Supplementary Materials for "Likelihood-based Dimension Folding on Tensor Data"

Ning Wang, Xin Zhang

Department of Statistics, Florida State University, Tallahassee, FL, 32306

and Bing Li

Department of Statistics, Pennsylvania State University, University Park, PA 16802

S.1 Primary biliary cirrhosis data analysis

In this section, we analyze the data from a Mayo Clinical trial on primary biliary cirrhosis (PBC). This data set is also used in Xue & Yin (2014) and Sheng & Yuan (2019). The data set contains laboratory results for 312 patients. Several biomarkers such as bilirubin, albumin and prothrombin time are adopted to diagnose PBC. We study this data set in the same way as Xue & Yin (2014) did. We treat the time baseline as one fold of the covariates and the multivariate predictors repeatedly measured over time as another fold of the covariates and thus we have predictors in a matrix form. We focus our attention on the measurements of bilirubin, albumin level, and prothrombin time at time points 6 months, 1 year, 2 years, and 3 years and thus the predictors are in the form of 3×4 matrix. The response is the survival time of a patient which is continuous. We partition the response variable into 8 slices, and treat each slice as one class, then apply FLAD and FELAD to get the estimation of the dimension folding subspace. To make a comparison with Xue & Yin (2014) and Sheng & Yuan (2019), we use $d_1 = d_2 = 1$.

The estimated basis matrices $\mathbf{\Gamma}_1$ and $\mathbf{\Gamma}_2$ are $(0.151, -0.989, 0.007)^T$ and $(0.274, 0.240, -0.009, 0.931)^T$ by FLAD, and $(0.025, -0.999, 0.023)^T$ and $(0.094, 0.469, -0.761, 0.438)^T$ by FELAD. To test the sig-

		$\mathbf{\Gamma}_{1,11}$	$\mathbf{\Gamma}_{1,21}$	$\mathbf{\Gamma}_{1,31}$	$\mathbf{\Gamma}_{2,11}$	$\mathbf{\Gamma}_{2,21}$	$\mathbf{\Gamma}_{2,31}$	$\mathbf{\Gamma}_{2,41}$
FLAD	Est.	0.151	-0.989	0.007	0.274	0.240	-0.009	0.931
	pvalue	0.002	0.006	0.904	0.484	0.216	0.736	0
FELAD	Est.	0.025	-0.999	0.023	0.094	0.469	-0.761	0.438
	pvalue	0.712	0.020	0.964	0.906	0.232	0.064	0.420
FSIR	Est.	0.018	-0.227	-0.974	0.623	0.204	0.727	0.208
	pvalue	0.002	0	0	0.606	0.596	0.598	0.616
FDR	Est.	0.057	-0.152	-0.987	-0.086	-0.241	-0.153	0.955
	pvalue	0.490	0	0.594	0.980	0.052	0.896	0.266
DCOV	Est.	0.989	-0.144	0.036	0.390	0.401	0.495	0.665
	pvalue	0	0	0.056	0	0	0	0
FMAVE	Est.	0.119	-0.989	-0.087	0.491	0.1811	0.159	0.837
	pvalue	0	0	0.366	0.170	0.200	0.586	0.146

Table S1: Bootstrapped p-values of coefficients estimated by different methods with bootstrap sample size 1000.

nificance of the coefficients, we compute the bootstrapped p-value for each element of $\mathbf{\Gamma}_1$ and $\mathbf{\Gamma}_2$. Table S1 indicates that, for FLAD, the bilirubin and albumin levels at year 3 significantly affects the length of the survival time at level 0.05, and for FELAD, the albumin level significantly affects the length of the survival time at level 0.05. By Table S1, all the methods indicate that albumin significantly affects the length of survival time at level 0.05, and the results of FLAD, FSIR, DCOV and FMAVE show that bilirubin significantly affects the survival time. DCOV is the most stable method: the subspace estimation does not change so much with respect to different bootstrapped samples.

In Figure S1, we show the smoothing spline plots for the logarithm of the response versus the projected predictor, and scatter plots for the fitted value versus residuals. The results of FLAD and FELAD are

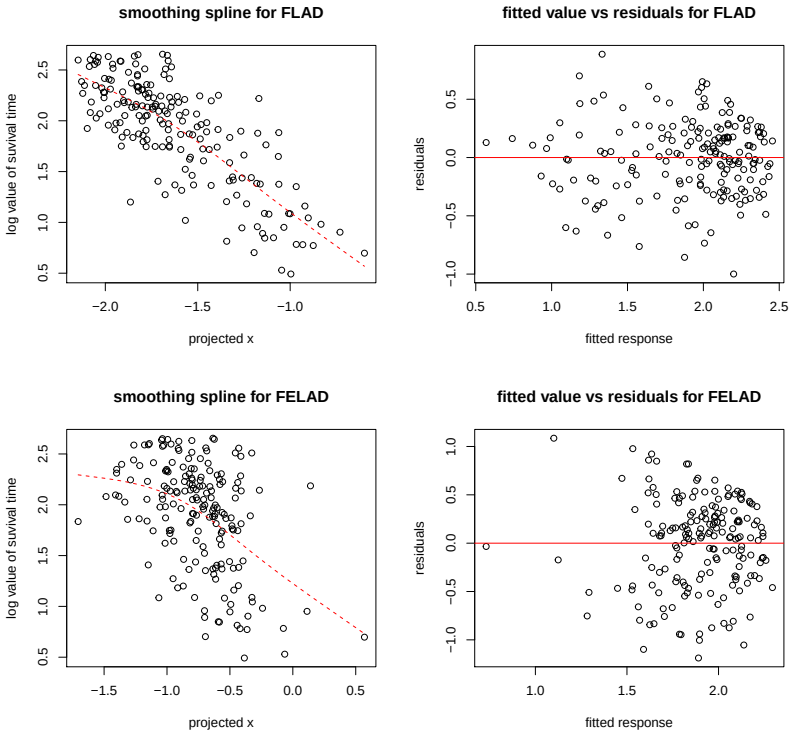

Figure S1: Left panels: Smoothing splines for the logarithm of response versus the projected predictor. Right panels: plots for the fitted values versus the residuals. The smoothing parameter is 0.076 for FLAD, and 0.033 for FELAD.

clear. It shows that the survival time has a negative relation with the projected predictor. Also, our analysis indicates that the survival time has a negative relationship with bilirubin level and a positive relationship with the albumin level, which is consistent with the medical outcome. The results are similar to those given in Xue & Yin (2014).

S.2 Proofs

S.2.1 Some Technical Lemmas

We first provide some lemmas that will be used in the proofs. The first three lemmas can be found in Appendix of Cook & Forzani (2009). Let $\mathbf{B}$ be a symmetric positive definite matrix, and at $(\boldsymbol{\alpha}, \boldsymbol{\alpha}_0) \in \mathbb{R}^{p\times p}$ be a full rank matrix with $\boldsymbol{\alpha}^T\boldsymbol{\alpha}_0 = 0$.

Lemma 1.

$$\boldsymbol{\alpha}(\boldsymbol{\alpha}^T\mathbf{B}\boldsymbol{\alpha})^{-1}\boldsymbol{\alpha}^T + \mathbf{B}^{-1}\boldsymbol{\alpha}_0(\boldsymbol{\alpha}_0^T\mathbf{B}^{-1}\boldsymbol{\alpha}_0)^{-1}\boldsymbol{\alpha}_0^T\mathbf{B}^{-1} = \mathbf{B}^{-1}. \tag{1}$$

As a consequence of Lemma 1, $\mathbf{I}_p - \mathbf{P}^T_{\boldsymbol{\alpha}(\mathbf{B})} = \mathbf{P}_{\boldsymbol{\alpha}_0(\mathbf{B}^{-1})}$. Additionally, if $(\boldsymbol{\alpha}, \boldsymbol{\alpha}_0)$ is an orthogonal matrix, then we have the following two lemmas.

Lemma 2.

$$|\boldsymbol{\alpha}_0^T\mathbf{B}\boldsymbol{\alpha}_0| = |\mathbf{B}||\boldsymbol{\alpha}^T\mathbf{B}^{-1}\boldsymbol{\alpha}|. \tag{2}$$

Lemma 3.

$$\begin{aligned}(\boldsymbol{\alpha}_0^T\mathbf{B}^{-1}\boldsymbol{\alpha}_0)^{-1} &= \boldsymbol{\alpha}_0^T\mathbf{B}\boldsymbol{\alpha}_0 - \boldsymbol{\alpha}_0^T\mathbf{B}\boldsymbol{\alpha}(\boldsymbol{\alpha}^T\mathbf{B}\boldsymbol{\alpha})^{-1}\boldsymbol{\alpha}^T\mathbf{B}\boldsymbol{\alpha}_0,\\ -(\boldsymbol{\alpha}_0^T\mathbf{B}^{-1}\boldsymbol{\alpha}_0)^{-1}(\boldsymbol{\alpha}_0^T\mathbf{B}^{-1}\boldsymbol{\alpha}) &= (\boldsymbol{\alpha}_0^T\mathbf{B}\boldsymbol{\alpha})(\boldsymbol{\alpha}^T\mathbf{B}\boldsymbol{\alpha})^{-1}.\end{aligned} \tag{3}$$

We will also use the following two lemmas about matrix derivatives.

