Supplementary Materials for "A Doubly-Enhanced EM Algorithm for Model-Based Tensor Clustering"

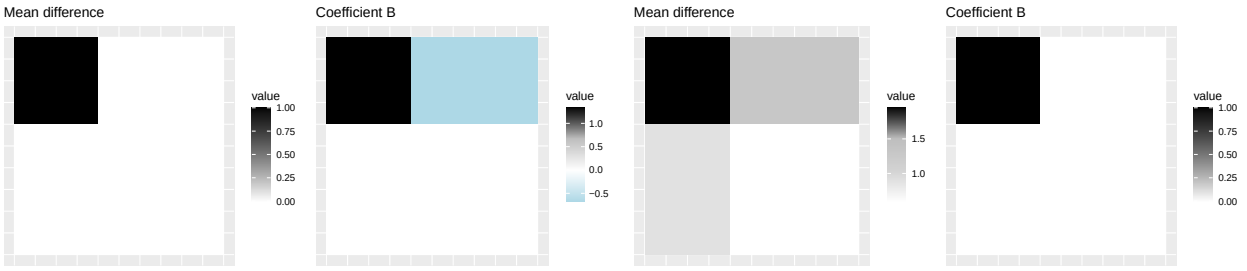
Qing Mai, Xin Zhang, Yuqing Pan and Kai Deng
Florida State University

A An example to demonstrate the effect of correlations

The following example demonstrates the effect of correlations as discussed in Section 2.3. We use the shorthand notation that, for any matrix $\mathbf{\Omega}$ and a scalar ρ, we say that $\mathbf{\Omega} = CS(\rho)$ if $\omega_{ii} = 1$ for all i and $\omega_{ij} = \rho$ for all $i \neq j$. The following Example S.1 reveals that ignoring the correlation structure could severely reduce the clustering and variable selection accuracy. Therefore, it is important to develop a tensor clustering method that takes into account of the correlation structure.

Example S.1. *Consider a two-way tensor* $\mathbf{X} \in \mathbb{R}^{5\times 5}$ *and the latent clustering label* $Y \in \{1, 2\}$ *generated from TNMM in* (2.3)*, with* $\pi_1^* = \pi_2^* = 0.5$ *and* $\boldsymbol{\mu}_1^* = 0$*. We consider the following two models.*

(a) $\boldsymbol{\mu}^*_{2,[1:2,1:2]} = 1$ *and* $\mu^*_{2,ij} = 0$ *for all other* i, j*,* $\mathbf{\Sigma}_1^* = \mathbf{I}_5$*,* $\mathbf{\Sigma}_2^* = CS(0.5)$*. Hence, only the*



(a) Example 1 (a) (b) Example 1 (b)

Figure S.1: Heatmaps of $\boldsymbol{\mu}_2^* - \boldsymbol{\mu}_1^*$ and $\mathbf{B}^*$. Each figure represents a 5×5 matrix and white spaces correspond to zero elements. Example 1 (a) shows that elements with no mean differences across clusters could still help with clustering, while Example 1 (b) shows that elements with mean differences may not help with clustering.

upper left 2×2 *corner belongs to* $\mathcal{A}$*, but*

$$\mathbf{B}^* = \begin{bmatrix} 4/3 & 4/3 & -2/3 & -2/3 & -2/3 \\ 4/3 & 4/3 & -2/3 & -2/3 & -2/3 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

has nonzero elements in the first two rows. See Figure S.1a for heatmaps of $\boldsymbol{\mu}_2^* - \boldsymbol{\mu}_1^*$ *and* $\mathbf{B}^*$*. Hence, the* $[1:2, 3:5]$ *block helps clustering even though they have constant means across clusters. Moreover, on a dataset with 10000 samples, the optimal rule based on* $\mathbf{B}^*$ *has an error rate of 12.98%, while the independence rule based on the set* $\mathcal{A}$ *has an error rate of 20.95%.*

(b) Set $\boldsymbol{\Sigma}_1^* = CS(0.5)$*,* $\boldsymbol{\Sigma}_2^* = CS(0.3)$ *and*

$$\boldsymbol{\mu}_2^* = \begin{bmatrix} 1.95 & 1.95 & 1.3 & 1.3 & 1.3 \\ 1.95 & 1.95 & 1.3 & 1.3 & 1.3 \\ 0.9 & 0.9 & 0.6 & 0.6 & 0.6 \\ 0.9 & 0.9 & 0.6 & 0.6 & 0.6 \\ 0.9 & 0.9 & 0.6 & 0.6 & 0.6 \end{bmatrix}.$$

Then the upper-left 2×2 *corner of* $\mathbf{B}^*$ *has nonzero elements all equal to 1, while all other elements in* $\mathbf{B}^*$ *are 0. See Figure S.1b for heatmaps of* $\boldsymbol{\mu}_2^* - \boldsymbol{\mu}_1^*$ *and* $\mathbf{B}^*$*. Even though all the elements are differentially distributed across clusters, most of them do not play a role in achieving the optimal clustering error. In this case, the clustering error goes up from 6.42% to 10.85% if we ignore the correlations.*

B A blockwise coordinate descent algorithm to solve (3.13)

To solve (3.13), we need the following result.

Proposition S.1. *For* $m = 1, \ldots, M$*, define* $\widehat{\boldsymbol{\Sigma}}_{m,\cdot j}^{(t)}$ *as the* j*-th column of* $\widehat{\boldsymbol{\Sigma}}_m^{(t)}$*. The solution of* $\mathbf{B}_{\cdot\mathcal{J}} = (b_{2\mathcal{J}}, \ldots, b_{K\mathcal{J}})$ *to* (3.13) *given* $b_{k\mathcal{J}'}$ *for* $k = 2, \ldots, K$ *and* $\mathcal{J}' \neq \mathcal{J}$ *is*

$$(\widehat{b}_{2\mathcal{J}}, \ldots, \widehat{b}_{K\mathcal{J}}) = \breve{\mathbf{B}}_{\cdot\mathcal{J}} \left(1 - \frac{\lambda^{(t)}}{\| \breve{\mathbf{B}}_{\cdot\mathcal{J}} \|}\right)_+, \tag{S.1}$$

where $\breve{b}_{k\mathcal{J}} = \dfrac{(\widehat{\mu}_{k\mathcal{J}} - \widehat{\mu}_{1\mathcal{J}}) - [\![\mathbf{B}_k^{\mathcal{J}}; \{\widehat{\boldsymbol{\Sigma}}_{m,\cdot j_1}^{(t)}\}^\intercal, \ldots, \{\widehat{\boldsymbol{\Sigma}}_{m,\cdot j_M}^{(t)}\}^\intercal]\!]}{\prod_{m=1}^M \widehat{\sigma}_{m,j_m j_m}^{(t)}}$ *with the elements in* $\mathbf{B}_k^{\mathcal{J}}$ *equals* $\mathbf{B}_{k\mathcal{J}'}$ *for all* $\mathcal{J}' \neq \mathcal{J}$ *and 0 otherwise.*

Proof of Proposition S.1. The desired conclusion can be proved in the same way as Lemma 5 in Pan et al. (2019) and is thus omitted. □

By Proposition S.1, we can solve (3.13) by iteratively updating $b_{k\mathcal{J}}$ while keeping all other elements fixed. Such an algorithm is summarized in Algorithm S.2.

Algorithm S.2 Algorithm to solve (3.13)

1. Input $\widehat{\mathbf{\Sigma}}_m^{(t)}, m = 1, \ldots, M$ and $\widehat{\boldsymbol{\mu}}_k^{(t)}, k = 1, \ldots, K$. Initialize $\widehat{\mathbf{B}}_k^{(t,0)} = 0$ for all $k = 2, \ldots, K$.
2. For $w = 1, 2, \ldots$, do the following until convergence:
   For each $\mathcal{J} = (j_1, \cdots, j_M)$,
   (a) Update $\breve{b}_{k,\mathcal{J}}^{(w)} \longleftarrow \dfrac{(\widehat{\mu}_{k,\mathcal{J}}^{(t)} - \widehat{\mu}_{1,\mathcal{J}}^{(t)}) - [\![\mathbf{B}_k^j; \{\widehat{\mathbf{\Sigma}}^{(t)}\}_{1,\cdot j_1}^{\intercal}, \cdots, \{\widehat{\mathbf{\Sigma}}^{(t)}\}_{M,\cdot j_M}^{\intercal}]\!]}{\prod_{m=1}^{M} \hat{\sigma}_{m,j_m j_m}}$;
   (b) Compute $\widehat{b}_{k,\mathcal{J}}^{(t,w)} \longleftarrow \breve{b}_{k,\mathcal{J}}^{(w)} \left(1 - \dfrac{\lambda^{(t)}}{\sqrt{\sum_{k=2}^{K} (\breve{b}_{k,\mathcal{J}}^{(w)})^2}}\right)_+$ for $k = 2, \ldots, K$.
3. Output $\widehat{\mathbf{B}}_k^{(t)} = \widehat{\mathbf{B}}_k^{(t,\infty)}$ at convergence.

C Estimation of the number of clusters

We present a working method for the estimation of K under TNMM in high dimensions, although a thorough study along this line is out of the scope of this manuscript. Recall that, in addition to K, we also have a tuning parameter λ that controls the amount of penalty. The tuning parameter λ is selected by a BIC-type measure, defined in (3.23). We slightly modify this criterion to estimate K along with λ. Recall that $\widehat{\mathcal{D}}^{\lambda} = \{\mathcal{J} : \widehat{b}_{k,\mathcal{J}}^{\lambda} \text{ is nonzero for at least one } k\}$. We let

$$\text{BIC}(\lambda, K) = -2 \sum_{i=1}^{n} \log\Big(\sum_{k=1}^{K} \widehat{\pi}_k^{\lambda} f_k(\mathbf{X}_i; \widehat{\boldsymbol{\theta}}_k^{\lambda})\Big) + \log(n) \cdot |\widehat{\mathcal{D}}^{\lambda}|. \tag{S.2}$$

Apparently, $\text{BIC}(\lambda, K)$ is equal to $BIC(\lambda)$ in (3.23) for any fixed K. Define $\mathbb{Z}^+$ as the set of positive integers. To estimate K, we let

$$(\widehat{K}, \widehat{\lambda}) = \underset{K \in \mathcal{K}, \lambda > 0}{\text{argmin}}\, \text{BIC}(\lambda, K), \tag{S.3}$$

Model	M1	M2	M3	M4	M5	M6
Correct Selection Rate	62%	80%	77%	58%	68%	22%

Table S.1: Estimation of number of clusters. Reported are the proportions of replicates in which we correctly selects K by the restricted BIC.

where $\mathcal{K} = \{K \in \mathbb{Z}^+ : \text{there are at least } 5\% \times n \text{ observations in each cluster}\}$. The set $\mathcal{K}$ contains all the K's with reasonably balanced clusters; if a K results in very small clusters (i.e, fewer than $5\% \times n$ observations in a cluster), it is excluded. We minimize $\text{BIC}(\lambda, K)$ with the constraint that $K \in \mathcal{K}$ because otherwise $\text{BIC}(\lambda, K)$ tends to overestimate K, with many small clusters. By restricting our attention to $\mathcal{K}$, we protect ourselves from overestimation. In this sense, we refer to the criterion in (S.3) as the restricted BIC.

We test the performance of the restricted BIC on Models M1–M6 in Section 5. In Table S.1 we report the proportion of replicates in which we correctly select K. It can be seen that the restricted BIC in general produces good selection results for K. On most models the correct selection rate is around 60% or higher. Model M6 seems to be the most challenging one, with a correct selection rate of 22%. This is likely due to the fact that Model 6 has relatively large number of clusters ($K = 6$) and small within-cluster sample size.

D Simulations on tensor block model

As suggested by a referee, we also consider DEEM under a special case of the tensor block model (TBM; Wang & Zeng (2019)). In our context, TBM is constructed with a tensor $\mathcal{X} \in \mathbb{R}^{n \times p_1 \times p_2 \times \cdots \times p_M}$. The tensor is clustered along all the modes. If an entry $X_{ij_1 \cdots j_M}$ is from the k-th cluster on the first mode and the k_m-th cluster on the $(m+1)$-th mode, then we assume that

$$X_{ij_1 \cdots j_M} = \mu^*_{k,k_1 \cdots k_M} + \epsilon_{ij_1 \cdots j_M},$$

where $\epsilon_{ij_1 \cdots j_M}$ are independent statistical noises with mean zero. To best align TBM with TNMM, in what follows we assume that $\epsilon_{ij_1 \cdots j_M}$ are normal, and we only consider clustering on the first mode.

We consider two TNMMs with the TBM structure. We consider $\mathbf{X}_i \in \mathbb{R}^{20 \times 20}$ for $i = 1, \ldots, 150$. The 150 observations are evenly split into two clusters. We assume that $\boldsymbol{\mu}^*_1 = 0$ and $\boldsymbol{\mu}^*_2$

	Optimal	K-means	SKM	DEEM	DTC	TBM	EM	AFPF	CHIME
TBM (a)	8.12 (0.20)	36.20 (0.87)	29.07 (1.18)	14.18 (1.11)	44.53 (0.44)	8.53 (0.29)	32.23 (0.84)	33.86 (0.82)	44.74 (0.42)
TBM (b)	7.74 (0.22)	35.70 (0.91)	23.13 (1.46)	12.04 (0.74)	43.64 (0.53)	10.74 (0.38)	31.78 (0.84)	32.82 (0.91)	44.79 (0.41)

Table S.2: Clustering results for TBMs (a) & (b). Reported are the means and standard errors (in parentheses) of clustering error rates based on 100 replicates.

is piecewise constant so that the model is a special case of TBM when we stack the observations $\mathbf{X}_i$ to form $\mathcal{X} \in \mathbb{R}^{150\times 20\times 20}$. We consider the following two settings for $\boldsymbol{\mu}_2^*$.

TBM (a): $\boldsymbol{\mu}^*_{2,[1:8,1]} = 1$ and all other entries are zero.

TBM (b): $\boldsymbol{\mu}^*_{2,[1:3,1]} = 1.5$, $\boldsymbol{\mu}^*_{2,[4:6,1]} = -0.5$, $\boldsymbol{\mu}^*_{2,[7:8,1]} = 0.5$ and all other entries are zero.

We continue to compare DEEM with all the competitors in Section 5.1 on these two models. The results are reported in Table S.2. Not surprisingly, the method TBM has the best performance on these two models, since it exploits all the model assumptions, such as the independence among variables and the piecewise-constant structure of $\boldsymbol{\mu}_k^*$. However, among all the other methods that do not assume TBM structure, DEEM is an obvious winner. It outperforms all the other methods by a large margin that assume independence among variables. It is slightly worse than TBM, likely because it does not use the information that the means are piece-wise constant, which greatly reduces the model complexity. Hence, it will be interesting in the future to extend DEEM to clustering problems with additional clustering structure beyond TNMM, such as TBM.

E Proofs of Lemmas 1, 2 and 3

We first list some simple propositions that will be repeatedly used.

Proposition S.2 (Kolda & Bader (2009) c.f Page 475)**.** *For any tensor* $\mathbf{C} \in \mathbb{R}^{p_1\times\cdots\times p_M}$ *and matrices* $\mathbf{A}_m \in \mathbb{R}^{p_m\times p_m}$*, we have that*

$$[\![\mathbf{C}; \mathbf{A}_1, \ldots \mathbf{A}_M]\!]_{(m)} = \mathbf{A}_m \mathbf{C}_{(m)} (\otimes_{j=M, j\neq m}^{j=1} \mathbf{A}_j). \tag{S.4}$$

Proposition S.3. *For matrices* $\boldsymbol{\Omega}_m, \widetilde{\boldsymbol{\Omega}}_m, m = 1, \ldots, M$*, we have*

$$\otimes_{m=M}^{m=1} \widetilde{\boldsymbol{\Omega}}_m - \otimes_{m=M}^{m=1} \boldsymbol{\Omega}_m = \sum_{j=1}^{M} \{\otimes_{m=M}^{m=j+1} \widetilde{\boldsymbol{\Omega}}_m\} \otimes \{\widetilde{\boldsymbol{\Omega}}_j - \boldsymbol{\Omega}_j\} \otimes \{\otimes_{m=j-1}^{m=1} \boldsymbol{\Omega}_m\}. \tag{S.5}$$

Proof of Proposition S.3. The desired conclusion can be proved by straightforward calculation. □

Proof of Lemma 1. The Q-function in (3.3) is equal to

$$Q_n(\boldsymbol{\theta} \mid \widetilde{\boldsymbol{\theta}}^{(t)}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\widetilde{\xi}_{ik}^{(t)}\{\log\pi_k - \frac{1}{2}\sum_{m=1}^{M} q_m \log|\boldsymbol{\Sigma}_m| - \frac{1}{2}\langle[\![\mathbf{X}_i-\boldsymbol{\mu}_k; \boldsymbol{\Sigma}_1^{-1},\ldots,\boldsymbol{\Sigma}_M^{-1}]\!], \mathbf{X}_i-\boldsymbol{\mu}_k\rangle\}. \tag{S.6}$$

By Proposition S.2 we have

$$[\![\mathbf{X}_i-\boldsymbol{\mu}_k; \boldsymbol{\Sigma}_1^{-1},\ldots,\boldsymbol{\Sigma}_M^{-1}]\!]_{(m)} = \boldsymbol{\Sigma}_m^{-1}(\mathbf{X}_i-\boldsymbol{\mu}_k)_{(m)}\boldsymbol{\Sigma}_{-m}^{-1}. \tag{S.7}$$

Hence,

$$\begin{aligned}
&\langle[\![\mathbf{X}_i-\boldsymbol{\mu}_k; \boldsymbol{\Sigma}_1^{-1},\ldots,\boldsymbol{\Sigma}_M^{-1}]\!], \mathbf{X}_i-\boldsymbol{\mu}_k\rangle && \text{(S.8)}\\
=\ &\langle[\![\mathbf{X}_i-\boldsymbol{\mu}_k; \boldsymbol{\Sigma}_1^{-1},\ldots,\boldsymbol{\Sigma}_M^{-1}]\!]_{(m)}, (\mathbf{X}_i-\boldsymbol{\mu}_k)_{(m)}\rangle && \text{(S.9)}\\
=\ &\mathrm{tr}(\boldsymbol{\Sigma}_{-m}^{-1}(\mathbf{X}_i-\boldsymbol{\mu}_k)_{(m)}^{\intercal}\boldsymbol{\Sigma}_m^{-1}(\mathbf{X}_i-\boldsymbol{\mu}_k)_{(m)}) && \text{(S.10)}\\
=\ &\mathrm{tr}\left(\boldsymbol{\Sigma}_m^{-1}(\mathbf{X}_i-\boldsymbol{\mu}_k)_{(m)}\boldsymbol{\Sigma}_{-m}^{-1}(\mathbf{X}_i-\boldsymbol{\mu}_k)_{(m)}^{\intercal}\right). && \text{(S.11)}
\end{aligned}$$

Consequently, when we fix all other parameters except for $\boldsymbol{\Sigma}_m$, the Q-function reduces to

$$\begin{aligned}
&Q_n^m(\boldsymbol{\Sigma}_m \mid \widetilde{\boldsymbol{\theta}}^{(t)})\\
=\ &Const - \frac{q_m}{2}\log|\boldsymbol{\Sigma}_m|\cdot\sum_{i=1}^{n}\sum_{k=1}^{K}\widetilde{\xi}_{ik}^{(t)} - \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\widetilde{\xi}_{ik}^{(t)}\mathrm{tr}\{\boldsymbol{\Sigma}_m^{-1}(\mathbf{X}_i-\boldsymbol{\mu}_k)_{(m)}\boldsymbol{\Sigma}_{-m}^{-1}(\mathbf{X}_i-\boldsymbol{\mu}_k)_{(m)}^{\intercal}\}\\
=\ &Const - \frac{nq_m}{2}\log|\boldsymbol{\Sigma}_m| - \frac{1}{2}\mathrm{tr}[\boldsymbol{\Sigma}_m^{-1}\{\sum_{i=1}^{n}\sum_{k=1}^{K}\widetilde{\xi}_{ik}^{(t)}(\mathbf{X}_i-\boldsymbol{\mu}_k)_{(m)}\boldsymbol{\Sigma}_{-m}^{-1}(\mathbf{X}_i-\boldsymbol{\mu}_k)_{(m)}^{\intercal}\}],
\end{aligned}$$

where we use the fact that $\sum_{k=1}^{K}\widetilde{\xi}_{ik}^{(t)} = 1$ for any i.

