Supplementary Materials: Proofs and Technical Details for "Parsimonious Tensor Response Regression"

Lexin Li and Xin Zhang

A Some preliminary results

We will apply the following two results repeatedly. For a positive definite matrix $\mathbf{B}$,

$$\mathbf{B} = \arg\min_{\mathbf{A}>0}\{\log|\mathbf{A}| + \mathrm{tr}(\mathbf{A}^{-1}\mathbf{B})\}.$$

In addition (Kolda, 2006; Proposition 3.7),

$$\mathbf{A} = \mathbf{B} \times_1 \mathbf{C}_1 \times_2 \cdots \times_N \mathbf{C}_N \quad \Leftrightarrow \quad \mathbf{A}_{(n)} = \mathbf{C}_n \mathbf{B}_{(n)} \left(\mathbf{C}_N \otimes \cdots \otimes \mathbf{C}_{n+1} \otimes \mathbf{C}_{n-1} \otimes \cdots \otimes \mathbf{C}_1\right)^{\mathsf{T}}.$$

For the envelope parameterization of $\mathbf{B} = [\![\mathbf{\Theta}; \mathbf{\Gamma}_1, \ldots, \mathbf{\Gamma}_m, \mathbf{I}_p]\!]$, we have

$$\mathbf{B}_{(m+1)} = \mathbf{\Theta}_{(m+1)}(\mathbf{\Gamma}_m^{\mathsf{T}} \otimes \cdots \otimes \mathbf{\Gamma}_1^{\mathsf{T}}), \text{ and } \mathbf{\Theta}_{(m+1)} = \mathbf{B}_{(m+1)}(\mathbf{\Gamma}_m \otimes \cdots \otimes \mathbf{\Gamma}_1).$$

As for $\mathbf{Y} = [\![\mathbf{Z}; \mathbf{\Gamma}_1, \ldots, \mathbf{\Gamma}_m]\!] \in \mathbb{R}^{r_1 \times \cdots \times r_m}$, we have the following results by treating $\mathbf{Y}$ as $(m+1)$-th order tensor with $r_{m+1} = 1$, and treating $\mathrm{vec}^{\mathsf{T}}(\mathbf{Y})$ as the mode-$(m+1)$ matricization.

$$\begin{aligned} \mathrm{vec}^{\mathsf{T}}(\mathbf{Z}) &= \mathrm{vec}^{\mathsf{T}}(\mathbf{Y} \times_1 \mathbf{\Gamma}_1^{\mathsf{T}} \times_2 \cdots \times_m \mathbf{\Gamma}_m^{\mathsf{T}}) = \mathrm{vec}^{\mathsf{T}}(\mathbf{Y})(\mathbf{\Gamma}_m \otimes \cdots \otimes \mathbf{\Gamma}_1), \\ \mathrm{vec}(\mathbf{Z}) &= (\mathbf{\Gamma}_m^{\mathsf{T}} \otimes \cdots \otimes \mathbf{\Gamma}_1^{\mathsf{T}})\mathrm{vec}(\mathbf{Y}). \end{aligned}$$

B Proof for Proposition 1 and 2

We first prove Proposition 1. Then Proposition 2 follows directly from the results of Proposition 1 and the definitions of reducing subspace and Tucker decomposition.

Under the tensor linear model $\mathbf{Y} = \mathbf{B}\bar{\times}_{(m+1)}\mathbf{X} + \boldsymbol{\epsilon}$, we can write

$$\mathbf{Y} \times_k \mathbf{Q}_k = (\mathbf{B}\bar{\times}_{(m+1)}\mathbf{X}) \times_k \mathbf{Q}_k + \boldsymbol{\epsilon} \times_k \mathbf{Q}_k = (\mathbf{B} \times_k \mathbf{Q}_k)\bar{\times}_{(m+1)}\mathbf{X} + \boldsymbol{\epsilon} \times_k \mathbf{Q}_k,$$

which implies that $\mathbf{Y} \times_k \mathbf{Q}_k | \mathbf{X} \sim \mathbf{Y} \times_k \mathbf{Q}_k$ is equivalent to $\mathbf{B} \times_k \mathbf{Q}_k = 0$.

