998 (0.002)	0.000 (0.000)	0.777 (0.002)
	PMD	0.772 (0.031)	0.713 (0.03)	0.758 (0.033)	0.703 (0.030)	0.971 (0.002)
	SCCA	0.570 (0.032)	0.520 (0.029)	0.510 (0.035)	0.523 (0.028)	0.988 (0.002)
200	GLAA	1.000 (0.000)	0.000 (0.000)	1.000 (0.000)	0.000 (0.000)	0.097 (0.003)
	ULA	0.946 (0.010)	0.001 (0.000)	0.936 (0.011)	0.002 (0.000)	0.875 (0.002)
	PMD	0.754 (0.033)	0.721 (0.030)	0.752 (0.032)	0.715 (0.030)	0.988 (0.001)
	SCCA	0.640 (0.028)	0.661 (0.027)	0.654 (0.031)	0.664 (0.027)	0.995 (0.001)
300	GLAA	0.972 (0.016)	0.001 (0.001)	0.97 (0.017)	0.001 (0.001)	0.129 (0.016)
	ULA	0.820 (0.014)	0.003 (0.000)	0.822 (0.014)	0.003 (0.000)	0.949 (0.002)
	PMD	0.774 (0.031)	0.761 (0.025)	0.802 (0.026)	0.760 (0.025)	0.992 (0.001)
	SCCA	0.672 (0.024)	0.678 (0.016)	0.662 (0.025)	0.681 (0.015)	0.997 (0.000)
400	GLAA	0.904 (0.029)	0.010 (0.004)	0.906 (0.029)	0.010 (0.004)	0.203 (0.028)
	ULA	0.652 (0.018)	0.004 (0.000)	0.630 (0.018)	0.005 (0.000)	0.979 (0.001)
	PMD	0.820 (0.030)	0.771 (0.029)	0.776 (0.032)	0.772 (0.029)	0.994 (0.001)
	SCCA	0.632 (0.023)	0.592 (0.012)	0.634 (0.024)	0.588 (0.012)	0.997 (0.000)
500	GLAA	0.850 (0.035)	0.007 (0.003)	0.850 (0.035)	0.008 (0.003)	0.254 (0.033)
	ULA	0.490 (0.021)	0.005 (0.000)	0.494 (0.018)	0.005 (0.000)	0.991 (0.001)
	PMD	0.784 (0.030)	0.768 (0.026)	0.772 (0.030)	0.767 (0.026)	0.996 (0.000)
	SCCA	0.530 (0.026)	0.525 (0.007)	0.502 (0.025)	0.524 (0.007)	0.997 (0.000)

Table S2: Simulation results for the f_k functions with $\xi = 5$. The rest is the same as Table S1.

select the active variables. This helps bridge our SVD problem to an equivalent higher-order orthogonal iteration (HOOI) problem. We then build on the recent results of Xu (2018) to establish the global and local algorithmic convergence for our method.

Recall the sample tensor estimator $\widetilde{\boldsymbol{\Delta}} = n^{-1}\sum_{i=1}^{n} \boldsymbol{X}_i \circ \boldsymbol{Y}_i \circ \boldsymbol{Z}_i \in \mathbb{R}^{p_1 \times p_2 \times p_3}$. Consider the sub-tensor $\widetilde{\boldsymbol{\Delta}}_s = \widetilde{\boldsymbol{\Delta}}_{[I_1,I_2,I_3]} \in \mathbb{R}^{s_1 \times s_2 \times s_3}$ that corresponds to the true active sets I_1, I_2, I_3 as defined in (4). For notational simplicity and without loss of generality, we assume all the active variables are located together at the beginning of each mode, i.e., $I_k = \{1, \ldots, s_k\}$, $k = 1, 2, 3$. Consider the optimization problem,

$$\text{maximize}_{\boldsymbol{\Lambda}_1,\boldsymbol{\Lambda}_2,\boldsymbol{\Lambda}_3} \|\widetilde{\boldsymbol{\Delta}}_s \times_1 \boldsymbol{P}_{\boldsymbol{\Lambda}_1} \times_2 \boldsymbol{P}_{\boldsymbol{\Lambda}_2} \times_3 \boldsymbol{P}_{\boldsymbol{\Lambda}_3}\|_F^2, \tag{S1}$$

where $\boldsymbol{\Lambda}_k \in \mathbb{R}^{s_k \times r_k}$ is a semi-orthogonal matrix, and $\boldsymbol{P}_{\boldsymbol{\Lambda}_k}$ is the projection onto the subspace $\text{span}(\boldsymbol{\Lambda}_k)$, $k = 1, 2, 3$. This optimization problem (S1) can be solved by the usual HOOI algorithm (De Lathauwer et al., 2000). Let $\widehat{\boldsymbol{\Lambda}}^{(t)} = (\widehat{\boldsymbol{\Lambda}}_1^{(t)}, \widehat{\boldsymbol{\Lambda}}_2^{(t)}, \widehat{\boldsymbol{\Lambda}}_3^{(t)})$ denote the HOOI estimate from the tth iteration for (S1). Besides, let $\widehat{\boldsymbol{\Gamma}}^{(t)} = (\widehat{\boldsymbol{\Gamma}}_1^{(t)}, \widehat{\boldsymbol{\Gamma}}_2^{(t)}, \widehat{\boldsymbol{\Gamma}}_3^{(t)})$ denote the estimate of our Algorithm 1 from the t-th iteration. The next proposition shows that our Algorithm 1 and the HOOI algorithm for solving the optimization problem (S1) produce equivalent results.

Proposition S1. *Suppose Assumptions (A1) to (A3) hold. Suppose $s \log p = o(n)$, and $t_{\max} = o(p)$. Then with probability tending to one, for the estimate $\widehat{\boldsymbol{\Gamma}}^{(t)}$ from Algorithm 1 and the HOOI estimate $\widehat{\boldsymbol{\Lambda}}^{(t)}$ from solving* (S1)*, we have $\widehat{\boldsymbol{\Gamma}}_k^{(t)} = ((\widehat{\boldsymbol{\Lambda}}_k^{(t)})^\top, \mathbf{0}^\top)^\top$, for $k = 1, 2, 3$, and $0 \leq t \leq t_{\max}$.*

Proposition S1 allows us to study the algorithmic convergence of our Algorithm 1 through the convergence of the HOOI algorithm, which was recently established by Xu (2018). Following Xu (2018), we introduce two definitions.

Definition S1 (Block-nondegeneracy)**.** *For a feasible solution $\widehat{\boldsymbol{\Lambda}} = (\widehat{\boldsymbol{\Lambda}}_1, \widehat{\boldsymbol{\Lambda}}_2, \widehat{\boldsymbol{\Lambda}}_3)$ of* (S1)*, let $\boldsymbol{\Omega}_1 = \widetilde{\boldsymbol{\Delta}}_{s,1}(\widehat{\boldsymbol{\Lambda}}_2 \otimes \widehat{\boldsymbol{\Lambda}}_3)$, $\boldsymbol{\Omega}_2 = \widetilde{\boldsymbol{\Delta}}_{s,2}(\widehat{\boldsymbol{\Lambda}}_3 \otimes \widehat{\boldsymbol{\Lambda}}_1)$, and $\boldsymbol{\Omega}_3 = \widetilde{\boldsymbol{\Delta}}_{s,3}(\widehat{\boldsymbol{\Lambda}}_1 \otimes \widehat{\boldsymbol{\Lambda}}_2)$, where $\widetilde{\boldsymbol{\Delta}}_{s,k}$ denotes the mode-k matricization of the tensor $\widetilde{\boldsymbol{\Delta}}_s$, $k = 1, 2, 3$. We say $\widehat{\boldsymbol{\Lambda}}$ is block-nondegenerative, if $\lambda_{r_k}(\boldsymbol{\Omega}_k) > \lambda_{r_k+1}(\boldsymbol{\Omega}_k)$, $k = 1, 2, 3$, where $\lambda_r(\cdot)$ denotes the rth singular value.*

Definition S2 (Block-wise maximizer)**.** *We say a solution $\widehat{\boldsymbol{\Lambda}} = (\widehat{\boldsymbol{\Lambda}}_1, \widehat{\boldsymbol{\Lambda}}_2, \widehat{\boldsymbol{\Lambda}}_3)$ is a block-wise maximizer of* (S1)*, if $\widehat{\boldsymbol{\Lambda}}_k$ is a solution to $\max_{\boldsymbol{\Lambda}_k} \|\widetilde{\boldsymbol{\Delta}}_s \times_k \boldsymbol{P}_{\boldsymbol{\Lambda}_k} \times_{j \neq k} \boldsymbol{P}_{\widehat{\boldsymbol{\Lambda}}_j}\|_F^2$, for $k = 1, 2, 3$.*

We next establish the global and local convergence of $\widehat{\boldsymbol{\Gamma}}^{(t)}$ from our Algorithm 1.

Theorem S1. *Suppose Assumptions (A1) to (A3) hold. Suppose $s \log p = o(n)$, and $t_{\max} = o(p)$. Then with probability tending to one, we have:*

(a) (Global convergence) If the HOOI sequence $\{\widehat{\boldsymbol{\Lambda}}^{(t)}\}_{t \geq 1}$ from (S1) *has a block-nondegenerative cluster point $\widehat{\boldsymbol{\Lambda}} = (\widehat{\boldsymbol{\Lambda}}_1, \widehat{\boldsymbol{\Lambda}}_2, \widehat{\boldsymbol{\Lambda}}_3)$, then $\widehat{\boldsymbol{\Lambda}}$ is a critical point and also a block-wise maximizer of* (S1)*, and $\lim_{t\to\infty} \boldsymbol{P}_{\widehat{\boldsymbol{\Lambda}}_k^{(t)}} = \boldsymbol{P}_{\widehat{\boldsymbol{\Lambda}}_k}$, $k = 1, 2, 3$. In addition, for the sequence $\{\widehat{\boldsymbol{\Gamma}}^{(t)}\}_{t \geq 1}$ from Algorithm 1, where $\widehat{\boldsymbol{\Gamma}}^{(t)} = (\widehat{\boldsymbol{\Gamma}}_1^{(t)}, \widehat{\boldsymbol{\Gamma}}_2^{(t)}, \widehat{\boldsymbol{\Gamma}}_3^{(t)})$, we have $\widehat{\boldsymbol{\Gamma}}_k^{(t)} = ((\widehat{\boldsymbol{\Lambda}}_k^{(t)})^\top, \mathbf{0}^\top)^\top$, and $\lim_{t\to\infty} \boldsymbol{P}_{\widehat{\boldsymbol{\Gamma}}_k^{(t)}} = \boldsymbol{P}_{\widehat{\boldsymbol{\Gamma}}_k}$, where $\widehat{\boldsymbol{\Gamma}}_k = (\widehat{\boldsymbol{\Lambda}}_k^\top, \mathbf{0}^\top)^\top$, and $k = 1, 2, 3$.*

(b) (Local convergence) If the initial estimator is sufficiently close to any block-nondegenerative local maximizer of (S1)*, then the sequence $\{\boldsymbol{P}_{\widehat{\boldsymbol{\Lambda}}_k^{(t)}}\}_{t \geq 1}$ must converge to some limiting point $\boldsymbol{P}_{\widetilde{\boldsymbol{\Lambda}}_k}$, $k = 1, 2, 3$, and the point $\widetilde{\boldsymbol{\Lambda}} = (\widetilde{\boldsymbol{\Lambda}}_1, \widetilde{\boldsymbol{\Lambda}}_2, \widetilde{\boldsymbol{\Lambda}}_3)$ is a local maximaizer of* (S1)*. In addition, for the sequence $\{\widehat{\boldsymbol{\Gamma}}^{(t)}\}_{t \geq 1}$, we have $\lim_{t\to\infty} \boldsymbol{P}_{\widehat{\boldsymbol{\Gamma}}_k^{(t)}} = \boldsymbol{P}_{\widetilde{\boldsymbol{\Gamma}}_k}$, where $\widetilde{\boldsymbol{\Gamma}}_k = ((\widetilde{\boldsymbol{\Lambda}}_k)^\top, \mathbf{0}^\top)^\top$, and $k = 1, 2, 3$.*

Statement (a) shows that, if there exists a cluster point $\widehat{\boldsymbol{\Lambda}}$ for $\{\widehat{\boldsymbol{\Lambda}}^{(t)}\}$, i.e., the point that a subsequence of $\{\widehat{\boldsymbol{\Lambda}}^{(t)}\}$ converges to, and this cluster point is block-nondegenerative, then this point is a critical point, where the derivative of the objective function $\|\widetilde{\boldsymbol{\Delta}} \times_1 \boldsymbol{P}_{\boldsymbol{\Lambda}_1} \times_2 \boldsymbol{P}_{\boldsymbol{\Lambda}_2} \times_3 \boldsymbol{P}_{\boldsymbol{\Lambda}_3}\|_F^2$ over $\boldsymbol{\Lambda}$ is

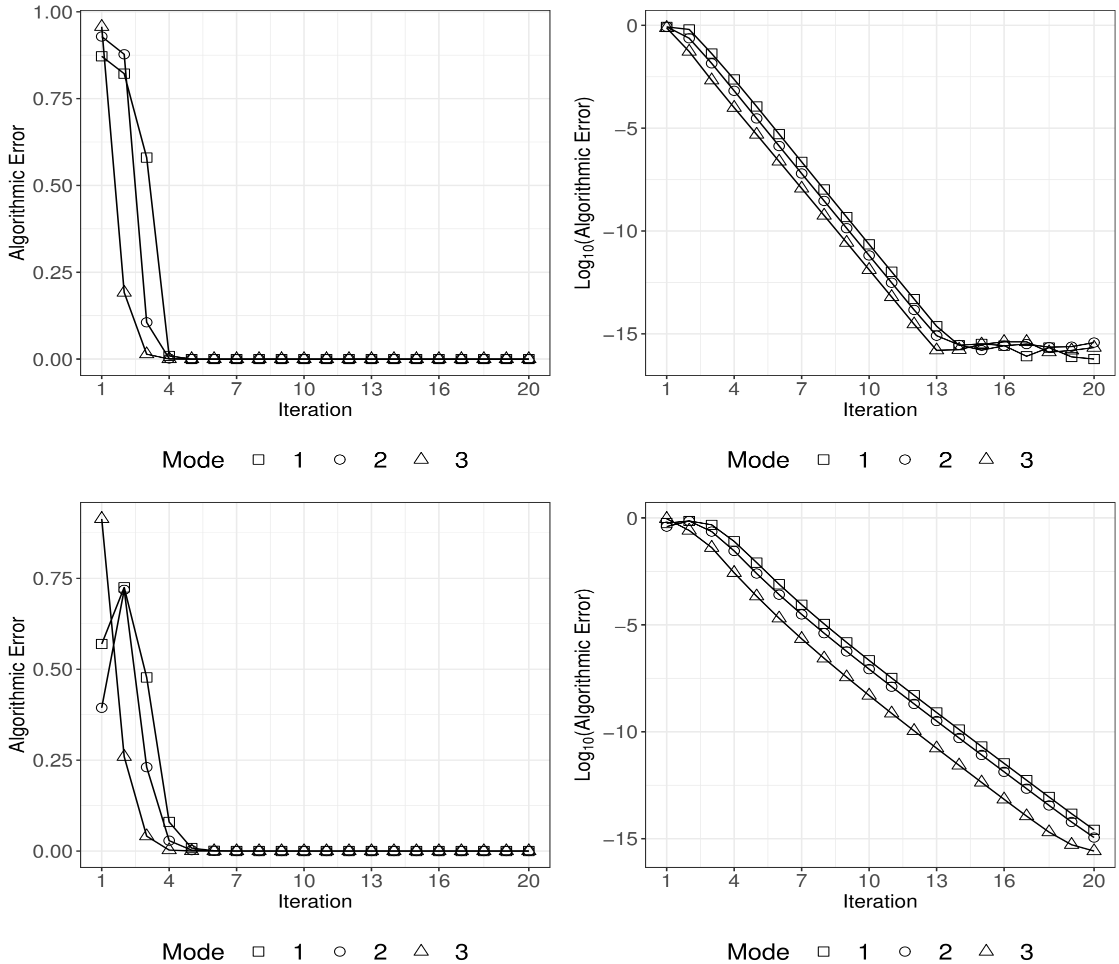

Figure S2: The algorithmic error versus the number of iterations. The top panels are based on a single data replication, and the bottom panels are averaged over 20 replications. The left panels are in the original scale, and the right panels are in the logarithmic scale.

zero. It is also a block-wise maximizer, to which the entire sequence $\lim_{t\to\infty} \boldsymbol{P}_{\widehat{\boldsymbol{\Gamma}}_k^{(t)}}$ converges. Xu (2018) argued that the block-nondegeneracy is a mild condition and generally holds in practice. A similar condition was also imposed in the orthogonal iteration method (Golub and Van Loan, 1996, Section 7.3.2) for computing the dominant invariant subspace of a matrix. It is needed to ensure the stability of the critical point, and in turn the convergence of $\{\widehat{\boldsymbol{\Lambda}}^{(t)}\}$ and $\{\widehat{\boldsymbol{\Gamma}}^{(t)}\}$.