By Petersen & Pedersen (2008), we have

$$\frac{\partial Q_n^m(\boldsymbol{\Sigma}_m \mid \widetilde{\boldsymbol{\theta}}^{(t)})}{\partial\boldsymbol{\Sigma}_m} = -\frac{nq_m}{2}\cdot\frac{1}{|\boldsymbol{\Sigma}_m|}\cdot|\boldsymbol{\Sigma}_m|\cdot\boldsymbol{\Sigma}_m - \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\widetilde{\xi}_{ik}^{(t)}(\mathbf{X}_i-\boldsymbol{\mu}_k)_{(m)}\boldsymbol{\Sigma}_{-m}^{-1}(\mathbf{X}_i-\boldsymbol{\mu}_k)_{(m)}^{\intercal}. \tag{S.12}$$

Set $\dfrac{\partial Q_n^m(\boldsymbol{\Sigma}_m \mid \widetilde{\boldsymbol{\theta}}^{(t)})}{\partial\boldsymbol{\Sigma}_m} = 0$ and we have

$$\boldsymbol{\Sigma}_m = \frac{1}{nq_m}\sum_{i=1}^{n}\sum_{k=1}^{K}\widetilde{\xi}_{ik}^{(t)}(\mathbf{X}_i-\boldsymbol{\mu}_k)_{(m)}\boldsymbol{\Sigma}_{-m}^{-1}(\mathbf{X}_i-\boldsymbol{\mu}_k)_{(m)}^{\intercal}. \tag{S.13}$$

By noting that $\mathbf{W}_{ik} = (\mathbf{X}_i-\boldsymbol{\mu}_k)_{(m)}\boldsymbol{\Sigma}_{-m}^{-1/2}$ we have the desired conclusion. □

Proof of Lemma 2. Recall that

$$\xi_{ik} = \Pr(Y_i = k \mid \mathbf{X}_i, \boldsymbol{\theta}^*) = \frac{\pi_k^* f_k(\mathbf{X}_i; \boldsymbol{\theta}^*)}{\sum_{j=1}^K \pi_j^* f_j(\mathbf{X}_i; \boldsymbol{\theta}^*)}, \tag{S.14}$$

where the density function f_k is given in (3.2). It follows that

$$\begin{aligned}
&\frac{\xi_{ik}}{\xi_{i1}} = \frac{\pi_k^* f_k(\mathbf{X}_i; \boldsymbol{\theta}^*)}{\pi_1^* f_1(\mathbf{X}_i; \boldsymbol{\theta}^*)} \\
= \;& \frac{\pi_k^*}{\pi_1^*} \exp\{-\frac{1}{2}(\langle [\![\mathbf{X}_i - \boldsymbol{\mu}_k^*; (\boldsymbol{\Sigma}_1^*)^{-1}, \ldots, (\boldsymbol{\Sigma}_M^*)^{-1}]\!], \mathbf{X}_i - \boldsymbol{\mu}_k^* \rangle \\
& - \langle [\![\mathbf{X}_i - \boldsymbol{\mu}_1^*; (\boldsymbol{\Sigma}^*)_1^{-1}, \ldots, (\boldsymbol{\Sigma}^*)_M^{-1}]\!], \mathbf{X}_i - \boldsymbol{\mu}_1^* \rangle)\}.
\end{aligned}$$

Note that $\langle [\![\mathbf{X}_i - \boldsymbol{\mu}_k^*; (\boldsymbol{\Sigma}^*)_1^{-1}, \ldots, (\boldsymbol{\Sigma}^*)_M^{-1}]\!], \mathbf{X}_i - \boldsymbol{\mu}_k^* \rangle = \mathrm{vec}^\intercal(\mathbf{X}_i - \boldsymbol{\mu}_k^*)(\boldsymbol{\Sigma}^*)^{-1}\mathrm{vec}(\mathbf{X}_i - \boldsymbol{\mu}_k^*)$, $\langle [\![\mathbf{X}_i - \boldsymbol{\mu}_1^*; (\boldsymbol{\Sigma}_1^*)^{-1}, \ldots, (\boldsymbol{\Sigma}_M^*)^{-1}]\!], \mathbf{X}_i - \boldsymbol{\mu}_1^* \rangle = \mathrm{vec}^\intercal(\mathbf{X}_i - \boldsymbol{\mu}_1^*)(\boldsymbol{\Sigma}^*)^{-1}\mathrm{vec}(\mathbf{X}_i - \boldsymbol{\mu}_1^*)$, where $\boldsymbol{\Sigma}^* = \otimes_{m=M}^{m=1} \boldsymbol{\Sigma}_m^*$. It follows that

$$\begin{aligned}
& \langle [\![\mathbf{X}_i - \boldsymbol{\mu}_k^*; (\boldsymbol{\Sigma}_1^*)^{-1}, \ldots, (\boldsymbol{\Sigma}_M^*)^{-1}]\!], \mathbf{X}_i - \boldsymbol{\mu}_k^* \rangle - \langle [\![\mathbf{X}_i - \boldsymbol{\mu}_1^*; (\boldsymbol{\Sigma}_1^*)^{-1}, \ldots, (\boldsymbol{\Sigma}_M^*)^{-1}]\!], \mathbf{X}_i - \boldsymbol{\mu}_1^* \rangle \\
= \;& \mathrm{vec}^\intercal(\mathbf{X}_i)(\boldsymbol{\Sigma}^*)^{-1}\mathrm{vec}(\mathbf{X}_i) - 2\mathrm{vec}^\intercal(\mathbf{X}_i)(\boldsymbol{\Sigma}^*)^{-1}\mathrm{vec}(\boldsymbol{\mu}_k^*) + \mathrm{vec}^\intercal(\boldsymbol{\mu}_k^*)(\boldsymbol{\Sigma}^*)^{-1}\mathrm{vec}(\boldsymbol{\mu}_k^*) \\
& - \mathrm{vec}^\intercal(\mathbf{X}_i)(\boldsymbol{\Sigma}^*)^{-1}\mathrm{vec}(\mathbf{X}_i) + 2\mathrm{vec}^\intercal(\mathbf{X}_i)(\boldsymbol{\Sigma}^*)^{-1}\mathrm{vec}(\boldsymbol{\mu}_k^*) - \mathrm{vec}^\intercal(\boldsymbol{\mu}_k^*)(\boldsymbol{\Sigma}^*)^{-1}\mathrm{vec}(\boldsymbol{\mu}_k^*) \\
= \;& -2\mathrm{vec}^\intercal\{\mathbf{X}_i - \frac{1}{2}(\boldsymbol{\mu}_k^* + \boldsymbol{\mu}_1^*)\}(\boldsymbol{\Sigma}^*)^{-1}\mathrm{vec}(\boldsymbol{\mu}_k^* - \boldsymbol{\mu}_1^*) \\
= \;& -2\langle \mathbf{X}_i - \frac{1}{2}(\boldsymbol{\mu}_k^* + \boldsymbol{\mu}_1^*), \mathbf{B}_k^* \rangle.
\end{aligned}$$

Hence,

$$\frac{\xi_{ik}}{\xi_{i1}} = \frac{\pi_k^*}{\pi_1^*} \exp\{\langle \mathbf{X}_i - \frac{1}{2}(\boldsymbol{\mu}_k^* + \boldsymbol{\mu}_1^*), \mathbf{B}_k^* \rangle\}, \quad k > 1. \tag{S.15}$$

Solve the equation system in (S.15) along with $\sum_{k=1}^K \xi_{ik} = 1$ and we have the desired conclusion.

□

To prove Lemma 3, we need the following proposition.

Proposition S.4. *[Pan et al. (2019) c.f. Lemma 2] If an M-way random tensor $\mathbf{W} \sim TN(0; \boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_M)$, then $\mathrm{E}\{\mathbf{W}_{(j)}\mathbf{W}_{(j)}^\intercal\} = \boldsymbol{\Omega}_j \cdot \prod_{l \neq j} \mathrm{tr}(\boldsymbol{\Omega}_l)$, where $\mathbf{W}_{(j)}$ is the mode-j matricization of $\mathbf{W}$.*

Now we prove Lemma 3.

Proof of Lemma 3. For any k, we have that

$$\begin{aligned}
& \frac{1}{q_m} E[\xi_{ik}\{\mathbf{X}_i - \boldsymbol{\mu}_k^*\}_{(m)}\{\mathbf{X}_i - \boldsymbol{\mu}_k^*\}_{(m)}^\intercal] && \text{(S.16)}\\
= & \frac{1}{q_m} E[\Pr(Y_i = k \mid \mathbf{X}_i, \boldsymbol{\theta}^*)(\mathbf{X}_i - \boldsymbol{\mu}_k^*)_{(m)}\{\mathbf{X}_i - \boldsymbol{\mu}_k^*\}_{(m)}^\intercal] && \text{(S.17)}\\
= & \frac{1}{q_m} E[E[\mathbf{1}(Y_i = k) \mid \mathbf{X}_i, \boldsymbol{\theta}^*](\mathbf{X}_i - \boldsymbol{\mu}_k^*)_{(m)}\{\mathbf{X}_i - \boldsymbol{\mu}_k^*\}_{(m)}^\intercal] && \text{(S.18)}\\
= & \frac{1}{q_m} E[E[\mathbf{1}(Y_i = k)(\mathbf{X}_i - \boldsymbol{\mu}_k^*)_{(m)}\{\mathbf{X}_i - \boldsymbol{\mu}_k^*\}_{(m)}^\intercal \mid \mathbf{X}_i, \boldsymbol{\theta}^*]] && \text{(S.19)}\\
= & \frac{1}{q_m} E_{\boldsymbol{\theta}^*}[\mathbf{1}(Y = k)(\mathbf{X}_i - \boldsymbol{\mu}_k^*)_{(m)}\{\mathbf{X}_i - \boldsymbol{\mu}_k^*\}_{(m)}^\intercal] && \text{(S.20)}\\
\propto & \pi_k^* \boldsymbol{\Sigma}_m^*, && \text{(S.21)}
\end{aligned}$$

where in the last equation we use Proposition S.4. Hence, $\frac{1}{q_m} E[\sum_{k=1}^K \xi_{ik}\{\mathbf{X}_i - \boldsymbol{\mu}_k^*\}_{(m)}\{\mathbf{X}_i - \boldsymbol{\mu}_k^*\}_{(m)}^\intercal] \propto \boldsymbol{\Sigma}_m^*$. □

F Proofs of Theorems 1 and 2

F.1 Notations and propositions

Some of the following notations are already defined in the main body of the paper. For better readability, we repeat them here. For two numbers $x, y \in \mathbb{R}$, we denote $x \wedge y = \min\{x, y\}, x \vee y = \max\{x, y\}$. For a vector $\mathbf{x} \in \mathbb{R}^p$, we denote

$$\|\mathbf{x}\|_0 = \sum_{j=1}^p 1(x_j \neq 0), \|\mathbf{x}\|_1 = \sum_{j=1}^p |x_j|, \|\mathbf{x}\|_2 = \sqrt{\sum_{j=1}^p x_j^2}, \|\mathbf{x}\|_\infty = \max_{j=1}^p |x_j|. \tag{S.22}$$

For a matrix $\mathbf{X} \in \mathbb{R}^{p\times q}$, we let $\|\mathbf{X}\|_1 = \max_{i=1}^p\{\sum_{j=1}^q |x_{ij}|\}$ and $\|\mathbf{X}\|_{\max} = \max_{i,j} |x_{ij}|$. For any positive integer $j \leq q$, $\mathbf{X}_{\cdot j}$ is the j-th column of $\mathbf{X}$. If $\mathbf{X}$ is symmetric, $\lambda_{\min}(\mathbf{X})$ is its smallest eigenvalue, and $\lambda_{\max}(\mathbf{X})$ is its largest eigenvalue. For a vector $\mathbf{u} \in \mathbb{R}^q$ and a matrix $\mathbf{A} \in \mathbb{R}^{p\times q}$, we denote

$$\|\mathbf{u}\|_{2,s} = \sup_{\|\mathbf{x}\|_2=1, \mathbf{x}\in\Gamma(s)} |\mathbf{u}^\intercal \mathbf{x}|, \quad \|\mathbf{A}\|_{2,s} = \sup_{\|\mathbf{x}\|_2=1, \mathbf{x}\in\Gamma(s)} \|\mathbf{A}\mathbf{x}\|_2. \tag{S.23}$$

For an M-way tensor $\mathbf{X} \in \mathbb{R}^{p_1\times p_2\times\cdots\times p_M}$, we denote $\|\mathbf{X}\| = \sqrt{\sum_{\mathcal{J}} X_{\mathcal{J}}^2}$ and $\|\mathbf{X}\|_0 = \sum_{\mathcal{J}} 1(X_{\mathcal{J}} \neq 0)$. We also need the following two propositions. We include the proof for the first one for completeness.

Proposition S.5. *For matrices $\mathbf{A} \in \mathbb{R}^{p\times p}, \mathbf{B} \in \mathbb{R}^{q\times q}$, we have $\|\mathbf{A}\otimes\mathbf{B}\|_1 \le \|\mathbf{A}\|_1 \cdot \|\mathbf{B}\|_1$.*

Proof of Proposition S.5. For any $j = 1, \ldots, pq$, there exist $l \in \{1, \ldots, p\}, t \in \{1, \ldots, q\}$ such that $(\mathbf{A}\otimes\mathbf{B})_{\cdot j} = (a_{1l}\mathbf{B}^\intercal_{\cdot t}, a_{2l}\mathbf{B}^\intercal_{\cdot t}, \ldots, a_{pl}\mathbf{B}^\intercal_{\cdot t})^\intercal$. It follows that

$$\begin{aligned}\|\mathbf{A}\otimes\mathbf{B}\|_1 &= \max_j \|(\mathbf{A}\otimes\mathbf{B})_{\cdot j}\|_1 \le \max_{l,t}\{|a_{1l}|\|\mathbf{B}_{\cdot t}\|_1 + |a_{2l}|\|\mathbf{B}_{\cdot t}\|_1 + \cdots + |a_{pl}|\|\mathbf{B}_{\cdot t}\|_1\} \\ &\le \max_l(\sum_v |a_{vl}|)\max_t \|\mathbf{B}_{\cdot t}\|_1 \le \|\mathbf{A}\|_1 \cdot \|\mathbf{B}\|_1.\end{aligned}$$

□

Proposition S.6 (Lyu et al. (2019) c.f Lemma S.1)**.** *Consider random matrices $\mathbf{W}_1, \ldots, \mathbf{W}_n \in \mathbb{R}^{p\times q}$ iid from $TN(0; \mathbf{\Omega}_R, \mathbf{\Omega}_L)$. If the eigenvalues of $\mathbf{\Omega}_R, \mathbf{\Omega}_L$ are universally bounded below and above, then for any symmetric and positive definite matrix $\mathbf{\Omega} \in \mathbb{R}^{q\times q}$, we have*

$$\|\frac{1}{nq}\sum_{i=1}^n \mathbf{W}_i\mathbf{\Omega}(\mathbf{W}_i)^\intercal - \frac{1}{q}E\mathbf{W}_i\mathbf{\Omega}(\mathbf{W}_i)^\intercal\|_{\max} = O_P(\sqrt{\frac{\log q}{np}}). \tag{S.24}$$

F.2 Properties for a generic penalized EM algorithm

We introduce a generic penalized EM algorithm (PEM) that includes DEEM as a special case. Then we present a proposition to guarantee its consistency under proper conditions. The proposition is later used to prove Theorems 1 & 2. The proofs for our proposition are closely related to those in Cai et al. (2019), but we reach a more general conclusion. We also list some properties related to the initialization algorithm that will be used to prove Lemmas S.8 & 4.

Consider iid observations $\{\mathbf{U}_i, Y_i\}_{i=1}^n$, where $\mathbf{U}_i \in \mathbb{R}^p$ and $Y_i \in \{1, 2\}$ is unobservable. We make the following model assumption for $\{\mathbf{U}_i, Y_i\}$:

(A1') $\Pr(Y_i = k) = \pi_k^*, \mathbf{U}_i \mid Y_i = k \sim N(\boldsymbol{\phi}_k^*, \mathbf{\Psi}^*)$, where $0 < \pi_k^* < 1$, $\boldsymbol{\phi}_k^* \in \mathbb{R}^p$, $\mathbf{\Psi}^* \in \mathbb{R}^{p\times p}$.