Similarly, $\mathbf{Y} \times_k \mathbf{Q}_k \perp\!\!\!\perp \mathbf{Y} \times_k \mathbf{P}_k | \mathbf{X}$ is equivalent to $\boldsymbol{\epsilon} \times_k \mathbf{Q}_k \perp\!\!\!\perp \boldsymbol{\epsilon} \times_k \mathbf{P}_k$, where the independence of two tensor-valued random variable is defined as independence of their vectorized forms: $\mathrm{vec}(\boldsymbol{\epsilon} \times_k \mathbf{Q}_k) \perp\!\!\!\perp \mathrm{vec}(\boldsymbol{\epsilon} \times_k \mathbf{P}_k)$. Because of the tensor normal distribution, the independence is equivalent to $\mathrm{cov}\{\mathrm{vec}(\boldsymbol{\epsilon} \times_k \mathbf{Q}_k), \mathrm{vec}(\boldsymbol{\epsilon} \times_k \mathbf{P}_k)\} = 0$, and hence to

$$\boldsymbol{\Sigma}_m \otimes \cdots \otimes \boldsymbol{\Sigma}_{k+1} \otimes \mathbf{Q}_k\boldsymbol{\Sigma}_k\mathbf{P}_k \otimes \boldsymbol{\Sigma}_{k-1} \otimes \cdots \otimes \boldsymbol{\Sigma}_1 = 0.$$

Therefore, we have shown that

$$\mathbf{Y} \times_k \mathbf{Q}_k \perp\!\!\!\perp \mathbf{Y} \times_k \mathbf{P}_k | \mathbf{X} \Leftrightarrow \mathbf{Q}_k\boldsymbol{\Sigma}_k\mathbf{P}_k = 0 \Leftrightarrow \boldsymbol{\Sigma}_k = \mathbf{P}_k\boldsymbol{\Sigma}_k\mathbf{P}_k + \mathbf{Q}_k\boldsymbol{\Sigma}_k\mathbf{Q}_k.$$

C Proof for Proposition 3

We first show that for subspaces $\mathcal{E} = \mathcal{E}_m \otimes \cdots \otimes \mathcal{E}_1$ and covariances $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_m \otimes \cdots \otimes \boldsymbol{\Sigma}_1$, where $\mathcal{E}_k \subseteq \mathbb{R}^{r_k}$ and $\boldsymbol{\Sigma}_k \in \mathbb{R}^{r_k \times r_k}$, $k = 1, \ldots, m$, the two conditions are equivalent: (1) $\mathcal{E}$ reduces $\boldsymbol{\Sigma}$; (2) $\mathcal{E}_k$ reduces $\boldsymbol{\Sigma}_k$ for all $k = 1, \ldots, m$.

The statement (2) implies (1) is straightforward. We need to show that (1), $\mathcal{E} = \mathcal{E}_m \otimes \cdots \otimes \mathcal{E}_1$ reduces $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_m \otimes \cdots \otimes \boldsymbol{\Sigma}_1$, implies (2) , $\mathcal{E}_k$ reduces $\boldsymbol{\Sigma}_k$ for all $k = 1, \ldots, m$. By definition, (1) implies that $\mathbf{P}_{\mathcal{E}_m \otimes \cdots \otimes \mathcal{E}_1}(\boldsymbol{\Sigma}_m \otimes \cdots \otimes \boldsymbol{\Sigma}_1)\mathbf{Q}_{\mathcal{E}_m \otimes \cdots \otimes \mathcal{E}_1} = 0$, where $\mathbf{Q}_{\mathcal{E}_m \otimes \cdots \otimes \mathcal{E}_1} = \mathbf{I} - \mathbf{P}_{\mathcal{E}_m \otimes \cdots \otimes \mathcal{E}_1} = \mathbf{I} - \mathbf{P}_{\mathcal{E}_m} \otimes \cdots \otimes \mathbf{P}_{\mathcal{E}_1}$. Expanding this, we see that