Lemma 4. *Suppose* $\mathbf{A}$,$\mathbf{B}$,$\mathbf{C}$, $\mathbf{X}$ *are full rank matrix,*$\mathbf{D}$ *is full rank symmetric matrix. Then we have:*

$$\frac{d(\mathrm{tr}((\mathbf{AXB}+\mathbf{C})\mathbf{D}^{-1}(\mathbf{AXB}+\mathbf{C})^T))}{d\mathbf{X}} = 2\mathbf{A}^T(\mathbf{AXB}+\mathbf{C})\mathbf{DB}^T. \tag{4}$$

Lemma 5. *Suppose* $\mathbf{A}$,$\mathbf{B}$ *are full rank matrices, and* $\mathbf{X}$ *is symmetric matrix. Then we have:*

$$\frac{d(\mathrm{tr}(\mathbf{AX}^{-1}\mathbf{B}))}{d\mathbf{X}} = -(\mathbf{X}^{-1}\mathbf{BAX}^{-1})^T. \tag{5}$$

S.2.2 Proof of Proposition 1

Proof. Let $\boldsymbol{\Gamma}_m$ be the basis matrix for $\mathcal{S}_m$, $\boldsymbol{\Gamma} = \bigotimes_{m=M}^{1} \boldsymbol{\Gamma}_m$, $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0)$ be a full rank matrix with $\boldsymbol{\Gamma}^T\boldsymbol{\Gamma}_0 = 0$, and $\mathbf{P}_{\boldsymbol{\Gamma}(\boldsymbol{\Sigma}_k)} = \boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_k\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_k$. It is easy to show that $Y \mid \mathbf{X} \sim Y \mid (\bigotimes_{m=M}^{1} \mathbf{P}_{\mathcal{S}_m})\mathrm{vec}(\mathbf{X})$ is equivalent to $\mathbf{X} \mid (\boldsymbol{\Gamma}^T\mathbf{X}, Y) \sim \mathbf{X} \mid \boldsymbol{\Gamma}^T\mathbf{X}$. By definition, $\mathrm{cov}(\mathrm{vec}(\mathbf{X}) \mid \boldsymbol{\Gamma}^T\mathbf{X}, Y = k) = (\mathbf{I}_{\prod_{m=1}^{M} p_m} - \mathbf{P}^T_{\boldsymbol{\Gamma}(\boldsymbol{\Sigma}_k)})\boldsymbol{\Sigma}_k$ and $\mathrm{E}(\mathrm{vec}(\mathbf{X}) \mid \boldsymbol{\Gamma}^T\mathrm{vec}(\mathbf{X}), Y = k) = \mathrm{vec}(\boldsymbol{\mu}) + \boldsymbol{\eta}\boldsymbol{\omega}_k + \mathbf{P}^T_{\boldsymbol{\Gamma}(\boldsymbol{\Sigma}_k)}(\mathrm{vec}(\mathbf{X}) - \mathrm{vec}(\boldsymbol{\mu}) - \boldsymbol{\eta}\boldsymbol{\omega}_k)$, where $\boldsymbol{\omega}_k$ is the vector in $\mathbb{R}^{dim(\mathcal{M})}$, and $\boldsymbol{\eta}$ is a basis matrix for $\mathcal{M}$. It is easy to show that $\mathrm{var}(\mathrm{vec}(\mathbf{X}) \mid \boldsymbol{\Gamma}^T\mathbf{X}, Y = k)$ and $\mathrm{E}(\mathrm{vec}\mathbf{X} \mid \boldsymbol{\Gamma}^T\mathrm{vec}(\mathbf{X}), Y = k)$ are constant if and only if $(\mathbf{I}_{\prod_{m=1}^{M} p_m} - \mathbf{P}^T_{\boldsymbol{\Gamma}(\boldsymbol{\Sigma}_k)})\boldsymbol{\Sigma}_k$ and $\mathbf{P}^T_{\boldsymbol{\Gamma}(\boldsymbol{\Sigma}_k)}$ are constant and $\mathbf{P}^T_{\boldsymbol{\Gamma}(\boldsymbol{\Sigma}_k)}\boldsymbol{\eta} = \boldsymbol{\eta}$.

By Lemma 1, $(\mathbf{I}_{\prod_{m=1}^{M} p_m} - \mathbf{P}^T_{\boldsymbol{\Gamma}(\boldsymbol{\Sigma}_k)})\boldsymbol{\Sigma}_k = \boldsymbol{\Gamma}_0(\boldsymbol{\Gamma}_0^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Gamma}_0)^{-1}\boldsymbol{\Gamma}_0^T = C_1$ for some constant C_1. Because $\mathbf{P}^T_{\boldsymbol{\Gamma}(\boldsymbol{\Sigma}_k)}$, $k = 1, \cdots, K$, are constant, we know that $\mathbf{I} - \mathbf{P}^T_{\boldsymbol{\Gamma}(\boldsymbol{\Sigma}_k)} = \boldsymbol{\Gamma}_0(\boldsymbol{\Gamma}_0^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Gamma}_0)^{-1}\boldsymbol{\Gamma}_0^T\boldsymbol{\Sigma}_k^{-1} = C_2$ for some constant C_2. So $\boldsymbol{\Gamma}_0\boldsymbol{\Sigma}_k^{-1}$ is a constant, which implies that $\mathbf{Q}_{\bigotimes_{m=M}^{1} \mathcal{S}_m}\boldsymbol{\Sigma}_k^{-1}$ does not change with k. Because $\mathbf{P}^T_{\boldsymbol{\Gamma}(\boldsymbol{\Sigma}_k)}$ is a constant, $\mathbf{P}^T_{\boldsymbol{\Gamma}(\boldsymbol{\Sigma}_k)}\boldsymbol{\eta} = \mathbf{P}^T_{\boldsymbol{\Gamma}(\boldsymbol{\Sigma})}\boldsymbol{\eta} = \boldsymbol{\eta}$ for any k. And $\mathbf{P}^T_{\boldsymbol{\Gamma}(\boldsymbol{\Sigma})}\boldsymbol{\eta} = \boldsymbol{\eta}$ if and only if $\mathbf{P}_{\boldsymbol{\Gamma}(\boldsymbol{\Sigma})}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}) = \boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}$. So we have $\boldsymbol{\Sigma}^{-1}\boldsymbol{\eta} \subseteq \bigotimes_{m=M}^{1} \mathcal{S}_m$, which is equivalent to $\boldsymbol{\Sigma}^{-1}\mathcal{M} \subseteq \bigotimes_{m=M}^{1} \mathcal{S}_m$. □

S.2.3 Proof of Proposition 2

Proof. From the proof of Proposition 1, condition "$\mathbf{Q}_{\bigotimes_{m=M}^{1} \mathcal{S}_m}\boldsymbol{\Sigma}_k^{-1}$ is a constant" of Proposition 1 is equivalent to $\mathbf{P}_{\boldsymbol{\Gamma}(\boldsymbol{\Sigma}_k)} = \mathbf{P}_{\boldsymbol{\Gamma}(\boldsymbol{\Sigma})}$ and $\boldsymbol{\Sigma}_k(\mathbf{I}_{\prod_{m=1}^{M} p_m} - \mathbf{P}_{\boldsymbol{\Gamma}(\boldsymbol{\Sigma}_k)}) = \boldsymbol{\Sigma}(\mathbf{I}_{\prod_{m=1}^{M} p_m} - \mathbf{P}_{\boldsymbol{\Gamma}(\boldsymbol{\Sigma})})$, which is equivalent to $\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma} = \mathbf{P}^T_{\boldsymbol{\Gamma}(\boldsymbol{\Sigma})}(\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma})\mathbf{P}_{\boldsymbol{\Gamma}(\boldsymbol{\Sigma})}$. Also $\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma} = \mathbf{P}^T_{\boldsymbol{\Gamma}(\boldsymbol{\Sigma})}(\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma})\mathbf{P}_{\boldsymbol{\Gamma}(\boldsymbol{\Sigma})}$ is equivalent to $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_k) \subseteq \bigotimes_{m=M}^{1} \mathcal{S}_m$. See Cook & Forzani (2009) for the proof of the equivalences.