Note that Model (A1') is equivalent to Model (A1) in Section G. We consider a generic PEM to recover Y_i. Denote $\boldsymbol{\theta}^{\text{vec}} = (\pi_k, \boldsymbol{\phi}_k, k = 1, 2; \mathbf{\Psi})$, $\xi_{ik}(\boldsymbol{\theta}^{\text{vec}}) = \Pr(Y_i = k \mid \mathbf{U}_i, \boldsymbol{\theta}^{\text{vec}})$ and $\boldsymbol{\xi}(\boldsymbol{\theta}^{\text{vec}}) \in \mathbb{R}^{n\times 2}$ with the i-th row being $(\xi_{i1}(\boldsymbol{\theta}^{\text{vec}}), \xi_{i2}(\boldsymbol{\theta}^{\text{vec}}))$. Given $\boldsymbol{\xi}(\boldsymbol{\theta}^{\text{vec}})$, we estimate the parameters in the M-step by

$$\widehat{\pi}_k(\boldsymbol{\theta}^{\text{vec}}) = g_{\pi_k}\{\boldsymbol{\xi}(\boldsymbol{\theta}^{\text{vec}}), \mathbf{U}\}, \widehat{\boldsymbol{\phi}}_k(\boldsymbol{\theta}^{\text{vec}}) = g_{\boldsymbol{\phi}_k}\{\boldsymbol{\xi}(\boldsymbol{\theta}^{\text{vec}}), \mathbf{U}\}, \widehat{\mathbf{\Psi}}(\boldsymbol{\theta}^{\text{vec}}) = g_{\mathbf{\Psi}}\{\boldsymbol{\xi}(\boldsymbol{\theta}^{\text{vec}}), \mathbf{U}\}, \tag{S.25}$$

where

$$g_{\pi_k}:\mathbb{R}^{n\times 2}\times\mathbb{R}^{n\times p}\mapsto\mathbb{R};\quad g_{\boldsymbol{\phi}_k}:\mathbb{R}^{n\times 2}\times\mathbb{R}^{n\times p}\mapsto\mathbb{R}^p;\quad g_{\boldsymbol{\Psi}}:\mathbb{R}^{n\times 2}\times\mathbb{R}^{n\times p}\mapsto\mathbb{R}^{p\times p} \tag{S.26}$$

are user-specified functions. We further use the shorthand notation

$$\begin{aligned} & g\{\boldsymbol{\xi}(\boldsymbol{\theta}^{\mathrm{vec}}),\mathbf{U}\}=\{\hat{\pi}_k(\boldsymbol{\theta}^{\mathrm{vec}}),\hat{\boldsymbol{\phi}}_k(\boldsymbol{\theta}^{\mathrm{vec}}),k=1,2;\widehat{\boldsymbol{\Psi}}(\boldsymbol{\theta}^{\mathrm{vec}})\} \\ = \;& [g_{\pi_k}\{\boldsymbol{\xi}(\boldsymbol{\theta}^{\mathrm{vec}}),\mathbf{U}\},g_{\boldsymbol{\phi}_k}\{\boldsymbol{\xi}(\boldsymbol{\theta}^{\mathrm{vec}}),\mathbf{U}\},k=1,2;g_{\boldsymbol{\Psi}}\{\boldsymbol{\xi}(\boldsymbol{\theta}^{\mathrm{vec}}),\mathbf{U}\}]. \end{aligned}$$

On the other hand, in the E-step, given the value of $\boldsymbol{\theta}^{\mathrm{vec}}=(\pi_k,\boldsymbol{\phi}_k,k=1,2;\boldsymbol{\Psi})$, $\xi_{ik}(\boldsymbol{\theta}^{\mathrm{vec}})$ is estimated by projecting $\mathbf{U}_i$ to $\boldsymbol{\beta}=\boldsymbol{\Psi}^{-1}(\boldsymbol{\phi}_2-\boldsymbol{\phi}_1)$. We obtain a sparse estimate for $\boldsymbol{\beta}$ through a penalized optimization problem. Such a PEM algorithm is summarized in Algorithm S.3. Note that PEM is generic since we do not restrict g to a specific form.

To derive the theoretical properties of the PEM algorithm, consider the parameter space

$$\begin{aligned} \boldsymbol{\Theta}^{\mathrm{vec}}(s,c_\pi,C_0,C_b,\Delta_0) \;=\; & \{\boldsymbol{\theta}^{\mathrm{vec}}=(\pi_k,\boldsymbol{\phi}_k,k=1,2;\boldsymbol{\Psi}): \\ & \boldsymbol{\phi}_k\in\mathbb{R}^p,\boldsymbol{\Psi}\in\mathbb{R}^p,C_0^{-1}\le\lambda_{\min}(\boldsymbol{\Psi})\le\lambda_{\max}(\boldsymbol{\Psi})\le C_0, \\ & \|\boldsymbol{\beta}\|_0\le s,\|\boldsymbol{\beta}\|_1\le C_b,c_\pi\le\pi_1,\pi_2\le 1-c_\pi,\Delta>\Delta_0\}, \end{aligned}$$

where s is a positive integer, $0<c_\pi<1,C_0,C_b>0$ and

$$\Delta=\sqrt{(\boldsymbol{\phi}_1-\boldsymbol{\phi}_2)^\intercal(\boldsymbol{\Psi})^{-1}(\boldsymbol{\phi}_1-\boldsymbol{\phi}_2)}.$$

Recall $\Gamma(s)$ defined in (4.2). For the true parameter $\boldsymbol{\theta}^{\mathrm{vec}*}\in\boldsymbol{\Theta}^{\mathrm{vec}}$, we define $\Delta^*=\sqrt{(\boldsymbol{\phi}_1^*-\boldsymbol{\phi}_2^*)^\intercal(\boldsymbol{\Psi}^*)^{-1}(\boldsymbol{\phi}_1^*-\boldsymbol{\phi}_2^*)}$ and the contraction basin of $\boldsymbol{\theta}^{\mathrm{vec}*}$ to be

$$\begin{aligned} \mathcal{B}_{con}(\boldsymbol{\theta}^{\mathrm{vec}*};a_\pi,a_\Delta,a_b,s)= \quad & \{\boldsymbol{\theta}^{\mathrm{vec}}=(\pi_k,\boldsymbol{\phi}_k,k=1,2;\boldsymbol{\Psi}): \\ & \boldsymbol{\phi}_k\in\mathbb{R}^p,\boldsymbol{\Psi}\in\mathbb{R}^{p\times p},\boldsymbol{\Psi}\succ 0,a_\pi<\pi_k<1-a_\pi, \\ & (1-a_\Delta)(\Delta^*)^2<|\delta_k(\boldsymbol{\beta})|,\sigma^2(\boldsymbol{\beta})<(1+a_\Delta)(\Delta^*)^2, \\ & \boldsymbol{\beta}-\boldsymbol{\beta}^*\in\Gamma(s),\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_1\le a_b\Delta^*,\|\boldsymbol{\phi}_k\|_{2,s}\le a_b\Delta^*\}, \end{aligned}$$

where $a_\pi,a_\Delta,a_b>0$, $a_\pi\le c_\pi<1$, $\delta_k(\boldsymbol{\beta})=\boldsymbol{\beta}^\intercal(\boldsymbol{\phi}_k^*-\frac{\boldsymbol{\phi}_1+\boldsymbol{\phi}_2}{2})$ and $\sigma^2(\boldsymbol{\beta})=\boldsymbol{\beta}^\intercal\boldsymbol{\Psi}^*\boldsymbol{\beta}$.

Now we introduce a few assumptions under which PEM is consistent. Define the distances:

$$d_2(\boldsymbol{\theta}^{\mathrm{vec}},\widetilde{\boldsymbol{\theta}}^{\mathrm{vec}}) = (\vee_{k=1}^2|\pi_k-\tilde{\pi}_k|)\vee(\vee_{k=1}^2\|\boldsymbol{\phi}_k-\tilde{\boldsymbol{\phi}}_k\|_2)\vee\|(\boldsymbol{\Psi}-\tilde{\boldsymbol{\Psi}})\tilde{\boldsymbol{\beta}}\|_2, \tag{S.32}$$

$$d_{2,s}(\boldsymbol{\theta}^{\mathrm{vec}},\widetilde{\boldsymbol{\theta}}^{\mathrm{vec}}) = (\vee_{k=1}^2|\pi_k-\tilde{\pi}_k|)\vee(\vee_{k=1}^2\|\boldsymbol{\phi}_k-\tilde{\boldsymbol{\phi}}_k\|_{2,s})\vee\|(\boldsymbol{\Psi}-\tilde{\boldsymbol{\Psi}})\tilde{\boldsymbol{\beta}}\|_{2,s}, \tag{S.33}$$

$$d_{2,s}^{(1)}(\boldsymbol{\theta}^{\mathrm{vec}},\widetilde{\boldsymbol{\theta}}^{\mathrm{vec}}) = (\vee_{k=1}^2|\pi_k-\tilde{\pi}_k|)\vee(\vee_{k=1}^2\|\boldsymbol{\phi}_k-\tilde{\boldsymbol{\phi}}_k\|_{2,s}). \tag{S.34}$$

Algorithm S.3 A generic PEM algorithm

1. Initialize $\theta^{\mathrm{vec}(0)}$, $0 < \kappa < 1$. Set

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\mathbb{R}^p}\{\frac{1}{2}\boldsymbol{\beta}^\intercal\widehat{\boldsymbol{\Psi}}\boldsymbol{\beta} - \boldsymbol{\beta}^\intercal(\widehat{\boldsymbol{\phi}}_2^{(0)} - \widehat{\boldsymbol{\phi}}_1^{(0)}) + \lambda^{(0)}\|\boldsymbol{\beta}\|_1\}, \tag{S.27}$$

   where $\lambda_n^{(0)} = C_d(|\hat{\pi}_2| \vee \|\widehat{\boldsymbol{\phi}}_1^{(0)} - \widehat{\boldsymbol{\phi}}_2^{(0)}\|_{2,s} \vee \|\widehat{\boldsymbol{\Psi}}^{(0)}\|_{2,s}) + C_\lambda\sqrt{\log p/n}$.

2. For $t = 0, \ldots,$ do

   (a) Update

$$(\hat{\pi}_k^{(t+1)}, \hat{\boldsymbol{\phi}}_k^{(t+1)}, k = 1, 2; \widehat{\boldsymbol{\Psi}}^{(t+1)}) = g(\boldsymbol{\xi}^{(t)}, \mathbf{U}); \tag{S.28}$$

   (b) Update $\hat{\xi}_{i1}^{(t+1)}, \hat{\xi}_{i1}^{(t+1)}$ by

$$\widehat{\xi}_{i1}^{(t+1)} = \frac{\pi_1^{(t+1)}}{\widehat{\pi}_1^{(t+1)} + \widehat{\pi}_2^{(t+1)}\exp[\{\mathbf{U}_i - (\widehat{\boldsymbol{\phi}}_1^{(t+1)} + \widehat{\boldsymbol{\phi}}_2^{(t+1)})/2\}^\intercal\widehat{\boldsymbol{\beta}}^{(t+1)}]}, \quad \hat{\xi}_{i2}^{(t+1)} = 1 - \hat{\xi}_{i1}^{(t+1)}, \tag{S.29}$$

   where

$$\widehat{\boldsymbol{\beta}}^{(t+1)} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\{\boldsymbol{\beta}^\intercal\widehat{\boldsymbol{\Psi}}^{(t+1)}\boldsymbol{\beta} - 2(\widehat{\boldsymbol{\phi}}_2^{(t+1)} - \widehat{\boldsymbol{\phi}}_1^{(t+1)})^\intercal\boldsymbol{\beta} + \lambda^{(t+1)}\|\boldsymbol{\beta}\|_1\}, \tag{S.30}$$

   with

$$\lambda^{(t+1)} = \kappa\lambda^{(t)} + C_\lambda\sqrt{\frac{\log p}{n}}. \tag{S.31}$$

3. Output $\hat{\xi}_{ik}$ at convergence.

Ideally, we hope that $g(\boldsymbol{\xi}, \mathbf{U})$ converges in probability as $n \to \infty$. We formally write this assumption as follows:

(A2) There exists $g^{lim}(\boldsymbol{\theta}^{\mathrm{vec}}) = \{\pi_k(\boldsymbol{\theta}^{\mathrm{vec}}), \boldsymbol{\phi}_k(\boldsymbol{\theta}^{\mathrm{vec}}), k = 1, 2; \boldsymbol{\Psi}(\boldsymbol{\theta}^{\mathrm{vec}})\}$ such that $|g_{\pi_k}\{\boldsymbol{\xi}(\boldsymbol{\theta}), \mathbf{U}\} - \pi_k(\boldsymbol{\theta}^{\mathrm{vec}})|$, $\|g_{\boldsymbol{\phi}_k}\{\boldsymbol{\xi}(\boldsymbol{\theta}), \mathbf{U}\} - \boldsymbol{\phi}_k(\boldsymbol{\theta}^{\mathrm{vec}})\|_{\max}$, $\|g_{\boldsymbol{\Psi}}\{\boldsymbol{\xi}(\boldsymbol{\theta}^{\mathrm{vec}}), \mathbf{U}\} - \boldsymbol{\Psi}(\boldsymbol{\theta}^{\mathrm{vec}})\|_{\max} \to 0$ in probability.

Under (A2), we further make the following assumptions concerning g and g^{lim}. For simplicity, we write the value of our estimate at $(t+1)$-th iteration, i.e., $g(\boldsymbol{\xi}(\widehat{\boldsymbol{\theta}}^{\mathrm{vec}(t+1)}), \mathbf{U})$, as $\widehat{\boldsymbol{\theta}}^{\mathrm{vec}(t+1)}$, and its limit $g^{lim}(\widehat{\boldsymbol{\theta}}^{\mathrm{vec}(t+1)})$ as $\boldsymbol{\theta}^{\mathrm{vec}(t+1)}$.

(A3) With a probability tending to 1, we have

$$\sup_{\boldsymbol{\theta}\in\mathcal{B}_{con}} d_{2,s}(\hat{\boldsymbol{\theta}}^{\mathrm{vec}(t+1)}, \boldsymbol{\theta}^{\mathrm{vec}(t+1)}) \leq C_{con}\sqrt{\frac{s\log p}{n}}, \tag{S.35}$$

$$\sup_{\boldsymbol{\theta}\in\mathcal{B}_{con}} \|(\widehat{\boldsymbol{\Psi}}^{(t+1)} - \boldsymbol{\Psi}^{(t+1)})\boldsymbol{\beta}^{\mathrm{vec}*}\|_\infty \leq C_{con}\sqrt{\frac{\log p}{n}}. \tag{S.36}$$

(A4) There exists a sufficiently large Δ_0 and $0 < \kappa_0 < \frac{1}{2\vee(16C_0)}$ that does not depend on n, p, such that for any $\boldsymbol{\theta}^{\mathrm{vec}} \in \mathcal{B}_{con}$ we have

$$d_{2,s}[g^{lim}\{\boldsymbol{\theta}\}, \boldsymbol{\theta}^{\mathrm{vec}*}] \leq \kappa_0 \cdot (d_{2,s}(\boldsymbol{\theta}^{\mathrm{vec}}, \boldsymbol{\theta}^{\mathrm{vec}*}) \vee \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2), \tag{S.37}$$

where g^{lim} is defined as in (A2).

(A5) For the initial value $\widehat{\boldsymbol{\theta}}^{\mathrm{vec}(0)}$, we have

$$d_{2,s}(\widehat{\boldsymbol{\theta}}^{\mathrm{vec}(0)}, \boldsymbol{\theta}^{\mathrm{vec}*}) \vee \|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|_2 \leq r\Delta^*, \quad \widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^* \in \Gamma(s), \tag{S.38}$$

with $r < \frac{|a_\pi - c_\pi|}{\Delta} \wedge \frac{\sqrt{9C_0+16c_\Delta}-\sqrt{9C_0}}{4} \wedge \frac{a_\Delta}{C_0} \wedge \frac{a_b}{5\sqrt{s}}$.

Proposition S.7. *Suppose that Assumptions (A1')–(A5) hold. If $\sqrt{s\log p/n} \ll r$, there exist $C_d, C_\lambda > 0, 0 < \kappa < 1/2$ such that with a probability greater than $1 - O(p^{-1})$, we have*

$$\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*\|_2 \lesssim \kappa^t d_{2,s}(\widehat{\boldsymbol{\theta}}^{\mathrm{vec}(0)}, \boldsymbol{\theta}^{\mathrm{vec}*}) + \sqrt{\frac{s\log p}{n}}. \tag{S.39}$$

Moreover, for $t \gtrsim (-\log(\kappa))^{-1}\log n \cdot d_{2,s}(\widehat{\boldsymbol{\theta}}^{\mathrm{vec}(0)}, \boldsymbol{\theta}^{\mathrm{vec}*})$, then

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \lesssim \sqrt{\frac{s\log p}{n}}, \tag{S.40}$$

$$R(PEM) - R(Opt) \lesssim \frac{s\log p}{n}. \tag{S.41}$$

Proof of Proposition S.7. Proposition S.7 can be proved following the same line of Theorems 3.1 & 3.2 in Cai et al. (2019) by replacing their Lemma 3.1 with Assumption (A4) and their Lemma 3.2 with Assumption (A3). □

The following two propositions are important for our proof as well.