We further carry out a simulation to study the algorithmic convergence of Algorithm 1 empirically. We adopt the second scenario in Section 5, where the three modalities are $\boldsymbol{X} \in \mathbb{R}^{p_1}$, $\boldsymbol{Y} \in \mathbb{R}^{p_2}$, and $\boldsymbol{Z} \in \mathbb{R}^{p_3}$, with $p_1 = p_2 = 100, p_3 = 20$, $s_1 = s_2 = s_3 = 5$, $r_1 = r_2 = 2, r_3 = 1$, and $n = 500$. We run Algorithm 1 for 20 iterations, as in all our simulations, we have found that the algorithm always converges within 20 iterations. We repeat for 20 data replications, and report

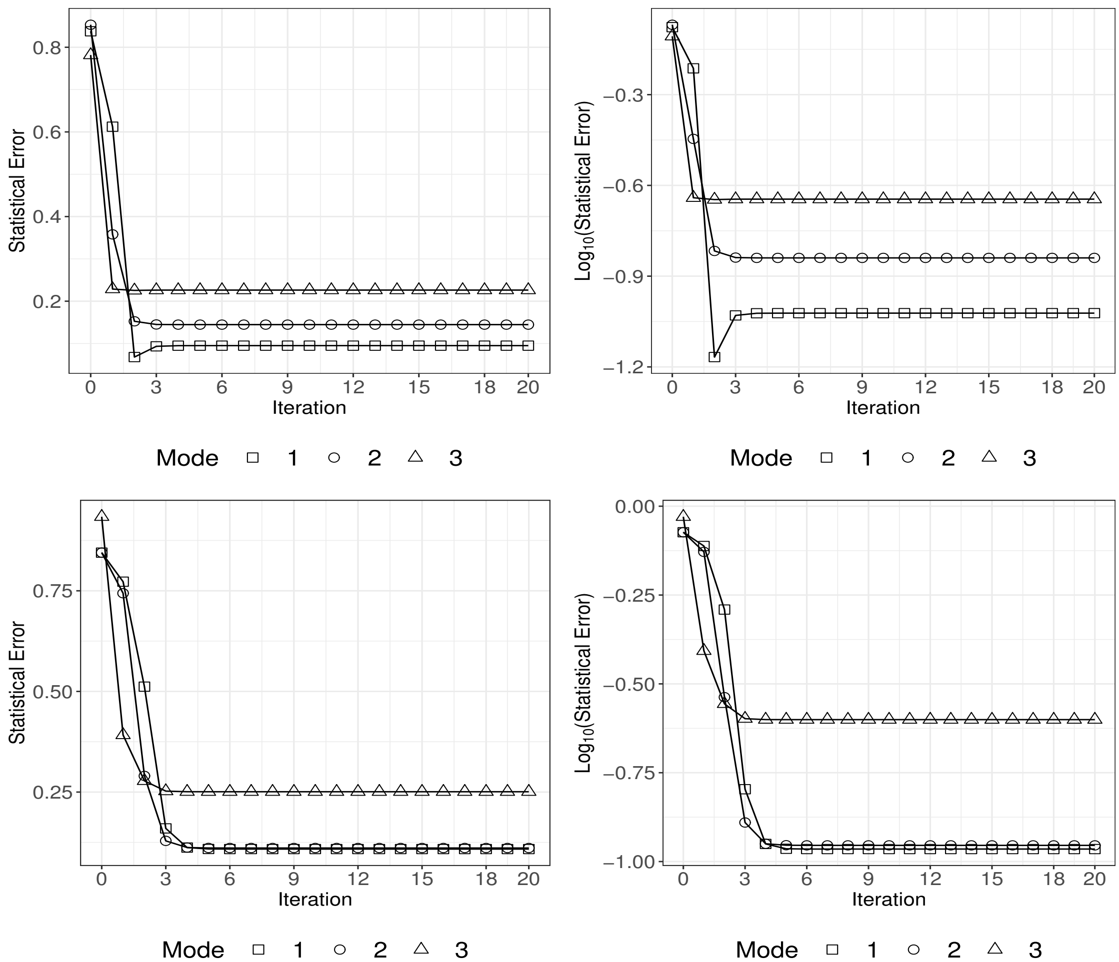

Figure S3: The statistical error versus the number of iterations. The top panels are based on a single data replication, and the bottom panels are averaged over 20 replications. The left panels are in the original scale, and the right panels are in the logarithmic scale.

the results from both a single replication and the averages from all 20 replications.

Figure S2 reports the algorithmic error,

$$D_a(\widehat{\mathbf{\Gamma}}_k^{(t)}, \widehat{\mathbf{\Gamma}}_k^{(t-1)}) = \|\boldsymbol{P}_{\widehat{\mathbf{\Gamma}}_k^{(t)}} - \boldsymbol{P}_{\widehat{\mathbf{\Gamma}}_k^{(t-1)}}\|_F/\sqrt{2r_k}, \text{ for mode } k = 1, 2, 3,$$

as the number of iterations t increases. The top panels are based on a single data replication, and the bottom panels are averaged over 20 replications. The left panels are in the original scale, and the right panels are in the logarithmic scale. It is seen that the sequence $\widehat{\mathbf{\Gamma}}^{(t)}$ converges quickly.

Figure S3 reports the statistical error,

$$\begin{aligned} D_s(\widehat{\mathbf{\Gamma}}_k^{(t)}, \mathbf{\Gamma}_k) &= \|\boldsymbol{P}_{\widehat{\mathbf{\Gamma}}_k^{(t)}} - \boldsymbol{P}_{\mathbf{\Gamma}_k}\|_F/\sqrt{2r_k}, \text{ for mode } k = 1, 2, \\ D_s(\widehat{\mathbf{\Gamma}}_3^{(t)}, \boldsymbol{G}_3') &= \|\boldsymbol{P}_{\widehat{\mathbf{\Gamma}}_3^{(t)}} - \boldsymbol{P}_{\boldsymbol{G}_3'}\|_F/\sqrt{2r_3}, \text{ for mode } 3, \end{aligned}$$

as the number of iterations t increases. The plot construction follows S2. It is seen that the statistical error decreases first, then becomes stable, as t increases. This agrees with our Theorem 2, which suggests that the statistical error is to decrease first, then is dominated by some constant.

Finally, we remark that, Theorem S1 establishes the algorithmic convergence of our Algorithm 1, but does not obtain the explicit convergence rate. Figure S2 suggests that, empirically, this rate is close to linear in the logarithmic scale. But deriving the explicit algorithmic convergence rate of Algorithm 1 theoretically is very challenging, and remains an open question in the general SVD optimization literature. We leave it as future research.

S3 Theory for Non-sparse Estimator

In some real applications, the modalities may be low-dimensional or have no sparsity structure. We can easily modify our Algorithm 1 to obtain a non-sparse version of the method. That is, we set $\eta_k = 0$ in Step 1, and $\widetilde{\eta}_k = 0$ in Step 2 of Algorithm 1, for $k = 1, 2, 3$. Then, $\widehat{I}_k^{(t)} = I_k$, for $t = 0, \ldots, t_{\max}$, and $k = 1, 2, 3$. We next establish the corresponding non-asymptotic error bounds for such non-sparse estimators.

Specifically, we consider the following non-sparse version of the dimension reduction model (1) for the three-way association analysis,

$$\boldsymbol{g}(\boldsymbol{z}) = \mathrm{E}(\boldsymbol{X}\boldsymbol{Y}^\top \mid \boldsymbol{Z} = \boldsymbol{z}) = \boldsymbol{\Gamma}_1^{\mathrm{ns}} \boldsymbol{f}\{(\boldsymbol{\Gamma}_3^{\mathrm{ns}})^\top \boldsymbol{z}\}(\boldsymbol{\Gamma}_2^{\mathrm{ns}})^\top, \tag{S2}$$

where $\boldsymbol{\Gamma}_k^{\mathrm{ns}} \in \mathbb{R}^{p_k \times r_k}, k = 1, 2, 3$, are semi-orthogonal basis matrices, and $\boldsymbol{f} : \mathbb{R}^{r_3} \mapsto \mathbb{R}^{r_1 \times r_2}$ is some latent function. The only difference between model (S2) and model (1) is that $\boldsymbol{\Gamma}_k^{\mathrm{ns}}$ is non-sparse, i.e., $s_k = p_k$. Let $\widehat{\boldsymbol{\Delta}}^{\mathrm{ns}}$ and $\widehat{\boldsymbol{\Gamma}}_k^{\mathrm{ns}}$, $k = 1, 2, 3$, denote the estimators returned from the non-sparse version of Algorithm 1. To establish their non-asymptotic error bounds, we modify Assumption (A2) accordingly as,

(A2$'$) Suppose $\lambda \geq \max\{C_4\sqrt{p/n^{1-b}}, C_5\}$ for some constants $C_4, C_5 > 0$ and $0 < b < 1$, where $\lambda \equiv \min\{\lambda_1, \lambda_2, \lambda_3\}$, and λ_k is the smallest non-zero singular value of $\boldsymbol{\Delta}_k$, $k = 1, 2, 3$.

Theorem S2. *Suppose Assumptions (A1), (A2$'$), and model* (S2) *hold. Suppose* $\log p \leq c_0 n^b$ *for some constant* $c_0 > 0$*, and* $0 < b < 1$ *is as defined in Assumption (A2$'$). Then, with probability at least* $1 - C'' \max\left[\exp(-n^b), \exp\{-(\sqrt{n^{1+b}(1+c_0)/2} - c_0 n^b)\}\right]$*, we have: (a)* $\|\widehat{\boldsymbol{\Delta}}^{\mathrm{ns}} - \boldsymbol{\Delta}\|_F \leq (c_5 + c_6 2^{-t_{\max}})\sqrt{p/n^{1-b}}$*; and (b)* $\max_{k=1,2} \|\boldsymbol{P}_{\widehat{\boldsymbol{\Gamma}}_k^{\mathrm{ns}}} - \boldsymbol{P}_{\boldsymbol{\Gamma}_k^{\mathrm{ns}}}\|_F \leq (c_7 + c_8 2^{-t_{\max}})\sqrt{p/n^{1-b}}$, $\|\boldsymbol{P}_{\widehat{\boldsymbol{\Gamma}}_3^{\mathrm{ns}}} - \boldsymbol{P}_{\boldsymbol{G}_3'}\|_F \leq (c_7 + c_8 2^{-t_{\max}})\sqrt{p/n^{1-b}}$*, where* C'', c_5, c_6, c_7, c_8 *are some positive constants.*

We make a few remarks. First, when the sample size n is large enough, such that $n \geq \{2(1 + c_0)\}^{1/(1-b)}$, the statements in Theorem S2 hold with probability at least $1 - C'' \exp(-n^b)$. Besides, the constants $c_5, \ldots, c_8$ depend on the constants M and σ as specified in Assumption (A1), and their explicit forms are given in the proof of Theorem S2. Second, the asymptotic consistency results for the non-sparse estimators can be obtained from Theorem S2 directly. If we allow p to diverge with n at the rate of $p = o(n^{1-b})$, then the estimated Tucker tensor $\widehat{\boldsymbol{\Delta}}^{\mathrm{ns}}$, and the estimated subspaces spanned by $\widehat{\boldsymbol{\Gamma}}_k^{\mathrm{ns}}$, $k = 1, 2, 3$, are consistent. Finally, comparing the non-sparse estimators with their sparse counterparts, the convergence rate of the non-sparse estimators from Theorem S2 is $O(\sqrt{p/n^{1-b}})$. It is slower than the convergence rate of the sparse estimators, which is $O(\sqrt{s \log p/n})$ from Theorem 2, when the true model is sparse.

S4 Auxiliary Lemmas

We first present a number of useful lemmas, then present the proofs of the propositions and theorems in the paper. We employ the following notation. Let $\mathbb{O}^{p \times q} = \{\boldsymbol{U} \in \mathbb{R}^{p \times q} \mid \boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}_q\}$, where $p \geq q$. For a matrix $\boldsymbol{U} \in \mathbb{O}^{p \times q}$, let $\sigma_i(\boldsymbol{U})$ denote its ith largest singular value, $i = 1, \ldots, q$. Let $\boldsymbol{U}_\perp \in \mathbb{O}^{p \times (p-q)}$ denote a basis matrix of the orthogonal complement subspace of $\mathrm{span}(\boldsymbol{U})$. For any two matrices $\boldsymbol{U}, \widetilde{\boldsymbol{U}} \in \mathbb{O}^{p \times q}$, let $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_q$ denote the singular values of $\boldsymbol{U}^\top \widetilde{\boldsymbol{U}}$ in a descending order. Define the diagonal matrix $\sin \Theta(\boldsymbol{U}, \widehat{\boldsymbol{U}}) = \mathrm{diag}(\sin \theta_1, \ldots, \sin \theta_q)$, where $\theta_i = \cos^{-1}(\sigma_i)$, $i = 1, \ldots, q$.

Lemma 1. *(Zhang and Han, 2019, Lemmas 2 and 6).*

(a) For any matrix $\boldsymbol{A} \in \mathbb{R}^{p \times q}$, with $\mathrm{rank}(\boldsymbol{A}) = r \leq \min\{p, q\}$, let $\boldsymbol{U} \in \mathbb{O}^{p \times r}$ denote the top-r left singular vectors of $\boldsymbol{A}$. Then, for any $\widetilde{\boldsymbol{U}} \in \mathbb{O}^{p \times r}$, $\|(\widetilde{\boldsymbol{U}}_\perp)^\top \boldsymbol{A}\|_F \geq \sigma_r(\boldsymbol{A}) \|\sin \Theta(\widetilde{\boldsymbol{U}}, \boldsymbol{U})\|$.

(b) For any matrices $\boldsymbol{A}, \widetilde{\boldsymbol{A}} \in \mathbb{R}^{p \times q}$, with $\mathrm{rank}(\widetilde{\boldsymbol{A}}) \leq r \leq \min\{p, q\}$, let $\boldsymbol{U} \in \mathbb{R}^{p \times r}$ denote the top-r left singular vectors of $\boldsymbol{A}$. Then, $\|\boldsymbol{U}_\perp^\top \widetilde{\boldsymbol{A}}\|_F \leq 2\|\boldsymbol{A} - \widetilde{\boldsymbol{A}}\|_F$.

Lemma 2. *(Cai and Zhang, 2018, Lemma 1). For $\boldsymbol{U}, \tilde{\boldsymbol{U}} \in \mathbb{O}^{p \times q}$, we have,*

$$\|\boldsymbol{P}_{\boldsymbol{U}} - \boldsymbol{P}_{\widetilde{\boldsymbol{U}}}\|_F = \sqrt{2}\|\sin \Theta(\boldsymbol{U}, \widetilde{\boldsymbol{U}})\|_F = \sqrt{2}\|\boldsymbol{U}_\perp^\top \widetilde{\boldsymbol{U}}\|_F.$$

A random variable X is sub-Gaussian, if $\mathrm{E}(X) = 0$, and there exists a positive constant σ, such that $\mathrm{E}\exp(sX) \leq \exp(s^2\sigma^2/2)$ for all $s \in \mathbb{R}$. It is denoted as $X \sim \mathrm{subG}(\sigma^2)$. A random variable Y is sub-exponential, if $\mathrm{E}(Y) = 0$, and there exist a pair of positive constants (ν^2, α), such that $\mathrm{E}\exp(sX) \leq \exp(s^2\nu^2/2)$ for $|s| < 1/\alpha$. It is denoted as $Y \sim \mathrm{subE}(\nu^2, \alpha)$.