By Proposition 1, we have $\boldsymbol{\Sigma}^{-1}\mathcal{M} \subseteq \bigotimes_{m=M}^{1} \mathcal{S}_m$ and $\boldsymbol{\Sigma}^{-1}\mathcal{M} \subseteq \bigotimes_{m=M}^{1} \widetilde{\mathcal{S}}_m$, so $\boldsymbol{\Sigma}^{-1}\mathcal{M} \subseteq \bigotimes_{m=M}^{1} \mathcal{S}_m \cap \bigotimes_{m=M}^{1} \widetilde{\mathcal{S}}_m$. Also, from Proposition 1 and the previous equivalences, we have $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}) \subseteq \bigotimes_{m=M}^{1} \mathcal{S}_m$ and $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}) \subseteq \bigotimes_{m=M}^{1} \widetilde{\mathcal{S}}_m$, and hence $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}) \subseteq \bigotimes_{m=M}^{1} \mathcal{S}_m \cap \bigotimes_{m=M}^{1} \widetilde{\mathcal{S}}_m$. If we can show $\bigotimes_{m=M}^{1} \mathcal{S}_m \cap \bigotimes_{m=M}^{1} \widetilde{\mathcal{S}}_m = \bigotimes_{m=M}^{1} (\mathcal{S}_m \cap \widetilde{\mathcal{S}}_m)$, then Proposition 2 is

proved.

We first show $\bigotimes_{m=M}^{1}(\mathcal{S}_m \cap \widetilde{\mathcal{S}}_m) \subseteq \bigotimes_{m=M}^{1}\mathcal{S}_m \cap \bigotimes_{m=M}^{1}\widetilde{\mathcal{S}}_m$. For $m = 1, \cdots, M$, and any $\boldsymbol{\delta}_m \in \mathcal{S}_m \cap \widetilde{\mathcal{S}}_m$, we have $\bigotimes_{m=M}^{1}\boldsymbol{\delta}_m \in \bigotimes_{m=M}^{1}\mathcal{S}_m$ and $\bigotimes_{m=M}^{1}\boldsymbol{\delta}_m \in \bigotimes_{m=M}^{1}\widetilde{\mathcal{S}}_m$. So $\bigotimes_{m=M}^{1}\boldsymbol{\delta}_m \in \bigotimes_{m=M}^{1}\mathcal{S} \cap \bigotimes_{m=M}^{1}\widetilde{\mathcal{S}}_m$. Then we know that $\bigotimes_{m=M}^{1}(\mathcal{S}_m \cap \widetilde{\mathcal{S}}_m) \subseteq \bigotimes_{m=M}^{1}\mathcal{S}_m \cap \bigotimes_{m=M}^{1}\widetilde{\mathcal{S}}_m$.

Next, we show $\bigotimes_{m=M}^{1}\mathcal{S}_m \cap \bigotimes_{m=M}^{1}\widetilde{\mathcal{S}}_m \subseteq \bigotimes_{m=M}^{1}(\mathcal{S}_m \cap \widetilde{\mathcal{S}}_m)$. For any vector $\boldsymbol{\nu} \in \bigotimes_{m=M}^{1}\mathcal{S}_m \cap \bigotimes_{m=M}^{1}\widetilde{\mathcal{S}}_m$, there exist $\boldsymbol{\alpha}_m \in \mathcal{S}_m$ and $\boldsymbol{\beta}_m \in \widetilde{\mathcal{S}}_m$, $m = 1, \cdots, M$, such that $\boldsymbol{\nu} = \bigotimes_{m=M}^{1}\boldsymbol{\alpha}_m = \bigotimes_{m=M}^{1}\boldsymbol{\beta}_m$. It can be shown that $\bigotimes_{m=M}^{1}\boldsymbol{\alpha}_m = \bigotimes_{m=M}^{1}\boldsymbol{\beta}_m$ if and only if $\boldsymbol{\alpha}_m = c_m\boldsymbol{\beta}_m$ for some constants c_m. This implies that $\boldsymbol{\alpha}_m \in \widetilde{\mathcal{S}}_m$ and hence $\boldsymbol{\alpha}_m \in \mathcal{S}_m \cap \widetilde{\mathcal{S}}_m$. Therefore, $\boldsymbol{\nu} = \bigotimes_{m=M}^{1}\boldsymbol{\alpha}_m \in \bigotimes_{m=M}^{1}(\mathcal{S}_m \cap \widetilde{\mathcal{S}}_m)$ and $\bigotimes_{m=M}^{1}\mathcal{S}_m \cap \bigotimes_{m=M}^{1}\widetilde{\mathcal{S}}_m \subseteq \bigotimes_{m=M}^{1}(\mathcal{S}_m \cap \widetilde{\mathcal{S}}_m)$. □

S.2.4 Proof of Proposition 3

Proof. Let $\bigotimes_{m=M}^{1}\boldsymbol{\Gamma}_m$ be the basis matrix for $\bigotimes_{m=M}^{1}\mathcal{S}_m$. If $\mathcal{S}_m$ is a mode-m dimension folding envelope subspace, it must be a dimension folding subspace. From Proposition 1, $\mathcal{S}_m$ is a mode-m dimension folding subspace if (a) $\boldsymbol{\Sigma}^{-1}\mathcal{M} \subseteq \bigotimes_{m=M}^{1}\mathcal{S}_m$ and (b) $\mathbf{Q}_{\bigotimes_{m=M}^{1}\mathcal{S}_m}\boldsymbol{\Sigma}_k^{-1}$ is a constant with respect to k. The envelope covariance structure naturally satisfies (b). So $\boldsymbol{\Sigma}^{-1}\mathcal{M} \subseteq \bigotimes_{m=M}^{1}\mathcal{S}_m$ and $\boldsymbol{\Sigma}_k = (\bigotimes_{m=M}^{1}\mathbf{P}_{\mathcal{S}_m})\boldsymbol{\Sigma}_k(\bigotimes_{m=M}^{1}\mathbf{P}_{\mathcal{S}_m}) + \mathbf{Q}_{\bigotimes_{m=M}^{1}\mathcal{S}_m}\boldsymbol{\Sigma}\mathbf{Q}_{\bigotimes_{m=M}^{1}\mathcal{S}_m}$ make $\mathcal{S}_m$ to be a mode-m dimension folding envelope subspace. □

S.2.5 Proof of Proposition 4

Proof. From the proof of Proposition 2 in Wang et al. (2019), we have $\bigotimes_{m=M}^{1}\mathcal{S}_m \cap \bigotimes_{m=M}^{1}\widetilde{\mathcal{S}}_m$ satisfies the two conditions in our Proposition 3. We also proved that $\bigotimes_{m=M}^{1}\mathcal{S}_m \cap \bigotimes_{m=M}^{1}\widetilde{\mathcal{S}}_m = \bigotimes_{m=M}^{1}(\mathcal{S}_m \cap \widetilde{\mathcal{S}}_m)$ in our Proposition 2. So the intersection of two mode-m dimension folding envelope subspace is a mode-m dimension folding envelope subspace. □

S.2.6 Proof of Proposition 5

Proof. Let $\boldsymbol{\Gamma} = \bigotimes_{m=M}^{1} \boldsymbol{\Gamma}_m$ be the basis matrix for $\mathcal{T}_{Y|\mathbf{X}}$, and $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0)$ be a full rank matrix with $\boldsymbol{\Gamma}^T\boldsymbol{\Gamma}_0 = 0$. By Proposition 1, $\mathrm{vec}(\boldsymbol{\mu} - \boldsymbol{\mu}_k) \subseteq \boldsymbol{\Sigma}\mathcal{T}_{Y|\mathbf{X}}$. So we have $\mathrm{vec}(\boldsymbol{\mu} - \boldsymbol{\mu}_k) = \boldsymbol{\Sigma}\boldsymbol{\Gamma}\boldsymbol{\nu}_k$ for some $\boldsymbol{\nu}_k \in \mathbb{R}^{\prod_{m=1}^{M} d_m}$. By definition, we have $\boldsymbol{\Gamma}_0\mathrm{vec}(\mathbf{X}) \mid (\boldsymbol{\Gamma}^T\mathrm{vec}(\mathbf{X}), Y = k) \sim N(\rho_k, \boldsymbol{\Theta}_k)$, with $\rho_k = \boldsymbol{\Gamma}_0^T\mathrm{vec}(\boldsymbol{\mu}) + \boldsymbol{\Gamma}_0^T\boldsymbol{\Sigma}\boldsymbol{\Gamma}\boldsymbol{\nu}_k + (\boldsymbol{\Gamma}_0^T\boldsymbol{\Sigma}_k\boldsymbol{\Gamma})(\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_k\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T(\mathrm{vec}(\mathbf{X}) - \mathrm{vec}(\boldsymbol{\mu}) - \boldsymbol{\Sigma}\boldsymbol{\Gamma}\boldsymbol{\nu}_k)$, and $\boldsymbol{\Theta}_k = \boldsymbol{\Gamma}_0^T\boldsymbol{\Sigma}_k\boldsymbol{\Gamma}_0 - \boldsymbol{\Gamma}_0^T\boldsymbol{\Sigma}_k\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_k\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_k\boldsymbol{\Gamma}_0$. By Lemma 3, we have $\boldsymbol{\Theta}_k = (\boldsymbol{\Gamma}_0^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Gamma}_0)^{-1}$. Since, by Proposition 1, $\boldsymbol{\Gamma}_0^T\boldsymbol{\Sigma}_k^{-1}$ is a constant, we have $\boldsymbol{\Theta}_k = (\boldsymbol{\Gamma}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Gamma}_0)^{-1} = \mathbf{D}$. Again by Lemma 3 and Proposition 1, we have $(\boldsymbol{\Gamma}_0^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Gamma})(\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_k\boldsymbol{\Gamma})^{-1} = (\boldsymbol{\Gamma}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Gamma})(\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}\boldsymbol{\Gamma}) = \mathbf{H}$. Hence $\rho_k = \mathbf{H}\boldsymbol{\Gamma}^T\mathrm{vec}(\mathbf{X}) + (\boldsymbol{\Gamma}_0^T - \mathbf{H}\boldsymbol{\Gamma}^T)\mathrm{vec}(\boldsymbol{\mu})$. □