Proposition S.8. *[Cai et al. (2019) c.f Lemmas 3.1 & 3.2] Consider data drawn from the model in (A1'). In the PEM, let $g_{\omega_k}(\boldsymbol{\xi}, \mathbf{U}) = \frac{1}{n}\sum_{i=1}^n \xi_{ik}$ and $g_{\boldsymbol{\phi}_k}(\boldsymbol{\xi}, \mathbf{U}) = \frac{\sum_{i=1}^n \xi_{ik}\mathbf{U}_i}{\sum_{i=1}^n \xi_{ik}}$. We have the following conclusions.*

1. *For any $\boldsymbol{\theta}^{\text{vec}*} \in \boldsymbol{\Theta}^{\text{vec}}$, there exists a constant $C_{con} > 0$ such that with probability at least $1 - O(p^{-1})$, we have*

$$\sup_{\boldsymbol{\theta} \in \mathcal{B}_{con}} d_{2,s}^{(1)}(g(\boldsymbol{\xi}(\boldsymbol{\theta}^{\text{vec}}), \mathbf{U}), g^{lim}(\boldsymbol{\theta}^{\text{vec}})) \leq C_{con} \sqrt{\frac{s \log p}{n}}. \tag{S.42}$$

2. *Under Assumption (A5), there exists $\kappa_0^{(1)}(\Delta)$ such that for any $\boldsymbol{\theta}^{\text{vec}} \in \mathcal{B}_{con}$,*

$$d_2^{(1)}[g^{lim}(\boldsymbol{\theta}^{\text{vec}}), \boldsymbol{\theta}^{\text{vec}*}] \leq \kappa_0^{(1)}(\Delta) \cdot \{d_{2,s}^{(1)}(\boldsymbol{\theta}^{\text{vec}}, \boldsymbol{\theta}^{\text{vec}*}) \vee \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2\}, \tag{S.43}$$

where $\kappa^{(1)}(\Delta) = o(1/\Delta^2)$ as $\Delta \to \infty$.

Proposition S.9 (Cai et al. (2019) c.f Proof of Theorem 3.3)**.** *Define*

$$\tilde{\mathcal{A}}_s = \{\mathbf{u} \in \{0,1\}^p : \mathbf{u}^{\intercal} \mathbf{e}_1 = 0, \|\mathbf{u}\|_0 = s\}, \tag{S.44}$$

where $e_{11} = 1$ and $e_{ij} = 0$ for $j > 1$. There exists

$$\boldsymbol{\theta}_i^{\text{vec}} = \{\pi_1 = \pi_2 = \frac{1}{2}, \boldsymbol{\phi}_1 = \epsilon \mathbf{u}_i + \eta \mathbf{e}_i, \boldsymbol{\phi}_2 = -\boldsymbol{\phi}_1, \boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_p\}, i = 1, \ldots, N, \tag{S.45}$$

where $\mathbf{u} \in \tilde{\mathcal{A}}_s, \epsilon = \sigma\sqrt{\log p/n}, \sigma^2 = O(1), \eta = O(1), c_0 \leq \sigma^2 \leq C_0$ and $(\boldsymbol{\phi}_2 - \boldsymbol{\phi}_1)^{\intercal} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\phi}_2 - \boldsymbol{\phi}_1) \geq c_L$ for some constant $c_L > 0$, such that

$$\inf_{\widehat{Y}} \sup_{i=1,\ldots,N} \{R(\widehat{Y} \mid \boldsymbol{\theta}_i^{\text{vec}}) - R(Opt \mid \boldsymbol{\theta}_i^{\text{vec}})\} \succeq \frac{\log p}{n}. \tag{S.46}$$

We further consider the following initialization condition.

(IC) For some permutation $\Pi : \{1,2\} \mapsto \{1,2\}$,

$$\max_{k=1,2} \{\|\widehat{\boldsymbol{\phi}}_k^{(0)} - \boldsymbol{\phi}_{\Pi(k)}\|_\infty\} \preceq \frac{1}{s}, \quad \|\widehat{\boldsymbol{\Psi}}^{(0)} - \boldsymbol{\Psi}\|_{\max} \preceq \frac{1}{s}. \tag{S.47}$$

Proposition S.10 (Cai et al. (2019) c.f Lemma 3.3)**.** *If the initial estimates $\widehat{\boldsymbol{\phi}}_k^{(0)}, \widehat{\boldsymbol{\Psi}}^{(0)}$ satisfy Condition (IC), then the resulting $\widehat{\boldsymbol{\theta}}^{\text{vec}(0)}$ produced by PEM satisfies Condition (A5).*

F.3 Proofs for Theorems 1 & 2

For simplicity, we present the proofs under the assumption that $\text{diag}(\boldsymbol{\Sigma}_m^*) = 1$ so that we do not need to rescale our covariance estimates. This assumption can be straightforwardly relaxed with more tedious calculation. To simplify the notation, we use $\mathbf{X}_{i(m)}$ to denote the matricization $(\mathbf{X}_i)_{(m)}$.

Lemma S.5. *Consider* $\mathbf{W}_1, \ldots, \mathbf{W}_n \in \mathbb{R}^{p_1 \times \cdots \times p_M}$ *iid from* $TN(0; \mathbf{\Omega}_1, \ldots, \mathbf{\Omega}_M)$*, where diagonal elements of* $\mathbf{\Omega}_m, m = 1, \ldots, M$ *are all ones, and* $\lambda_{\max}(\mathbf{\Omega}_m)$ *are universally bounded below and above. Define*

$$\widehat{\mathbf{\Omega}}_m = \frac{1}{np_{-m}} \sum_{i=1}^{n} (\mathbf{W}_i)_{(m)} (\mathbf{W}_i)_{(m)}^{\intercal}. \tag{S.48}$$

Then there exists a constant C *such that with a probability tending to 1, we have*

$$\|\widehat{\mathbf{\Omega}}_m - \mathbf{\Omega}_m\|_{\max} \leq C \sqrt{\frac{\log p_m}{nq_m}}. \tag{S.49}$$

Proof of Lemma S.5. Because $\mathbf{W}_i \sim TN(0; \mathbf{\Omega}_1, \ldots, \mathbf{\Omega}_M)$, there exists $\mathbf{Z} \sim TN(0; \mathbf{I}_{p_1}, \ldots, \mathbf{I}_{p_M})$ such that

$$\mathbf{W}_i = [\![\mathbf{Z}; \mathbf{\Omega}_1^{1/2}, \ldots, \mathbf{\Omega}_M^{1/2}]\!]. \tag{S.50}$$

It follows that $\mathbf{W}_{i(m)} = [\![\mathbf{Z}_{(m)}; \mathbf{\Omega}_m, \otimes_{j=M, j\neq m}^{j=1} \mathbf{\Omega}_M]\!]$ and hence

$$\mathbf{W}_{i(m)} \sim TN(0; \mathbf{\Omega}_m, \otimes_{j=M, j\neq m}^{j=1} \mathbf{\Omega}_j).$$

By Proposition S.4, we have that $\frac{1}{q_m} E(\mathbf{W}_{i(m)})(\mathbf{W}_{i(m)})^{\intercal} = \mathbf{\Omega}_m$ since $\prod_{j\neq m} \operatorname{tr}(\mathbf{\Omega}_j) = q_m$. Then the desired conclusion follows from Proposition S.6. □

Recall that we have data $\{\mathbf{X}_i, Y_i\}_{i=1}^n$, where $\Pr(Y_i = k) = \pi_k^*, \mathbf{X}_i \mid Y_i = k \sim TN(\boldsymbol{\mu}_k^*; \mathbf{\Sigma}_1^*, \ldots, \mathbf{\Sigma}_M^*)$. We consider the parameter space $\mathbf{\Theta}(c_\pi, C_b, s, \{C_m\}_{m=1}^M, C_b, \Delta_0)$ in (4.1). For ease of presentation, we write $\mathbf{\Theta}$ instead of $\mathbf{\Theta}(c_\pi, C_b, s, \{C_m\}_{m=1}^M, C_b, \Delta_0)$ when there is no confusion.

Let $\mathbf{U}_i = \operatorname{vec}(\mathbf{X}_i), \boldsymbol{\phi}_k = \operatorname{vec}(\boldsymbol{\mu}_k), \mathbf{\Psi} = \otimes_{m=M}^{m=1} \mathbf{\Sigma}_m$. For any $\boldsymbol{\theta} = (\pi_k, \boldsymbol{\mu}_k, k = 1, 2; \mathbf{\Sigma}_m, m = 1, \ldots, M)$, we denote its vectorized counterpart as $\boldsymbol{\theta}^{\text{vec}} = (\pi_k, \boldsymbol{\phi}_k, k = 1, 2, \mathbf{\Psi})$. It is easy to see that, if $\boldsymbol{\theta}^* \in \mathbf{\Theta}$, we must have $\boldsymbol{\theta}^{\text{vec}*} \in \mathbf{\Theta}^{\text{vec}}$, with $C_0 = \prod_{m=1}^M C_m$ and $p = \prod_{m=1}^M p_m$. Moreover,

$$\operatorname{vec}(\widehat{\mathbf{B}}^{(t+1)}) = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left[\frac{1}{2} \boldsymbol{\beta}^{\intercal} \{ \bigotimes_{m=M}^{m=1} \widehat{\mathbf{\Sigma}}^{(t+1)} \} \boldsymbol{\beta} - \boldsymbol{\beta}^{\intercal} (\widehat{\boldsymbol{\mu}}_1^{(t+1)} - \widehat{\boldsymbol{\mu}}_2^{(t+1)}) + \lambda_n^{(t+1)} \|\boldsymbol{\beta}\|_1 \right]. \tag{S.51}$$

Hence, DEEM is a special case of PEM with

$$g_{\pi_k}(\boldsymbol{\xi}, \mathbf{U}) = \frac{1}{n} \sum_{i=1}^{n} \xi_{ik}, \quad g_{\boldsymbol{\phi}_k}(\boldsymbol{\xi}, \mathbf{U}) = \frac{\sum_{i=1}^n \xi_{ik} \mathbf{U}_i}{\sum_{i=1}^n \xi_{ik}}, \tag{S.52}$$

$$g_{\mathbf{\Psi}}(\boldsymbol{\xi}, \mathbf{U}) = \otimes \left\{ \frac{1}{np_{-m}} \sum_{i=1}^{n} \xi_{ik} (\mathbf{X}_i - g_{\boldsymbol{\phi}_k}(\boldsymbol{\xi}, \mathbf{U}))_{(m)} (\mathbf{X}_i - g_{\boldsymbol{\phi}_k}(\boldsymbol{\xi}, \mathbf{U}))_{(m)}^{\intercal} \right\}. \tag{S.53}$$

With a little abuse of notation, we define

$$d_{2,s}(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) = (\vee_k |\pi_k - \tilde{\pi}_k|) \vee (\vee_k \|\mathrm{vec}(\boldsymbol{\mu}_k - \tilde{\boldsymbol{\mu}}_k)\|_{2,s}) \vee \|(\boldsymbol{\Psi} - \widetilde{\boldsymbol{\Psi}})\widetilde{\boldsymbol{\beta}}\|_{2,s}. \tag{S.54}$$

We have two lemmas concerning the theoretical properties of this particular choice of g.

Lemma S.6. *Suppose $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ with $0 < c_\pi < 1$ and g defined as in DEEM. We have the following conclusions.*

1. *If $\frac{\log p}{n} = o(1)$, then we must have*

$$g^{lim}(\boldsymbol{\theta}) = \{\pi_k(\boldsymbol{\theta}), \boldsymbol{\mu}_k(\boldsymbol{\theta}), k = 1, 2; \boldsymbol{\Psi}(\boldsymbol{\theta})\}, \tag{S.55}$$

   where

$$\pi_k(\boldsymbol{\theta}) = E[\xi_{ik}(\boldsymbol{\theta})], \quad \boldsymbol{\mu}_k(\boldsymbol{\theta}) = \frac{E\xi_{ik}\mathbf{U}_i}{E\xi_{ik}}, \tag{S.56}$$

$$\boldsymbol{\Psi}(\boldsymbol{\theta}) = \bigotimes_{m=M}^{m=1} [\frac{1}{p_{-m}} E\{\xi_{i1}(\mathbf{X}_i - \boldsymbol{\mu}_k(\boldsymbol{\theta}))_{(m)}(\mathbf{X}_i - \boldsymbol{\mu}_k(\boldsymbol{\theta}))_{(m)}^\intercal\}]. \tag{S.57}$$

2. *There exists a constant $C_{con} > 0$, such that with a probability greater than $1 - O(p^{-1})$, we have*

$$\sup_{\boldsymbol{\theta} \in \mathcal{B}_{con}} d_{2,s}(g\{\boldsymbol{\xi}(\boldsymbol{\theta}), \mathbf{U}\}, g^{\lim}(\boldsymbol{\theta})) \leq C_{con}\sqrt{\frac{s \log p}{n}}, \tag{S.58}$$

$$\sup_{\boldsymbol{\theta} \in \mathcal{B}_{con}} \|[g_{\boldsymbol{\Psi}}\{\boldsymbol{\xi}(\boldsymbol{\theta}), \mathbf{U}\} - \boldsymbol{\Psi}(\boldsymbol{\theta})]\boldsymbol{\beta}^*\|_\infty \leq C_{con}\sqrt{\frac{\log p}{n}}. \tag{S.59}$$

Proof of Lemma S.6. By the first conclusion in Proposition S.8, we have that

$$\sup_{\boldsymbol{\theta} \in \mathcal{B}_{con}} d_{2,s}^{(1)}(g\{\boldsymbol{\xi}(\boldsymbol{\theta}), \mathbf{U}\}, g^{lim}\{\boldsymbol{\theta}\}) \leq C_{con}\sqrt{\frac{s \log p}{n}}. \tag{S.60}$$

Hence, to prove the desired conclusions, it suffices to focus on $\boldsymbol{\Psi}$ and verify that

$$\sup_{\boldsymbol{\theta} \in \mathcal{B}_{con}} \|g_{\boldsymbol{\Psi}}\{\boldsymbol{\xi}(\boldsymbol{\theta}), \mathbf{U}\} - \boldsymbol{\Psi}(\boldsymbol{\theta})\|_{\max} = o_P(1), \tag{S.61}$$

$$\sup_{\boldsymbol{\theta} \in \mathcal{B}_{con}} \|[g_{\boldsymbol{\Psi}}\{\boldsymbol{\xi}(\boldsymbol{\theta}), \mathbf{U}\} - \boldsymbol{\Psi}(\boldsymbol{\theta})]\boldsymbol{\beta}^*\|_{2,s} \leq C_{con}\sqrt{\frac{s \log p}{n}}, \tag{S.62}$$

$$\sup_{\boldsymbol{\theta} \in \mathcal{B}_{con}} \|[g_{\boldsymbol{\Psi}}\{\boldsymbol{\xi}(\boldsymbol{\theta}), \mathbf{U}\} - \boldsymbol{\Psi}(\boldsymbol{\theta})]\boldsymbol{\beta}^*\|_\infty \leq C_{con}\sqrt{\frac{\log p}{n}}. \tag{S.63}$$

For simplicity, we write $\widehat{\boldsymbol{\Psi}}(\boldsymbol{\theta}) = g_{\boldsymbol{\Psi}}\{\boldsymbol{\xi}(\boldsymbol{\theta}), \mathbf{U}\}$. We start by bounding $\sup_{\boldsymbol{\theta}\in\mathcal{B}_{con}} \|\widehat{\boldsymbol{\Psi}}(\boldsymbol{\theta}) - \boldsymbol{\Psi}(\boldsymbol{\theta})\|_{\max}$. Note that $\widehat{\boldsymbol{\Psi}}(\boldsymbol{\theta})$ and $\boldsymbol{\Psi}(\boldsymbol{\theta})$ can be written as

$$\widehat{\boldsymbol{\Psi}}(\boldsymbol{\theta}) = \otimes_{m=M}^{m=1} \widehat{\boldsymbol{\Sigma}}_m(\boldsymbol{\theta}), \quad \boldsymbol{\Psi}(\boldsymbol{\theta}) = \otimes_{m=M}^{m=1} \boldsymbol{\Sigma}_m(\boldsymbol{\theta}). \tag{S.64}$$

By Proposition S.3 we have

$$\begin{aligned}
& \|\widehat{\boldsymbol{\Psi}}(\boldsymbol{\theta}) - \boldsymbol{\Psi}(\boldsymbol{\theta})\|_{\max} \\
= \; & \| \otimes_{m=M}^{m=1} \widehat{\boldsymbol{\Sigma}}_m(\boldsymbol{\theta}) - \otimes_{m=M}^{m=1} \boldsymbol{\Sigma}_m(\boldsymbol{\theta})\|_{\max} \\
\leq \; & \sum_{j=1}^{M} [\prod_{m=M}^{m=j+1} (\|\widehat{\boldsymbol{\Sigma}}_m(\boldsymbol{\theta}) - \boldsymbol{\Sigma}_m(\boldsymbol{\theta})\|_{\max} + \|\boldsymbol{\Sigma}_m(\boldsymbol{\theta})\|_{\max}) \cdot \|\widehat{\boldsymbol{\Sigma}}_j(\boldsymbol{\theta}) - \boldsymbol{\Sigma}_j(\boldsymbol{\theta})\|_{\max} \cdot \prod_{m=j-1}^{m=1} \|\boldsymbol{\Sigma}_m(\boldsymbol{\theta})\|_{\max}] \\
\leq \; & \sum_{j=1}^{M} [\prod_{m=M}^{m=j+1} (\|\widehat{\boldsymbol{\Sigma}}_m(\boldsymbol{\theta}) - \boldsymbol{\Sigma}_m(\boldsymbol{\theta})\|_{\max} + \|\boldsymbol{\Sigma}_m(\boldsymbol{\theta})\|_{\max}) \cdot \|\widehat{\boldsymbol{\Sigma}}_j(\boldsymbol{\theta}) - \boldsymbol{\Sigma}_j(\boldsymbol{\theta})\|_{\max}].
\end{aligned}$$