Lemma 3 (Hoeffding's lemma)**.** *Let X be a random variable, such that $\mathrm{E}(X) = 0$ and $X \in [a, b]$ almost surely. Then, $X \sim \mathrm{subG}\{(b-a)^2/4\}$.*

Lemma 4. *(Wainwright, 2019).*

(a) *Let* $X \sim \mathrm{subG}\,(\sigma^2)$*, then the random variable* $Z = X^2 - \mathbb{E}\,(X^2)$ *is sub-exponential:* $Z \sim \mathrm{subE}(256\sigma^4, 16\sigma^2)$

(b) *Let* $X \sim \mathrm{subG}(\sigma^2)$ *and* $Y \sim \mathrm{subG}(\tau^2)$*, then* $X + Y \sim \mathrm{subG}\{(\sigma+\tau)^2\}$*. Additionally, if* X *is independent of* Y*, then* $X + Y \sim \mathrm{subG}(\sigma^2 + \tau^2)$*.*

(c) *Let* $X_1, \ldots, X_n$ *be* n *independent random variables such that* $X_i \sim \mathrm{subE}(\nu_i^2, \alpha_i)$*,* $i = 1, \ldots, n$*, then* $\sum_{i=1}^n X_i \sim \mathrm{subE}\left(\sum_{i=1}^n \nu_i^2, \max_{i\in\{1,\ldots,n\}}\{\alpha_i\}\right)$*.*

The next lemma establishes the concentration inequality for each element in $\widetilde{\boldsymbol{\Delta}}$.

Lemma 5. *Suppose the random variables* X *and* Y *are bounded, such that* $|XY| \leq M$ *for some constant* $M > 0$*, and* $Z \sim \mathrm{subG}(\sigma^2)$*. Let* (X_i, Y_i, Z_i) *denote independent copies of* (X, Y, Z)*. Then, for any* $\epsilon > 0$*, we have*

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n X_iY_iZ_i - \mathrm{E}(XYZ)\right| \geq \epsilon\right\} \leq 6\max\left\{\exp\left(-\frac{n\epsilon^2}{512(M+\sigma)^4}\right), \exp\left(-\frac{n\epsilon}{32(M+\sigma)^2}\right)\right\}.$$

Proof of Lemma 5: Consider the decomposition,

$$\begin{aligned}
&\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n X_iY_iZ_i - \mathrm{E}(XYZ)\right| \geq \epsilon\right\}\\
&= \mathbb{P}\left[\left|\sum_{i=1}^n [\{X_iY_i - \mathrm{E}(XY)\}Z_i + \mathrm{E}(XY)Z_i - \mathrm{E}[\{XY - \mathrm{E}(XY)\}Z] - \mathrm{E}(XY)\mathrm{E}(Z)]\right| \geq n\epsilon\right]\\
&\leq \mathbb{P}\left[\left|\sum_{i=1}^n [\{X_iY_i - \mathrm{E}(XY)\}Z_i - \mathrm{E}[\{XY - \mathrm{E}(XY)\}Z]]\right| \geq \frac{n\epsilon}{2}\right] + \mathbb{P}\left\{\left|\mathrm{E}(XY)\sum_{i=1}^n Z_i\right| \geq \frac{n\epsilon}{2}\right\}\\
&\equiv B_1 + B_2,
\end{aligned} \tag{S3}$$

where the inequality is due to that $\mathrm{E}(Z) = 0$. We next bound B_1 and B_2, respectively.

For B_1, write $F_i = X_iY_i - \mathrm{E}(XY)$. Besides, $F_iZ_i = \{(F_i + Z_i)^2 - (F_i - Z_i)^2\}/4$, and $\mathrm{E}(FZ) = \{\mathrm{E}(F_i + Z_i)^2 - \mathrm{E}(F_i - Z_i)^2\}/4$. Then,

$$\begin{aligned}
B_1 &= \mathbb{P}\left[\left|\sum_{i=1}^n \{F_iZ_i - \mathrm{E}(FZ)\}\right| \geq \frac{n\epsilon}{2}\right]\\
&= \mathbb{P}\left[\left|\sum_{i=1}^n \{(F_i + Z_i)^2 - \mathrm{E}(F_i + Z_i)^2\} - \sum_{i=1}^n \{(F_i - Z_i)^2 - \mathrm{E}(F_i - Z_i)^2\}\right| \geq 2n\epsilon\right]\\
&\leq \mathbb{P}\left[\left|\sum_{i=1}^n \{(F_i + Z_i)^2 - \mathrm{E}(F_i + Z_i)^2\}\right| \geq n\epsilon\right] + \mathbb{P}\left[\left|\sum_{i=1}^n \{(F_i - Z_i)^2 - \mathrm{E}(F_i - Z_i)^2\}\right| \geq n\epsilon\right]\\
&\equiv B_{1,1} + B_{1,2}.
\end{aligned}$$

For $B_{1,1}$, by Lemma 3, the bounded variable $F_i \sim \text{subG}(M^2)$. Recall that $Z_i \sim \text{subG}(\sigma^2)$. By Lemma 4(b), $F_i + Z_i \sim \text{subG}\{(M+\sigma)^2\}$. By Lemma 4(a) and Lemma 4(c), it follows that

$$(F_i + Z_i)^2 - \mathrm{E}(F_i + Z_i)^2 \sim \text{subE}\{256(M+\sigma)^4, 16(M+\sigma)^2\},$$
$$\sum_{i=1}^n \{(F_i + Z_i)^2 - \mathrm{E}(F_i + Z_i)^2\} \sim \text{subE}\{256n(M+\sigma)^4, 16(M+\sigma)^2\}.$$

Then, by the tail inequality of sub-exponential random variables, we have

$$B_{1,1} \le 2\max\left\{\exp\left(-\frac{n\epsilon^2}{512(M+\sigma)^4}\right), \exp\left(-\frac{n\epsilon}{32(M+\sigma)^2}\right)\right\}.$$

For $B_{1,2}$, since $-Z_i \sim \text{subG}(\sigma^2)$, by Lemma 4(b), $F_i - Z_i \sim \text{subG}\{(M+\sigma)^2\}$. Following a similar argument for the bound of $B_{1,1}$, we have

$$B_{1,2} \le 2\max\left\{\exp\left(-\frac{n\epsilon^2}{512(M+\sigma)^4}\right), \exp\left(-\frac{n\epsilon}{32(M+\sigma)^2}\right)\right\}.$$

Combining the bounds for $B_{1,1}$ and $B_{1,2}$, we obtain,

$$B_1 \le 4\max\left\{\exp\left(-\frac{n\epsilon^2}{512(M+\sigma)^4}\right), \exp\left(-\frac{n\epsilon}{32(M+\sigma)^2}\right)\right\}.$$

For B_2, since $|XY| \le M$, $|\mathrm{E}(XY)| \le M$. Combined with the fact that $Z \sim \text{subG}(\sigma^2)$, and by the tail inequality of sub-Gaussian random variable, we have

$$B_2 \le \mathbb{P}\left(|\sum_{i=1}^n Z_i| \ge \frac{n\epsilon}{2M}\right) \le 2\exp\left(-\frac{n\epsilon^2}{8M^2\sigma^2}\right).$$

Because $512(M+\sigma)^4 \ge 8M^2\sigma^2$, we can further bound B_2 as,

$$B_2 \le 2\exp\{-\frac{n\epsilon^2}{512(M+\sigma)^4}\} \le 2\max\left\{\exp\left(-\frac{n\epsilon^2}{512(M+\sigma)^4}\right), \exp\left(-\frac{n\epsilon}{32(M+\sigma)^2}\right)\right\}.$$

Combining the bounds for B_1 and B_2 completes the proof of Lemma 5. □

S5 Proofs

S5.1 Proof of Proposition 1

For X_j, $j = 1, \dots, p_1$, and Y_k, $k = 1, \dots, p_2$, we have that,

$$\begin{aligned} & \mathrm{E}\left\{\frac{d}{d\boldsymbol{Z}}\mathrm{E}(X_jY_k \mid \boldsymbol{Z})\right\} = \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathrm{cov}\{\boldsymbol{Z}, \mathrm{E}(X_jY_k \mid \boldsymbol{Z})\} \\ = \; & \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathrm{E}\{\boldsymbol{Z}\mathrm{E}(X_jY_k \mid \boldsymbol{Z})\} = \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\mathrm{E}(X_jY_k\boldsymbol{Z}), \end{aligned}$$

where the first equality is by definition, the second equality is by the multivariate version of Stein's lemma (Liu, 1994, Lemma 1). For the tensor $\boldsymbol{\Phi} = \mathrm{GLA}(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{Z}) \in \mathbb{R}^{p_1 \times p_2 \times p_3}$, let $\mathrm{GLA}_{jk\cdot}(\boldsymbol{X}, \boldsymbol{Y} | \boldsymbol{Z}) \in \mathbb{R}^{p_3}$ denote its (j,k)th fiber, $j = 1, \ldots, p_1$, $k = 1, \ldots, p_2$. By definition,

$$\mathrm{GLA}_{jk\cdot}(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{Z}) = \mathrm{E}\left\{\frac{d}{d\boldsymbol{Z}}\mathrm{E}(X_j Y_k \mid \boldsymbol{Z})\right\}.$$

Stacking all the fibers back to the tensor $\boldsymbol{\Phi}$ completes the proof of Proposition 1. $\square$

S5.2 Proof of Theorem 1

First, we prove part (a) of the theorem. By model (1) and Definition 1, we have,

$$\boldsymbol{\Phi} = \mathrm{E}\left\{\frac{d}{d\boldsymbol{Z}}\mathrm{E}(\boldsymbol{X}\boldsymbol{Y}^\top \mid \boldsymbol{Z})\right\} = \mathrm{E}\left[\frac{d}{d\boldsymbol{Z}}\left\{\boldsymbol{\Gamma}_1 \boldsymbol{f}(\boldsymbol{\Gamma}_3^\top \boldsymbol{Z})\boldsymbol{\Gamma}_2^\top\right\}\right],$$

which implies that,

$$\boldsymbol{\Phi} = \boldsymbol{\Phi} \times_1 \boldsymbol{P}_{\boldsymbol{\Gamma}_1} = \boldsymbol{\Phi} \times_2 \boldsymbol{P}_{\boldsymbol{\Gamma}_2} = \boldsymbol{\Phi} \times_1 \boldsymbol{P}_{\boldsymbol{\Gamma}_1} \times_2 \boldsymbol{P}_{\boldsymbol{\Gamma}_2}. \tag{S4}$$

Similar to the proof of Proposition 1, the (j,k)th fiber of $\boldsymbol{\Phi}$, $j = 1, \ldots, p_1$, $k = 1, \ldots, p_2$, satisfies,

$$\boldsymbol{\Phi}_{jk\cdot} = \mathrm{E}\left\{\frac{d}{d\boldsymbol{Z}}\mathrm{E}(X_j Y_k \mid \boldsymbol{Z})\right\} = \mathrm{E}\left\{\frac{d}{d\boldsymbol{Z}}\mathrm{E}(X_j Y_k \mid \boldsymbol{\Gamma}_3^\top \boldsymbol{Z})\right\} = \boldsymbol{\Gamma}_3 \mathrm{E}\left\{\frac{d}{d(\boldsymbol{\Gamma}_3^\top \boldsymbol{Z})}\mathrm{E}(X_j Y_k \mid \boldsymbol{\Gamma}_3^\top \boldsymbol{Z})\right\}.$$

The last equation implies that $\boldsymbol{\Phi}_{jk\cdot} = \boldsymbol{P}_{\boldsymbol{\Gamma}_3}\boldsymbol{\Phi}_{jk\cdot}$. Stacking all the (j,k)-th fibers to form the GLA tensor, we obtain that,

$$\boldsymbol{\Phi} = \boldsymbol{\Phi} \times_3 \boldsymbol{P}_{\boldsymbol{\Gamma}_3}. \tag{S5}$$

The optimization problem in (3),

$$(\boldsymbol{G}_1'', \boldsymbol{G}_2'', \boldsymbol{G}_3'') = \arg\min_{\boldsymbol{G}_1, \boldsymbol{G}_2, \boldsymbol{G}_3} \|\boldsymbol{\Phi} - \boldsymbol{\Phi} \times_1 \boldsymbol{P}_{\boldsymbol{G}_1} \times_2 \boldsymbol{P}_{\boldsymbol{G}_2} \times_3 \boldsymbol{P}_{\boldsymbol{G}_3}\|_F^2,$$

where $\boldsymbol{G}_k \in \mathbb{R}^{p_k \times r_k}$ is the semi-orthogonal matrix, is equivalent to

$$(\boldsymbol{G}_1'', \boldsymbol{G}_2'', \boldsymbol{G}_3'') = \arg\max_{\boldsymbol{G}_1, \boldsymbol{G}_2, \boldsymbol{G}_3} \|\boldsymbol{\Phi} \times_1 \boldsymbol{P}_{\boldsymbol{G}_1} \times_2 \boldsymbol{P}_{\boldsymbol{G}_2} \times_3 \boldsymbol{P}_{\boldsymbol{G}_3}\|_F^2.$$

By (S4) and (S5), the solutions to the above problems satisfy $\mathrm{span}(\boldsymbol{\Gamma}_k) = \mathrm{span}(\boldsymbol{G}_k'')$, $k = 1, 2, 3$.

Analogously, model (1) directly implies that,

$$\boldsymbol{\Delta} = \mathrm{E}(\boldsymbol{X} \circ \boldsymbol{Y} \circ \boldsymbol{Z}) = \mathrm{E}\{\mathrm{E}(\boldsymbol{X}\boldsymbol{Y}^\top \mid \boldsymbol{Z}) \circ \boldsymbol{Z}\} = \mathrm{E}\left\{\boldsymbol{\Gamma}_1 \boldsymbol{f}(\boldsymbol{\Gamma}_3^\top \boldsymbol{Z})\boldsymbol{\Gamma}_2^\top \circ \boldsymbol{Z}\right\}.$$

It is then straightforward to see that,

$$\boldsymbol{\Delta} = \boldsymbol{\Delta} \times_1 \boldsymbol{P}_{\boldsymbol{\Gamma}_1} = \boldsymbol{\Delta} \times_2 \boldsymbol{P}_{\boldsymbol{\Gamma}_2} = \boldsymbol{\Delta} \times_1 \boldsymbol{P}_{\boldsymbol{\Gamma}_1} \times_2 \boldsymbol{P}_{\boldsymbol{\Gamma}_2}.$$

We hence obtain that, the solution to the optimization problem (2),

$$(\boldsymbol{G}_1', \boldsymbol{G}_2', \boldsymbol{G}_3') = \arg \min_{\boldsymbol{G}_1, \boldsymbol{G}_2, \boldsymbol{G}_3} \|\boldsymbol{\Delta} - \boldsymbol{\Delta} \times_1 \boldsymbol{P}_{\boldsymbol{G}_1} \times_2 \boldsymbol{P}_{\boldsymbol{G}_2} \times_3 \boldsymbol{P}_{\boldsymbol{G}_3}\|_F^2,$$

must satisfy that $\mathrm{span}(\boldsymbol{\Gamma}_k) = \mathrm{span}(\boldsymbol{G}_k')$, $k = 1, 2$.

Next, we prove part (b) of the theorem. Since $\boldsymbol{Z}$ follows the normal distribution, by the multivariate Stein's Lemma, the (j, k)th fiber of $\boldsymbol{\Phi}$, $j = 1, \dots, p_1$, $k = 1, \dots, p_2$, satisfies that,

$$\begin{aligned} \boldsymbol{\Phi}_{jk\cdot} &= \boldsymbol{\Gamma}_3 \mathrm{E}\left\{ \frac{d}{d(\boldsymbol{\Gamma}_3^\top \boldsymbol{Z})} \mathrm{E}(X_j Y_k \mid \boldsymbol{\Gamma}_3^\top \boldsymbol{Z}) \right\} = \boldsymbol{\Gamma}_3 \{\mathrm{Cov}(\boldsymbol{\Gamma}_3^\top \boldsymbol{Z})\}^{-1} \mathrm{E}(X_j Y_k \boldsymbol{\Gamma}_3^\top \boldsymbol{Z}) \\ &= \boldsymbol{\Gamma}_3 (\boldsymbol{\Gamma}_3^\top \boldsymbol{\Sigma}_{\boldsymbol{Z}} \boldsymbol{\Gamma}_3)^{-1} \boldsymbol{\Gamma}_3^\top \mathrm{E}(X_j Y_k \boldsymbol{Z}), \end{aligned}$$

which implies that $\boldsymbol{\Phi}_{jk\cdot} = \boldsymbol{\Gamma}_3 (\boldsymbol{\Gamma}_3^\top \boldsymbol{\Sigma}_{\boldsymbol{Z}} \boldsymbol{\Gamma}_3)^{-1} \boldsymbol{\Gamma}_3^\top \boldsymbol{\Delta}_{jk\cdot}$. Stacking all the (j, k)-th fibers to form the GLA tensor, we have

$$\boldsymbol{\Phi} = \boldsymbol{\Delta} \times_3 \{\boldsymbol{\Gamma}_3 (\boldsymbol{\Gamma}_3^\top \boldsymbol{\Sigma}_{\boldsymbol{Z}} \boldsymbol{\Gamma}_3)^{-1} \boldsymbol{\Gamma}_3^\top\}. \tag{S6}$$

It follows from (S6) that

$$\mathrm{span}(\boldsymbol{\Gamma}_3) = \mathrm{span}(\boldsymbol{G}_3'') = \mathrm{span}\{\boldsymbol{\Gamma}_3 (\boldsymbol{\Gamma}_3^\top \boldsymbol{\Sigma}_{\boldsymbol{Z}} \boldsymbol{\Gamma}_3)^{-1} \boldsymbol{\Gamma}_3^\top \boldsymbol{G}'\}.$$

Finally, the equality between $\mathrm{span}(\boldsymbol{\Gamma}_3)$ and $\mathrm{span}(\boldsymbol{G}')$ follows from the last equality and the fact that all these subspaces have the same dimension r_3.