$$\mathbf{P}_{\mathcal{E}_m}\boldsymbol{\Sigma}_m \otimes \cdots \otimes \mathbf{P}_{\mathcal{E}_1}\boldsymbol{\Sigma}_1 - \mathbf{P}_{\mathcal{E}_m}\boldsymbol{\Sigma}_m\mathbf{P}_{\mathcal{E}_m} \otimes \cdots \otimes \mathbf{P}_{\mathcal{E}_1}\boldsymbol{\Sigma}_1\mathbf{P}_{\mathcal{E}_1} = 0.$$

If we right-multiply $\mathbf{P}_{\mathcal{E}_m} \otimes \cdots \otimes \mathbf{P}_{\mathcal{E}_2} \otimes \mathbf{I}_{r_1}$ to both sides of the above equation, we obtain

$$\mathbf{P}_{\mathcal{E}_m}\boldsymbol{\Sigma}_m\mathbf{P}_{\mathcal{E}_m} \otimes \cdots \otimes \mathbf{P}_{\mathcal{E}_2}\boldsymbol{\Sigma}_2\mathbf{P}_{\mathcal{E}_2} \otimes (\mathbf{P}_{\mathcal{E}_1}\boldsymbol{\Sigma}_1 - \mathbf{P}_{\mathcal{E}_1}\boldsymbol{\Sigma}_1\mathbf{P}_{\mathcal{E}_1}) = 0,$$

which implies that $\mathbf{P}_{\mathcal{E}_1}\mathbf{\Sigma}_1 - \mathbf{P}_{\mathcal{E}_1}\mathbf{\Sigma}_1\mathbf{P}_{\mathcal{E}_1} = 0$. Since $\mathbf{P}_{\mathcal{E}_1}\mathbf{\Sigma}_1 = \mathbf{P}_{\mathcal{E}_1}\mathbf{\Sigma}_1(\mathbf{P}_{\mathcal{E}_1} + \mathbf{Q}_{\mathcal{E}_1})$, we conclude that $\mathbf{P}_{\mathcal{E}_1}\mathbf{\Sigma}_1\mathbf{Q}_{\mathcal{E}_1} = 0$ and hence $\mathcal{E}_1$ reduces $\mathbf{\Sigma}_1$. Using the similar argument, we can see that $\mathcal{E}_k$ reduces $\mathbf{\Sigma}_k$, for all $k = 1, \ldots, m$.

We next show that $\mathcal{E} = \mathcal{E}_m \otimes \cdots \otimes \mathcal{E}_1$ contains $\mathrm{span}(\mathbf{B}_{(m+1)}^\intercal)$ implies $\mathcal{E}_k$ contains $\mathrm{span}(\mathbf{B}_{(k)})$. Let $\mathbf{G}_k \in \mathbb{R}^{r_k \times q_k}$ be semi-orthogonal matrix such that $\mathrm{span}(\mathbf{G}_k) = \mathrm{span}(\mathbf{B}_{(k)})$, then we have Tucker decomposition of $\mathbf{B}$ as $\mathbf{B} = [\![\boldsymbol{\eta}; \mathbf{G}_1, \ldots, \mathbf{G}_m, \mathbf{I}_p]\!]$ for some $\boldsymbol{\eta} \in \mathbb{R}^{q_1 \times \cdots \times q_m \times p}$. Therefore we can write $\mathbf{B}_{(m+1)}^\intercal = (\mathbf{G}_m \otimes \cdots \otimes \mathbf{G}_1)\boldsymbol{\eta}_{(m+1)}^\intercal$. Hence $\mathcal{E} = \mathcal{E}_m \otimes \cdots \otimes \mathcal{E}_1$ contains $\mathrm{span}(\mathbf{B}_{(m+1)}^\intercal)$ implies that $\mathcal{E}_k$ contains $\mathrm{span}(\mathbf{G}_k) = \mathrm{span}(\mathbf{B}_{(k)})$.