It can be seen that a bound on $\|\widehat{\boldsymbol{\Psi}}(\boldsymbol{\theta}) - \boldsymbol{\Psi}(\boldsymbol{\theta})\|_{\max}$ is implied by bounds on $\|\widehat{\boldsymbol{\Sigma}}_m(\boldsymbol{\theta}) - \boldsymbol{\Sigma}_m(\boldsymbol{\theta})\|_{\max}$. We provide bounds on $\|\widehat{\boldsymbol{\Sigma}}_m(\boldsymbol{\theta}) - \boldsymbol{\Sigma}_m(\boldsymbol{\theta})\|_{\max}$ in what follows. Note that

$$\begin{aligned}
\widehat{\boldsymbol{\Sigma}}_m(\boldsymbol{\theta}) &= \frac{1}{nq_m} \sum_{i=1}^{n} \sum_{k=1}^{2} \xi_{ik} (\mathbf{X}_i - \widehat{\boldsymbol{\mu}}_k)_{(m)} (\mathbf{X}_i - \widehat{\boldsymbol{\mu}}_k)_{(m)}^{\intercal} \\
&= \frac{1}{nq_m} \sum_{i=1}^{n} \sum_{k=1}^{2} \xi_{ik} (\mathbf{X}_i)_{(m)} (\mathbf{X}_i)_{(m)}^{\intercal} - \frac{1}{nq_m} \sum_{i=1}^{n} \sum_{k=1}^{2} \xi_{ik} \{\widehat{\boldsymbol{\mu}}_k(\boldsymbol{\theta})\}_{(m)} \{\mathbf{X}_{i(m)}\}^{\intercal} \\
&\quad - \frac{1}{nq_m} \sum_{i=1}^{n} \sum_{k=1}^{2} \xi_{ik} \{\mathbf{X}_{i(m)}\} \{\widehat{\boldsymbol{\mu}}_k(\boldsymbol{\theta})\}_{(m)}^{\intercal} + \frac{1}{nq_m} \sum_{i=1}^{n} \sum_{k=1}^{2} \xi_{ik} \{\widehat{\boldsymbol{\mu}}_k(\boldsymbol{\theta})\}_{(m)} \{\widehat{\boldsymbol{\mu}}_k(\boldsymbol{\theta})\}_{(m)}^{\intercal}.
\end{aligned}$$

For any i, we have $\sum_{k=1}^{2} \xi_{ik} = 1$. Hence,

$$\frac{1}{nq_m} \sum_{i=1}^{n} \sum_{k=1}^{2} \xi_{ik} (\mathbf{X}_i)_{(m)} (\mathbf{X}_i)_{(m)}^{\intercal} = \frac{1}{nq_m} \sum_{i=1}^{n} (\mathbf{X}_i)_{(m)} (\mathbf{X}_i)_{(m)}^{\intercal}.$$

Moreover, for any k, we have $\frac{1}{n}\sum_{i=1}^{n} \xi_{ik} = \hat{\pi}_k(\boldsymbol{\theta})$, $\sum_{i=1}^{n} \xi_{ik} \mathbf{X}_i = \{\sum_{i=1}^{n} \xi_{ik}\} \widehat{\boldsymbol{\mu}}_k(\boldsymbol{\theta}) = n\widehat{\pi}_k(\boldsymbol{\theta}) \widehat{\boldsymbol{\mu}}_k(\boldsymbol{\theta})$ by the definitions of $\widehat{\boldsymbol{\mu}}_k(\boldsymbol{\theta})$ and $\widehat{\pi}_k(\boldsymbol{\theta})$, which implies

$$\begin{aligned}
& \frac{1}{nq_m} \sum_{i=1}^{n} \sum_{k=1}^{2} \xi_{ik} \{\widehat{\boldsymbol{\mu}}_k(\boldsymbol{\theta})\}_{(m)} \{\mathbf{X}_{i(m)}\}^{\intercal} = \frac{1}{nq_m} \sum_{k=1}^{2} \{\sum_{i=1}^{n} \xi_{ik}\} \{\widehat{\boldsymbol{\mu}}_k(\boldsymbol{\theta})\}_{(m)} \{\widehat{\boldsymbol{\mu}}_k(\boldsymbol{\theta})\}_{(m)}^{\intercal} \\
= \; & \frac{1}{q_m} \sum_{k=1}^{2} \widehat{\pi}_k(\boldsymbol{\theta}) \{\widehat{\boldsymbol{\mu}}_k(\boldsymbol{\theta})\}_{(m)} \{\widehat{\boldsymbol{\mu}}_k(\boldsymbol{\theta})\}_{(m)}^{\intercal}, \\
& \frac{1}{nq_m} \sum_{i=1}^{n} \sum_{k=1}^{2} \xi_{ik} \{\widehat{\boldsymbol{\mu}}_k(\boldsymbol{\theta})\}_{(m)} \{\widehat{\boldsymbol{\mu}}_k(\boldsymbol{\theta})\}_{(m)}^{\intercal} = \frac{1}{q_m} \sum_{k=1}^{2} \widehat{\pi}_k(\boldsymbol{\theta}) \{\widehat{\boldsymbol{\mu}}_k(\boldsymbol{\theta})\}_{(m)} \{\widehat{\boldsymbol{\mu}}_k(\boldsymbol{\theta})\}_{(m)}^{\intercal}.
\end{aligned}$$

Consequently,

$$\widehat{\boldsymbol{\Sigma}}_m(\boldsymbol{\theta}) = \frac{1}{nq_m}\sum_{i=1}^{n}\mathbf{X}_{i(m)}\{\mathbf{X}_{i(m)}\}^{\intercal} - \frac{1}{q_m}\sum_{k=1}^{2}\widehat{\pi}_k(\boldsymbol{\theta})\{\widehat{\boldsymbol{\mu}}_k(\boldsymbol{\theta})\}_{(m)}\{\widehat{\boldsymbol{\mu}}_k(\boldsymbol{\theta})\}_{(m)}^{\intercal}.$$

Similarly, we have

$$\boldsymbol{\Sigma}_m(\boldsymbol{\theta}) = \frac{1}{q_m}E\{\mathbf{X}_{i(m)}\}\{\mathbf{X}_{i(m)}\}^{\intercal} - \frac{1}{q_m}\sum_{k=1}^{2}\pi_k(\boldsymbol{\theta})\{\boldsymbol{\mu}_k(\boldsymbol{\theta})\}_{(m)}\{\boldsymbol{\mu}_k(\boldsymbol{\theta})\}_{(m)}^{\intercal}.$$

As a result,

$$\begin{aligned}
&\|\widehat{\boldsymbol{\Sigma}}_m(\boldsymbol{\theta}) - \boldsymbol{\Sigma}_m(\boldsymbol{\theta})\|_{\max} \\
\leq\ & \|\frac{1}{nq_m}\sum_{i=1}^{n}\mathbf{X}_{i(m)}\{\mathbf{X}_i\}_{(m)}^{\intercal} - \frac{1}{q_m}E\{\mathbf{X}_{i(m)}\}\{\mathbf{X}_{i(m)}\}^{\intercal}\|_{\max} \\
&+\frac{1}{q_m}\sum_{k=1}^{2}\|\widehat{\pi}_k(\boldsymbol{\theta})\{\widehat{\boldsymbol{\mu}}_k(\boldsymbol{\theta})\}_{(m)}\{\widehat{\boldsymbol{\mu}}_k(\boldsymbol{\theta})\}_{(m)}^{\intercal} - \pi_k(\boldsymbol{\theta})\{\boldsymbol{\mu}_k(\boldsymbol{\theta})\}_{(m)}\{\boldsymbol{\mu}_k(\boldsymbol{\theta})\}_{(m)}^{\intercal}\|_{\max} \\
\equiv\ & L_1 + \sum_{k=1}^{2}L_2^{(k)}.
\end{aligned}$$

For L_1, we note that

$$\begin{aligned}
&\frac{1}{nq_m}\sum_{i=1}^{n}\mathbf{X}_{i(m)}(\mathbf{X}_{i(m)})^{\intercal} && \text{(S.65)}\\
=\ & \frac{1}{nq_m}\sum_{k=1}^{2}\sum_{Y_i=k}(\mathbf{X}_i - \boldsymbol{\mu}_k^*)_{(m)}(\mathbf{X}_i - \boldsymbol{\mu}_k^*)_{(m)}^{\intercal} + \frac{1}{q_m}\sum_{k=1}^{2}\{\tilde{\pi}_k(\boldsymbol{\mu}_k^*)_{(m)}(\widetilde{\boldsymbol{\mu}}_k)_{(m)}^{\intercal} \\
&+\tilde{\pi}_k(\widetilde{\boldsymbol{\mu}}_k)_{(m)}(\boldsymbol{\mu}_k^*)_{(m)}^{\intercal} - \tilde{\pi}_k(\boldsymbol{\mu}_k^*)_{(m)}(\boldsymbol{\mu}_k^*)_{(m)}^{\intercal}\}, && \text{(S.66)}
\end{aligned}$$

where $\tilde{\pi}_k = \frac{1}{n}\sum_{i=1}^{n}1(Y_i = k)$, $\tilde{\boldsymbol{\mu}}_k = \frac{1}{n\tilde{\pi}_k}\sum_{Y_i=k}\mathbf{X}_i$.

By Proposition S.6, we have that

$$\begin{aligned}
&\|\frac{1}{nq_m}\sum_{k=1}^{2}\sum_{Y^i=k}(\mathbf{X}_i - \boldsymbol{\mu}_k^*)_{(m)}(\mathbf{X}_i - \boldsymbol{\mu}_k^*)_{(m)}^{\intercal} - \frac{1}{nq_m}\sum_{k=1}^{2}\sum_{Y^i=k}E(\mathbf{X}_i - \boldsymbol{\mu}_k^*)_{(m)}(\mathbf{X}_i - \boldsymbol{\mu}_k^*)_{(m)}^{\intercal}\|_{\max} \\
=\ & O_P\left(\sqrt{\frac{\log p_m}{nq_m}}\right).
\end{aligned}$$

For the remaining three terms in (S.66), we focus on the term $\frac{\tilde{\pi}_k}{q_m}(\boldsymbol{\mu}_k^*)_{(m)}(\tilde{\boldsymbol{\mu}}_k)_{(m)}^{\intercal}$, as all the other terms can be bounded similarly. Note that

$$\begin{aligned}
&\frac{\tilde{\pi}_k}{q_m}(\boldsymbol{\mu}_k^*)_{(m)}(\tilde{\boldsymbol{\mu}}_k)_{(m)}^{\intercal} - \frac{\pi_k^*}{p_{-m}}(\boldsymbol{\mu}_k^*)_{(m)}(\boldsymbol{\mu}_k^*)_{(m)}^{\intercal} && \text{(S.67)}\\
=\ & \frac{\tilde{\pi}_k}{q_m}(\boldsymbol{\mu}_k^*)_{(m)}(\tilde{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k^*)_{(m)}^{\intercal} + \frac{1}{q_m}(\tilde{\pi}_k - \pi_k)(\boldsymbol{\mu}_k^*)_{(m)}(\boldsymbol{\mu}_k^*)_{(m)}^{\intercal} && \text{(S.68)}\\
\equiv\ & L_3 + L_4. && \text{(S.69)}
\end{aligned}$$

By Hoeffding's inequality, we have $|\tilde{\pi}_k - \pi_k| = O_P(\sqrt{\frac{1}{n}})$ and thus $\|L_4\|_{\max} = O_P(\sqrt{\frac{1}{n}})$.

For L_3, we note that by the Chernoff bound $\|\tilde{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k^*\|_{\max} = O_P(\sqrt{\frac{\log p}{n}})$ and thus

$$\|L_3\|_{\max} = \|\frac{\tilde{\pi}_k}{q_m}(\boldsymbol{\mu}_k^*)_{(m)}(\tilde{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k^*)_{(m)}^\intercal\|_{\max} \leq \frac{1}{q_m}\|(\boldsymbol{\mu}_k^*)_{(m)}\|_\infty \cdot \|\tilde{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k^*\|_{\max} \quad \text{(S.70)}$$

$$= \frac{1}{q_m} \cdot O_P(\sqrt{\frac{\log p}{n}}). \quad \text{(S.71)}$$

It follows that $\|\widehat{\boldsymbol{\Sigma}}_m(\boldsymbol{\theta}) - \boldsymbol{\Sigma}_m(\boldsymbol{\theta})\|_{\max} = O_P(\sqrt{\frac{\log p}{n}})$ and

$$\|\widehat{\boldsymbol{\Psi}}(\boldsymbol{\theta}) - \boldsymbol{\Psi}(\boldsymbol{\theta})\|_{\max} = O_P(\sqrt{\frac{\log p}{n}}). \quad \text{(S.72)}$$

To verify (S.61)–(S.63), we note that

$$\begin{aligned} & \|(\widehat{\boldsymbol{\Psi}}(\boldsymbol{\theta}) - \boldsymbol{\Psi}(\boldsymbol{\theta}))\boldsymbol{\beta}^*\|_{2,s} = \arg\max_{\mathbf{u}\in\Gamma(s)\cap\mathbb{S}^{p-1}} |\langle \mathbf{u}, (\widehat{\boldsymbol{\Psi}}(\boldsymbol{\theta}) - \boldsymbol{\Psi}(\boldsymbol{\theta}))\boldsymbol{\beta}^*\rangle| \\ \leq\ & \|\mathbf{u}\|_1\|\boldsymbol{\beta}^*\|_1\|\widehat{\boldsymbol{\Psi}}(\boldsymbol{\theta}) - \boldsymbol{\Psi}(\boldsymbol{\theta})\|_{\max} \leq 7C\sqrt{s}\|\widehat{\boldsymbol{\Psi}}(\boldsymbol{\theta}) - \boldsymbol{\Psi}(\boldsymbol{\theta})\|_{\max} \\ & \|(\widehat{\boldsymbol{\Psi}}(\boldsymbol{\theta}) - \boldsymbol{\Psi}(\boldsymbol{\theta}))\boldsymbol{\beta}^*\|_\infty \leq \|\widehat{\boldsymbol{\Psi}}(\boldsymbol{\theta}) - \boldsymbol{\Psi}(\boldsymbol{\theta})\|_{\max}\|\boldsymbol{\beta}^*\|_1 \leq C_b\|\widehat{\boldsymbol{\Psi}}(\boldsymbol{\theta}) - \boldsymbol{\Psi}(\boldsymbol{\theta})\|_{\max}. \end{aligned}$$

By (S.72) we have (S.61)–(S.63).

□

Lemma S.7. *Assume that $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$. For sufficiently large Δ_0 there exists $0 < \kappa_0 < \frac{1}{2\vee(16C_0)}$ and $\Delta_0 > 0$ that does not depend on n, p such that for any $\boldsymbol{\theta}^{\text{vec}} \in \mathcal{B}_{con}$, we have*

$$d_2(g^{lim}(\boldsymbol{\theta}^{\text{vec}}), \boldsymbol{\theta}^{*\text{vec}}) \leq \kappa_0 \cdot (d_{2,s}(\boldsymbol{\theta}^{\text{vec}}, \boldsymbol{\theta}^{*\text{vec}}) \vee \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2). \quad \text{(S.73)}$$

Proof of Lemma S.7. By the first conclusion in Proposition S.8, we have that

$$d_2^{(1)}(g^{lim}(\boldsymbol{\theta}^{\text{vec}}), \boldsymbol{\theta}^{*\text{vec}}) \leq \kappa_0^{(1)}(\Delta) \cdot (d_{2,s}(\boldsymbol{\theta}^{\text{vec}}, \boldsymbol{\theta}^{*\text{vec}}) \vee \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2). \quad \text{(S.74)}$$

Thus, it only remains to verify that

$$\|(\boldsymbol{\Psi}(\boldsymbol{\theta}) - \boldsymbol{\Psi}^*)\boldsymbol{\beta}^*\|_2 \leq \kappa_0 \cdot (d_{2,s}(\boldsymbol{\theta}^{\text{vec}}, \boldsymbol{\theta}^{*\text{vec}}) \vee \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2), \quad \text{(S.75)}$$

where $\boldsymbol{\Psi}(\boldsymbol{\theta}) = \bigotimes_{m=M}^{m=1} \boldsymbol{\Sigma}_m(\boldsymbol{\theta})$. We start by comparing $\boldsymbol{\Sigma}_m(\boldsymbol{\theta})$ with $\boldsymbol{\Sigma}_m^*$. For any k, we have that

$$\begin{aligned}
& \frac{1}{q_m} E[\xi_{ik}(\boldsymbol{\theta}^*)\{\mathbf{X} - \boldsymbol{\mu}_k^*\}_{(m)}\{\mathbf{X} - \boldsymbol{\mu}_k^*\}_{(m)}^\intercal] && \text{(S.76)} \\
= & \frac{1}{q_m} E[\Pr(Y = k \mid \mathbf{X}, \boldsymbol{\theta}^*)(\mathbf{X} - \boldsymbol{\mu}_k^*)_{(m)}\{\mathbf{X} - \boldsymbol{\mu}_k^*\}_{(m)}^\intercal] && \text{(S.77)} \\
= & \frac{1}{q_m} E[E[\mathbf{1}(Y = k) \mid \mathbf{X}, \boldsymbol{\theta}^*](\mathbf{X} - \boldsymbol{\mu}_k^*)_{(m)}\{\mathbf{X} - \boldsymbol{\mu}_k^*\}_{(m)}^\intercal] && \text{(S.78)} \\
= & \frac{1}{q_m} E[E[\mathbf{1}(Y = k)(\mathbf{X} - \boldsymbol{\mu}_k^*)_{(m)}\{\mathbf{X} - \boldsymbol{\mu}_k^*\}_{(m)}^\intercal \mid \mathbf{X}, \boldsymbol{\theta}^*]] && \text{(S.79)} \\
= & \frac{1}{q_m} E_{\boldsymbol{\theta}^*}[\mathbf{1}(Y = k)(\mathbf{X} - \boldsymbol{\mu}_k^*)_{(m)}\{\mathbf{X} - \boldsymbol{\mu}_k^*\}_{(m)}^\intercal] && \text{(S.80)} \\
= & \pi_k^* \boldsymbol{\Sigma}_m^*, && \text{(S.81)}
\end{aligned}$$

where the last equation follows from Proposition S.4. Hence, $\boldsymbol{\Sigma}_m(\boldsymbol{\theta}^*) = \sum_k \pi_k^* \boldsymbol{\Sigma}_m^* = \boldsymbol{\Sigma}_m^*$. It follows that

$$\begin{aligned}
& q_m\{\boldsymbol{\Sigma}_m(\boldsymbol{\theta}) - \boldsymbol{\Sigma}_m^*\} \\
= & \pi_1(\boldsymbol{\theta})[\boldsymbol{\mu}_1(\boldsymbol{\theta})]_{(m)}[\boldsymbol{\mu}_1(\boldsymbol{\theta})]_{(m)}^\intercal - \pi_1^*[\boldsymbol{\mu}_1^*]_{(m)}[\boldsymbol{\mu}_1^*]_{(m)}^\intercal + \pi_2(\boldsymbol{\theta})\{\boldsymbol{\mu}_2(\boldsymbol{\theta})\}_{(m)}\{\boldsymbol{\mu}_2(\boldsymbol{\theta})\}_{(m)}^\intercal - \pi_2^*\{\boldsymbol{\mu}_2^*\}_{(m)}\{\boldsymbol{\mu}_2^*\}_{(m)}^\intercal.
\end{aligned}$$