This completes the proof of Theorem 1. $\square$

S5.3 Proof of Theorem 2

We introduce two events. The first event provides an upper bound for the discrepancy between $\widetilde{\boldsymbol{\Delta}}$ and its population counterpart restricted on the true active sets. The second event implies that the inactive variables are not falsely selected for each mode k and iteration t.

(E1) $\max_{k=1,2,3} \|(\widetilde{\boldsymbol{\Delta}}_k - \boldsymbol{\Delta}_k)_{[I_k, I_{-k}]}\|_F \le \sqrt{\alpha s \log p / n}$, where $\alpha = 513(M + \sigma)^4$.

(E2) $\widehat{I}_k^{(t)} \subseteq I_k$, $k = 1, 2, 3$, in each iteration $t = 0, 1, \dots, t_{\max}$.

We prove Theorem 2 in two main steps. In the first step, we show that the two events (E1) and (E2) hold with a high probability under Assumption (A1). Then statement (c) follows directly. In the second step, we show statements (a) & (b) hold under events (E1), (E2) and Assumption (A2).

Step 1: In this step, we show the two events (E1) and (E2) hold with probability at least $1 - C \max\left[p^{-\gamma}, p^{-\left\{\sqrt{n(\gamma+1)/(2\log p)} - 1\right\}}\right]$, for some constants $C, \gamma > 0$.

Let $\widetilde{\Delta}_{ijk}$ and Δ_{ijk} denote the (i,j,k)-th elements of $\widetilde{\boldsymbol{\Delta}}$ and $\boldsymbol{\Delta}$. By Assumption (A1) and Lemma 5, with the constant $\alpha = 513(M+\sigma)^4$, we have,

$$\begin{aligned}
&\mathbb{P}\left(|\widetilde{\Delta}_{ijk} - \Delta_{ijk}| \geq \sqrt{\alpha \log p/n}\right) \\
&\leq 6\max\left\{\exp\left(-\frac{\alpha \log p}{512(M+\sigma)^4}\right), \exp\left(-\frac{\sqrt{n\alpha \log p}}{32(M+\sigma)^2}\right)\right\} \\
&= 6\max\{p^{-c}, p^{-\sqrt{nc/(2\log p)}}\},
\end{aligned} \tag{S7}$$

where the constant $c = 513/512$. For event (E1),

$$\|(\widetilde{\boldsymbol{\Delta}}_k - \boldsymbol{\Delta}_k)_{[I_k, I_{-k}]}\|_F \leq \sqrt{s}\|(\widetilde{\boldsymbol{\Delta}}_k - \boldsymbol{\Delta}_k)_{[I_k, I_{-k}]}\|_{\max},$$

for $k = 1,2,3$. Then,

$$\begin{aligned}
&\mathbb{P}\left\{\|(\widetilde{\boldsymbol{\Delta}}_k - \boldsymbol{\Delta}_k)_{[I_k, I_{-k}]}\|_F \geq \sqrt{\alpha s \log p/n}\right\} \leq \ \mathbb{P}\left\{\|(\widetilde{\boldsymbol{\Delta}}_k - \boldsymbol{\Delta}_k)_{[I_k, I_{-k}]}\|_{\max} \geq \sqrt{\alpha \log p/n}\right\} \\
&\leq s\,\mathbb{P}(|\widetilde{\Delta}_{ijk} - \Delta_{ijk}| \geq \sqrt{\alpha \log p/n}) \leq 6s\max\{p^{-c}, p^{-\sqrt{nc/(2\log p)}}\}.
\end{aligned}$$

Let $\kappa(c,p,n) = \max\left\{p^{-c}, p^{-\sqrt{nc/(2\log p)}}\right\}$. Then we have,

$$\mathbb{P}\left\{\|(\widetilde{\boldsymbol{\Delta}}_k - \boldsymbol{\Delta}_k)_{[I_k, I_{-k}]}\|_F \leq \sqrt{\alpha s \log p/n},\ k=1,2,3\right\} \geq 1 - 18s\kappa(c,p,n). \tag{S8}$$

For event (E2), we begin with the case when the iteration $t = 0$, i.e., $\left\{\widehat{I}_k^{(0)} \subseteq I_k, k = 1,2,3\right\}$. If $\widehat{I}_k^{(0)} \not\subseteq I_k$, then there exists an $i \in I_k^c$, such that $i \in \widehat{I}_k^{(0)}$, where I_k^c is the complement set of I_k. Therefore, by the definition of $\widehat{I}_k^{(0)}$, we have,

$$\begin{aligned}
&\mathbb{P}(\widehat{I}_k^{(0)} \not\subseteq I_k) \leq \sum_{i\in I_k^c} \mathbb{P}\left\{\|(\widetilde{\boldsymbol{\Delta}}_k)_{[i,:]}\|_\infty \geq \sqrt{\alpha \log p/n}\right\} \\
&\leq \sum_{i\in I_k^c} \mathbb{P}\left\{\|(\widetilde{\boldsymbol{\Delta}}_k - \boldsymbol{\Delta}_k)_{[i,:]}\|_\infty \geq \sqrt{\alpha \log p/n}\right\} \leq 6p_k p_{-k}\kappa(c,p,n) = 6p\kappa(c,p,n),
\end{aligned}$$

where the second inequality follows from the fact that $(\boldsymbol{\Delta}_k)_{[i,:]} = \mathbf{0}$ for $i \in I_k^c$. Henceforth, $\mathbb{P}(\widehat{I}_k^{(0)} \subseteq I_k) \geq 1 - 6p\kappa(c,p,n)$. It follows that,

$$\mathbb{P}(\widehat{I}_k^{(0)} \subseteq I_k, k=1,2,3) \geq 1 - 18p\kappa(c,p,n). \tag{S9}$$

Let $A^{(0)}$ denote the event $\{\widehat{I}_k^{(0)} \subseteq I_k, k = 1,2,3\}$. Next, we derive the probability $\mathbb{P}\big[A^{(0)} \cap \{\widehat{I}_1^{(1)} \subseteq I_1\}\big]$. We decompose this probability into two parts,

$$\mathbb{P}[A^{(0)} \cap \{\widehat{I}_1^{(1)} \subseteq I_1\}] = \mathbb{P}\{A^{(0)}\} - \mathbb{P}[A^{(0)} \cap \{\widehat{I}_1^{(1)} \not\subseteq I_1\}]. \tag{S10}$$

We derive the upper bound of $\mathbb{P}(A^{(0)} \cap \{\widehat{I}_1^{(1)} \not\subseteq I_1\})$ as,

$$
\begin{aligned}
&\mathbb{P}\left[A^{(0)} \cap \left\{\widehat{I}_1^{(1)} \not\subseteq I_1\right\}\right] \leq \sum_{i \in I_1^c} \mathbb{P}\left[A^{(0)} \cap \left\{i \in \widehat{I}_1^{(1)}\right\}\right] \\
=& \sum_{i \in I_1^c} \mathbb{P}\left[A^{(0)} \cap \left\{\|(\widetilde{\mathbf{\Delta}}_1 - \mathbf{\Delta}_1)_{[i,:]} \widehat{\mathbf{\Gamma}}_{-1}^{(1)}\|_2^2 \geq \alpha s_{-1} \log p/n\right\}\right] \\
\leq& \sum_{i \in I_1^c} \mathbb{P}\left\{\|(\widetilde{\mathbf{\Delta}}_1 - \mathbf{\Delta}_1)_{[i,I_{-1}]}\|_2^2 \geq \alpha s_{-1} \log p/n\right\} \qquad \text{(S11)} \\
\leq& \sum_{i \in I_1^c} \mathbb{P}\left\{\max_{j \in I_{-1}} |(\widetilde{\mathbf{\Delta}}_1 - \mathbf{\Delta}_1)_{[i,j]}| \geq \sqrt{\alpha \log p/n}\right\} \leq 6 p_1 s_{-1} \kappa(c,p,n) \leq 6 p \kappa(c,p,n), \qquad \text{(S12)}
\end{aligned}
$$

where (S11) follows from the fact that $\widehat{I}_{-1}^{(1)} \subseteq I_{-1}$ under event $A^{(0)}$.

Combining (S9), (S10), (S12), we obtain that,

$$
\mathbb{P}\left[A^{(0)} \cap \left\{\widehat{I}_1^{(1)} \subseteq I_1\right\}\right] \geq 1 - 24 p \kappa(c,p,n).
$$

Following similar arguments, we have that,

$$
\mathbb{P}\left[A^{(0)} \cap \left\{\widehat{I}_k^{(1)} \subseteq I_k, k = 1,2,3\right\}\right] \geq 1 - 36 p \kappa(c,p,n).
$$

After $t_{\max}$ iterations, we obtain that, for some constant $C > 0$,

$$
\mathbb{P}\left\{\widehat{I}_k^{(t)} \subseteq I_k,\ k = 1,2,3,\ t = 0,1,\ldots,t_{\max}\right\} \geq 1 - 18(t_{\max}+1) p \kappa(c,p,n). \qquad \text{(S13)}
$$

Combining (S8) and (S13), we obtain that,

$$
\mathbb{P}\left(E_1 \cap E_2\right) \geq 1 - C t_{\max} p \kappa(c,p,n) = 1 - C \max\left[p^{-\gamma}, p^{-\left\{\sqrt{n(\gamma+1)/(2\log p)} - 1\right\}}\right], \qquad \text{(S14)}
$$

where we have used the fact that $t_{\max}$ is a constant and $\gamma = c - 1 = 1/512$. This completes Step 1. Moreover, statement (c) follows directly.

Step 2: In this step, we establish the non-asymptotic error bounds in statements (a) & (b) of Theorem 2 under events (E1), (E2), and Assumption (A2).

For $t = 0,1,\ldots,t_{\max}$ and $k = 1,2,3$, define

$$
\begin{aligned}
E_k^{(t)} &= \|(\widehat{\mathbf{\Gamma}}_{k\perp}^{(t)})^\top \mathbf{\Delta}_k\|_F,\ k = 1,2,3, \quad & E^{(t)} &= \max_{k=1,2,3} E_k^{(t)}, \\
F_j^{(t)} &= \|(\widehat{\mathbf{\Gamma}}_{j\perp}^{(t)})^\top \mathbf{\Gamma}_j\|_F,\ j = 1,2, & F_3^{(t)} &= \|(\widehat{\mathbf{\Gamma}}_{3\perp}^{(t)})^\top \boldsymbol{G}_3'\|_F,
\end{aligned} \qquad \text{(S15)}
$$

The quantity $F_k^{(t)}$ measures the distance between $\mathrm{span}(\mathbf{\Gamma}_k)$ and $\mathrm{span}(\widehat{\mathbf{\Gamma}}_k^{(t)})$. The quantity $E_k^{(t)}$ is the scaled upper bound of $F_k^{(t)}$. This is because, $\mathrm{span}(\mathbf{\Gamma}_k)$ is actually the left singular space of $\mathbf{\Delta}_k$, and thus for each $k = 1, 2, 3$, by Lemma 1 and Lemma 2, we have,

$$E_k^{(t)} \geq \lambda_k \| \sin \Theta(\widehat{\mathbf{\Gamma}}_k^{(t)}, \mathbf{\Gamma}_k) \|_F = \lambda_k F_k^{(t)}. \tag{S16}$$

The non-asymptotic error bounds of Theorem 2 rely on the upper bound of $E^{(t)}$. Next, we derive such a bound in four steps. In Step 2.1, we derive the upper bound of the initialization $E^{(0)}$. In Step 2.2, we obtain the inequality that connects the consecutive $E_k^{(t)}$ and $F_k^{(t)}$. In Step 2.3, we derive the upper bound of $E^{(t)}$. In Step 2.4, we put the three results together to complete Step 2.

Step 2.1: We first show that, under events (E1) and (E2),

$$E_k^{(0)} = \left\| (\widehat{\mathbf{\Gamma}}_{k\perp}^{(0)})^\top \mathbf{\Delta}_k \right\|_F \leq 8 \sqrt{\frac{\alpha s \log p}{n}}. \tag{S17}$$

Toward that goal, for each k, we decompose $E_k^{(0)}$ into two parts,

$$E_k^{(0)} \leq \left\| (\widehat{\mathbf{\Gamma}}_{k\perp}^{(0)})^\top \boldsymbol{D}_{\widehat{I}_k^{(0)}} \mathbf{\Delta}_k \boldsymbol{D}_{\widehat{I}_{-k}^{(0)}} \right\|_F + \left\| (\widehat{\mathbf{\Gamma}}_{k\perp}^{(0)})^\top \left\{ \mathbf{\Delta}_k - \boldsymbol{D}_{\widehat{I}_k^{(0)}} \mathbf{\Delta}_k \boldsymbol{D}_{\widehat{I}_{-k}^{(0)}} \right\} \right\|_F \equiv B_3 + B_4.$$

For B_3, according to Step 1 of Algorithm 1, $\widehat{\mathbf{\Gamma}}_k^{(0)}$ consists of the top-r_k left singular vectors of $\boldsymbol{D}_{\widehat{I}_k^{(0)}} \widetilde{\mathbf{\Delta}}_k \boldsymbol{D}_{\widehat{I}_{-k}^{(0)}}$. By Lemma 1, we have that,

$$B_3 \leq 2 \left\| (\widetilde{\mathbf{\Delta}}_k - \mathbf{\Delta}_k)_{[\widehat{I}_k^{(0)}, \widehat{I}_{-k}^{(0)}]} \right\|_F \leq 2 \left\| (\widetilde{\mathbf{\Delta}}_k - \mathbf{\Delta}_k)_{[I_k, I_{-k}]} \right\|_F,$$

where the last inequality comes from the fact that $\widehat{I}_k^{(0)} \subseteq I_k$ under event (E2). Moreover, under event (E1), we have that,

$$B_3 \leq 2 \max_{k=1,2,3} \| (\widetilde{\mathbf{\Delta}}_k - \mathbf{\Delta}_k)_{[I_k, I_{-k}]} \|_F \leq 2 \sqrt{\alpha s \log p / n}.$$

For B_4, we have that,

$$\begin{aligned} B_4 \leq & \left\| \mathbf{\Delta}_k - \boldsymbol{D}_{\widehat{I}_k^{(0)}} \mathbf{\Delta}_k \boldsymbol{D}_{\widehat{I}_{-k}^{(0)}} \right\|_F \leq \sqrt{\sum_{k=1}^{3} \left\| (\mathbf{\Delta}_k)_{[I_k \setminus \widehat{I}_k^{(0)}, I_{-k}]} \right\|_F^2} \leq \sum_{k=1}^{3} \left\| (\mathbf{\Delta}_k)_{[I_k \setminus \widehat{I}_k^{(0)}, I_{-k}]} \right\|_F \\ \leq & \sum_{k=1}^{3} \left\| (\widetilde{\mathbf{\Delta}}_k)_{[I_k \setminus \widehat{I}_k^{(0)}, I_{-k}]} \right\|_F + \sum_{k=1}^{3} \left\| (\widetilde{\mathbf{\Delta}}_k - \mathbf{\Delta}_k)_{[I_k \setminus \widehat{I}_k^{(0)}, I_{-k}]} \right\|_F \equiv B_{41} + B_{42}. \end{aligned}$$

For B_{41}, by the definition of $\widehat{I}_k^{(0)}$, for any $i \notin \widehat{I}_k^{(0)}$, $\|(\widetilde{\mathbf{\Delta}}_k)_{[i,:]}\|_\infty < \eta_k = \sqrt{\alpha \log p / n}$. It then follows that $\|(\widetilde{\mathbf{\Delta}}_k)_{[i, I_{-k}]}\|_2 < \sqrt{s_{-k} \alpha \log p / n}$. Therefore,

$$B_{41} \leq \sum_{k=1}^{3} \sqrt{s_k s_{-k} \alpha \log p / n} = 3 \sqrt{\alpha s \log p / n}.$$

For B_{42}, under event (E1), we have,

$$\left\|(\widetilde{\mathbf{\Delta}}_k - \mathbf{\Delta}_k)_{[I_k \setminus \widehat{I}_k^{(0)}, I_{-k}]}\right\|_F \leq \left\|(\widetilde{\mathbf{\Delta}}_k - \mathbf{\Delta}_k)_{[I_k, I_{-k}]}\right\|_F \leq \sqrt{\alpha s \log p/n}.$$

Therefore, $B_{42} \leq 3\sqrt{\alpha s \log p/n}$.