S.2.7 Proof of Proposition 6

Proof. By the equality $f(\boldsymbol{\Gamma}^T\mathrm{vec}(\mathbf{X}), \boldsymbol{\Gamma}_0^T\mathrm{vec}(\mathbf{X}) \mid Y = k) = f(\boldsymbol{\Gamma}^T\mathrm{vec}(\mathbf{X}) \mid Y = k) f(\boldsymbol{\Gamma}_0^T\mathrm{vec}(\mathbf{X}) \mid \boldsymbol{\Gamma}^T\mathrm{vec}(\mathbf{X}), Y = k)$ and Proposition 2, the log-likelihood function is

$$
\begin{aligned}
\mathrm{L}(\boldsymbol{\Gamma}) &= \sum_{i=1}^{n} \log f(\boldsymbol{\Gamma}^T\mathrm{vec}(\mathbf{X}_i), \boldsymbol{\Gamma}_0^T\mathrm{vec}(\mathbf{X}_i) \mid Y_i = k) \\
&= \sum_{i=1}^{n} \log f(\boldsymbol{\Gamma}^T\mathrm{vec}(\mathbf{X}_i) \mid Y_i = k) + \sum_{i=1}^{n} \log f(\boldsymbol{\Gamma}_0^T\mathrm{vec}(\mathbf{X}_i) \mid \boldsymbol{\Gamma}^T\mathrm{vec}(\mathbf{X}_i), Y_i = k) \\
&= -\frac{npq}{2}\log(2\pi) - \frac{n}{2}\log|\mathbf{D}| - \frac{1}{2}\sum_k n_k log|\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_k\boldsymbol{\Gamma}| \\
&\quad - \frac{1}{2}\sum_k n_k(\mathrm{vec}(\overline{\mathbf{X}}_k) - \mathrm{vec}(\boldsymbol{\mu}) - \boldsymbol{\Sigma}\boldsymbol{\Gamma}\nu_k)^T\boldsymbol{\Gamma}^T(\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_k\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}(\mathrm{vec}(\overline{\mathbf{X}}_k) - \mathrm{vec}(\boldsymbol{\mu}) - \boldsymbol{\Sigma}\boldsymbol{\Gamma}\boldsymbol{\nu}_k)) \\
&\quad - \frac{1}{2}\sum_k n_k(\mathrm{vec}(\overline{\mathbf{X}}_k) - \mathrm{vec}(\boldsymbol{\mu}))^T\mathbf{K}\mathbf{D}^{-1}\mathbf{K}^T(\mathrm{vec}(\overline{\mathbf{X}}_k) - \mathrm{vec}(\boldsymbol{\mu})) \\
&\quad - \frac{1}{2}\sum_k n_k\mathrm{tr}\{\boldsymbol{\Gamma}^T\tilde{\boldsymbol{\Sigma}}_k\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_k\boldsymbol{\Gamma})^{-1}\} - \frac{1}{2}\sum_k n_k\mathrm{tr}(\mathbf{K}\mathbf{D}^{-1}\mathbf{K}^T\widetilde{\boldsymbol{\Sigma}}_k),
\end{aligned}
$$

where $\widetilde{\mathbf{\Sigma}}_k$ is the sample covariance matrix of $\mathbf{X}_k$, and $\mathbf{K} = (\mathbf{\Gamma}_0 - \mathbf{\Gamma}\mathbf{H}^T)$. Let $f_k = n_k/n$. The only term in $\mathrm{L}(\mathbf{\Gamma})$ that involves $\boldsymbol{\nu}_k$ is

$$T = \sum_k n_k(\mathrm{vec}(\overline{\mathbf{X}}_k) - \mathrm{vec}(\boldsymbol{\mu}) - \mathbf{\Sigma}\mathbf{\Gamma}\boldsymbol{\nu}_k)^T\mathbf{\Gamma}^T(\mathbf{\Gamma}^T\mathbf{\Sigma}_k\mathbf{\Gamma})^{-1}\mathbf{\Gamma}(\mathrm{vec}(\overline{\mathbf{X}}_y) - \mathrm{vec}(\boldsymbol{\mu}) - \mathbf{\Sigma}\mathbf{\Gamma}\boldsymbol{\nu}_k)).$$

We need to minimize it subject to $\sum_k f_k\nu_k = 0$. Let $\mathbf{B}_k = \mathbf{\Gamma}^T\mathbf{\Sigma}_k\mathbf{\Gamma}$, and $\mathbf{Z}_k = \Gamma^T(\mathrm{vec}(\overline{\mathbf{X}}_k) - \mathrm{vec}(\boldsymbol{\mu}))$. For any quantity a_k, let $\overline{a} = \sum_k f_k a_k$. We use the Lagrange multiplier method to minimize $T/n = \sum_k f_k(\mathbf{Z}_k - \overline{\mathbf{B}}\boldsymbol{\nu}_k)^T\mathbf{B}_k^{-1}(\mathbf{Z}_k - \overline{B}\boldsymbol{\nu}_k) + \lambda\overline{\boldsymbol{\nu}}$. Differentiating it with respect to $\boldsymbol{\nu}_k$, we get

$$-2f_k\mathbf{Z}_k + 2f_k\overline{\mathbf{B}}\boldsymbol{\nu}_k + f_k\mathbf{B}_k\overline{\mathbf{B}}^{-1}\lambda = 0. \tag{6}$$

By summation over the classes k of the above equation, we get $-2\overline{\mathbf{Z}} + \lambda = 0$. Finally, substituting it back into (6), we have $\boldsymbol{\nu}_k = \overline{\mathbf{B}}^{-1}(\mathbf{Z}_k - \mathbf{B}_k\overline{\mathbf{B}}^{-1}\overline{\mathbf{Z}})$. Therefore

$$\begin{aligned} T &= n(\mathbf{\Gamma}^T\mathrm{vec}(\overline{\mathbf{X}}) - \mathbf{\Gamma}^T\mathrm{vec}(\boldsymbol{\mu}))^T\overline{\mathbf{B}}^{-1}(\mathbf{\Gamma}^T\mathrm{vec}(\overline{\mathbf{X}}) - \mathbf{\Gamma}^T\mathrm{vec}(\boldsymbol{\mu})) \\ &= n(\mathrm{vec}(\overline{\mathbf{X}}) - \mathrm{vec}(\boldsymbol{\mu}))^T\mathbf{\Gamma}\overline{\mathbf{B}}^{-1}\mathbf{\Gamma}^T(\mathrm{vec}(\overline{\mathbf{X}}) - \mathrm{vec}(\boldsymbol{\mu})). \end{aligned}$$

Notice that, in $\mathrm{L}(\mathbf{\Gamma})$, the term

$$\begin{aligned} &\frac{1}{2}\sum_k n_k(\mathrm{vec}(\overline{\mathbf{X}}_k) - \mathrm{vec}(\boldsymbol{\mu}))^T\mathbf{K}\mathbf{D}^{-1}\mathbf{K}^T(\mathrm{vec}(\overline{\mathbf{X}}_k) - \mathrm{vec}(\boldsymbol{\mu})) \\ =&\frac{1}{2}\sum_k n_k(\mathrm{vec}(\overline{\mathbf{X}}) - \mathrm{vec}(\boldsymbol{\mu}))^T\mathbf{K}\mathbf{D}^{-1}\mathbf{K}^T(\mathrm{vec}(\overline{\mathbf{X}}) - \mathrm{vec}(\boldsymbol{\mu})) \\ &+ \frac{1}{2}\sum_k n_k\mathrm{tr}(\mathbf{K}\mathbf{D}^{-1}\mathbf{K}^T((\mathrm{vec}(\overline{\mathbf{X}}_k) - \mathrm{vec}(\overline{\mathbf{X}}))(\mathrm{vec}(\overline{\mathbf{X}}_k) - \mathrm{vec}(\overline{\mathbf{X}}))^T)). \end{aligned}$$