We note that $\pi_1(\boldsymbol{\theta})\boldsymbol{\mu}_1(\boldsymbol{\theta}) + \pi_2(\boldsymbol{\theta})\boldsymbol{\mu}_2(\boldsymbol{\theta}) = E\mathbf{X} = \pi_1^*\boldsymbol{\mu}_1^* + \pi_2^*\boldsymbol{\mu}_2^*$. Hence, let $\boldsymbol{\mu}^* = \pi_1^*\boldsymbol{\mu}_1^* + \pi_2^*\boldsymbol{\mu}_2^*$ and we have

$$\begin{aligned}
& \pi_1(\boldsymbol{\theta})\{\boldsymbol{\mu}_1(\boldsymbol{\theta})\}_{(m)}\{\boldsymbol{\mu}_1(\boldsymbol{\theta})\}_{(m)}^\intercal + \pi_2(\boldsymbol{\theta})\{\boldsymbol{\mu}_2(\boldsymbol{\theta})\}_{(m)}\{\boldsymbol{\mu}_2(\boldsymbol{\theta})\}_{(m)}^\intercal \\
= & \pi_1(\boldsymbol{\theta})\{\boldsymbol{\mu}_1(\boldsymbol{\theta}) - \boldsymbol{\mu}^*\}_{(m)}\{\boldsymbol{\mu}_1(\boldsymbol{\theta}) - \boldsymbol{\mu}^*\}_{(m)}^\intercal + \pi_1(\boldsymbol{\theta})\{\boldsymbol{\mu}_1(\boldsymbol{\theta})\}_{(m)}\{\boldsymbol{\mu}_{(m)}^*\}^\intercal - \pi_1(\boldsymbol{\theta})\boldsymbol{\mu}_{(m)}^*\{\boldsymbol{\mu}_{(m)}^*\}^\intercal \\
& + \pi_1(\boldsymbol{\theta})\boldsymbol{\mu}_{(m)}^*\{\boldsymbol{\mu}_1(\boldsymbol{\theta})\}_{(m)}^\intercal \\
& + \pi_2(\boldsymbol{\theta})\{\boldsymbol{\mu}_2(\boldsymbol{\theta}) - \boldsymbol{\mu}^*\}_{(m)}\{\boldsymbol{\mu}_2(\boldsymbol{\theta}) - \boldsymbol{\mu}^*\}_{(m)}^\intercal + \pi_2(\boldsymbol{\theta})\{\boldsymbol{\mu}_2(\boldsymbol{\theta})\}_{(m)}\{\boldsymbol{\mu}_{(m)}^*\}^\intercal - \pi_2(\boldsymbol{\theta})\boldsymbol{\mu}_{(m)}^*\{\boldsymbol{\mu}_{(m)}^*\}^\intercal \\
& + \pi_2(\boldsymbol{\theta})\boldsymbol{\mu}_{(m)}^*\{\boldsymbol{\mu}_2(\boldsymbol{\theta})\}_{(m)}^\intercal \\
= & \pi_1(\boldsymbol{\theta})\{\boldsymbol{\mu}_1(\boldsymbol{\theta}) - \boldsymbol{\mu}^*\}_{(m)}\{\boldsymbol{\mu}_1(\boldsymbol{\theta}) - \boldsymbol{\mu}^*\}_{(m)}^\intercal + \pi_2(\boldsymbol{\theta})\{\boldsymbol{\mu}_2(\boldsymbol{\theta}) - \boldsymbol{\mu}^*\}_{(m)}\{\boldsymbol{\mu}_2(\boldsymbol{\theta}) - \boldsymbol{\mu}^*\}_{(m)}^\intercal \\
& + \{\boldsymbol{\mu}_{(m)}^*\}\{\boldsymbol{\mu}_{(m)}^*\}^\intercal.
\end{aligned}$$

Similarly,

$$\begin{aligned}
& \pi_1^*[\boldsymbol{\mu}_1^*]_{(m)}[\boldsymbol{\mu}_1^*]_{(m)}^\intercal + \pi_2^*\{\boldsymbol{\mu}_2^*\}_{(m)}\{\boldsymbol{\mu}_2^*\}_{(m)}^\intercal \\
= & \pi_1^*\{\boldsymbol{\mu}_1^* - \boldsymbol{\mu}^*\}_{(m)}\{\boldsymbol{\mu}_1^* - \boldsymbol{\mu}^*\}_{(m)}^\intercal + \pi_2^*\{\boldsymbol{\mu}_2^* - \boldsymbol{\mu}^*\}_{(m)}\{\boldsymbol{\mu}_2^* - \boldsymbol{\mu}^*\}_{(m)}^\intercal + \{\boldsymbol{\mu}_{(m)}^*\}\{\boldsymbol{\mu}_{(m)}^*\}^\intercal.
\end{aligned}$$

It follows that

$$
\begin{aligned}
& q_m\{\boldsymbol{\Sigma}_m(\boldsymbol{\theta})-\boldsymbol{\Sigma}_m^*\} \\
= \;& \pi_1(\boldsymbol{\theta})\{\boldsymbol{\mu}_1(\boldsymbol{\theta})-\boldsymbol{\mu}^*\}_{(m)}\{\boldsymbol{\mu}_1(\boldsymbol{\theta})-\boldsymbol{\mu}^*\}_{(m)}^{\intercal}+\pi_2(\boldsymbol{\theta})\{\boldsymbol{\mu}_2(\boldsymbol{\theta})-\boldsymbol{\mu}^*\}_{(m)}\{\boldsymbol{\mu}_2(\boldsymbol{\theta})-\boldsymbol{\mu}^*\}_{(m)}^{\intercal} \\
& -\pi_1^*\{\boldsymbol{\mu}_1^*-\boldsymbol{\mu}^*\}_{(m)}\{\boldsymbol{\mu}_1^*-\boldsymbol{\mu}^*\}_{(m)}^{\intercal}-\pi_2^*\{\boldsymbol{\mu}_2^*-\boldsymbol{\mu}^*\}_{(m)}\{\boldsymbol{\mu}_2^*-\boldsymbol{\mu}^*\}_{(m)}^{\intercal} \\
= \;& [\pi_1(\boldsymbol{\theta})\{\boldsymbol{\mu}_1(\boldsymbol{\theta})-\boldsymbol{\mu}^*\}_{(m)}\{\boldsymbol{\mu}_1(\boldsymbol{\theta})-\boldsymbol{\mu}^*\}_{(m)}^{\intercal}-\pi_1^*\{\boldsymbol{\mu}_1^*-\boldsymbol{\mu}^*\}_{(m)}\{\boldsymbol{\mu}_1^*-\boldsymbol{\mu}^*\}_{(m)}^{\intercal}] \\
& +[\pi_2(\boldsymbol{\theta})\{\boldsymbol{\mu}_2(\boldsymbol{\theta})-\boldsymbol{\mu}^*\}_{(m)}\{\boldsymbol{\mu}_2(\boldsymbol{\theta})-\boldsymbol{\mu}^*\}_{(m)}^{\intercal}-\pi_2^*\{\boldsymbol{\mu}_2^*-\boldsymbol{\mu}^*\}_{(m)}\{\boldsymbol{\mu}_2^*-\boldsymbol{\mu}^*\}_{(m)}^{\intercal}] \\
= \;& q_m\cdot L_5+q_m\cdot L_6.
\end{aligned}
$$

To bound $\|\boldsymbol{\Sigma}_m(\boldsymbol{\theta})-\boldsymbol{\Sigma}_m^*\|_2$, we only provide a bound on L_5, as L_6 can be bounded similarly.

$$
\begin{aligned}
L_5 \;=\;& \frac{1}{q_m}\|\pi_1(\boldsymbol{\theta})\{\boldsymbol{\mu}_1(\boldsymbol{\theta})-\boldsymbol{\mu}^*\}_{(m)}\{\boldsymbol{\mu}_1(\boldsymbol{\theta})-\boldsymbol{\mu}^*\}_{(m)}^{\intercal}-\pi_1^*\{\boldsymbol{\mu}_1^*-\boldsymbol{\mu}^*\}_{(m)}\{\boldsymbol{\mu}_1^*-\boldsymbol{\mu}^*\}_{(m)}^{\intercal}\|_2 \\
\le\;& \frac{1}{q_m}|\pi_1^*-\pi_1(\boldsymbol{\theta})|\cdot\|(\boldsymbol{\mu}_1^*-\boldsymbol{\mu}^*)_{(m)}(\boldsymbol{\mu}_1^*-\boldsymbol{\mu}^*)_{(m)}^{\intercal}\|_2 \\
& +\frac{1}{q_m}|\pi_1^*|\cdot\|(\boldsymbol{\mu}_1(\boldsymbol{\theta})-\boldsymbol{\mu}^*)_{(m)}(\boldsymbol{\mu}_1(\boldsymbol{\theta})-\boldsymbol{\mu}_1^*)_{(m)}^{\intercal}\|_2 \\
& +\frac{1}{q_m}|\pi_1(\boldsymbol{\theta})|\cdot\|(\boldsymbol{\mu}_1^*-\boldsymbol{\mu}^*)_{(m)}(\boldsymbol{\mu}_1(\boldsymbol{\theta})-\boldsymbol{\mu}_1^*)_{(m)}^{\intercal}\|_2 \\
\equiv\;& L_7+L_8+L_9.
\end{aligned}
$$

For L_7, we note that $\|\mathrm{vec}(\boldsymbol{\mu}_1^*-\boldsymbol{\mu}^*)\|_2^2\le C_0\Delta^2$ by the definition of Δ. By the second conclusion in Proposition S.8, there exists $\kappa_0^{(1)}(\Delta)=o(\frac{1}{\Delta^2})$ such that

$$
\begin{aligned}
L_7 \;\le\;& \frac{1}{q_m}|\pi_1^*-\pi_1(\boldsymbol{\theta})|\cdot\|\mathrm{vec}(\boldsymbol{\mu}_1^*-\boldsymbol{\mu}^*)\|_2^2\le\frac{C_0}{q_m}|\pi_1^*-\pi_1(\boldsymbol{\theta})|\cdot\Delta^2 \\
\le\;& \frac{C_0}{q_m}\Delta^2\cdot\kappa_0^{(1)}(\Delta)\cdot d_{2,s}^{(1)}(\boldsymbol{\theta},\boldsymbol{\theta}^*)\vee\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
L_8 \;\le\;& \frac{1}{q_m}\|[\{\boldsymbol{\mu}_1(\boldsymbol{\theta})-\boldsymbol{\mu}_1^*\}_{(m)}+\{\boldsymbol{\mu}_1^*-\boldsymbol{\mu}^*\}_{(m)}]\{\boldsymbol{\mu}_1(\boldsymbol{\theta})-\boldsymbol{\mu}_1^*\}_{(m)}^{\intercal}\|_2 \\
\le\;& \frac{1}{q_m}\left[\|\{\boldsymbol{\mu}_1(\boldsymbol{\theta})-\boldsymbol{\mu}_1^*\}_{(m)}\|_2+\|\{\boldsymbol{\mu}_1^*-\boldsymbol{\mu}^*\}_{(m)}\|_2\right]\cdot\|\{\boldsymbol{\mu}_1(\boldsymbol{\theta})-\boldsymbol{\mu}_1^*\}_{(m)}^{\intercal}\|_2.
\end{aligned}
$$

We note that

$$
\begin{aligned}
\|\{\boldsymbol{\mu}_1(\boldsymbol{\theta})-\boldsymbol{\mu}_1^*\}_{(m)}\|_2 \;\le\;& \|\mathrm{vec}\{\boldsymbol{\mu}_1(\boldsymbol{\theta})-\boldsymbol{\mu}_1^*\}\|_F=\|\mathrm{vec}\{\boldsymbol{\mu}_1(\boldsymbol{\theta})-\boldsymbol{\mu}_1^*\}\|_2 \\
\le\;& \kappa_0^{(1)}(\Delta)\cdot\{d_{2,s}^{(1)}(\boldsymbol{\theta},\boldsymbol{\theta}^*)\vee\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\}, \\
\|\{\boldsymbol{\mu}_1^*-\boldsymbol{\mu}^*\}_{(m)}\|_2 \;\le\;& \|\mathrm{vec}\{\boldsymbol{\mu}_1^*-\boldsymbol{\mu}^*\}\|_2\le\sqrt{C_0}\Delta.
\end{aligned}
$$

Further noting that for any $\boldsymbol{\theta} \in \mathcal{B}_{con}$, we have $d_{2,s}^{(1)}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \vee \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq (a_b + \sqrt{C_0})\Delta$. As a result,

$$\begin{aligned} L_8 &\leq \frac{1}{q_m}\{\kappa^{(1)} \cdot d_{2,s}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) + \sqrt{C_0}\Delta\}\kappa^{(1)} \cdot \{d_{2,s}^{(1)}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \vee \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2\} && \text{(S.82)} \\ &\leq \frac{1}{q_m}\{a_b + \sqrt{C_0}\}\Delta \cdot \kappa_0^{(1)}(\Delta) \cdot \{d_{2,s}^{(1)}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \vee \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2\}. && \text{(S.83)} \end{aligned}$$

It is easy to see that L_9 has a similar bound. Therefore, let $\kappa_{\boldsymbol{\Sigma},0} = \frac{1}{q_m}\{a_b + \sqrt{C_0}\}\Delta \cdot \kappa^{(1)}(\Delta)$ and we have $\kappa_{\boldsymbol{\Sigma},0} = o(1/\Delta)$ and

$$\|\boldsymbol{\Sigma}_m(\boldsymbol{\theta}) - \boldsymbol{\Sigma}_m^*\|_2 \leq \frac{1}{p_{-m}}\kappa_{\boldsymbol{\Sigma},0} \cdot \{d_{2,s}^{(1)}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \vee \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2\}. \tag{S.84}$$

By Proposition S.3 we have

$$\begin{aligned} &\| \otimes_{m=M}^{m=1} \boldsymbol{\Sigma}_m(\boldsymbol{\theta}) - \otimes_{m=M}^{m=1} \boldsymbol{\Sigma}_m^* \|_2 \\ \leq\ & \sum_{j=1}^{M}[\prod_{m=M}^{m=j}(\|\boldsymbol{\Sigma}_m(\boldsymbol{\theta}) - \boldsymbol{\Sigma}_m^*\|_2 + \|\boldsymbol{\Sigma}_m^*\|_2) \cdot \|\boldsymbol{\Sigma}_j(\boldsymbol{\theta}) - \boldsymbol{\Sigma}_j^*\|_2 \cdot \prod_{m=j-1}^{m=1}\|\boldsymbol{\Sigma}_m^*\|_2] \\ \leq\ & M\{\frac{\kappa_{\boldsymbol{\Sigma}_0}}{p_{-m}} \cdot d_{2,s}^{(1)}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \vee \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 + C_0\}^{M-1} \cdot \frac{\kappa_{\boldsymbol{\Sigma}_0}}{p_{-m}} \cdot \{d_{2,s}^{(1)}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \vee \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2\} \\ \leq\ & M\{\frac{\kappa_{\boldsymbol{\Sigma},0}}{p_{-m}}(a_b + \sqrt{c_0})\Delta + C_0\}^{M-1} \cdot \frac{\kappa_{\boldsymbol{\Sigma},0}}{p_{-m}} \cdot \{d_{2,s}^{(1)}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \vee \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2\}. \end{aligned}$$

Moreover,

$$\begin{aligned} &\|\{\boldsymbol{\Sigma}(\boldsymbol{\theta}) - \boldsymbol{\Sigma}^*\}\boldsymbol{\beta}^*\|_2 \leq \|\boldsymbol{\Sigma}(\boldsymbol{\theta}) - \boldsymbol{\Sigma}^*\|_2 \cdot \|\boldsymbol{\beta}^*\|_2 && \text{(S.85)} \\ \leq\ & M\{\frac{\kappa_{\boldsymbol{\Sigma},0}}{p_{-m}}(a_b + \sqrt{c_0})\Delta + C_0\}^{M-1} \cdot \frac{\kappa_{\boldsymbol{\Sigma}_0}}{p_{-m}} \cdot d_{2,s}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \cdot \sqrt{C_0}\Delta && \text{(S.86)} \\ \equiv\ & \kappa_{\boldsymbol{\Sigma}} \cdot d_{2,s}(\boldsymbol{\theta}, \boldsymbol{\theta}^*). && \text{(S.87)} \end{aligned}$$

Since $\kappa_{\boldsymbol{\Sigma},0} = o(1/\Delta)$, we have $\kappa_{\boldsymbol{\Sigma}} = o(1)$. It follows that there exists $\Delta_0 > 0$ such that when $\Delta > \Delta_0$, we have that $\kappa_0 = \min\{\kappa_0^{(1)}(\Delta_0), \kappa_{\boldsymbol{\Sigma}}\} < \frac{1}{2\vee(16C_0)}$ and

$$d_{2,s}(g^{lim}(\boldsymbol{\theta}), \boldsymbol{\theta}^*) \leq \kappa_0 \cdot d_{2,s}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \vee \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2. \tag{S.88}$$

□

We are now ready to prove Theorems 1 & 2. Recall that we generate the sequence of tuning parameters $\lambda_n^{(t)}$ as follows. For some constant $C_d, C_\lambda > 0$, we let

$$\lambda^{(0)} = C_d \cdot (|\hat{\pi}_2| \vee \|\mathrm{vec}(\hat{\boldsymbol{\mu}}_1^{(0)} - \hat{\boldsymbol{\mu}}_2^{(0)})\|_{2,s} \vee \| \otimes_{m=M}^{m=1} \widehat{\boldsymbol{\Sigma}}_m^{(0)}\|_{2,s})/\sqrt{s} + C_\lambda\sqrt{\sum_{m=1}^{M}\log p_m/n}, \tag{S.89}$$

and $\lambda^{(t+1)} = \kappa\lambda^{(t)} + C_\lambda\sqrt{\frac{\log p}{n}}$. Note this choice coincides with that in the generic PEM algorithm.