Combining the bounds for B_{41} and B_{42}, we obtain that $B_4 \leq 6\sqrt{\alpha s \log p/n}$.

Combining the bounds for B_3 and B_4 completes Step 2.1.

Step 2.2: We next establish the following inequalities, under events (E1) and (E2), and for each iteration $t = 1, 2, \ldots, t_{\max}$, that

$$\begin{aligned}
\left\{1 - \sqrt{(F_2^{(t-1)})^2 + (F_3^{(t-1)})^2}\right\} E_1^{(t)} &\leq 4\sqrt{\alpha s \log p/n}, \\
\left\{1 - \sqrt{(F_1^{(t)})^2 + (F_3^{(t-1)})^2}\right\} E_2^{(t)} &\leq 4\sqrt{\alpha s \log p/n}, \\
\left\{1 - \sqrt{(F_1^{(t)})^2 + (F_2^{(t)})^2}\right\} E_3^{(t)} &\leq 4\sqrt{\alpha s \log p/n}.
\end{aligned} \tag{S18}$$

We only prove the first inequality, while the other two inequalities can be derived similarly.

Toward that end, we bound $E_1^{(t)}$ by three terms as follows,

$$\begin{aligned}
E_1^{(t)} = {} & \left\|(\widehat{\mathbf{\Gamma}}_{1\perp}^{(t)})^\top \mathbf{\Delta}_1 \left(\boldsymbol{P}_{\widehat{\mathbf{\Gamma}}_{-1}^{(t)}} + \boldsymbol{P}_{\widehat{\mathbf{\Gamma}}_{-1\perp}^{(t)}}\right)\right\|_F \leq \left\|(\widehat{\mathbf{\Gamma}}_{1\perp}^{(t)})^\top \mathbf{\Delta}_1 \widehat{\mathbf{\Gamma}}_{-1}^{(t)}\right\|_F + \left\|(\widehat{\mathbf{\Gamma}}_{1\perp}^{(t)})^\top \mathbf{\Delta}_1 \widehat{\mathbf{\Gamma}}_{-1\perp}^{(t)}\right\|_F \\
\leq {} & \left\|(\widehat{\mathbf{\Gamma}}_{1\perp}^{(t)})^\top \boldsymbol{D}_{\widehat{I}_1^{(t)}} \mathbf{\Delta}_1 \widehat{\mathbf{\Gamma}}_{-1}^{(t)}\right\|_F + \left\|(\widehat{\mathbf{\Gamma}}_{1\perp}^{(t)})^\top \boldsymbol{D}_{I_1 \setminus \widehat{I}_1^{(t)}} \mathbf{\Delta}_1 \widehat{\mathbf{\Gamma}}_{-1}^{(t)}\right\|_F + \left\|(\widehat{\mathbf{\Gamma}}_{1\perp}^{(t)})^\top \mathbf{\Delta}_1 \widehat{\mathbf{\Gamma}}_{-1\perp}^{(t)}\right\|_F \\
\equiv {} & B_5 + B_6 + B_7.
\end{aligned}$$

For B_5, according to Step 2b of Algorithm 1, $\widehat{\mathbf{\Gamma}}_1^{(t)}$ consists of the top-r_1 left singular vectors of $\boldsymbol{D}_{\widehat{I}_1^{(t)}} \widetilde{\mathbf{\Delta}}_1 \widehat{\mathbf{\Gamma}}_{-1}^{(t)}$. By Lemma 1, we have,

$$B_5 \leq 2\left\|\boldsymbol{D}_{\widehat{I}_1^{(t)}} (\widetilde{\mathbf{\Delta}}_1 - \mathbf{\Delta}_1) \widehat{\mathbf{\Gamma}}_{-1}^{(t)}\right\|_F.$$

Under events (E1) and (E2), we further bound B_5 as

$$B_5 \leq 2\left\|(\widetilde{\mathbf{\Delta}}_1 - \mathbf{\Delta}_1)_{[I_1, I_{-1}]}\right\|_F \leq 2\sqrt{\alpha s \log p/n}.$$

For B_6, we have that,

$$B_6 \leq \left\|(\widehat{\mathbf{\Gamma}}_{1\perp}^{(t)})^\top \boldsymbol{D}_{I_1 \setminus \widehat{I}_1^{(t)}} \widetilde{\mathbf{\Delta}}_1 \widehat{\mathbf{\Gamma}}_{-1}^{(t)}\right\|_F + \left\|(\widehat{\mathbf{\Gamma}}_{1\perp}^{(t)})^\top \boldsymbol{D}_{I_1 \setminus \widehat{I}_1^{(t)}} (\widetilde{\mathbf{\Delta}}_1 - \mathbf{\Delta}_1) \widehat{\mathbf{\Gamma}}_{-1}^{(t)}\right\|_F \equiv B_{61} + B_{62}.$$

For B_{61}, by definition of $\widehat{I}_1^{(t)}$ in Step 2a of Algorithm 1, for any $i \notin \widehat{I}_1^{(t)}$, $\|(\widetilde{\mathbf{\Delta}}_1)_{[i,:]} \widehat{\mathbf{\Gamma}}_{-1}^{(t)}\|_2^2 \leq \widetilde{\eta}_1$. Recall that $\widetilde{\eta}_k = \alpha s_{-k} \log p/n$, $k = 1, 2, 3$. Therefore,

$$B_{61} \leq \left\|\boldsymbol{D}_{I_1 \setminus \widehat{I}_1^{(t)}} \widetilde{\mathbf{\Delta}}_1 \widehat{\mathbf{\Gamma}}_{-1}^{(t)}\right\|_F \leq \sqrt{s_1 \widetilde{\eta}_1} = \sqrt{\alpha s \log p/n}.$$

For B_{62}, under event (E1), we have,

$$B_{62} \leq \|(\widetilde{\boldsymbol{\Delta}}_1 - \boldsymbol{\Delta}_1)_{[I_1, I_{-1}]}\|_F \leq \sqrt{\alpha s \log p / n}.$$

Combining the bounds for B_{61} and B_{62}, we obtain the upper bound of B_6 as,

$$B_6 \leq 2\sqrt{\alpha s \log p / n}.$$

For B_7, by the Tucker decomposition structure of $\boldsymbol{\Delta}$, we have $\boldsymbol{\Delta}_1 = \boldsymbol{\Delta}_1 \boldsymbol{P}_{\boldsymbol{\Gamma}_{-1}}$. Then,

$$B_7 = \left\| (\widehat{\boldsymbol{\Gamma}}_{1\perp}^{(t)})^\top \boldsymbol{\Delta}_1 \boldsymbol{\Gamma}_{-1} \boldsymbol{\Gamma}_{-1}^\top \widehat{\boldsymbol{\Gamma}}_{-1\perp}^{(t)} \right\|_F \leq \left\| (\widehat{\boldsymbol{\Gamma}}_{1\perp}^{(t)})^\top \boldsymbol{\Delta}_1 \right\|_F \|\boldsymbol{\Gamma}_{-1}^\top \widehat{\boldsymbol{\Gamma}}_{-1\perp}^{(t)}\| \leq E_1^{(t)} \left\| \sin \Theta(\boldsymbol{\Gamma}_{-1}, \widehat{\boldsymbol{\Gamma}}_{-1}^{(t)}) \right\|.$$

Recall that $\sin \Theta(\boldsymbol{\Gamma}_{-1}, \widehat{\boldsymbol{\Gamma}}_{-1}^{(t)}) = \operatorname{diag}(\sin \theta_1, \ldots, \sin \theta_{r_{-1}})$, where $\cos \theta_i = \sigma_i(\boldsymbol{\Gamma}_{-1}^\top \widehat{\boldsymbol{\Gamma}}_{-1}^{(t)})$ is the ith largest singular value of the matrix $\boldsymbol{\Gamma}_{-1}^\top \widehat{\boldsymbol{\Gamma}}_{-1}^{(t)}$. Henceforth,

$$\left\| \sin \Theta(\boldsymbol{\Gamma}_{-1}, \widehat{\boldsymbol{\Gamma}}_{-1}^{(t)}) \right\| = \sqrt{1 - \sigma_{\min}^2(\boldsymbol{\Gamma}_{-1}^\top \widehat{\boldsymbol{\Gamma}}_{-1}^{(t)})},$$

where $\sigma_{\min}(\boldsymbol{\Gamma}_{-1}^\top \widehat{\boldsymbol{\Gamma}}_{-1}^{(t)})$ denotes the smallest singular value of $\boldsymbol{\Gamma}_{-1}^\top \widehat{\boldsymbol{\Gamma}}_{-1}^{(t)}$. Using the fact that $\sigma_{\min}(\boldsymbol{A} \otimes \widetilde{\boldsymbol{A}}) \geq \sigma_{\min}(\boldsymbol{A}) \cdot \sigma_{\min}(\widetilde{\boldsymbol{A}})$ for any dimension compatible matrices $\boldsymbol{A}$, $\widetilde{\boldsymbol{A}}$, we have that,

$$\begin{aligned} \sigma_{\min}(\boldsymbol{\Gamma}_{-1}^\top \widehat{\boldsymbol{\Gamma}}_{-1}^{(t)}) &= \sigma_{\min}\{(\boldsymbol{\Gamma}_2^\top \widehat{\boldsymbol{\Gamma}}_2^{(t-1)}) \otimes (\boldsymbol{\Gamma}_3^\top \widehat{\boldsymbol{\Gamma}}_3^{(t-1)})\} \geq \sigma_{\min}(\boldsymbol{\Gamma}_2^\top \widehat{\boldsymbol{\Gamma}}_2^{(t-1)}) \cdot \sigma_{\min}(\boldsymbol{\Gamma}_3^\top \widehat{\boldsymbol{\Gamma}}_3^{(t-1)}) \\ &= \sqrt{\{1 - \|\sin \Theta(\boldsymbol{\Gamma}_2, \widehat{\boldsymbol{\Gamma}}_2^{(t-1)})\|^2\}\{1 - \|\sin \Theta(\boldsymbol{\Gamma}_3, \widehat{\boldsymbol{\Gamma}}_3^{(t-1)})\|^2\}}. \end{aligned} \tag{S19}$$

Therefore, we have that,

$$\begin{aligned} &\|\sin \Theta(\boldsymbol{\Gamma}_{-1}, \widehat{\boldsymbol{\Gamma}}_{-1}^{(t)})\| \leq \sqrt{1 - \{1 - \|\sin \Theta(\boldsymbol{\Gamma}_2, \widehat{\boldsymbol{\Gamma}}_2^{(t-1)})\|^2\}\{1 - \|\sin \Theta(\boldsymbol{\Gamma}_3, \widehat{\boldsymbol{\Gamma}}_3^{(t-1)})\|^2\}} \\ &\leq \sqrt{\|\sin \Theta(\boldsymbol{\Gamma}_2, \widehat{\boldsymbol{\Gamma}}_2^{(t-1)})\|^2 + \|\sin \Theta(\boldsymbol{\Gamma}_3, \widehat{\boldsymbol{\Gamma}}_3^{(t-1)})\|^2} \leq \sqrt{(F_2^{(t-1)})^2 + (F_3^{(t-1)})^2}. \end{aligned} \tag{S20}$$

Consequently, we have that,

$$B_7 \leq E_1^{(t)} \sqrt{(F_2^{(t-1)})^2 + (F_3^{(t-1)})^2}.$$

Combining the bounds for B_5, B_6, B_7, we obtain the inequality connecting $E_1^{(t)}$ with $F_2^{(t-1)}$ and $F_3^{(t-1)}$ in (S18), we obtain that,

$$E_1^{(t)} \leq 4\sqrt{\alpha s \log p / n} + E_1^{(t)} \sqrt{(F_2^{(t-1)})^2 + (F_3^{(t-1)})^2}.$$

This completes Step 2.2.

Step 2.3: We next establish the upper bound for $E^{(t)}$ under Assumption (A2) and events (E1)(E2), for $t = 0, 1, \ldots, t_{\max}$, as

$$E^{(t)} \leq \left(7 + \frac{1}{2^t}\right) \sqrt{\frac{\alpha s \log p}{n}}. \tag{S21}$$

We prove (S21) by induction. For $t = 0$, by (S17), we have $E^{(0)}$ satisfies (S21). Suppose (S21) holds for $E^{(t')}$, $t' = 0, 1, \ldots, t-1$, we next show that (S21) holds for $E^{(t)}$.

In Assumption (A2), we set the constant $C_1 = 56\sqrt{2\alpha}/3$ in the lower bound of λ, such that $\lambda \geq (56\sqrt{2}/3)\sqrt{\alpha s \log p/n}$. Then, since (S21) holds for $E^{(t-1)}$, $F_2^{(t-1)} \leq E_2^{(t-1)}/\lambda_2$, and $F_3^{(t-1)} \leq E_3^{(t-1)}/\lambda_3$ by (S16), we have that,

$$F_k^{(t-1)} \leq \frac{3}{56\sqrt{2}} \left(7 + \frac{1}{2^{t-1}}\right), \quad k = 2, 3. \tag{S22}$$

Moreover, we have that,

$$\left\{1 - \sqrt{(F_2^{(t-1)})^2 + (F_3^{(t-1)})^2}\right\} > \frac{4}{7}.$$

Therefore,

$$E_1^{(t)} \leq 4\left\{1 - \sqrt{(F_2^{(t-1)})^2 + (F_3^{(t-1)})^2}\right\}^{-1} \sqrt{\frac{\alpha s \log p}{n}} \leq \left(7 + \frac{1}{2^t}\right) \sqrt{\frac{\alpha s \log p}{n}}.$$

Using similar arguments, we obtain the same upper bounds for $E_2^{(t)}$ and $E_3^{(t)}$. Putting them together gives the upper bound for $E^{(t)}$ as in (S21). Then, by induction, we have that (S21) holds for all $t = 0, 1, \ldots, t_{\max}$, which completes Step 2.3.

Then, after $t_{\max}$ iterations, we have that

$$E^{(t_{\max})} = \max_{k=1,2,3} E_k^{(t_{\max})} \leq \left(7 + \frac{1}{2^{t_{\max}}}\right) \sqrt{\frac{\alpha s \log p}{n}}. \tag{S23}$$

Let $\widehat{\boldsymbol{\Gamma}}_k = \widehat{\boldsymbol{\Gamma}}_k^{(t_{\max})}$, $k = 1, 2, 3$. Then by (S16) and Assumption (A2), we have that,

$$\|\sin\Theta(\widehat{\boldsymbol{\Gamma}}_k, \boldsymbol{\Gamma}_k)\|_F \leq \frac{E_k^{(t_{\max})}}{\lambda_k} \leq \left(7 + \frac{1}{2^{t_{\max}}}\right) \frac{\sqrt{\alpha}}{\lambda} \sqrt{\frac{s \log p}{n}} \leq \left(7 + \frac{1}{2^{t_{\max}}}\right) \frac{\sqrt{\alpha}}{C_2} \sqrt{\frac{s \log p}{n}},$$

for $k = 1, 2, 3$. By Lemma 2, we obtain that,

$$\|\boldsymbol{P}_{\widehat{\boldsymbol{\Gamma}}_k} - \boldsymbol{P}_{\boldsymbol{\Gamma}_k}\|_F = \sqrt{2}\|\sin\Theta(\widehat{\boldsymbol{\Gamma}}_k, \boldsymbol{\Gamma}_k)\|_F \leq (c_3 + c_4 2^{-t_{\max}})\sqrt{\frac{s \log p}{n}}, \quad k = 1, 2, 3,$$

where $c_3 = 7\sqrt{1026}(M + \sigma)^2/C_2$, and $c_4 = \sqrt{1026}(M + \sigma)^2/C_2$. Thus, statement (b) is proved.