D Proof for Lemma 1

Here we aim to show that $\mathbf{\Sigma}_k^{(0)}$, the Kronecker covariance estimator from Manceur and Dutilleul (2013), is a $\sqrt{n}$-consistent estimator for $\mathbf{\Sigma}_k$, for all k, if $\boldsymbol{\varepsilon}_i$, $i = 1, \ldots, n$, are i.i.d. from tensor normal distribution with mean 0 and covariances $\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_m$.

First we note that the mode-k matricization $\boldsymbol{\varepsilon}_{i(k)} \in \mathbb{R}^{r_k \times (\prod_{j \neq k} r_j)}$, $i = 1, \ldots, n$, are i.i.d. from a matrix normal distribution with mean 0 and covariances $\mathbf{\Sigma}_k$ (the row covariance matrix) and $\mathbf{\Delta}_k \equiv \mathbf{\Sigma}_m \otimes \cdots \otimes \mathbf{\Sigma}_{k+1} \otimes \mathbf{\Sigma}_{k-1} \otimes \cdots \otimes \mathbf{\Sigma}_1$ (the column covariance matrix). Then following Gupta and Nagar (1999; Chapter 2, Theorem 2.3.5), we have the following second-order moments:

$$\mathrm{E}(\boldsymbol{\varepsilon}_{i(k)}\boldsymbol{\varepsilon}_{i(k)}^\intercal) \quad = \quad \mathbf{\Sigma}_k \mathrm{tr}(\mathbf{\Delta}_k),$$

where $\mathrm{tr}(\cdot)$ is the trace operator of matrices.

In the iterative algorithm of obtaining $\mathbf{\Sigma}_k^{(0)}$, the starting value for $\mathbf{\Sigma}_k^{(0)}$ is

$$\frac{1}{n\prod_{j \neq k} r_j} \sum_{i=1}^{n} \mathbf{e}_{i(k)}\mathbf{e}_{i(k)}^\intercal,$$

which is thus $\sqrt{n}$-consistent for $\mathbf{\Sigma}_k$ up to a scalar difference due to $\mathrm{tr}(\mathbf{\Delta}_k)$. As we discussed in Section 4.1 regarding the identifiability of $\mathbf{\Sigma}_k$'s, the scalar difference can be resolved after normalizing $\mathbf{\Sigma}^{(0)} = \tau\mathbf{\Sigma}_m^{(0)} \otimes \cdots \otimes \mathbf{\Sigma}_1^{(0)}$ according to the scalar $\tau = (n\prod_j r_j)^{-1} \sum_{i=1}^n \mathrm{vec}(\mathbf{e}_i)\{(\mathbf{\Sigma}_m^{(0)})^{-1} \otimes \cdots \otimes (\mathbf{\Sigma}_1^{(0)})^{-1}\}\mathrm{vec}^\intercal(\mathbf{e}_i)$ at the end of each iteration.

Therefore, after the first iteration, we have obtained $\sqrt{n}$-consistent estimators $\boldsymbol{\Sigma}_k^{(0)}$ for $\boldsymbol{\Sigma}_k$. In the iterations that follow, the updating equations of $\boldsymbol{\Sigma}_k^{(0)}$ for $k = 1, \ldots, m$, are obtained by maximizing the tensor normal likelihood function. Hence it is guaranteed that the final estimators of $\boldsymbol{\Sigma}_k^{(0)}$ from the algorithm are also $\sqrt{n}$-consistent for $\boldsymbol{\Sigma}_k$.