Proof of Theorem 1. We prove Theorem 1 by verifying the five assumptions for Proposition S.7. Assumption (A1') apparently holds when $\mathbf{X}_i$ is drawn from the mixture tensor normal model. Condition (C1) implies (A5), while (A2) & (A3) are implied by Lemma S.6. By Lemma S.7, (A4) is also true. Hence, the desired conclusion follows from the first conclusion in Proposition S.7. □

Proof of Theorem 2. Similar to the proof of Theorem 1, (A1')–(A5) holds. By the second conclusion in Proposition S.7, we have the first desired conclusion. To show the second conclusion, we consider

$$\boldsymbol{\theta}_i = \{\pi_1 = \pi_2, \boldsymbol{\mu}_1 = \epsilon\mathbf{U}_i + \eta\mathbf{E}_i, \boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1 = \sigma^2\mathbf{I}_{p_1}, \boldsymbol{\Sigma}_m = \mathbf{I}_{p_m}, m > 1\}, i = 1, \ldots, N \tag{S.90}$$

such that the corresponding $\boldsymbol{\theta}_i^{\text{vec}}$ is equal to that defined in (S.45). Note that this is possible because $\bigotimes_{m=M}^{m=1} \boldsymbol{\Sigma}_m = \sigma^2\mathbf{I}_p$. It immediately follows from Proposition S.9 that

$$\inf_{\widehat{Y}} \sup_{i=1}^{N} \{R(\widehat{Y} \mid \boldsymbol{\theta}_i) - R(Opt \mid \boldsymbol{\theta}_i)\} \succeq \frac{\log p}{n}. \tag{S.91}$$

Since $\boldsymbol{\theta}_i \in \boldsymbol{\Theta}$, we have

$$\inf_{\widehat{Y}} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \{R(\widehat{Y} \mid \boldsymbol{\theta}_i) - R(Opt \mid \boldsymbol{\theta}_i)\} \succeq \frac{\log p}{n}. \tag{S.92}$$

Combine this result with the first conclusion and we have that DEEM is minimax optimal. □

G An Initialization Algorithm

We propose an initialization algorithm that satisfies Condition (C1) with a high probability. For ease of presentation, we refer to Algorithm $\mathcal{B}$ in Hardt & Price (2015) as the HP Algorithm. The HP Algorithm considers $\mathbf{U}_i, i = 1, \ldots, n$, where $\mathbf{U}_i \in \mathbb{R}^p$ is drawn from Gaussian mixtures. In other words, we assume that

(A1) $\mathbf{U}_i \sim \pi_1^* N(\boldsymbol{\phi}_1^*, \boldsymbol{\Psi}^*) + \pi_2^* N(\boldsymbol{\phi}_1^*, \boldsymbol{\Psi}^*)$, where $0 < \pi_k^* < 1$, $\boldsymbol{\phi}_k^* \in \mathbb{R}^p$, $\boldsymbol{\Psi}^* \in \mathbb{R}^{p\times p}$.

The HP Algorithm finds estimates for $\boldsymbol{\phi}_k^*, \boldsymbol{\Psi}^*$. We have the following result for the HP Algorithm.

Proposition S.11 (Hardt & Price (2015), Cai et al. (2019) c.f Proposition 3.1)**.** *Suppose that we observe n i.i.d samples $\mathbf{U}_i$ from Model (A1). Given $\epsilon, \delta > 0$, if $\frac{1}{\epsilon^6}\log\left(\frac{p}{\delta}\log\left(\frac{1}{\epsilon}\right)\right) = o(n)$, then with probability at least $1-\delta$, the HP algorithm produces estimates $\widehat{\boldsymbol{\phi}}_k^{(0)}$ and $\widehat{\boldsymbol{\Psi}}^{(0)}$ such that for some permutation $\Pi : \{1,2\} \mapsto \{1,2\}$, we have*

$$\max\{\|\widehat{\boldsymbol{\phi}}_1^{(0)} - \boldsymbol{\phi}^*_{\Pi(1)}\|_\infty^2, \|\widehat{\boldsymbol{\phi}}_2^{(0)} - \boldsymbol{\phi}^*_{\Pi(2)}\|_\infty^2, \|\widehat{\boldsymbol{\Psi}}^{(0)} - \boldsymbol{\Psi}^*\|_{\max}\} \le \epsilon\left(\frac{1}{4}\|\boldsymbol{\phi}_1^* - \boldsymbol{\phi}_2^*\|_\infty^2 + \|\boldsymbol{\Psi}^*\|_{\max}\right). \quad \text{(S.93)}$$

Obviously, a main difference in our problem of interest is that we consider clustering for tensors instead of vectors. But we make two observations to connect tensor clustering to the HP Algorithm. First, for any two indices $\mathcal{J} = (j_1, \cdots, j_M)$ and $\mathcal{J}' = (j_1', \cdots, j_M')$, let $\mathbf{U}_i = (X_{i,\mathcal{J}}, X_{i,\mathcal{J}'})^\intercal \in \mathbb{R}^2$. Then $\mathbf{U}_i$ follows Model (A1), and the HP algorithm can be applied to $\mathbf{U}_i$ to estimate the within-cluster means of $X_{i,\mathcal{J}}$ and $X_{i,\mathcal{J}'}$. Repeat this procedure for all the indices and we can find an initial value for $\widehat{\boldsymbol{\mu}}_k$. We do note that the HP Algorithm could be applied to each $X_{\mathcal{J}}$ as well, but doing so could potentially lead to mismatched mean estimates; see our discussion after we propose our algorithm. Hence, we instead consider a pair of variables at a time to match the estimates.

Second, for $\boldsymbol{\Sigma}_m^*$, we average $\mathbf{X}_i$ along the m-th mode to reduce them to vectors that follow Model (A1). More explicitly, let

$$\mathbf{X}_i^m = \frac{1}{\sqrt{q_m}}\mathbf{X}_{(m)}\mathbf{1}_{q_m} \in \mathbb{R}^{p_m}. \quad \text{(S.94)}$$

We have the following lemma.

Lemma S.8. *For $\mathbf{X}_i$ following TNMM and $\mathbf{X}_i^m$ defined in* (S.94)*, we have $\mathbf{X}_i^m \sim \pi_1^* N(\boldsymbol{\mu}_1^m, a_m\boldsymbol{\Sigma}_m^*) + \pi_2^* N(\boldsymbol{\mu}_2^m, a_m\boldsymbol{\Sigma}_m^*)$, where $\boldsymbol{\mu}_k^m = (\boldsymbol{\mu}_k^*)_{(m)} \cdot \frac{1}{\sqrt{q_m}}\mathbf{1}_{q_m}$ and $a_m = \frac{1}{q_m}\mathbf{1}_{q_m}^\intercal\boldsymbol{\Sigma}_{-m}\mathbf{1}_{q_m}$ with $\boldsymbol{\Sigma}_{-m} = \boldsymbol{\Sigma}_M^* \otimes \cdots \boldsymbol{\Sigma}_{m+1}^* \otimes \boldsymbol{\Sigma}_{m-1}^* \otimes \cdots \otimes \boldsymbol{\Sigma}_1^*$.*

The proof of Lemma S.8 is presented in Section G.1. Lemma S.8 indicates that the within-cluster covariance of $\mathbf{X}_i^m$ is proportional to $\boldsymbol{\Sigma}_m^*$. Moreover, recall that we assume that $\sigma^*_{m,11} = 1, m > 1$. Hence, with proper scaling, we can apply the HP algorithm to $\mathbf{X}_i^m$ to find an initial estimate for $\boldsymbol{\Sigma}_m^*$.

With these two observations, we propose the initialization algorithm for tensor data in Algorithm S.4. Algorithm S.4 finds $\widehat{\boldsymbol{\mu}}_k^{(0)}, \widehat{\boldsymbol{\Sigma}}_m^{(0)}$ in two mostly separate steps. The step for finding $\widehat{\boldsymbol{\Sigma}}_m^{(0)}$ (i.e, Step 3) is relatively more straightforward; it converts our tensor observations in the same

way suggested in Lemma S.8 to obtain estimates for $a_m\mathbf{\Sigma}_m^*$. Then these estimates are rescaled according to our identifiability assumptions in Section 2.2.

The step for finding $\widehat{\boldsymbol{\mu}}_k^{(0)}$ (i.e, Step 2) is more involved. As briefly mentioned earlier, a major obstacle in this step is that, if we estimate $\mu_{k\mathcal{J}}^*$ separately, the estimates could be mismatched. For example, if we consider two indices $\mathcal{J}$ and $\mathcal{J}'$ separately to obtain $\widehat{\xi}_{k\mathcal{J}}$ and $\widehat{\xi}_{k\mathcal{J}'}$, it is possible that $\widehat{\xi}_{1\mathcal{J}}$ is close to $\mu_{1\mathcal{J}}^*$, while $\widehat{\xi}_{1\mathcal{J}'}$ is close to $\mu_{2\mathcal{J}'}^*$, as the cluster labels are not identifiable themselves. Consequently, when we put them together, $(\widehat{\xi}_{1\mathcal{J}}, \widehat{\xi}_{1\mathcal{J}'})$ is neither close to $\boldsymbol{\mu}_{1,(\mathcal{J},\mathcal{J}')}^*$, nor $\boldsymbol{\mu}_{2,(\mathcal{J},\mathcal{J}')}^*$. To resolve this issue, in Step 2(a) we find an element $X_{\mathcal{J}_0}$ such that the clusters are best separated so that estimates for all the other elements can be matched according to this element. Indeed, in Step 2(b), we repeatedly apply the HP Algorithm to $(X_{\mathcal{J}_0}, X_{\mathcal{J}})$ to obtain $(\widehat{\xi}_{k\mathcal{J}_0}^{\mathcal{J}}, \widehat{\xi}_{k\mathcal{J}})$. The estimate $\widehat{\xi}_{k\mathcal{J}_0}^{\mathcal{J}}$ is compared to $\widetilde{\mu}_{k\mathcal{J}_0}$ to match $\widehat{\xi}_{k\mathcal{J}}$ in Step 2(c).

In principle, one could also vectorize the whole tensor $\mathbf{X}_i$ to $\mathbf{U}_i \in \mathbb{R}^{\prod_{m=1}^M p_m}$ and directly apply the HP Algorithm to $\mathbf{U}_i$ to find initial values for $\boldsymbol{\mu}_k^*$. Such an approach does not have the potential issue of mismatched estimates, but it will result in drastic increase in storage. This is because the HP algorithm produces estimates for the mean and the covariance matrix simultaneously. Hence, in order to estimate $\boldsymbol{\mu}_k^*$, the HP Algorithm needs to estimate $\mathrm{cov}(\mathbf{U}_i \mid Y_i)$ at the same time. The latter has an intimidating number of parameters at the order of $O(\prod_{m=1}^M p_m^2)$, and significantly inflates the storage and computation costs. However, $\mathrm{cov}(\mathbf{U}_i \mid Y_i)$ is not needed for later tensor clustering, and thus the additional costs do not facilitate our analysis in any way. In contrast, Algorithm S.4 avoids estimating $\mathrm{cov}(\mathbf{U}_i \mid Y_i)$ and keeps the storage and computation costs at a low level. This fact demonstrates the potential benefit of exploiting the tensor structure in initialization as well.

G.1 Proof of Lemma S.8

Define Y_i as the latent variable such that $\mathbf{X}_i \mid (Y_i = k) \sim TN(\boldsymbol{\mu}_k^*; \mathbf{\Sigma}_1^*, \ldots, \mathbf{\Sigma}_M^*)$. It suffices to show that $\mathbf{X}_i^m \mid (Y_i = k) \sim N(\boldsymbol{\mu}_k^m, a_m\mathbf{\Sigma}_m^*)$. By the definition of tensor normal distribution, there exists $\mathbf{Z} \in \mathbb{R}^{p_1 \times \cdots \times p_M}$ with iid $N(0,1)$ elements such that, when $Y_i = k$,

$$\mathbf{X}_i = \boldsymbol{\mu}_k^* + [\![\mathbf{Z}; (\mathbf{\Sigma}_1^*)^{1/2}, \ldots, (\mathbf{\Sigma}_M^*)^{1/2}]\!]. \tag{S.95}$$

By Proposition S.2, we have that, conditional on $Y_i = k$,

$$\mathbf{X}_i^m = (\mathbf{X}_i)_{(m)} \cdot \frac{1}{\sqrt{q_m}}\mathbf{1}_{q_m} = (\boldsymbol{\mu}_k)_{(m)} \cdot \frac{1}{q_m}\mathbf{1}_{q_m} + (\mathbf{\Sigma}_m^*)^{1/2}\mathbf{Z}_{(m)}(\frac{1}{\sqrt{q_m}}\mathbf{\Sigma}_{-m}\mathbf{1}_{q_m}). \tag{S.96}$$

Algorithm S.4 An initialization algorithm for tensor data.

1. Input $\mathbf{X}_i, i = 1, \ldots, n$.

2. Find $\widehat{\boldsymbol{\mu}}_k^{(0)}$ as follows:

   (a) For each $\mathcal{J}$, apply the HP algorithm to $\{X_{i,\mathcal{J}}\}_{i=1}^n$ to obtain $\widetilde{\mu}_{k\mathcal{J}}$ as estimates of $\mu_{k\mathcal{J}}$. Find $\mathcal{J}_0 = \arg\max_{\mathcal{J}} |\widetilde{\mu}_{2\mathcal{J}} - \widetilde{\mu}_{1\mathcal{J}}|$. Let $\widehat{\mu}_{k\mathcal{J}_0}^{(0)} = \widetilde{\mu}_{k\mathcal{J}_0}$. Also record $\widehat{\mathrm{var}}(X_{i,1\cdots 1} \mid Y)$ obtained from the HP Algorithm on $\{X_{i,1\cdots 1}\}_{i=1}^n$.

   (b) For each $\mathcal{J}$, apply the HP Algorithm to $(X_{i,\mathcal{J}_0}, X_{i,\mathcal{J}})_{i=1}^n$ to find $\widehat{\boldsymbol{\nu}}_k^{\mathcal{J}} = (\widehat{\nu}_{k\mathcal{J}_0}^{\mathcal{J}}, \widehat{\nu}_{k\mathcal{J}})$ such that $\widehat{\nu}_{k\mathcal{J}_0}^{\mathcal{J}}$ is an estimate for the within-cluster mean of $X_{\mathcal{J}_0}$ and $\widehat{\nu}_{k\mathcal{J}}$ estimates the within-cluster mean of $X_{\mathcal{J}}$.

   (c) For each $\mathcal{J} \neq \mathcal{J}_0$, find a permutation $\Lambda_{\mathcal{J}} : \{1, 2\} \mapsto \{1, 2\}$ to minimize $\max_k\{|\widehat{\nu}_{\Lambda_{\mathcal{J}}(k),\mathcal{J}_0} - \widetilde{\mu}_{k\mathcal{J}_0}|\}$ and set $\widehat{\mu}_{k,\mathcal{J}}^{(0)} = \widehat{\nu}_{\Lambda_{\mathcal{J}}(k),\mathcal{J}}$.

3. Find $\widehat{\boldsymbol{\Sigma}}_m^{(0)}$ as follows:

   (a) For $m = 1, \ldots, M$, do

      i. Let $\mathbf{X}_i^m = \frac{1}{\sqrt{p_{-m}}} \mathbf{X}_{(m)} \mathbf{1}_{p_{-m}}$. Apply the HP algorithm to $\mathbf{X}_i^m$ to estimate the within-cluster covariance of $\mathbf{X}_i^m$, denoted as $\widehat{\boldsymbol{\Psi}}_m^{(0)}$.

      ii. Set $\widetilde{\boldsymbol{\Psi}}_m^{(0)} \propto \widehat{\boldsymbol{\Psi}}_m^{(0)}$ with $\tilde{\Psi}_{m,11}^{(0)} = 1$.

   (b) Let $\widehat{\boldsymbol{\Sigma}}_1^{(0)} = \widehat{\mathrm{var}}(X_{i,1\cdots 1} \mid Y) \cdot \widehat{\boldsymbol{\Psi}}_1^{(0)}$, where $\widehat{\mathrm{var}}(X_{i,1\cdots 1} \mid Y)$ is found in Step 2(a). For $m > 1$, let $\widehat{\boldsymbol{\Sigma}}_m^{(0)} = \widehat{\boldsymbol{\Psi}}_m^{(0)}$.

4. Output $\widehat{\boldsymbol{\mu}}_k^{(0)}, \widehat{\boldsymbol{\Sigma}}_m^{(0)}$.

Note that $\mathbf{X}_i^m$ is a vector and $\mathrm{vec}\{(\mathbf{\Sigma}_m^*)^{1/2}\mathbf{Z}_{(m)}(\frac{1}{\sqrt{q_m}}\mathbf{\Sigma}_{-m}\mathbf{1}_{q_m})\} = \{(\frac{1}{\sqrt{q_m}}\mathbf{\Sigma}_{-m}\mathbf{1}_{q_m})^\intercal \otimes (\mathbf{\Sigma}_m^*)^{1/2}\}\mathrm{vec}(\mathbf{Z}_{(m)})$. It follows that $\mathbf{X}_i^m \mid (Y_i = k) \sim N(\boldsymbol{\mu}_k^m, a_m\mathbf{\Sigma}_m^*)$.

G.2 Proof of Lemma 4

We start by verifying that the estimates $\widehat{\boldsymbol{\phi}}_k^{(0)} = \mathrm{vec}(\widehat{\boldsymbol{\mu}}_k^{(0)})$, $\widehat{\mathbf{\Psi}}^{(0)} = \otimes_{m=M}^{m=1}\widehat{\mathbf{\Sigma}}_m^{(0)}$ satisfy Condition (IC). We consider $\widehat{\boldsymbol{\mu}}_k^{(0)}$ and $\otimes_{m=M}^{m=1}\widehat{\mathbf{\Sigma}}_m^{(0)}$ separately.