Step 2.4: Finally, we put together the results in Steps 2.1 to 2.3 to derive the non-asymptotic error bound for $\widehat{\boldsymbol{\Delta}}$ in statement (a).

We decompose the upper bound of the estimation error for $\widehat{\boldsymbol{\Delta}} = \widetilde{\boldsymbol{\Delta}} \times_1 \boldsymbol{P}_{\widehat{\boldsymbol{\Gamma}}_1} \times_2 \boldsymbol{P}_{\widehat{\boldsymbol{\Gamma}}_2} \times_3 \boldsymbol{P}_{\widehat{\boldsymbol{\Gamma}}_3}$ as,

$$\begin{aligned}\|\widehat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}\|_F &\leq \|\boldsymbol{\Delta} \times_1 \boldsymbol{P}_{\widehat{\boldsymbol{\Gamma}}_1} \times_2 \boldsymbol{P}_{\widehat{\boldsymbol{\Gamma}}_2} \times_3 \boldsymbol{P}_{\widehat{\boldsymbol{\Gamma}}_3} - \boldsymbol{\Delta}\|_F + \|(\widetilde{\boldsymbol{\Delta}} - \boldsymbol{\Delta}) \times_1 \boldsymbol{P}_{\widehat{\boldsymbol{\Gamma}}_1} \times_2 \boldsymbol{P}_{\widehat{\boldsymbol{\Gamma}}_2} \times_3 \boldsymbol{P}_{\widehat{\boldsymbol{\Gamma}}_3}\|_F \\ &\equiv B_8 + B_9. \end{aligned} \tag{S24}$$

For B_8, we have that,

$$\begin{aligned} B_8 &= \|(\boldsymbol{\Delta} - \boldsymbol{\Delta} \times_1 \boldsymbol{P}_{\widehat{\boldsymbol{\Gamma}}_1}) + (\boldsymbol{\Delta} \times_1 \boldsymbol{P}_{\widehat{\boldsymbol{\Gamma}}_1} - \boldsymbol{\Delta} \times_1 \boldsymbol{P}_{\widehat{\boldsymbol{\Gamma}}_1} \times_2 \boldsymbol{P}_{\widehat{\boldsymbol{\Gamma}}_2}) \\ &\quad + (\boldsymbol{\Delta} \times_1 \boldsymbol{P}_{\widehat{\boldsymbol{\Gamma}}_1} \times_2 \boldsymbol{P}_{\widehat{\boldsymbol{\Gamma}}_2} - \boldsymbol{\Delta} \times_1 \boldsymbol{P}_{\widehat{\boldsymbol{\Gamma}}_1} \times_2 \boldsymbol{P}_{\widehat{\boldsymbol{\Gamma}}_2} \times_3 \boldsymbol{P}_{\widehat{\boldsymbol{\Gamma}}_3})\|_F \\ &= \|\boldsymbol{\Delta} \times_1 \boldsymbol{P}_{\widehat{\boldsymbol{\Gamma}}_{1\perp}} + \boldsymbol{\Delta} \times_1 \boldsymbol{P}_{\widehat{\boldsymbol{\Gamma}}_1} \times_2 \boldsymbol{P}_{\widehat{\boldsymbol{\Gamma}}_{2\perp}} + \boldsymbol{\Delta} \times_1 \boldsymbol{P}_{\widehat{\boldsymbol{\Gamma}}_1} \times_2 \boldsymbol{P}_{\widehat{\boldsymbol{\Gamma}}_2} \times_3 \boldsymbol{P}_{\widehat{\boldsymbol{\Gamma}}_{3\perp}}\|_F \\ &\leq \sum_{k=1}^{3} \|\boldsymbol{\Delta} \times_k (\widehat{\boldsymbol{\Gamma}}_{k\perp})^\top\|_F = 3E^{(t_{\max})} \leq 3(7 + 2^{-t_{\max}})\sqrt{\alpha s \log p/n}, \end{aligned} \tag{S25}$$

where the last inequality follows from (S23) in Step 2.3.

For B_9, under events (E1)(E2), we have that,

$$B_9 \leq \|(\widetilde{\boldsymbol{\Delta}} - \boldsymbol{\Delta}) \times_1 \widehat{\boldsymbol{\Gamma}}_1^\top \times_2 \widehat{\boldsymbol{\Gamma}}_2^\top \times_3 \widehat{\boldsymbol{\Gamma}}_3^\top\|_F \leq \|(\widetilde{\boldsymbol{\Delta}}_1 - \boldsymbol{\Delta}_1)_{[I_1, I_{-1}]}\|_F \leq \sqrt{\alpha s \log p/n}. \tag{S26}$$

Combining the bounds for B_8 and B_9, we obtain that,

$$\|\widehat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}\|_F \leq (c_1 + c_2 2^{-t_{\max}})\sqrt{\frac{s \log p}{n}},$$

where $c_1 = 22\sqrt{513}(M + \sigma)^2$ and $c_2 = 3\sqrt{513}(M + \sigma)^2$. Thus, statement (a) is proved.

This completes the proof of Theorem 2. □

S5.4 Proof of Corollary 1

As $n \gg 2(\gamma + 1) \log p$, all three statements in Theorem 2 hold with probability at least $1 - Cp^{-\gamma}$ for some constant $\gamma > 0$. The probability $1 - Cp^{-\gamma}$ approaches to one as n, p diverges. In addition, since $s \log p = o(n)$, as n, p, s diverge, all three statements in Corollary 1 hold with probability approaching to one. □

S5.5 Proof of Theorem 3

We introduce the following event.

(E3) $I_k^{(t)} \subseteq \widehat{I}_k^{(t)}$, $k = 1, 2, 3$ and $t = 0, 1, \ldots, t_{\max}$, where $I_k^{(t)}$ are the oracle active set that contains the most important variables, defined as $I_k^{(0)} := \{i : \|(\boldsymbol{\Delta}_k)_{[i,:]}\|_\infty > 2\sqrt{\alpha \log p/n}\}$ and $I_k^{(t)} = \left\{i : \|(\boldsymbol{\Delta}_k)_{[i,:]}\|_2^2 > 4\alpha s_{-k} \log p/n\right\}$ for $t \geq 1$.

First, we show both events (E1) to (E3) hold under Assumptions (A1) and (A2), with probability at least $1 - C' \max\left[p^{-\gamma}, p^{-\left\{\sqrt{n(\gamma+1)/(2\log p)}-1\right\}}\right]$, for some constants $C', \gamma > 0$.

We derive the probability of $\left\{I_k^{(0)} \subseteq \widehat{I}_k^{(0)}, k = 1, 2, 3\right\}$. Note that,

$$\mathbb{P}\left\{I_k^{(0)} \subseteq \widehat{I}_k^{(0)}\right\} = \mathbb{P}\left[\bigcap_{i \in I_k^{(0)}} \left\{i \in \widehat{I}_k^{(0)}\right\}\right].$$

For any $i \in I_k^{(0)}$, we have $\|(\mathbf{\Delta}_k)_{[i,:]}\|_\infty \geq 2\sqrt{\alpha \log p/n}$. If $\|(\widetilde{\mathbf{\Delta}}_k - \mathbf{\Delta}_k)_{[i,:]}\|_\infty \leq \sqrt{\alpha \log p/n}$, then

$$\|(\widetilde{\mathbf{\Delta}}_k)_{[i,:]}\|_\infty \geq \|(\mathbf{\Delta}_k)_{[i,:]}\|_\infty - \|(\widetilde{\mathbf{\Delta}}_k - \mathbf{\Delta}_k)_{[i,:]}\|_\infty \geq \sqrt{\alpha \log p/n}.$$

By the definition of $\widehat{I}_k^{(0)}$, it follows that $i \in \widehat{I}_k^{(0)}$. Recall that $\kappa(c, p, n) = \max\left\{p^{-c}, p^{-\sqrt{nc/(2\log p)}}\right\}$. Then, for any $i \in I_k^{(0)}$, according to (S7), we have,

$$\mathbb{P}\left\{i \in \widehat{I}_k^{(0)}\right\} \geq \mathbb{P}\left\{\|(\widetilde{\mathbf{\Delta}}_k - \mathbf{\Delta}_k)_{[i,:]}\|_\infty \leq \sqrt{\alpha \log p/n}\right\} \geq 1 - 6p_{-k}\kappa(c, p, n).$$

Moreover, by definition, we have $I_k^{(0)} \subseteq I_k$ for $k = 1, 2, 3$. Then $|I_k^{(0)}| \leq s_k$. Therefore,

$$\mathbb{P}\left\{I_k^{(0)} \subseteq \widehat{I}_k^{(0)}\right\} \geq 1 - 6s_k p_{-k}\kappa(c, p, n) \geq 1 - 6p\kappa(c, p, n).$$

It follows that,

$$\mathbb{P}\left\{I_k^{(0)} \subseteq \widehat{I}_k^{(0)},\ k = 1, 2, 3\right\} \geq 1 - 18p\kappa(c, p, n). \tag{S27}$$

Next, we derive the probability of $\left\{I_k^{(t)} \subseteq \widehat{I}_k^{(t)}, k = 1, 2, 3, t = 1, \ldots, t_{\max}\right\}$, which relies on events (E1) and (E2). We aim to prove the lower bound of

$$\mathbb{P}\left(E_1 \cap E_2 \cap \left[\bigcap_{t=1}^{t_{\max}} \bigcap_{k=1}^{3} \left\{I_k^{(t)} \in \widehat{I}_k^{(t)}\right\}\right]\right) = \mathbb{P}\left(\bigcap_{t=1}^{t_{\max}} \bigcap_{k=1}^{3} \left[E_1 \cap E_2 \cap \left\{I_k^{(t)} \in \widehat{I}_k^{(t)}\right\}\right]\right) \tag{S28}$$

For $k = 1, 2, 3$ and $t = 1, \ldots, t_{\max}$, we have that,

$$E_1 \cap E_2 \cap \left\{I_k^{(t)} \in \widehat{I}_k^{(t)}\right\} = \bigcap_{i \in I_k^{(t)}} \left[E_1 \cap E_2 \cap \left\{i \in \widehat{I}_k^{(t)}\right\}\right]. \tag{S29}$$

We next prove that, for any $i \in I_k^{(t)}$, the following subset relation holds,

$$\left[E_1 \cap E_2 \cap \left\{\|(\widetilde{\mathbf{\Delta}}_k - \mathbf{\Delta}_k)_{[i,I_{-k}]}\|_2^2 \leq \frac{\alpha s_{-k} \log p}{n}\right\}\right] \subseteq \left[E_1 \cap E_2 \cap \left\{i \in \widehat{I}_k^{(t)}\right\}\right]. \tag{S30}$$

Note that (S30) is used to bound (S29) from below. It suffices to show that,

$$\left[E_1 \cap E_2 \cap \left\{\|(\widetilde{\boldsymbol{\Delta}}_k - \boldsymbol{\Delta}_k)_{[i,I_{-k}]}\|_2^2 \le \frac{\alpha s_{-k}\log p}{n}\right\}\right] \subseteq \left\{i \in \widehat{I}_k^{(t)}\right\}, \quad \text{for any } i \in I_k^{(t)}.$$

By the definition of $\widehat{I}_k^{(t)}$ in Step 2a of Algorithm 1, for $t \ge 1$, $i \in \widehat{I}_k^{(t)}$ if $\|(\widetilde{\boldsymbol{\Delta}}_k)_{[i,:]}\widehat{\boldsymbol{\Gamma}}_{-k}^{(t)}\|_2^2 > \alpha s_{-k}\log p/n$. For $i \in I_k^{(t)}$, we decompose the lower bound of $\|(\widetilde{\boldsymbol{\Delta}}_k)_{[i,:]}\widehat{\boldsymbol{\Gamma}}_{-k}^{(t)}\|_2^2$ into two parts,

$$\|(\widetilde{\boldsymbol{\Delta}}_k)_{[i,:]}\widehat{\boldsymbol{\Gamma}}_{-k}^{(t)}\|_2^2 \ge \|(\boldsymbol{\Delta}_k)_{[i,:]}\widehat{\boldsymbol{\Gamma}}_{-k}^{(t)}\|_2^2 - \|(\widetilde{\boldsymbol{\Delta}}_k - \boldsymbol{\Delta}_k)_{[i,:]}\widehat{\boldsymbol{\Gamma}}_{-k}^{(t)}\|_2^2 \equiv B_{10} - B_{11}.$$

For B_{10}, since $\boldsymbol{\Delta}_k = \boldsymbol{\Delta}_k \boldsymbol{P}_{\boldsymbol{\Gamma}_{-k}}$,

$$B_{10} = \|(\boldsymbol{\Delta}_k)_{[i,:]}\boldsymbol{\Gamma}_{-k}\boldsymbol{\Gamma}_{-k}^\top\widehat{\boldsymbol{\Gamma}}_{-k}^{(t)}\|_2^2 \ge \sigma_{\min}^2(\boldsymbol{\Gamma}_{-k}^\top\widehat{\boldsymbol{\Gamma}}_{-k}^{(t)})\|(\boldsymbol{\Delta}_k)_{[i,:]}\boldsymbol{\Gamma}_{-k}\|_2^2. \tag{S31}$$

Moreover, for $\|(\boldsymbol{\Delta}_k)_{[i,:]}\boldsymbol{\Gamma}_{-k}\|_2^2$, since we consider $i \in I_k^{(t)}$, it follows that,

$$\|(\boldsymbol{\Delta}_k)_{[i,:]}\boldsymbol{\Gamma}_{-k}\|_2^2 = (\boldsymbol{\Delta}_k)_{[i,:]}\boldsymbol{P}_{\boldsymbol{\Gamma}_{-k}}(\boldsymbol{\Delta}_k)_{[i,:]}^\top = \|(\boldsymbol{\Delta}_k)_{[i,:]}\|_2^2 > 4\alpha s_{-k}\log p/n, \tag{S32}$$

where the last inequality follows from the definition of $I_k^{(t)}$. For $\sigma_{\min}^2(\boldsymbol{\Gamma}_{-k}^\top\widehat{\boldsymbol{\Gamma}}_{-k}^{(t)})$, when $k=1$,

$$\begin{aligned}
\sigma_{\min}^2(\boldsymbol{\Gamma}_{-1}^\top\widehat{\boldsymbol{\Gamma}}_{-1}^{(t)}) &= \sigma_{\min}^2\{(\boldsymbol{\Gamma}_2^\top\widehat{\boldsymbol{\Gamma}}_2^{(t-1)}) \otimes (\boldsymbol{\Gamma}_3^\top\widehat{\boldsymbol{\Gamma}}_3^{(t-1)})\} \\
&\ge \sigma_{\min}^2(\boldsymbol{\Gamma}_2^\top\widehat{\boldsymbol{\Gamma}}_2^{(t-1)}) \cdot \sigma_{\min}^2(\boldsymbol{\Gamma}_3^\top\widehat{\boldsymbol{\Gamma}}_3^{(t-1)}) \\
&= (1 - \|\sin\Theta(\boldsymbol{\Gamma}_2, \widehat{\boldsymbol{\Gamma}}_2^{(t-1)})\|^2)(1 - \|\sin\Theta(\boldsymbol{\Gamma}_3, \widehat{\boldsymbol{\Gamma}}_3^{(t-1)})\|^2) \\
&\ge 1 - (F_2^{(t-1)})^2 - (F_3^{(t-1)})^2,
\end{aligned}$$

where $F_k^{(t)}$ is as defined in (S15), and the last inequality follows from Lemma 2. Under Assumption (A2) and events (E1)(E2), according to (S22), $F_k^{(t-1)} \le 3/(7\sqrt{2}) \le 1/2$ for $k = 2, 3$. Thus, $\sigma_{\min}^2(\boldsymbol{\Gamma}_{-1}^\top\widehat{\boldsymbol{\Gamma}}_{-1}^{(t)})$ is bounded from below as,

$$\sigma_{\min}^2(\boldsymbol{\Gamma}_{-1}^\top\widehat{\boldsymbol{\Gamma}}_{-1}^{(t)}) \ge 1/2. \tag{S33}$$

Similarly, we obtain the same lower bound for $\sigma_{\min}^2(\boldsymbol{\Gamma}_{-k}^\top\widehat{\boldsymbol{\Gamma}}_{-k}^{(t)})$, $k = 2, 3$. Combining (S31), (S32), and (S33), we obtain the lower bound for B_{10} as

$$B_{10} \ge 2\alpha s_{-k}\log p/n.$$

For B_{11}, since for any sample $\omega \in E_1 \cap E_2 \cap \{\|(\widetilde{\boldsymbol{\Delta}}_k - \boldsymbol{\Delta}_k)_{[i,I_{-k}]}\|_2^2 \le \alpha s_{-k}\log p/n\}$, we have,

$$B_{11} \le \|(\widetilde{\boldsymbol{\Delta}}_k - \boldsymbol{\Delta}_k)_{[i,I_{-k}]}\|_2^2 \le \alpha s_{-k}\log p/n,$$

where the first inequality follows from the fact that $\widehat{I}_k^{(t)} \subseteq I_k$ under event (E2).