E Proof for Theorem 1

E.1 Consistency of the one-step estimator

From Cook and Zhang (2016; Proposition 6) we know that if $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{U}}$ are $\sqrt{n}$-consistent estimators for $\mathbf{M} > 0$ and $\mathbf{U} \geq 0$, then the 1D algorithm for minimizing $J_n(\mathbf{G}) = \log|\mathbf{G}^\intercal\widehat{\mathbf{M}}\mathbf{G}| + \log|\mathbf{G}^\intercal(\widehat{\mathbf{M}} + \widehat{\mathbf{U}})^{-1}\mathbf{G}|$ produce $\sqrt{n}$-consistent estimator for the projection onto the envelope $\mathcal{E}_{\mathbf{M}}(\mathbf{U})$. We use this result to prove the $\sqrt{n}$-consistency of $\mathbf{P}^{os}_{\boldsymbol{\Gamma}_k}$, noting that $f_k^{*(0)}$ and Algorithm 2 are special instances of J_n and the 1D algorithm.

First $\boldsymbol{\Sigma}_k^{(0)}$, the Kronecker covariance estimator from Manceur and Dutilleul (2013), is $\sqrt{n}$-consistent estimator for $\boldsymbol{\Sigma}_k$. Next, by comparing $f_k^{*(0)}(\mathbf{G}_k)$ to $J_n(\mathbf{G})$ we see that $\mathbf{N}_k^{(0)} = (n\prod_{j\neq k} r_j)^{-1}\sum_{i=1}^n \mathbf{Y}_{i(k)}\boldsymbol{\Sigma}_{-k}^{(0)}\mathbf{Y}_{i(k)}^\intercal$ is analogous to $(\widehat{\mathbf{M}} + \widehat{\mathbf{U}})$ in $J_n(\mathbf{G})$, where we have defined

$$\boldsymbol{\Sigma}_{-k}^{(0)} \equiv \left((\boldsymbol{\Sigma}_m^{(0)})^{-1} \otimes \cdots \otimes (\boldsymbol{\Sigma}_{k+1}^{(0)})^{-1} \otimes (\boldsymbol{\Sigma}_{k-1}^{(0)})^{-1} \otimes \cdots \otimes (\boldsymbol{\Sigma}_1^{(0)})^{-1}\right).$$

Therefore we focus on $\mathbf{N}_k^{(0)} - \boldsymbol{\Sigma}_k^{(0)}$, which is analogous to $\widehat{\mathbf{U}}$ in $J_n(\mathbf{G})$. Recall that $\mathbf{Y}_i = \mathbf{B}^{(0)}\bar{\times}_{(m+1)}\mathbf{X}_i^\intercal + \mathbf{e}_i$ and that $\boldsymbol{\Sigma}_k^{(0)}$ is obtained based on $\mathbf{e}_i$ as $\boldsymbol{\Sigma}_k^{(0)} = (n\prod_{j\neq k} r_j)^{-1}\sum_{i=1}^n \mathbf{e}_{i(k)}\boldsymbol{\Sigma}_{-k}^{(0)}\mathbf{e}_{i(k)}^\intercal$. Hence,

$$\mathbf{N}_k^{(0)} - \boldsymbol{\Sigma}_k^{(0)} = (n\prod_{j\neq k} r_j)^{-1}\sum_{i=1}^n (\mathbf{Y}_i - \mathbf{e}_i)_{(k)}\boldsymbol{\Sigma}_{-k}^{(0)}(\mathbf{Y}_i - \mathbf{e}_i)_{(k)}^\intercal,$$

where we recognize $\mathbf{Y}_i - \mathbf{e}_i = \mathbf{B}^{(0)}\bar{\times}_{(m+1)}\mathbf{X}_i^\intercal$. Define the following "scaled" regression coefficient tensor

$$\mathbf{B}^{\{k\}} = [\![\mathbf{B}; \boldsymbol{\Sigma}_1^{-1/2}, \ldots, \boldsymbol{\Sigma}_{k-1}^{-1/2}, \mathbf{I}_{r_k}, \boldsymbol{\Sigma}_{k+1}^{-1/2}, \ldots, \boldsymbol{\Sigma}_m^{-1/2}, \boldsymbol{\Sigma}_{\mathbf{X}}^{1/2}]\!].$$