The bound on $\widehat{\boldsymbol{\mu}}_k^{(0)}$. For $\widehat{\boldsymbol{\mu}}_k^{(0)}$ and some constant $C_{\text{init}} > 0$, we consider the event

$$\begin{aligned}\mathcal{A} \;=\; &\{\text{For each } \mathcal{J}, \text{there exist a permutations } \Pi_{\mathcal{J}}, \Upsilon_{\mathcal{J}} \text{ such that } |\tilde{\mu}_{\Pi_{\mathcal{J}(k)},\mathcal{J}_0} - \mu^*_{k\mathcal{J}}| \le \frac{C_{\text{init}}}{s} \\ &\text{and } \|\widehat{\boldsymbol{\nu}}^{\mathcal{J}}_{\Upsilon_{\mathcal{J}}(k)} - \boldsymbol{\mu}^*_{k,(\mathcal{J},\mathcal{J}_0)}\|_{\max} \le \frac{C_{\text{init}}}{s}\}, \end{aligned} \tag{S.97}$$

where $\widetilde{\mu}_{k\mathcal{J}_0}$ and $\widehat{\boldsymbol{\nu}}_k^{\mathcal{J}}$ are defined in Step 2(a) & 2(b) of Algorithm S.4, respectively.

We claim that $\mathcal{A}$ implies the existence of a permutation Π such that $\|\widehat{\boldsymbol{\mu}}_{\Pi(k)} - \boldsymbol{\mu}_k^*\|_{\max} \preceq \frac{1}{s}$. We now prove this claim. We consider two cases determined by the magnitude of $\|\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*\|_{\max}$.

Case 1: $\|\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*\|_{\max} \le \frac{8C_{\text{init}}}{s}$. For the permutation $\Pi_{\mathcal{J}}$ in $\mathcal{A}$, we have

$$\begin{aligned} &|\widehat{\nu}_{1\mathcal{J}} - \widehat{\nu}_{2\mathcal{J}}| = |\widehat{\nu}_{\Pi_{\mathcal{J}}(1),\mathcal{J}} - \widehat{\nu}_{\Pi_{\mathcal{J}}(2),\mathcal{J}}| && \text{(S.98)}\\ \le\; &|\widehat{\nu}_{\Pi_{\mathcal{J}}(1),\mathcal{J}} - \mu^*_{1\mathcal{J}}| + |\widehat{\nu}_{\Pi_{\mathcal{J}}(2),\mathcal{J}} - \mu^*_{2\mathcal{J}}| + \|\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*\|_{\max} && \text{(S.99)}\\ \le\; &\frac{10C_{\text{init}}}{s}. && \text{(S.100)}\end{aligned}$$

It follows that, for any permutation Π, we have

$$\begin{aligned} |\widehat{\nu}_{\Pi(k),\mathcal{J}} - \mu_{k\mathcal{J}^*}| \;\le\; &|\widehat{\nu}_{\Pi(k),\mathcal{J}} - \widehat{\nu}_{\Pi_{\mathcal{J}}(k),\mathcal{J}}| + |\widehat{\nu}_{\Pi_{\mathcal{J}}(k),\mathcal{J}} - \mu^*_{k\mathcal{J}}| && \text{(S.101)}\\ \le\; &|\widehat{\nu}_{1\mathcal{J}} - \widehat{\nu}_{2\mathcal{J}}| + |\widehat{\nu}_{\Pi_{\mathcal{J}}(k),\mathcal{J}} - \mu^*_{k\mathcal{J}}| && \text{(S.102)}\\ \le\; &\frac{11C_{\text{init}}}{s}. && \text{(S.103)}\end{aligned}$$

And thus our claim holds in this case with an arbitrary Π.

Case 2: $\|\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*\|_{\max} > \frac{8C_{\text{init}}}{s}$. Without loss of generality, assume that there exists a unique $\mathcal{J}_0^*$ such that $|\mu_{1\mathcal{J}_0^*} - \mu_{2\mathcal{J}_0^*}| = \|\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*\|_{\max}$ and for any $\mathcal{J} \neq \mathcal{J}_0^*$, we have $|\mu_{1\mathcal{J}_0^*} - \mu_{2\mathcal{J}_0^*}| - |\mu_{1\mathcal{J}} - \mu_{2\mathcal{J}}| > \frac{4C_{\text{init}}}{s}$.

Under $\mathcal{A}$, for any $\mathcal{J} \neq \mathcal{J}_0^*$, we have

$$\begin{aligned} |\tilde{\mu}_{1\mathcal{J}} - \tilde{\mu}_{2\mathcal{J}}| &\leq |\tilde{\mu}_{\Pi_{\mathcal{J}}(1),\mathcal{J}} - \mu^*_{1\mathcal{J}}| + |\mu^*_{1\mathcal{J}} - \mu^*_{2\mathcal{J}}| + |\tilde{\mu}_{\Pi_{\mathcal{J}}(2),\mathcal{J}} - \mu^*_{2\mathcal{J}}| && \text{(S.104)} \\ &\leq |\mu^*_{1\mathcal{J}} - \mu^*_{2\mathcal{J}}| + \frac{2C_{\text{init}}}{s} && \text{(S.105)} \\ &< |\mu_{1\mathcal{J}_0^*} - \mu_{2\mathcal{J}_0^*}| - \frac{2C_{\text{init}}}{s}. && \text{(S.106)} \end{aligned}$$

Meanwhile,

$$|\tilde{\mu}_{1\mathcal{J}_0^*} - \tilde{\mu}_{2\mathcal{J}_0^*}| \geq |\mu_{1\mathcal{J}_0^*} - \mu_{2\mathcal{J}_0^*}| - \sum_k |\tilde{\mu}_{\Pi_{\mathcal{J}_0^*}(k),\mathcal{J}_0^*} - \mu_{k\mathcal{J}_0^*}| \geq |\mu_{1\mathcal{J}_0^*} - \mu_{2\mathcal{J}_0^*}| - \frac{2C_{\text{init}}}{s} \quad \text{(S.107)}$$

It follows that $\mathcal{J}_0 = \mathcal{J}_0^*$. Also note that $|\widehat{\mu}_{\Pi_{\mathcal{J}_0}(k),\mathcal{J}_0} - \mu^*_{k\mathcal{J}_0}| \leq \frac{C_{\text{init}}}{s}$.

For $\mathcal{J} \neq \mathcal{J}_0$, we have

$$\begin{aligned} |\widehat{\mu}^{(0)}_{k\mathcal{J}} - \mu^*_{k\mathcal{J}}| &= |\widehat{\nu}_{\Lambda_{\mathcal{J}}(k),\mathcal{J}} - \mu_{k\mathcal{J}^*}| && \text{(S.108)} \\ &\leq |\widehat{\nu}_{\Lambda_{\mathcal{J}}(k),\mathcal{J}} - \widehat{\nu}_{\Upsilon_{\mathcal{J}}(k),\mathcal{J}}| + |\widehat{\nu}_{\Upsilon_{\mathcal{J}}(k),\mathcal{J}} - \mu^*_{k\mathcal{J}}| && \text{(S.109)} \\ &\equiv L_1 + L_2. && \text{(S.110)} \end{aligned}$$

Under $\mathcal{A}$, we have $L_2 \leq \frac{C_{\text{init}}}{s}$. We further show that L_1 by proving that $\Lambda_{\mathcal{J}} = \Upsilon_{\mathcal{J}}$. Because $\Lambda_{\mathcal{J}}$ minimizes $\max_k\{|\widehat{\nu}_{\Lambda_{\mathcal{J}}(k),\mathcal{J}_0} - \tilde{\mu}_{k\mathcal{J}_0}|\}$, it suffices to show that

$$\max_k\{|\widehat{\nu}_{\Lambda_{\mathcal{J}}(k),\mathcal{J}_0} - \tilde{\mu}_{k\mathcal{J}_0}|\} < \max_k\{|\widehat{\nu}_{\bar{\Lambda}_{\mathcal{J}}(k),\mathcal{J}_0} - \tilde{\mu}_{k\mathcal{J}_0}|\}, \quad \text{(S.111)}$$

where $\bar{\Lambda}_{\mathcal{J}} = 3 - \Lambda_{\mathcal{J}}$ is the only other permutation from $\{1,2\}$ to $\{1,2\}$. Now note that (S.111) is implied by

$$|\widehat{\nu}_{\Lambda_{\mathcal{J}}(k),\mathcal{J}_0} - \tilde{\mu}_{k\mathcal{J}_0}| < |\widehat{\nu}_{\bar{\Lambda}_{\mathcal{J}}(k),\mathcal{J}_0} - \tilde{\mu}_{k\mathcal{J}_0}| \text{ for } k = 1,2. \quad \text{(S.112)}$$

We show that (S.112) is true. Without loss of generality, we assume that there is no label switching in $\tilde{\mu}_{k\mathcal{J}}$, i.e, $|\tilde{\mu}_{k\mathcal{J}} - \mu^*_{k\mathcal{J}}| < |\tilde{\mu}_{k'\mathcal{J}} - \mu^*_{k\mathcal{J}}|$ for $k' \neq k$. Note that the left-hand side

$$|\widehat{\nu}_{\Lambda_{\mathcal{J}}(k),\mathcal{J}_0} - \tilde{\mu}_{k\mathcal{J}_0}| \leq |\widehat{\nu}_{\Lambda_{\mathcal{J}}(k),\mathcal{J}_0} - \mu^*_{k\mathcal{J}_0}| + |\mu^*_{k\mathcal{J}_0} - \tilde{\mu}_{k\mathcal{J}_0}| \leq \frac{2C_{\text{init}}}{s}. \quad \text{(S.113)}$$

For the right-hand side, we note that for any $k' \neq k$, we must have $\bar{\Lambda}_{\mathcal{J}}(k) = \Lambda_{\mathcal{J}}(k')$. Hence,

$$\begin{aligned} & |\widehat{\nu}_{\bar{\Lambda}_{\mathcal{J}}(k),\mathcal{J}_0} - \tilde{\mu}_{k\mathcal{J}_0}| && \text{(S.114)} \\ = \; & |\widehat{\nu}_{\Lambda_{\mathcal{J}}(k')} - \tilde{\mu}_{k\mathcal{J}_0}| = |(\widehat{\nu}_{\Lambda_{\mathcal{J}},\mathcal{J}_0} - \mu^*_{k'\mathcal{J}_0}) + (\mu^*_{k'\mathcal{J}_0} - \mu_{k\mathcal{J}_0}) + (\mu^*_{k\mathcal{J}_0} - \tilde{\mu}_{k\mathcal{J}_0})| && \text{(S.115)} \\ \geq \; & |\mu_{1\mathcal{J}_0^*} - \mu_{2\mathcal{J}_0^*}| - \sum_k |\widehat{\nu}_{\Lambda_{\mathcal{J}}(k),\mathcal{J}_0} - \mu^*_{k\mathcal{J}_0}| && \text{(S.116)} \\ < \; & \frac{2C_{\text{init}}}{s}. && \text{(S.117)} \end{aligned}$$

As a result, $\Lambda_{\mathcal{J}} = \Upsilon_{\mathcal{J}}$, and our claim is true in this case, too.

By Proposition S.11, we have $\Pr(\mathcal{A}) \geq 1 - \frac{1}{\prod_m p_m}$. Consequently, there exists a permutation Π such that

$$\Pr(\|\widehat{\boldsymbol{\mu}}^{(0)}_{\Pi(k)} - \boldsymbol{\mu}^*_k\|_{\max} \preceq \frac{1}{s}) \geq 1 - \frac{1}{\prod_m p_m}. \tag{S.118}$$

The bound on $\widehat{\boldsymbol{\Sigma}}^{(0)}_m$. By Proposition S.3, we have that

$$\|\bigotimes_{m=M}^{m=1} \widehat{\boldsymbol{\Sigma}}^{(0)}_m - \bigotimes_{m=M}^{m=1} \boldsymbol{\Sigma}^*_m\|_{\max} \preceq O(\|\widehat{\boldsymbol{\Sigma}}^{(0)}_m - \boldsymbol{\Sigma}_m\|_{\max}). \tag{S.119}$$

Hence, it suffices to find a bound for the probability that $\|\widehat{\boldsymbol{\Sigma}}_m - \boldsymbol{\Sigma}_m\|_{\max} \preceq \frac{1}{s}$. Consider the event

$$\mathcal{B} = \{|\widehat{\text{var}}(X_{i,1\cdots 1} \mid Y) - \text{var}(X_{1\cdots 1} \mid Y)| \preceq \frac{1}{s}, \|\widehat{\boldsymbol{\Psi}}^{(0)}_m - a_m \boldsymbol{\Sigma}^*_m\|_{\max} \preceq \frac{1}{s}\}, \tag{S.120}$$

where $a_m = \frac{1}{q_m} \mathbf{1}^\intercal_{q_m} \boldsymbol{\Sigma}^*_{-m} \mathbf{1}_{q_m}$. By Lemma S.8,

$$\frac{1}{q_m} \mathbf{X}_{(m)} \mathbf{1}_{q_m} \sim \pi^*_1 N(\boldsymbol{\phi}_1, a_m \boldsymbol{\Sigma}^*_m) + \pi^*_2 N(\boldsymbol{\phi}_2, a_m \boldsymbol{\Sigma}^*_m), \tag{S.121}$$

where $a_m = \frac{1}{q_m} \mathbf{1}^\intercal_{q_m} \boldsymbol{\Sigma}^*_{-m} \mathbf{1}_{q_m}$. In other words, $\frac{1}{q_m} \mathbf{X}_{(m)} \mathbf{1}_{q_m}$ follows Model (A1). By letting $\epsilon = \frac{C}{s^2}$ and $\delta = \prod_m p_m^{-3}$ in Proposition S.11, with probability greater than $1 - C \prod_m p_m^{-1}$, we have $\Pr(\mathcal{B}) \preceq \frac{1}{s}$.

We claim that event $\mathcal{B}$ implies that $\|\widehat{\boldsymbol{\Sigma}}_m - \boldsymbol{\Sigma}^*_m\|_{\max} \preceq \frac{1}{s}$. We proceed to prove this claim. Because eigenvalues of $\boldsymbol{\Sigma}^*_m$ are bounded below and above, a_m is also bounded below and above. For ease of presentation, we let $\boldsymbol{\Psi}_m = a_m \boldsymbol{\Sigma}^*_m$. When we rescale $\widehat{\Psi}^{(0)}_m$ to $\widetilde{\Psi}_m$ such that the upper-left corner is equal to 1, we must have

$$\begin{aligned}
& \|\widetilde{\Psi}^{(0)}_m - (\Psi_{m,11})^{-1} \boldsymbol{\Psi}_m\|_{\max} && \text{(S.122)} \\
= \; & \|(\hat{\Psi}_{m,11})^{-1} \widehat{\boldsymbol{\Psi}}^{(0)}_m - (\Psi_{m,11})^{-1} \boldsymbol{\Psi}_m\|_{\max} && \text{(S.123)} \\
\leq \; & \|(\hat{\Psi}^{(0)}_{m,11})^{-1} (\widehat{\boldsymbol{\Psi}}^{(0)}_m - \boldsymbol{\Psi}_m)\|_{\max} + |\widehat{\Psi}^{(0)}_{m,11} - \Psi^{-1}_{m,11}| \cdot \|\boldsymbol{\Psi}_m\|_{\max} && \text{(S.124)} \\
\preceq \; & \|\widehat{\boldsymbol{\Psi}}^{(0)}_m - \boldsymbol{\Psi}_m\|_{\max} \preceq \frac{1}{s}. && \text{(S.125)}
\end{aligned}$$

Because $\widehat{\boldsymbol{\Sigma}}^{(0)}_m = (\hat{\Psi}^{(0)}_{m,11})^{-1} \widehat{\boldsymbol{\Psi}}^{(0)}_m$, $\boldsymbol{\Sigma}^*_m = (\Psi_{m,11})^{-1} \boldsymbol{\Psi}_m$ for $m > 1$, we have $\|\widehat{\boldsymbol{\Sigma}}_m - \boldsymbol{\Sigma}^*_m\|_{\max} \preceq \frac{1}{s}$ for $m > 1$.

For $m = 1$, we let $\sigma = \mathrm{var}(X_{i,1\cdots 1} \mid Y)$ and $\widehat{\sigma} = \widehat{\mathrm{var}}(X_{i,1\cdots 1} \mid Y)$. Then

$$\|\widehat{\mathbf{\Sigma}}_1^{(0)} - \mathbf{\Sigma}_1^*\|_{\max} = \|\widehat{\sigma}\widetilde{\mathbf{\Psi}}_1^{(0)} - \mathbf{\Sigma}_1^*\|_{\max} \tag{S.126}$$

$$\leq \ |\hat{\sigma} - \sigma| \cdot \|\widetilde{\mathbf{\Psi}}_1^{(0)}\|_{\max} + \sigma\|\widetilde{\mathbf{\Psi}}_1^{(0)} - \sigma^{-1}\mathbf{\Sigma}_1^*\|_{\max} \tag{S.127}$$

$$\preceq \ \max\{|\widehat{\sigma} - \sigma|, \|\widetilde{\mathbf{\Psi}}_1^{(0)} - \sigma^{-1}\mathbf{\Sigma}_1^*\|_{\max}\} \preceq \frac{1}{s}. \tag{S.128}$$

Therefore, with a probability greater than $1 - O(\prod_m p_m^{-1})$, we have that $\widehat{\boldsymbol{\mu}}_k^{(0)}, \bigotimes_{m=M}^{m=1} \widehat{\mathbf{\Sigma}}_m^{(0)}$ satisfy Condition (IC). By Proposition S.10 we have the desired conclusion.

References

Cai, T. T., Ma, J. & Zhang, L. (2019), 'Chime: Clustering of high-dimensional gaussian mixtures with em algorithm and its optimality', *The Annals of Statistics* **47**(3), 1234–1267.

Hardt, M. & Price, E. (2015), Tight bounds for learning a mixture of two gaussians, *in* 'Proceedings of the forty-seventh annual ACM symposium on Theory of computing', ACM, pp. 753–760.

Kolda, T. G. & Bader, B. W. (2009), 'Tensor decompositions and applications', *SIAM Review* **51**(3), 455–500.

Lyu, X., Sun, W. W., Wang, Z., Liu, H., Yang, J. & Cheng, G. (2019), 'Tensor graphical model: Non-convex optimization and statistical inference', *IEEE transactions on pattern analysis and machine intelligence* .

Pan, Y., Mai, Q. & Zhang, X. (2019), 'Covariate-adjusted tensor classification in high dimensions', *Journal of the American statistical association* **114**(527), 1305–1319.

Petersen, K. B. & Pedersen, M. S. (2008), 'The matrix cookbook', *Technical University of Denmark* **7**(15), 510.

Wang, M. & Zeng, Y. (2019), Multiway clustering via tensor block models, *in* 'Advances in Neural Information Processing Systems', pp. 715–725.