The only two terms in $\mathrm{L}(\mathbf{\Gamma})$ that involve $\mathrm{vec}(\boldsymbol{\mu})$ are one from $\frac{1}{2}T$ and the other one from above equation. Hence we can solve the equation $\frac{\partial \mathrm{L}}{\partial \mathrm{vec}(\boldsymbol{\mu})} = 0$, and obtain $\mathbf{\Gamma}\overline{\mathbf{B}}^{-1}\mathbf{\Gamma}^T(\mathrm{vec}(\overline{\mathbf{X}}) - \mathrm{vec}(\boldsymbol{\mu})) + \mathbf{K}\mathbf{D}^{-1}\mathbf{K}^T(\mathrm{vec}(\overline{\mathbf{X}}) - \mathrm{vec}(\boldsymbol{\mu})) = 0$, where $\overline{\mathbf{B}} = \mathbf{\Gamma}^T\mathbf{\Sigma}\mathbf{\Gamma}$. By Lemma 1, $\mathbf{K}^T = \mathbf{\Gamma}_0^T(\mathbf{I} - \mathbf{\Sigma}\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Sigma}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T) = (\mathbf{\Gamma}_0^T\mathbf{\Sigma}^{-1}\mathbf{\Gamma}_0)^{-1}\mathbf{\Gamma}_0^T\mathbf{\Sigma}^{-1}$, and $\mathbf{K}\mathbf{D}^{-1}\mathbf{K}^T = \mathbf{\Sigma}^{-1}\mathbf{\Gamma}_0(\mathbf{\Gamma}_0^T\mathbf{\Sigma}^{-1}\mathbf{\Gamma}_0)^{-1}\mathbf{\Gamma}_0^T\mathbf{\Sigma}^{-1}$. Again, using Lemma 1, $\mathbf{\Sigma}^{-1}\mathbf{\Gamma}_0(\mathbf{\Gamma}_0^T\mathbf{\Sigma}^{-1}\mathbf{\Gamma}_0)^{-1}\mathbf{\Gamma}_0^T\mathbf{\Sigma}^{-1} + \mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Sigma}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T = \mathbf{\Sigma}^{-1}$. Hence we have $n\mathbf{\Sigma}^{-1}(\mathrm{vec}(\overline{\mathbf{X}}) - \mathrm{vec}(\boldsymbol{\mu})) = 0$.

Then $\mathrm{L}(\boldsymbol{\Gamma})$ is maximized with respect to $\mathrm{vec}(\boldsymbol{\mu})$ when $\mathrm{vec}(\boldsymbol{\mu}) = \overline{\mathbf{X}}$. Substituting $\overline{\mathbf{X}}$ for $\mathrm{vec}(\boldsymbol{\mu})$, the likelihood function becomes

$$\begin{aligned}\mathrm{L}(\boldsymbol{\Gamma}) = &-\frac{npq}{2}\log 2(\pi) - \frac{n}{2}\log|\mathbf{D}| - \frac{1}{2}\sum_k n_k \log|\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_k\boldsymbol{\Gamma}| \\ &- \frac{1}{2}\sum_k n_k \mathrm{tr}\{\boldsymbol{\Gamma}^T\widetilde{\boldsymbol{\Sigma}}_k\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_k\boldsymbol{\Gamma})^{-1}\} - \frac{1}{2}\sum_y n_y \mathrm{tr}(\mathbf{K}\mathbf{D}^{-1}\mathbf{K}^T\widetilde{\boldsymbol{\Delta}}_k),\end{aligned}$$

where $\widetilde{\boldsymbol{\Delta}}_k = \widetilde{\boldsymbol{\Sigma}}_k + (\overline{\mathbf{X}}_k - \overline{\mathbf{X}})(\overline{\mathbf{X}}_k - \overline{\mathbf{X}})^T$. Recall that $\mathbf{K} = (\boldsymbol{\Gamma}_0 - \boldsymbol{\Gamma}\mathbf{H}^T)$. So the only term that involves $\mathbf{H}$ is $\frac{1}{2}\sum_k n_k \mathrm{tr}((\boldsymbol{\Gamma}_0 - \boldsymbol{\Gamma}\mathbf{H}^T)\mathbf{D}^{-1}(\boldsymbol{\Gamma}_0 - \boldsymbol{\Gamma}\mathbf{H}^T)^T\widetilde{\boldsymbol{\Delta}}_y)$.

By solving the equality $\frac{\partial \mathrm{L}}{\partial \mathbf{H}} = 0$ and applying Lemma 4, we get $\sum_k n_k\mathbf{D}^{-1}\boldsymbol{\Gamma}_0^T\widetilde{\boldsymbol{\Delta}}_k\boldsymbol{\Gamma} + \sum_k n_k\mathbf{D}^{-1}\mathbf{H}\boldsymbol{\Gamma}^T\widetilde{\boldsymbol{\Delta}}_k\boldsymbol{\Gamma} = 0$. So $\widehat{\mathbf{H}} = (\sum_k n_k\boldsymbol{\Gamma}_0^T\widetilde{\boldsymbol{\Delta}}_k\boldsymbol{\Gamma})(\sum_k n_k\boldsymbol{\Gamma}^T\widetilde{\boldsymbol{\Delta}}_k\boldsymbol{\Gamma})^{-1} = (\boldsymbol{\Gamma}_0^T\widetilde{\boldsymbol{\Sigma}}_{\mathbf{X}}\boldsymbol{\Gamma})(\boldsymbol{\Gamma}^T\widetilde{\boldsymbol{\Sigma}}_{\mathbf{X}}\boldsymbol{\Gamma})^{-1}$ is the minimizer of $\mathrm{L}(\boldsymbol{\Gamma})$.

Next we solve the equation $\frac{\partial \mathrm{L}}{\partial \mathbf{D}} = 0$. By Lemma 5, the solution is

$$\begin{aligned}\widehat{\mathbf{D}} &= \widehat{\mathbf{K}}^T\widetilde{\boldsymbol{\Sigma}}_{\mathbf{X}}\widehat{\mathbf{K}} \\ &= (\boldsymbol{\Gamma}_0 - \widehat{\mathbf{H}}\boldsymbol{\Gamma}^T)\widetilde{\boldsymbol{\Sigma}}_{\mathbf{X}}(\boldsymbol{\Gamma}_0 - \widehat{\mathbf{H}}\boldsymbol{\Gamma}^T)^T \\ &= ((\boldsymbol{\Gamma}_0^T\widetilde{\boldsymbol{\Sigma}}_{\mathbf{X}}^{-1}\boldsymbol{\Gamma}_0)^{-1}\boldsymbol{\Gamma}_0^T\widetilde{\boldsymbol{\Sigma}}_{\mathbf{X}}^{-1})\widetilde{\boldsymbol{\Sigma}}_{\mathbf{X}}((\boldsymbol{\Gamma}_0^T\widetilde{\boldsymbol{\Sigma}}_{\mathbf{X}}^{-1}\boldsymbol{\Gamma}_0)^{-1}\boldsymbol{\Gamma}_0^T\widetilde{\boldsymbol{\Sigma}}_{\mathbf{X}}^{-1})^T \\ &= (\boldsymbol{\Gamma}_0^T\widetilde{\boldsymbol{\Sigma}}_{\mathbf{X}}^{-1}\boldsymbol{\Gamma})^{-1}.\end{aligned}$$

Finally by solving $\frac{\partial \mathrm{L}}{\partial \boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_k\boldsymbol{\Gamma}} = 0$, we get $\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_k\boldsymbol{\Gamma} = \boldsymbol{\Gamma}^T\widehat{\boldsymbol{\Sigma}}_k\boldsymbol{\Gamma}$. After plugging back $\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_k\boldsymbol{\Gamma} = \boldsymbol{\Gamma}^T\widehat{\boldsymbol{\Sigma}}_k\boldsymbol{\Gamma}$ into $\mathrm{L}(\boldsymbol{\Gamma})$, we prove the asserted result. □

S.2.8 Proof of Proposition 7

Proof. Let $\boldsymbol{\Gamma} = \bigotimes_{m=M}^1 \boldsymbol{\Gamma}_m$ be the basis matrix of $\mathcal{E}_{Y|X}$, and $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0)$ be orthonormal matrix, $\boldsymbol{\Gamma}^T\boldsymbol{\Gamma}_0 = 0$. Under FELAD model assumption, we have $\boldsymbol{\Sigma}_k = (\bigotimes_{m=M}^1 \boldsymbol{\Gamma}_m)\boldsymbol{\Omega}_k(\bigotimes_{m=M}^1 \boldsymbol{\Gamma}_m^T) + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$, by which we can show that $|\boldsymbol{\Sigma}_k| = |\boldsymbol{\Omega}_k||\boldsymbol{\Omega}_0|$ and $\boldsymbol{\Sigma}_k^{-1} = (\bigotimes_{m=M}^1 \boldsymbol{\Gamma}_m)\boldsymbol{\Omega}_k^{-1}(\bigotimes_{m=M}^1 \boldsymbol{\Gamma}_m^T) + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T$. By Proposition 1, $\mathrm{vec}(\boldsymbol{\mu}_k - \boldsymbol{\mu}) = \boldsymbol{\Gamma}\boldsymbol{\Sigma}\boldsymbol{\nu}_k$ for some vector $\boldsymbol{\nu}_k$. For the FELAD model, $\boldsymbol{\Sigma}\boldsymbol{\Gamma}\boldsymbol{\nu}_k =$