Then we see that $\mathbf{N}_k^{(0)} - \boldsymbol{\Sigma}_k^{(0)} = (\prod_{j\neq k} r_j)^{-1}\widehat{\mathbf{B}}^{\{k\}}(\widehat{\mathbf{B}}^{\{k\}})^\intercal$ for

$$\widehat{\mathbf{B}}^{\{k\}} = [\![\mathbf{B}; \left(\boldsymbol{\Sigma}_1^{(0)}\right)^{-1/2}, \ldots, \left(\boldsymbol{\Sigma}_{k-1}^{(0)}\right)^{-1/2}, \mathbf{I}_{r_k}, \left(\boldsymbol{\Sigma}_{k+1}^{(0)}\right)^{-1/2}, \ldots, \left(\boldsymbol{\Sigma}_m^{(0)}\right)^{-1/2}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}^{1/2}]\!],$$

where $\widehat{\mathbf{\Sigma}}_{\mathbf{X}} = n^{-1}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^{\intercal}$. We claim that $\mathbf{N}_k^{(0)} - \mathbf{\Sigma}_k^{(0)}$ is a $\sqrt{n}$-consistent estimator for $\mathbf{B}_{(k)}^{\{k\}}(\mathbf{B}_{(k)}^{\{k\}})^{\intercal}$ up to an upfront scaling constant, $(\prod_{j\neq k} r_j)^{-1}$, based on the fact that the sample matrices $\mathbf{\Sigma}_j^{(0)}$ and $\widehat{\mathbf{\Sigma}}_{\mathbf{X}}$ are asymptotically independent of the OLS estimator $\mathbf{B}^{(0)}$. By Cook and Zhang (2016; Proposition 6) we now see that $\mathbf{P}_{\mathbf{\Gamma}_k}^{os}$ is a $\sqrt{n}$-consistent estimator for the projection onto the envelope $\mathcal{E}_{\mathbf{\Sigma}_k}\left(\mathbf{B}_{(k)}^{\{k\}}(\mathbf{B}_{(k)}^{\{k\}})^{\intercal}\right) = \mathcal{E}_{\mathbf{\Sigma}_k}(\mathbf{B}_{(k)}^{\{k\}})$. By definition of $\mathbf{B}^{\{k\}}$ and the property of Tucker operator, we have

$$\mathbf{B}_{(k)}^{\{k\}} = \mathbf{B}_{(k)}\left(\mathbf{\Sigma}_1^{-1/2}\otimes\cdots\otimes\mathbf{\Sigma}_{k-1}^{-1/2}\otimes\mathbf{\Sigma}_{k+1}^{-1/2}\otimes\cdots\otimes\mathbf{\Sigma}_m^{-1/2}\otimes\mathbf{\Sigma}_{\mathbf{X}}^{1/2}\right),$$

which implies $\mathrm{span}(\mathbf{B}_{(k)}^{\{k\}}) = \mathrm{span}(\mathbf{B}_{(k)})$, and hence $\mathcal{E}_{\mathbf{\Sigma}_k}(\mathbf{B}_{(k)}^{\{k\}}) = \mathcal{E}_{\mathbf{\Sigma}_k}(\mathbf{B}_{(k)})$.