Combining the bounds for B_{10} and B_{11}, we obtain that,

$$\|(\widetilde{\mathbf{\Delta}}_k)_{[i,:]}\widehat{\mathbf{\Gamma}}_{-k}^{(t)}\|_2^2 \geq \alpha s_{-k}\log p/n,$$

and (S30) is established.

Combining (S29) and (S30), we obtain that,

$$\begin{aligned}
&\mathbb{P}\left[E_1\cap E_2\cap\left[\bigcap_{t=1}^{t_{\max}}\bigcap_{k=1}^{3}\left\{I_k^{(t)}\in\widehat{I}_k^{(t)}\right\}\right]\right]\\
\geq&\mathbb{P}\left(\bigcap_{t=1}^{t_{\max}}\bigcap_{k=1}^{3}\bigcap_{i\in I_k^{(t)}}\left[E_1\cap E_2\cap\left\{\|(\widetilde{\mathbf{\Delta}}_k-\mathbf{\Delta}_k)_{[i,I_{-k}]}\|_2^2\leq\frac{\alpha s_{-k}\log p}{n}\right\}\right]\right)\\
=&\mathbb{P}\left(E_1\cap E_2\cap\left[\bigcap_{t=1}^{t_{\max}}\bigcap_{k=1}^{3}\bigcap_{i\in I_k^{(t)}}\left\{\|(\widetilde{\mathbf{\Delta}}_k-\mathbf{\Delta}_k)_{[i,I_{-k}]}\|_2^2\leq\frac{\alpha s_{-k}\log p}{n}\right\}\right]\right). \qquad \text{(S34)}
\end{aligned}$$

Under Assumption (A1), and by the result from Step 1 of the proof of Theorem 2, we have,

$$\mathbb{P}\left(E_1\cap E_2\right)\geq 1-Ct_{\max}p\kappa(c,p,n). \qquad \text{(S35)}$$

On the other hand, by (S7),

$$\begin{aligned}
\mathbb{P}\left\{\|(\widetilde{\mathbf{\Delta}}_k-\mathbf{\Delta}_k)_{[i,I_{-k}]}\|_2^2>\alpha s_{-k}\log p/n\right\}&\leq\mathbb{P}\left\{\|(\widetilde{\mathbf{\Delta}}_k-\mathbf{\Delta}_k)_{[i,I_{-k}]}\|_\infty>\sqrt{\alpha\log p/n}\right\}\\
&\leq 6s_{-k}\kappa(c,p,n).
\end{aligned}$$

Since $I_k^{(t)}\subseteq I_k$ for $k=1,2,3$ and $t=1,\ldots$, then

$$\begin{aligned}
\mathbb{P}\left[\bigcap_{t=1}^{t_{\max}}\bigcap_{k=1}^{3}\bigcap_{i\in I_k^{(t)}}\left\{\|(\widetilde{\mathbf{\Delta}}_k-\mathbf{\Delta}_k)_{[i,I_{-k}]}\|_2^2\leq\alpha s_{-k}\log p/n\right\}\right]&\geq 1-18t_{\max}s_ks_{-k}\kappa(c,p,n)\\
&=1-18t_{\max}s\kappa(c,p,n). \qquad \text{(S36)}
\end{aligned}$$

Combining (S34), (S35), and (S36), we obtain the lower bound that,

$$\mathbb{P}\left(E_1\cap E_2\cap\left[\bigcap_{t=1}^{t_{\max}}\bigcap_{k=1}^{3}\left\{I_k^{(t)}\in\widehat{I}_k^{(t)}\right\}\right]\right)\geq 1-(18+C)t_{\max}p\kappa(c,p,n). \qquad \text{(S37)}$$

Combining (S27) and (S37), we obtain that,

$$\begin{aligned}
\mathbb{P}\left(E_1\cap E_2\cap\left[\bigcap_{t=0}^{t_{\max}}\bigcap_{k=1}^{3}\left\{I_k^{(t)}\in\widehat{I}_k^{(t)}\right\}\right]\right)&\geq 1-C't_{\max}p\kappa(c,p,n)\\
&=1-C't_{\max}\max\left[p^{-\gamma},p^{-\left\{\sqrt{n(\gamma+1)/(2\log p)}-1\right\}}\right],
\end{aligned}$$

where $\gamma = c - 1$. We have thus shown that both events (E1) to (E3) hold simultaneously under Assumptions (A1) and (A2) with a high probability.

By events (E2) and (E3), we have that,

$$I_k^{(t)} \subseteq \widehat{I}_k^{(t)} \subseteq I_k, \quad k = 1, 2, 3,\ t = 0, \ldots, t_{\max}. \tag{S38}$$

Under Assumption (A3), for $k = 1, 2, 3$, and $i \in I_k$, by taking $C_3 \geq \sqrt{4\alpha}$, we have that,

$$\|(\boldsymbol{\Delta}_k)_{[i,:]}\|_2 \geq \delta_{\min} \geq C_3\sqrt{s \log p/n} > \sqrt{4\alpha s_{-k} \log p/n}.$$

By the definition of $I_k^{(t)}$ for $t \geq 1$, we obtain that,

$$I_k \subseteq I_k^{(t)}, \quad k = 1, 2, 3,\ t = 1, \ldots, t_{\max}. \tag{S39}$$

Combining (S38) and (S39), we have that,

$$\widehat{I}_k^{(t)} = I_k^{(t)} = I_k, \quad k = 1, 2, 3,\ t = 1, \ldots, t_{\max}. \tag{S40}$$

Moreover, by the definition of $I_k^{(0)}$ and $I_k^{(t)}$ for $t \geq 1$, we have $I_k^{(t)} \subseteq I_k^{(0)} \subseteq I_k$, for $k = 1, 2, 3$ and $t = 1, \ldots, t_{\max}$. Thus, by (S40), it follows that $I_k^{(0)} = I_k$. By (S38), we obtain that $\widehat{I}_k^{(0)} = I_k$, $k = 1, 2, 3$.

Therefore, we have proved that,

$$\widehat{I}_k^{(t)} = I_k, \quad k = 1, 2, 3,\ t = 0, \ldots, t_{\max}, \tag{S41}$$

with probability at least $1 - C' \max\left[p^{-\gamma}, p^{-\left\{\sqrt{n(\gamma+1)/(2\log p)}-1\right\}}\right]$, by treating $t_{\max}$ as a constant not diverging with n, p. Let $\widehat{I}_k$ denote the estimated active sets after $t_{\max}$ iterations. Then we have $\widehat{I}_k = I_k$, $k = 1, 2, 3$, which completes the proof of Theorem 3. $\square$

S5.6 Proof of Proposition S1

We introduce the following event.

(E4) $\widehat{I}_k^{(t)} = I_k$ for $k = 1, 2, 3$ and $t = 0, 1, \ldots$

Recall $\widehat{\boldsymbol{\Lambda}}^{(t)} = (\widehat{\boldsymbol{\Lambda}}_1^{(t)}, \widehat{\boldsymbol{\Lambda}}_2^{(t)}, \widehat{\boldsymbol{\Lambda}}_3^{(t)})$ is the HOOI estimate from the tth iteration for the optimization,

$$\text{maximize}_{\boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2, \boldsymbol{\Lambda}_3} \|\widetilde{\boldsymbol{\Delta}}_s \times_1 \boldsymbol{P}_{\boldsymbol{\Lambda}_1} \times_2 \boldsymbol{P}_{\boldsymbol{\Lambda}_2} \times_3 \boldsymbol{P}_{\boldsymbol{\Lambda}_3}\|_F^2,$$

where $\boldsymbol{\Lambda}_k \in \mathbb{R}^{s_k \times r_k}$ is a semi-orthogonal matrix, and $\widetilde{\boldsymbol{\Delta}}_s = \widetilde{\boldsymbol{\Delta}}_{[I_1, I_2, I_3]}$. Also, recall we assume all the active variables are located together at the beginning of each mode.

For $t = 0$, the initial basis matrix from Algorithm 1 is,

$$\begin{aligned} \widehat{\mathbf{\Gamma}}_k^{(0)} &= (\boldsymbol{V}_k^\top, \mathbf{0}^\top)^\top, \quad \text{where} \\ \boldsymbol{V}_k &= \mathrm{SVD}_{r_k}\left\{(\widetilde{\mathbf{\Delta}}_k)_{[\widehat{I}_k^{(0)}, \widehat{I}_{-k}^{(0)}]}\right\} = \mathrm{SVD}_{r_k}\left\{(\widetilde{\mathbf{\Delta}}_k)_{[I_k, I_{-k}]}\right\} = \mathrm{SVD}_{r_k}(\widetilde{\mathbf{\Delta}}_{s,k}), \quad k = 1, 2, 3. \end{aligned}$$

The above second equality holds under event (E4), and the last equality follows from the fact that $(\widetilde{\mathbf{\Delta}}_k)_{[I_k, I_{-k}]} = \widetilde{\mathbf{\Delta}}_{s,k}$. Since $\mathrm{SVD}_{r_k}(\widetilde{\mathbf{\Delta}}_{s,k}) = \widehat{\mathbf{\Lambda}}_k^{(0)}$, we have that,

$$\widehat{\mathbf{\Gamma}}_k^{(0)} = \{(\widehat{\mathbf{\Lambda}}_k^{(0)})^\top, \mathbf{0}^\top\}^\top, \quad k = 1, 2, 3. \tag{S42}$$

For $t = 1$, we update the matrix $\widehat{\mathbf{\Gamma}}_1^{(1)}$ in Algorithm 1 as,

$$\begin{aligned} \widehat{\mathbf{\Gamma}}_1^{(1)} &= (\boldsymbol{W}^\top, \mathbf{0}^\top)^\top, \quad \text{where} \\ \boldsymbol{W} &= \mathrm{SVD}_{r_1}\left\{(\widetilde{\mathbf{\Delta}}_1)_{[\widehat{I}_1^{(1)}, :]}\widehat{\mathbf{\Gamma}}_{-1}^{(1)}\right\} = \mathrm{SVD}_{r_1}\left\{(\widetilde{\mathbf{\Delta}}_1)_{[I_1, :]}\widehat{\mathbf{\Gamma}}_{-1}^{(1)}\right\} = \mathrm{SVD}_{r_1}\left\{(\widetilde{\mathbf{\Delta}}_1)_{[I_1, :]}(\widehat{\mathbf{\Gamma}}_2^{(0)} \otimes \widehat{\mathbf{\Gamma}}_3^{(0)})\right\} \\ &= \mathrm{SVD}_{r_1}\left\{(\widetilde{\mathbf{\Delta}}_1)_{[I_1, I_{-1}]}(\widehat{\mathbf{\Lambda}}_2^{(0)} \otimes \widehat{\mathbf{\Lambda}}_3^{(0)})\right\} = \mathrm{SVD}_{r_1}\left\{\widetilde{\mathbf{\Delta}}_{s,1}(\widehat{\mathbf{\Lambda}}_2^{(0)} \otimes \widehat{\mathbf{\Lambda}}_3^{(0)})\right\}. \end{aligned}$$

The above second equality holds under event (E4), and the fourth equality follows from (S42). Since $\mathrm{SVD}_{r_1}\left\{\widetilde{\mathbf{\Delta}}_{s,1}(\widehat{\mathbf{\Lambda}}_2^{(0)} \otimes \widehat{\mathbf{\Lambda}}_3^{(0)})\right\} = \widehat{\mathbf{\Lambda}}_1^{(1)}$, we have that,

$$\widehat{\mathbf{\Gamma}}_1^{(1)} = \{(\widehat{\mathbf{\Lambda}}_1^{(1)})^\top, \mathbf{0}^\top\}^\top.$$

Following similar arguments, we are able to prove that,

$$\widehat{\mathbf{\Gamma}}_k^{(t)} = \{(\widehat{\mathbf{\Lambda}}_k^{(t)})^\top, \mathbf{0}^\top\}^\top, \quad \text{for } k = 1, 2, 3, \ t = 1, 2, \ldots \tag{S43}$$

Combining (S42) and (S43), under event (E4), we obtain that,

$$\widehat{\mathbf{\Gamma}}_k^{(t)} = \{(\widehat{\mathbf{\Lambda}}_k^{(t)})^\top, \mathbf{0}^\top\}^\top, \quad \text{for } k = 1, 2, 3, \ t = 0, 1, \ldots$$

According to (S41) in the proof of Theorem 3, under Assumptions (A1) to (A3), we have the variable selection consistency, i.e., $\widehat{I}_k^{(t)} = I_k$, for $k = 1, 2, 3$ and $t = 0, \ldots, t_{\max}$, with probability at least $1 - C\max\left[p^{-\gamma}, p^{-\left\{\sqrt{n(\gamma+1)/(2\log p)}-1\right\}}\right]$. Since $s\log p = o(n)$ and $t_{\max} = o(p)$, $\widehat{I}_k^{(t)} = I_k$, for $k = 1, 2, 3$ and $t \geq 0$ with probability approaching to one. Therefore, event (E4) holds with probability approaching to one. This completes the proof of Proposition S1. □

S5.7 Proof of Theorem S1

We first prove statement (a). According to Proposition S1, under Assumptions (A1) to (A3), and that $s\log p = o(n)$ and $t_{\max} = o(p)$, with probability approaching to one, we have $\widehat{\mathbf{\Gamma}}_k^{(t)} =$

$\{(\widehat{\boldsymbol{\Lambda}}_k^{(t)})^\top, \mathbf{0}^\top\}^\top$, for $k = 1, 2, 3$, and $t = 0, 1, \ldots$. Note that $\widehat{\boldsymbol{\Lambda}}^{(t)} = (\widehat{\boldsymbol{\Lambda}}_1^{(t)}, \widehat{\boldsymbol{\Lambda}}_2^{(t)}, \widehat{\boldsymbol{\Lambda}}_3^{(t)})$ is the HOOI estimate from the tth iteration for the optimization,

$$\text{maximize}_{\boldsymbol{\Lambda}_1,\boldsymbol{\Lambda}_2,\boldsymbol{\Lambda}_3} \|\widetilde{\boldsymbol{\Delta}}_s \times_1 \boldsymbol{P}_{\boldsymbol{\Lambda}_1} \times_2 \boldsymbol{P}_{\boldsymbol{\Lambda}_2} \times_3 \boldsymbol{P}_{\boldsymbol{\Lambda}_3}\|_F^2,$$

where $\boldsymbol{\Lambda}_k \in \mathbb{R}^{s_k \times r_k}$ is a semi-orthogonal matrix, $k = 1, 2, 3$, and $\widetilde{\boldsymbol{\Delta}}_s = \widetilde{\boldsymbol{\Delta}}_{[I_1,I_2,I_3]}$. By Xu (2018, Theorem 1.1 (i)), assuming $\{\widehat{\boldsymbol{\Lambda}}^{(t)}\}$ has a block-nondegenerative cluster point $\widehat{\boldsymbol{\Lambda}} = (\widehat{\boldsymbol{\Lambda}}_1, \widehat{\boldsymbol{\Lambda}}_2, \widehat{\boldsymbol{\Lambda}}_3)$, then $\widehat{\boldsymbol{\Lambda}}$ is a critical point and a block-wise maximizer of (S1). Also, $\lim_{t\to\infty} \boldsymbol{P}_{\widehat{\boldsymbol{\Lambda}}_k^{(t)}} = \boldsymbol{P}_{\widehat{\boldsymbol{\Lambda}}_k}$, for $k = 1, 2, 3$. Since

$$\boldsymbol{P}_{\widehat{\boldsymbol{\Gamma}}_k^{(t)}} = \begin{bmatrix} \boldsymbol{P}_{\widehat{\boldsymbol{\Lambda}}_k^{(t)}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

it follows that $\lim_{t\to\infty} \boldsymbol{P}_{\widehat{\boldsymbol{\Gamma}}_k^{(t)}} = \boldsymbol{P}_{\widehat{\boldsymbol{\Gamma}}_k}$, where $\widehat{\boldsymbol{\Gamma}}_k = (\widehat{\boldsymbol{\Lambda}}_k^\top, \mathbf{0}^\top)^\top$, and $k = 1, 2, 3$.