$\mathbf{\Gamma}(\sum_{k=1}^{K}\pi_k\mathbf{\Omega}_k)\boldsymbol{\nu}_k$. Let $\boldsymbol{\alpha}_k = (\sum_{k=1}^{K}\pi_k\mathbf{\Omega}_k)\boldsymbol{\nu}_k$. Then

$$\begin{aligned}\mathrm{L}(\mathbf{\Gamma}) = &-\frac{np}{2}\log(2\pi) - \frac{1}{2}\sum_{k=1}^{K} n_k \log|\mathbf{\Omega}_k| - \frac{n}{2}\log|\mathbf{\Omega}_0| \\ &- \frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(\mathbf{\Gamma}^T\mathrm{vec}(\mathbf{X}_{ki} - \overline{\boldsymbol{\mu}}) - \boldsymbol{\alpha}_k)^T\mathbf{\Omega}_k^{-1}(\mathbf{\Gamma}^T\mathrm{vec}(\mathbf{X}_{ki} - \overline{\boldsymbol{\mu}}) - \boldsymbol{\alpha}_k) \\ &- \frac{n}{2}\sum_{i=1}^{n}(\mathbf{\Gamma}_0^T\mathrm{vec}(\mathbf{X}_i - \overline{\boldsymbol{\mu}}))^T\mathbf{\Omega}_0^{-1}(\mathbf{\Gamma}_0^T\mathrm{vec}(\mathbf{X}_i - \overline{\boldsymbol{\mu}})).\end{aligned}$$

By solving the equation $\frac{\partial \mathrm{L}}{\partial \boldsymbol{\alpha}_k} = 0$, we have

$$\mathbf{\Omega}_k^{-1}(\mathbf{\Gamma}^T\mathrm{vec}(\mathbf{X}_{ki} - \overline{\boldsymbol{\mu}}) - \boldsymbol{\alpha}_k) = 0,$$

$$\boldsymbol{\alpha}_k = \mathbf{\Gamma}^T\mathrm{vec}(\overline{\mathbf{X}}_k - \overline{\boldsymbol{\mu}}).$$

Hence

$$\mathrm{L}(\mathbf{\Gamma}) = -\frac{np}{2}\log(2\pi) - \frac{1}{2}\sum_{k=1}^{K} n_k \log|\mathbf{\Omega}_k| - \frac{n}{2}\log|\mathbf{\Omega}_0| - \frac{1}{2}\sum_{k=1}^{K} n_k \mathrm{tr}(\mathbf{\Gamma}^T\widetilde{\mathbf{\Sigma}}_k\mathbf{\Gamma}^T\mathbf{\Omega}_k^{-1}) - \frac{n}{2}\mathrm{tr}\{\mathbf{\Gamma}_0^T\widetilde{\mathbf{\Sigma}}_{\mathbf{X}}\mathbf{\Gamma}_0\mathbf{\Omega}_0^{-1}\}.$$

By solving the equation $\frac{\partial \mathrm{L}}{\partial \mathbf{\Omega}_0} = 0$ and applying Lemma 4, we have

$$\mathbf{\Omega}_0 = \mathbf{\Gamma}_0^T\widetilde{\mathbf{\Sigma}}_{\mathbf{X}}\mathbf{\Gamma}_0.$$

By solving $\frac{\partial \mathrm{L}}{\partial \mathbf{\Omega}_k} = 0$ and using Lemma 4 again, we get

$$\mathbf{\Omega}_k = \mathbf{\Gamma}^T\widetilde{\mathbf{\Sigma}}_k\mathbf{\Gamma}.$$

Plugging the above relation into $\mathrm{L}(\mathbf{\Gamma})$, we have

$$\mathrm{L}(\mathbf{\Gamma}) = -\frac{np}{2}\log(2\pi) - \frac{np}{2} - \frac{n}{2}\log|\mathbf{\Gamma}_0^T\widetilde{\mathbf{\Sigma}}_{\mathbf{X}}\mathbf{\Gamma}_0| - \frac{1}{2}\sum_{k=1}^{K} n_k \log|\mathbf{\Gamma}^T\widetilde{\mathbf{\Sigma}}_k\mathbf{\Gamma}|.$$

By Lemma 2, we have $\log|\mathbf{\Gamma}_0^T\widetilde{\mathbf{\Sigma}}_{\mathbf{X}}\mathbf{\Gamma}_0| = \log|\mathbf{\Gamma}^T\widetilde{\mathbf{\Sigma}}_{\mathbf{X}}^{-1}\mathbf{\Gamma}| + \log|\widetilde{\mathbf{\Sigma}}_{\mathbf{X}}|$. Finally, we have

$$\mathrm{L}(\mathbf{\Gamma}) = -\frac{np}{2}\log(2\pi + 1) - \frac{n}{2}\log|\widetilde{\mathbf{\Sigma}}_{\mathbf{X}}| - \frac{n}{2}\log|\mathbf{\Gamma}^T\widetilde{\mathbf{\Sigma}}_{\mathbf{X}}^{-1}\mathbf{\Gamma}| - \frac{1}{2}\sum_{k=1}^{K} n_k \log|\mathbf{\Gamma}^T\widetilde{\mathbf{\Sigma}}_k\mathbf{\Gamma}|.$$

. □

S.2.9 Proof of Proposition 8

Proof. For the FLAD model, the free parameter $\boldsymbol{\phi}^T$ can be written as $(\mathrm{vec}^T(\boldsymbol{\mu}), \mathrm{vec}^T(\boldsymbol{\alpha}_1), \cdots, \mathrm{vec}^T(\boldsymbol{\alpha}_{K-1})$,$\mathrm{vec}^T(\boldsymbol{\Gamma}_1), \cdots, \mathrm{vec}^T(\boldsymbol{\Gamma}_M), \mathrm{vech}^T(\boldsymbol{\Sigma}), \mathrm{vech}^T(\mathbf{M}_1), \cdots, \mathrm{vech}^T(\mathbf{M}_K))^T = (\boldsymbol{\phi}_1^T, \cdots, \boldsymbol{\phi}_{2K+M}^T)^T$ with parameter space $\boldsymbol{\Theta}$ being the Cartesian product of the parameter space for the individual components. Because the elements of $\mathbf{h}$ are analytic, they are twice continuously differentiable over $\boldsymbol{\Theta}$ and every point in $\boldsymbol{\Theta}$ is regular (Shapiro 1986, Definition 2.1), except on a set of Lebesgue measure 0.

Let $\widehat{\mathbf{x}}^T = (\mathrm{vec}^T(\widehat{\boldsymbol{\mu}}_1), \cdots, \mathrm{vec}^T(\widehat{\boldsymbol{\mu}}_K), \mathrm{vech}^T(\widetilde{\boldsymbol{\Sigma}}_1), \cdots, \mathrm{vech}^T(\widetilde{\boldsymbol{\Sigma}}_K))^T$ be the MLE under the full model. The FLAD discrepancy function is defined as $\mathrm{F}_{FLAD}(\widehat{\mathbf{x}}, \mathbf{h}) = \mathrm{L}_p(\widehat{\mathbf{x}} \mid \widehat{\mathbf{x}}) - \mathrm{L}_d(h \mid \widehat{\mathbf{x}})$, where

$$\mathrm{L}_d(\mathbf{h} \mid \widehat{\mathbf{x}}) = -\sum_{k=1}^{K} \frac{n_k}{2} \{\log|\boldsymbol{\Sigma}_k| + \mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}_k \boldsymbol{\Sigma}_k^{-1}) + (\overline{\mathbf{X}}_k - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\overline{\mathbf{X}}_k - \boldsymbol{\mu}_k)\}.$$

F_{FLAD} is an analytic function of $\widehat{\mathbf{x}}$ and $\mathbf{h}$. Also note that $\mathrm{F}_{FLAD} \geq 0$ for all $\widehat{x}$ and $\mathbf{h}$, and $\mathrm{F}_{FLAD} = 0$ if and only if $\widehat{\mathbf{x}} = \mathbf{h}$. So F_{FLAD} satisfies the necessary condition for a discrepancy function (Shapiro 1986).