So far we have shown that $\mathbf{P}_{\mathbf{\Gamma}_k}^{os}$ is a $\sqrt{n}$-consistent estimator for the projection onto the envelope $\mathcal{E}_{\mathbf{\Sigma}_k}(\mathbf{B}_{(k)})$. The second part of the proposition is based on the $\sqrt{n}$-consistency of $\mathbf{B}^{(0)} = \mathbf{B}_{\mathrm{OLS}}$ and $\mathbf{P}_{\mathbf{\Gamma}_k}^{os}$, $k = 1,\ldots,m$, and the definition of $\mathbf{B}^{os}$:

$$\mathbf{B}^{os} = [\![\mathbf{B}^{(0)}; \mathbf{P}_{\mathbf{\Gamma}_1}^{os},\ldots,\mathbf{P}_{\mathbf{\Gamma}_m}^{os}, \mathbf{I}_p]\!].$$

E.2 Consistency of the likelihood-based estimator

The $\sqrt{n}$-consistency of the likelihood-based estimator from minimizing $\ell(\mathbf{B},\mathbf{\Sigma})$ relies on Shapiro's (1986) results on the asymptotics of over-parameterized structural models. The proof is parallel to the proof of Proposition 4 in Cook and Zhang (2015) and is thus ommitted.

F Proof for Theorem 2 (including derivations for the updating equations in Section 4.1)

Following the discussion in Section 5.2 of the paper, the iterative estimator is not guaranteed to be necessarily the maximum likelihood estimator (MLE), due to the existence of multiple local minima. However, it is asymptotically equivalent to the MLE. This is because, under the tensor normal distribution, the initialization of Algorithm 1 is built upon $\sqrt{n}$-consistent estimators, while each parameter in Algorithm 1 is iteratively obtained along the partial derivative of

the log-likelihood. From the classical theory of point estimation, we know that one Newton-Raphson step from the starting value provides an estimator that is asymptotically equivalent to the MLE even in the presence of multiple local minima (Lehmann and Casella, 1998, p. 454). Consequently, to prove Theorem 2, we only focus on the asymptotic properties of the theoretical MLE.

Specifically, we need to derive all the objective functions and updating equations in Section 4.1 from the negative normal log-likelihood function:

$$\ell(\mathbf{B}, \mathbf{\Sigma}) = \log|\mathbf{\Sigma}| + n^{-1}\sum_{i=1}^{n}\{\mathrm{vec}(\mathbf{Y}_i) - \mathbf{B}_{(m+1)}^{\intercal}\mathbf{X}_i\}\mathbf{\Sigma}^{-1}\{\mathrm{vec}(\mathbf{Y}_i) - \mathbf{B}_{(m+1)}^{\intercal}\mathbf{X}_i\},$$

which is to be optimized over $\mathbf{\Sigma} = \mathbf{\Sigma}(\{\mathbf{\Gamma}_k, \mathbf{\Omega}_k, \mathbf{\Omega}_{0k}\}_{k=1}^{m})$ and $\mathbf{B} = \mathbf{B}(\{\mathbf{\Gamma}_k\}_{k=1}^{m}, \mathbf{\Theta})$ for semi-orthogonal $\mathbf{\Gamma}_k \in \mathbb{R}^{r_k \times u_k}$, positive definite and symmetric $\mathbf{\Omega}_k \in \mathbb{S}^{u_k}$ and $\mathbf{\Omega}_{0k} \in \mathbb{S}^{r_k - u_k}$, and $\mathbf{\Theta} \in \mathbb{R}^{u_m \times \cdots \times u_1 \times p}$. Since it is impossible to obtain explicit forms of MLEs, we show that the series of equations used in Algorithm 1 come from partially minimizing ℓ with specified parameters fixed. We summarize our findings in the following statements and give detailed derivations immediately after. Note that the updating equations in Section 4.1 is obtained from the following equations by superscripting $^{(t)}$ on the left hand sides and superscripting $^{(t+1)}$ on the right hand sides of equations.