We next prove statement (b). By Xu (2018, Theorem 1.1 (ii)), if the starting point $\widehat{\boldsymbol{\Lambda}}^{(0)}$ is sufficiently close to any block-nondegenerative local maximizer of (S1), then $\lim_{t\to\infty} \boldsymbol{P}_{\widehat{\boldsymbol{\Lambda}}_k^{(t)}} = \boldsymbol{P}_{\widetilde{\boldsymbol{\Lambda}}_k}$, where $\widetilde{\boldsymbol{\Lambda}} = (\widetilde{\boldsymbol{\Lambda}}_1, \widetilde{\boldsymbol{\Lambda}}_2, \widetilde{\boldsymbol{\Lambda}}_3)$ is some local maximizer of (S1). Therefore, we have $\lim_{t\to\infty} \boldsymbol{P}_{\widehat{\boldsymbol{\Gamma}}_k^{(t)}} = \boldsymbol{P}_{\widetilde{\boldsymbol{\Gamma}}_k}$, where $\widetilde{\boldsymbol{\Gamma}}_k = (\widetilde{\boldsymbol{\Lambda}}_k^\top, \mathbf{0}^\top)^\top$, and $k = 1, 2, 3$.

This completes the proof of Theorem S1. $\square$

S5.8 Proof of Theorem S2

The proof of Theorem S2 follows a similar structure as that of Theorem 2. Recall that $\log p \leq c_0 n^b$ for some constants $c_0 > 0$ and $0 < b < 1$. We first introduce the following event.

(E5) $\|\widetilde{\boldsymbol{\Delta}} - \boldsymbol{\Delta}\|_F \leq \sqrt{\beta p / n^{1-b}}$, where $\beta = 512(1 + c_0)(M + \sigma)^4$.

We prove Theorem S2 in two main steps. In the first step, we prove that event (E5) holds with a high probability under Assumption (A1). In the second step, we derive the non-asymptotic error bounds in Theorem S2 under event (E5) and Assumption (A2′).

Step 1: Under Assumption (A1), by Lemma 5, we have that,

$$\begin{aligned}
&\mathbb{P}\left(\|\widetilde{\boldsymbol{\Delta}} - \boldsymbol{\Delta}\|_F \geq \sqrt{\beta p/n^{1-b}}\right) \leq \mathbb{P}\left(\|\widetilde{\boldsymbol{\Delta}} - \boldsymbol{\Delta}\|_{\max} \geq \sqrt{\beta/n^{1-b}}\right) \\
&\leq p\, \mathbb{P}\left(|\widetilde{\Delta}_{ijk} - \Delta_{ijk}| > \sqrt{\beta/n^{1-b}}\right) \leq 6p \max\left[\exp\{-(1+c_0)n^b\}, \exp\{-\sqrt{n^{1+b}(1+c_0)/2}\}\right] \\
&\leq 6 \max\left[\exp(-n^b), \exp\{-(\sqrt{n^{1+b}(1+c_0)/2} - c_0 n^b)\}\right],
\end{aligned}$$

where the last step is due to that $\log p \leq p \leq c_0 n^{1/2}$ for some positive constant c_0. Henceforth,

$$\mathbb{P}(\|\widetilde{\boldsymbol{\Delta}} - \boldsymbol{\Delta}\|_F \leq \sqrt{\beta p/n^{1-b}}) \geq 1 - C'' \max\left[\exp(-n^b), \exp\{-(\sqrt{n^{1+b}(1+c_0)/2} - c_0 n^b)\}\right].$$

Step 2: Let $\widehat{\boldsymbol{\Gamma}}_k^{(t)}$ denote the updated basis matrix at the tth iteration of the non-sparse version of Algorithm 1. For $t = 0, 1, \ldots, t_{\max}$ and $k = 1, 2, 3$, define

$$E_k^{(t)} = \left\|(\widehat{\boldsymbol{\Gamma}}_{k\perp}^{(t)})^\top \boldsymbol{\Delta}_k\right\|_F, \quad F_k^{(t)} = \left\|(\widehat{\boldsymbol{\Gamma}}_{k\perp}^{(t)})^\top \boldsymbol{\Gamma}_k^{\mathrm{ns}}\right\|_F, \quad E^{(t)} = \max_{k=1,2,3} E_k^{(t)}.$$

Parallel to Step 2 in the proof of Theorem 2, we derive the error bounds in four steps.

Step 2.1: When $t = 0$, since $\widehat{\boldsymbol{\Gamma}}_k^{(0)}$ is composed of the top-r_k left singular vectors of $\widetilde{\boldsymbol{\Delta}}_k$, by Lemma 1, we have that,

$$E_k^{(0)} = \|(\widehat{\boldsymbol{\Gamma}}_{k\perp}^{(0)})^\top \boldsymbol{\Delta}_k\|_F \leq 2\|\widetilde{\boldsymbol{\Delta}}_k - \boldsymbol{\Delta}_k\|_F = 2\|\widetilde{\boldsymbol{\Delta}} - \boldsymbol{\Delta}\|_F, \quad k = 1, 2, 3.$$

Under event (E5), we have that,

$$E_k^{(0)} \leq 2\sqrt{\beta p / n^{1-b}}, \quad k = 1, 2, 3. \tag{S44}$$

This completes Step 2.1.

Step 2.2: We next establish the following inequalities, under event (E5), and for each iteration $t = 1, 2, \ldots, t_{\max}$, that

$$\begin{aligned}
\left[1 - \sqrt{(F_2^{(t-1)})^2 + (F_3^{(t-1)})^2}\right] E_1^{(t)} &\leq 2\sqrt{\beta p / n^{1-b}}, \\
\left[1 - \sqrt{(F_1^{(t)})^2 + (F_3^{(t-1)})^2}\right] E_2^{(t)} &\leq 2\sqrt{\beta p / n^{1-b}}, \\
\left[1 - \sqrt{(F_1^{(t)})^2 + (F_2^{(t)})^2}\right] E_3^{(t)} &\leq 2\sqrt{\beta p / n^{1-b}}.
\end{aligned} \tag{S45}$$

We only prove the first inequality, while the other two inequalities can be derived similarly.

Toward that end, we bound $E_1^{(t)}$ by three terms as follows,

$$\begin{aligned}
E_1^{(t)} &= \left\|(\widehat{\boldsymbol{\Gamma}}_{1\perp}^{(t)})^\top \boldsymbol{\Delta}_1 \left(\boldsymbol{P}_{\widehat{\boldsymbol{\Gamma}}_{-1}^{(t)}} + \boldsymbol{P}_{\widehat{\boldsymbol{\Gamma}}_{-1\perp}^{(t)}}\right)\right\|_F \leq \left\|(\widehat{\boldsymbol{\Gamma}}_{1\perp}^{(t)})^\top \boldsymbol{\Delta}_1 \widehat{\boldsymbol{\Gamma}}_{-1}^{(t)}\right\|_F + \left\|(\widehat{\boldsymbol{\Gamma}}_{1\perp}^{(t)})^\top \boldsymbol{\Delta}_1 \widehat{\boldsymbol{\Gamma}}_{-1\perp}^{(t)}\right\|_F \\
&\equiv B_{12} + B_{13}.
\end{aligned} \tag{S46}$$

For B_{12}, since $\widehat{\boldsymbol{\Gamma}}_1^{(t)}$ consists of the top-r_k left singular vectors of $\widetilde{\boldsymbol{\Delta}}_1 \widehat{\boldsymbol{\Gamma}}_{-1}^{(t)}$, by Lemma 1, we have,

$$B_{12} \leq 2\|(\widetilde{\boldsymbol{\Delta}}_1 - \boldsymbol{\Delta}_1)\widehat{\boldsymbol{\Gamma}}_{-1}^{(t)}\|_F.$$

Under events (E5), we further bound B_{12} as

$$B_{12} \leq 2\|\widetilde{\boldsymbol{\Delta}} - \boldsymbol{\Delta}\|_F \leq 2\sqrt{\beta p / n^{1-b}}.$$

For B_{13}, since $\mathbf{\Delta}_1 = \mathbf{\Delta}_1 \boldsymbol{P}_{\mathbf{\Gamma}^{\rm ns}_{-1}}$, we have,

$$B_{13} = \|(\widehat{\mathbf{\Gamma}}^{(t)}_{1\perp})^\top \mathbf{\Delta}_1 \mathbf{\Gamma}^{\rm ns}_{-1} (\mathbf{\Gamma}^{\rm ns}_{-1})^\top \widehat{\mathbf{\Gamma}}^{(t)}_{-1\perp}\|_F \le \|(\widehat{\mathbf{\Gamma}}^{(t)}_{1\perp})^\top \mathbf{\Delta}_1\|_F \|(\mathbf{\Gamma}^{\rm ns}_{-1})^\top \widehat{\mathbf{\Gamma}}^{(t)}_{-1\perp}\| \le E_1^{(t)} \|\sin\Theta(\mathbf{\Gamma}^{\rm ns}_{-1}, \widehat{\mathbf{\Gamma}}^{(t)}_{-1})\|.$$

By (S20), we have that,

$$\|\sin\Theta(\mathbf{\Gamma}^{\rm ns}_{-1}, \widehat{\mathbf{\Gamma}}^{(t)}_{-1})\| \le \sqrt{(F_2^{(t-1)})^2 + (F_3^{(t-1)})^2}.$$

Then, we obtain that,

$$B_{13} \le E_1^{(t)} \sqrt{(F_2^{(t-1)})^2 + (F_3^{(t-1)})^2}.$$

Combining the bounds for B_{12} and B_{12} in (S46), we obtain that,

$$E_1^{(t)} \le 2\sqrt{\beta p / n^{1-b}} + E_1^{(t)} \sqrt{(F_2^{(t-1)})^2 + (F_3^{(t-1)})^2}.$$

This completes Step 2.2.

Step 2.3: We next establish the upper bound for $E^{(t)}$ under Assumption (A2′) and event (E5), for $t = 0, 1, \ldots, t_{\max}$, as

$$E^{(t)} \le \left(3 + \frac{1}{2^t}\right) \sqrt{\beta p / n^{1-b}}. \tag{S47}$$

We prove (S47) by induction. For $t = 0$, by (S44), $E^{(0)}$ satisfies (S47). Suppose (S47) holds for $E^{(t')}$ where $t' = 0, 1, \ldots, t-1$, we next show that (S47) holds for $E^{(t)}$.

By Lemma 1 and Lemma 2, we have,

$$E_k^{(t-1)} \ge \lambda_k \|\sin\Theta(\widehat{\mathbf{\Gamma}}^{(t)}_k, \mathbf{\Gamma}^{\rm ns}_k)\|_F = \lambda_k F_k^{(t-1)}. \tag{S48}$$

By taking the constant $C_4 = 12\sqrt{2\beta}$ in Assumption (A2′), such that $\lambda \ge 12\sqrt{2}\sqrt{\beta p/n^{1-b}}$, we have that,

$$F_k^{(t-1)} \le \frac{1}{12\sqrt{2}} \left(3 + \frac{1}{2^{t-1}}\right), \quad k = 2, 3. \tag{S49}$$

Then we have that,

$$\left\{1 - \sqrt{(F_2^{(t-1)})^2 + (F_3^{(t-1)})^2}\right\} > \frac{2}{3}.$$

Therefore, by (S45), we have that,

$$E_1^{(t)} \le 2\left\{1 - \sqrt{(F_2^{(t-1)})^2 + (F_3^{(t-1)})^2}\right\}^{-1} \sqrt{\beta p/n^{1-b}} \le \left(3 + \frac{1}{2^t}\right) \sqrt{\beta p / n^{1-b}}.$$

Following similar arguments, we obtain the same upper bounds for $E_2^{(t)}$ and $E_3^{(t)}$. Then, by induction, (S21) holds for all $t = 0, 1, \ldots, t_{\max}$.

After $t_{\max}$ iterations, we obtain that,

$$E^{(t_{\max})} = \max_{k=1,2,3} E_k^{(t_{\max})} \leq \left(3 + \frac{1}{2^{t_{\max}}}\right) \sqrt{\beta p / n^{1-b}}.$$

Let $\widehat{\mathbf{\Gamma}}_k^{\mathrm{ns}} = \widehat{\mathbf{\Gamma}}_k^{(t_{\max})}$, $k = 1, 2, 3$. Then by (S48) and Lemma 2, under Assumption (A2′), for $k = 1, 2, 3$, we have that,

$$\|\boldsymbol{P}_{\widehat{\mathbf{\Gamma}}_k^{\mathrm{ns}}} - \boldsymbol{P}_{\mathbf{\Gamma}_k^{\mathrm{ns}}}\|_F \leq \sqrt{2}\frac{E_k^{(t_{\max})}}{\lambda_k} \leq \left(3 + \frac{1}{2^{t_{\max}}}\right) \frac{\sqrt{2\beta}}{\lambda} \sqrt{p/n^{1-b}} = (c_7 + c_8 2^{-t_{\max}})\sqrt{p/n^{1-b}},$$

where $c_7 = 3\sqrt{1024(1 + c_0)}(M + \sigma)^2/C_5$, and $c_8 = \sqrt{1024(1 + c_0)}(M + \sigma)^2/C_5$. Thus, statement (b) is proved.

Step 2.4: Finally, we put together the results in Steps 2.1 to 2.3 to derive the non-asymptotic error bound in statement (a). That is, we bound $\|\widehat{\mathbf{\Delta}}^{\mathrm{ns}} - \mathbf{\Delta}\|_F$ as,

$$\begin{aligned}\|\widehat{\mathbf{\Delta}}^{\mathrm{ns}} - \mathbf{\Delta}\|_F &\leq \|\mathbf{\Delta} \times_1 \boldsymbol{P}_{\widehat{\mathbf{\Gamma}}_1^{\mathrm{ns}}} \times_2 \boldsymbol{P}_{\widehat{\mathbf{\Gamma}}_2^{\mathrm{ns}}} \times_3 \boldsymbol{P}_{\widehat{\mathbf{\Gamma}}_3^{\mathrm{ns}}} - \mathbf{\Delta}\|_F + \|(\widetilde{\mathbf{\Delta}} - \mathbf{\Delta}) \times_1 \boldsymbol{P}_{\widehat{\mathbf{\Gamma}}_1^{\mathrm{ns}}} \times_2 \boldsymbol{P}_{\widehat{\mathbf{\Gamma}}_2^{\mathrm{ns}}} \times_3 \boldsymbol{P}_{\widehat{\mathbf{\Gamma}}_3^{\mathrm{ns}}}\|_F \\ &\leq 3E^{(t_{\max})} + \|\widetilde{\mathbf{\Delta}} - \mathbf{\Delta}\|_F.\end{aligned}$$

Under event (E5), by (S47), we have that,

$$\|\widehat{\mathbf{\Delta}}^{\mathrm{ns}} - \mathbf{\Delta}\|_F \leq \left(10 + \frac{3}{2^{t_{\max}}}\right) \sqrt{\beta p/n^{1-b}} = (c_5 + c_6 2^{-t_{\max}})\sqrt{p/n^{1-b}},$$

where $c_5 = 10\sqrt{512(1 + c_0)}(M + \sigma)^2$, and $c_6 = 3\sqrt{512(1 + c_0)}(M + \sigma)^2$. Thus, statement (a) is proved.

This completes the proof of Theorem S2.

References

Cai, T. T. and Zhang, A. (2018). Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics*, 46(1):60–89.

De Lathauwer, L., De Moor, B., and Vandewalle, J. (2000). On the best rank-1 and rank-(r1, r2,..., rn) approximation of higher-order tensors. *SIAM journal on Matrix Analysis and Applications*, 21(4):1324–1342.

Golub, G. H. and Van Loan, C. F. (1996). Matrix computations. johns hopkins studies in the mathematical sciences.

Liu, J. S. (1994). Siegel's formula via stein's identities. *Statistics & Probability Letters*, 21(3):247–251.

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

Xu, Y. (2018). On the convergence of higher-order orthogonal iteration. *Linear and Multilinear Algebra*, 66(11):2247–2265.

Zhang, A. and Han, R. (2019). Optimal sparse singular value decomposition for high-dimensional high-order data. *Journal of the American Statistical Association*, pages 1–34.