Let $\mathbf{W} = \frac{1}{2} \frac{\partial^2 \mathrm{F}_{FLAD}}{\partial \mathbf{h} \partial \mathbf{h}^T}$ evaluated at $(\mathbf{h}_0, \mathbf{h}_0)$, where $\mathbf{h}_0$ is the true value of $\mathbf{h}$. It is easy to check $\mathbf{W} = \mathbf{J}$, which is also a sufficient condition for FLAD to give asymptotic efficient estimators (Shapiro 1986)(eqn.5.1). Then, using Proposition 4.1 in (Shapiro 1986), we have $\sqrt{n}(\widehat{\mathbf{h}} - \mathbf{h}) \xrightarrow{D} N(0, \mathbf{V})$, where $\mathbf{V} = \mathbf{H}(\mathbf{H}^T\mathbf{J}\mathbf{H})^\dagger \mathbf{H}^T \mathbf{J}\mathbf{J}^{-1} \mathbf{J}\mathbf{H}(\mathbf{H}^T\mathbf{J}\mathbf{H})^\dagger \mathbf{H}^T = \mathbf{H}(\mathbf{H}^T\mathbf{J}\mathbf{H})^\dagger \mathbf{H}^T$. By calculation, we have $\mathbf{V}_0^{-1/2}(\mathbf{V}_0 - \mathbf{V})\mathbf{V}_0^{-1/2} = \mathbf{I} - \mathbf{J}^{1/2}\mathbf{H}(\mathbf{H}^T\mathbf{J}\mathbf{H})^\dagger \mathbf{H}^T \mathbf{J}^{1/2} = \mathbf{Q}_{\mathbf{J}^{1/2}\mathbf{H}} \geq 0$.

The proof of the situation of FELAD is similar to that of FLAD, so we omit the details here.

Let $\mathbf{P}_0 = \mathbf{V}_0^{-1/2}\mathbf{V}\mathbf{V}_0^{-1/2}$, and $\mathbf{P}_1 = \mathbf{V}_0^{-1/2}\mathbf{V}_1\mathbf{V}_0^{-1/2}$. We have $\mathbf{P}_0 = \mathbf{J}^{1/2}\mathbf{H}(\mathbf{H}^T\mathbf{J}\mathbf{H})^\dagger \mathbf{H}^T \mathbf{J}^{1/2}$, $\mathbf{P}_1 = \mathbf{J}^{1/2}\mathbf{H}\mathbf{G}_1(\mathbf{G}_1^T\mathbf{H}^T\mathbf{J}\mathbf{H}\mathbf{G}_1)^\dagger \mathbf{G}_1^T\mathbf{H}^T\mathbf{J}^{1/2}$, and $\mathbf{P}_0\mathbf{P}_1 = \mathbf{J}^{1/2}\{\mathbf{H}(\mathbf{H}^T\mathbf{J}\mathbf{H})^\dagger \mathbf{H}^T\mathbf{J}\mathbf{H}\}\mathbf{G}_1(\mathbf{G}_1^T\mathbf{H}^T\mathbf{J}\mathbf{H}\mathbf{G}_1)^\dagger \mathbf{G}_1^T\mathbf{H}^T\mathbf{J}^{1/2}$. Using matrix identity (Rao & Mitra 1971)(Theo 2.4(c))

$$\mathbf{H}(\mathbf{H}^T\mathbf{J}\mathbf{H})^\dagger \mathbf{H}^T\mathbf{J}\mathbf{H} = \mathbf{H},$$

and

$$(\mathbf{H}^T\mathbf{J}\mathbf{H})(\mathbf{H}^T\mathbf{J}\mathbf{H})^\dagger \mathbf{H}^T = \mathbf{H}^T,$$

we get $\mathbf{P}_0\mathbf{P}_1 = \mathbf{P}_1$. Similarly, we also have $\mathbf{P}_1\mathbf{P}_0 = \mathbf{P}_1$. Then $\mathbf{P}_0 - \mathbf{P}_1 = \mathbf{P}_0 - \mathbf{P}_0\mathbf{P}_1 = \mathbf{P}_0(\mathbf{I} - \mathbf{P}_1) = \mathbf{P}_{\mathbf{J}^{1/2}\mathbf{H}}\mathbf{Q}_{\mathbf{J}^{1/2}\mathbf{H}\mathbf{G}_1}$. Also $\mathbf{P}_0 - \mathbf{P}_1 = \mathbf{P}_0 - \mathbf{P}_1\mathbf{P}_0 = (\mathbf{I} - \mathbf{P}_1)\mathbf{P}_0 = \mathbf{Q}_{\mathbf{J}^{1/2}\mathbf{H}\mathbf{G}_1}\mathbf{P}_{\mathbf{J}^{1/2}\mathbf{H}}$.

□

S.2.10 Proof of Proposition 9

The proof of this proposition is parallel to the proof of Proposition 8. However, the Fisher information $\mathbf{J}$ is changed and may not equal to $\mathbf{W} = \frac{1}{2}\frac{\partial^2 \mathrm{F}_{FLAD}}{\partial \mathbf{h}\partial \mathbf{h}^T}$ at $(\mathbf{h}_0, \mathbf{h}_0)$. In this case, by Proposition 4.1 in Shapiro (1986), the estimator $\widehat{\mathbf{h}}$ is still $\sqrt{n}$-consistent, but may not be the most efficient one.

S.2.11 Proof of Proposition 10

Proof. The population objective function of FLAD is

$$\begin{aligned}
\mathrm{L}(\mathcal{S}) &= \frac{1}{2}\log|\boldsymbol{\Gamma}_0^T\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\boldsymbol{\Gamma}_0| - \frac{1}{2}\sum_{k=1}^{K}\pi_k\log|\boldsymbol{\Gamma}_0^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Gamma}_0| - \frac{1}{2}\log|\boldsymbol{\Sigma}_{\mathbf{X}}| + \frac{1}{2}\sum_{k=1}^{K}\pi_k\log|\boldsymbol{\Sigma}_k| \\
&\leq \frac{1}{2}\log|\boldsymbol{\Gamma}_0^T(\sum_{k=1}^{K}\pi_k\boldsymbol{\Sigma}_k)^{-1}\boldsymbol{\Gamma}_0| - \frac{1}{2}\sum_{k=1}^{K}\pi_k\log|\boldsymbol{\Gamma}_0^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Gamma}_0| - \frac{1}{2}\log|\boldsymbol{\Sigma}_{\mathbf{X}}| + \frac{1}{2}\sum_{k=1}^{K}\pi_k\log|\boldsymbol{\Sigma}_k| \\
&\leq -\frac{1}{2}\log|\boldsymbol{\Sigma}_{\mathbf{X}}| + \frac{1}{2}\sum_{k=1}^{K}\pi_k\log|\boldsymbol{\Sigma}_k|.
\end{aligned}$$

where $\boldsymbol{\Gamma}_m \in \mathbb{R}^{p_m \times d_m}$, $\boldsymbol{\Gamma}_0$ is the orthogonal complement of $\boldsymbol{\Gamma} = \bigotimes_{m=M}^{1}\boldsymbol{\Gamma}_m$, and $\mathcal{S} = \mathrm{span}(\boldsymbol{\Gamma})$. The first inequality follows since $\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \leq (\sum_{k=1}^{K}\pi_k\boldsymbol{\Sigma}_k)^{-1}$ and the second inequality follows since the function $\log|\boldsymbol{\Gamma}_0^T\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}_0|$ is convex in $\boldsymbol{\Delta}$ on the space formed by symmetric and positive definite matrices (See Section A.6 of Cook & Forzani (2009)). Define $\mathcal{S}_{FLAD} = \mathrm{argmax}\,\mathrm{L}(\mathcal{S})$. If we find an orthogonal matrix $(\boldsymbol{\eta}, \boldsymbol{\eta}_0)$, where $\boldsymbol{\eta}_m \in \mathbb{R}^{p_m \times d_m}$ and $\boldsymbol{\eta} = \bigotimes_{m=M}^{1}\boldsymbol{\eta}_m$, such that $\frac{1}{2}\log|\boldsymbol{\eta}_0^T\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\boldsymbol{\eta}_0| = \frac{1}{2}\sum_{k=1}^{K}\pi_k\log|\boldsymbol{\eta}_0^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\eta}_0|$, then $\mathcal{S}_{FLAD} = \mathrm{span}(\boldsymbol{\eta})$. Next, we show how to get the desired $(\boldsymbol{\eta}, \boldsymbol{\eta}_0)$ such that $\mathrm{span}(\boldsymbol{\eta}) \subseteq \mathcal{T}_{Y|\mathbf{X}}$ and $\frac{1}{2}\log|\boldsymbol{\eta}_0^T\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\boldsymbol{\eta}_0| = \frac{1}{2}\sum_{k=1}^{K}\pi_k\log|\boldsymbol{\eta}_0^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\eta}_0|$.

We first introduce the definition of "Kronecker envelope" proposed by Li et al. (2010).

Definition 1. *Let* $\mathbf{U} \in \mathbb{R}^{\prod_{m=1}^{M} p_M \times l}$ *be a random matrix. There are subspaces* $\mathcal{S}_1 \subseteq \mathbb{P}^{p_m}$, $m = 1, \cdots, M$ *such that (i)* $\mathrm{span}(\mathbf{U}) \subseteq \bigotimes_{m=M}^{1} \mathcal{S}_m$ *almost surely; and (ii) If there exists other subspaces* $\mathcal{S}_1^{'} \subseteq \mathbb{P}^{p_m}$, $m = 1, \cdots, M$, *that satisfies Condition (i), then* $\bigotimes_{m=M} \mathcal{S}_m \subseteq \bigotimes_{m=M} \mathcal{S}_m^{'}$. *Then the subspace* $\bigotimes_{m=M} \mathcal{S}_m$ *is called the Kronecker envelope of* $\mathbf{U}$, *denoted as* $\mathcal{E}^{\otimes}(\mathbf{U})$.

We choose the random matrix $\mathbf{U}$ to be $\mathbf{\Sigma}_{\mathbf{X}}^{-1}(\mathbf{\Sigma}_{\mathbf{X}} - \mathrm{cov}(\mathbf{X} \mid Y))$, where Y take values in $\{1, \cdots, K\}$. Under the condition that $\mathrm{E}(\mathrm{vec}(\mathbf{X}) \mid \boldsymbol{\beta}^T \mathrm{vec}(\mathbf{X}))$ is linear in $\boldsymbol{\beta}^T \mathrm{vec}(\mathbf{X})$, and $\mathrm{var}(\mathrm{vec}(\mathbf{X}) \mid \boldsymbol{\beta}^T \mathrm{vec}(\mathbf{X}))$ is nonrandom, Cook & Weisberg (1991) showed that the subspace spanned by $\mathbf{U} = \mathbf{\Sigma}_{\mathbf{X}}^{-1}(\mathbf{\Sigma}_{\mathbf{X}} - \mathrm{cov}(\mathbf{X} \mid Y))$ is subspace of $\mathcal{S}_{Y|\mathrm{vec}(\mathbf{X})}$. By Theorem 5 of Li et al. (2010), we have $\mathcal{E}^{\otimes}(\mathbf{U}) \subseteq \mathcal{T}_{Y|\mathbf{X}}$. Let $\boldsymbol{\eta} = \bigotimes_{m=M} \boldsymbol{\eta}_m$ be a basis matrix for $\mathcal{E}^{\otimes}(\mathbf{U})$. Then $\mathbf{\Sigma}_{\mathbf{X}}^{-1}(\mathbf{\Sigma}_{\mathbf{X}} - \mathbf{\Sigma}_k) = \boldsymbol{\eta}\mathbf{w}_k$ for some matrix $\mathbf{w}_k$, $k = 1, \cdots, K$. Consequently, $\mathbf{\Sigma}_{\mathbf{X}}^{-1}(\mathbf{\Sigma}_{\mathbf{X}} - \mathbf{\Sigma}_k) = \mathbf{P}_{\boldsymbol{\eta}(\mathbf{\Sigma}_{\mathbf{X}})} \mathbf{\Sigma}_{\mathbf{X}}^{-1}(\mathbf{\Sigma}_{\mathbf{X}} - \mathbf{\Sigma}_k)$. Thus, $\mathbf{\Sigma}_{\mathbf{X}} = \mathbf{\Sigma}_k + \mathbf{P}_{\boldsymbol{\eta}(\mathbf{\Sigma}_{\mathbf{X}})}^T (\mathbf{\Sigma}_{\mathbf{X}} - \mathbf{\Sigma}_k) = \mathbf{\Sigma}_k + \mathbf{P}_{\boldsymbol{\eta}(\mathbf{\Sigma}_{\mathbf{X}})}^T (\mathbf{\Sigma}_{\mathbf{X}} - \mathbf{\Sigma}_k) \mathbf{P}_{\boldsymbol{\eta}(\mathbf{\Sigma}_{\mathbf{X}})}$. Then, by direct multiplication, we have $\mathbf{\Sigma}_k^{-1} = \mathbf{\Sigma}_{\mathbf{X}}^{-1} + \boldsymbol{\eta}\{(\boldsymbol{\eta}^T \mathbf{\Sigma}_k \boldsymbol{\eta})^{-1} - (\boldsymbol{\eta}^T \mathbf{\Sigma}_{\mathbf{X}} \boldsymbol{\eta})^{-1}\}\boldsymbol{\eta}^T$, which implies $\frac{1}{2}\log|\boldsymbol{\eta}_0^T \mathbf{\Sigma}_{\mathbf{X}}^{-1} \boldsymbol{\eta}_0| = \frac{1}{2} \sum_{k=1}^{K} \pi_k \log|\boldsymbol{\eta}_0^T \mathbf{\Sigma}_k^{-1} \boldsymbol{\eta}_0|$. As such, $\mathcal{S}_{FLAD} = \mathrm{span}(\boldsymbol{\eta}) \subseteq \mathcal{T}_{Y|\mathbf{X}}$. The $\sqrt{n}$-consistent property of $\mathcal{S}_{FLAD}$ can be obtained from Proposition 4.1 of Shapiro (1986), the proof is similar to that of Proposition 8 and 9, we omit the details here. □

S.3 Closed-Form derivatives of Algorithm 1

The objective functions (3.8) in the paper can be solved by standard Stiefel or Grassmann manifold optimization packages, where we can plug in the closed-form derivatives to speed up the computation. Here

we give the closed-form derivatives when $M = 2$.

$$\begin{aligned}
&\frac{\partial(\log|(\mathbf{I}_{d_2}\otimes\mathbf{\Gamma}_1^T)\mathbf{M}(\mathbf{I}_{d_2}\otimes\mathbf{\Gamma}_1)|)}{\partial\mathrm{vec}(\mathbf{\Gamma}_1)}\\
= \quad & 2\mathrm{vec}\left\{\mathbf{M}(\mathbf{I}_{d_2}\otimes\mathbf{\Gamma}_1)\{(\mathbf{I}_{d_2}\otimes\mathbf{\Gamma}_1^T)\mathbf{M}(\mathbf{I}_{d_2}\otimes\mathbf{\Gamma}_1)\}^{-1}\right\}(\mathbf{I}_{d_2}\otimes\mathbf{T}_{d_1d_2}\otimes\mathbf{I}_{p_1})\{\mathrm{vec}(\mathbf{I}_{d_2})\otimes\mathbf{I}_{p_1d_1}\},\\
&\frac{\partial(\log|(\mathbf{\Gamma}_2^T\otimes\mathbf{I}_{d_1})\mathbf{M}(\mathbf{\Gamma}_2^T\otimes\mathbf{I}_{d_1})|)}{\partial\mathrm{vec}(\mathbf{\Gamma}_2)}\\
= \quad & 2\mathrm{vec}\left\{\mathbf{M}(\mathbf{\Gamma}_2\otimes\mathbf{I}_{d_1})\{(\mathbf{\Gamma}_2^T\otimes\mathbf{I}_{d_1})\mathbf{M}(\mathbf{\Gamma}_2\otimes\mathbf{I}_{d_1})\}^{-1}\right\}(\mathbf{I}_{d_2}\otimes\mathbf{T}_{d_1p_2}\otimes\mathbf{I}_{d_1})\{\mathbf{I}_{p_2d_2}\otimes\mathrm{vec}(\mathbf{I}_{d_1})\},
\end{aligned}$$

where $\mathbf{T}_{mn}$ is $mn\times mn$ permutation matrix whose ij-th element is 1 if $j = 1+m(i-1)-(mn-1)\lfloor\frac{i-1}{n}\rfloor$ or 0 otherwise.

References

Cook, R. D. & Forzani, L. (2009), 'Likelihood-based sufficient dimension reduction', *Journal of the American Statistical Association* **104**(485), 197–208.

Cook, R. D. & Weisberg, S. (1991), 'Comment: Sliced inverse regression for dimension reduction', *Journal of the American Statistical Association* **86**(414), 328–332.

Li, B., Kim, M. K., Altman, N. et al. (2010), 'On dimension folding of matrix-or array-valued statistical objects', *The Annals of Statistics* **38**(2), 1094–1121.

Rao, C. & Mitra, S. K. (1971), 'Generalized inverse of a matrix and its applications', *Berkeley Symposium on Mathematical Statistics and Probability* **1**, 601–620.

Shapiro, A. (1986), 'Asymototic theory of overparameterized structual models', *Journal of the American Statistical Association* **81**, 142–149.

Sheng, W. & Yuan, Q. (2019), 'Sufficient dimension folding in regression via distance covariance for matrix-valued predictors', *Statistical Analysis and Data Mining* .

Wang, W., Zhang, X. & Li, L. (2019), 'Common reducing subspace model and network alternation analysis', *Biometrics* **75**(4), 1109–1120.

Xue, Y. & Yin, X. (2014), 'Sufficient dimension folding for regression mean fuction', *Journal of Computatioanl and Graphical Statistics* **23**, 1028–1043.