Under the tensor normal assumption, the MLEs satisfy the following equations,

$$\begin{aligned}
\mathbf{\Theta} &= \mathbb{Z} \times_{(m+1)} \{(\mathbb{X}\mathbb{X}^{\intercal})^{-1}\mathbb{X}\}, \\
\mathbf{B} &= \mathbb{Y} \times_1 \mathbf{P}_{\mathbf{\Gamma}_1} \times_2 \cdots \times_m \mathbf{P}_{\mathbf{\Gamma}_m} \times_{(m+1)} \{(\mathbb{X}\mathbb{X}^{\intercal})^{-1}\mathbb{X}\}, \\
&= [\![\widehat{\mathbf{B}}_{\mathrm{OLS}}; \mathbf{P}_{\mathbf{\Gamma}_1}, \cdots, \mathbf{P}_{\mathbf{\Gamma}_m}, \mathbf{I}_p]\!] \\
\mathbf{\Omega}_k &= \frac{1}{n\prod_{j\neq k} r_j}\sum_{i=1}^{n} \mathbf{s}_{i(k)}\{\mathbf{\Omega}_m^{-1} \otimes \cdots \otimes \mathbf{\Omega}_{k+1}^{-1} \otimes \mathbf{\Omega}_{k-1}^{-1} \otimes \cdots \otimes \mathbf{\Omega}_1^{-1}\}\mathbf{s}_{i(k)}^{\intercal}, \\
\mathbf{\Omega}_{0k} &= \frac{1}{n\prod_{j\neq k} r_j}\sum_{i=1}^{n} \mathbf{\Gamma}_{0k}^{\intercal}\mathbf{Y}_{i(k)}\{\mathbf{\Sigma}_m^{-1} \otimes \cdots \otimes \mathbf{\Sigma}_{k+1}^{-1} \otimes \mathbf{\Sigma}_{k-1}^{-1} \otimes \cdots \otimes \mathbf{\Sigma}_1^{-1}\}\mathbf{Y}_{i(k)}^{\intercal}\mathbf{\Gamma}_{0k},
\end{aligned}$$

where the data tensors $\mathbf{Z}_i$ and $\mathbb{Z}$ are defined according to $\mathbf{Z} = [\![\mathbf{Y}; \mathbf{\Gamma}_1^{\intercal}, \ldots, \mathbf{\Gamma}_m^{\intercal}]\!]$, and the residual tensor $\mathbf{s}_i = \mathbf{Z}_i - \mathbf{\Theta}\bar{\times}_{(m+1)}\mathbf{X}_i$. Under the tensor normal assumption, the MLE of $\{\mathbf{\Gamma}_k\}_{k=1}^{m}$ can be

obtained as minimizer of the following objective function

$$f_k(\mathbf{\Gamma}_k) \quad = \quad \log|\mathbf{\Gamma}_k^\intercal \mathbf{M}_k \mathbf{\Gamma}_k| + \log|\mathbf{\Gamma}_k^\intercal \mathbf{N}_k^{-1} \mathbf{\Gamma}_k|,$$

where $\mathbf{M}_k$, $\mathbf{N}_k$ are defined as

$$\mathbf{M}_k = (n\prod_{j\neq k} r_j)^{-1} \sum_{i=1}^{n} \boldsymbol{\delta}_{i(k)} (\mathbf{\Sigma}_m^{-1} \otimes \cdots \otimes \mathbf{\Sigma}_{k+1}^{-1} \otimes \mathbf{\Sigma}_{k-1}^{-1} \otimes \cdots \otimes \mathbf{\Sigma}_1^{-1}) \boldsymbol{\delta}_{i(k)}^\intercal,$$

$$\mathbf{N}_k = (n\prod_{j\neq k} r_j)^{-1} \sum_{i=1}^{n} \mathbf{Y}_{i(k)} (\mathbf{\Sigma}_m^{-1} \otimes \cdots \otimes \mathbf{\Sigma}_{k+1}^{-1} \otimes \mathbf{\Sigma}_{k-1}^{-1} \otimes \cdots \otimes \mathbf{\Sigma}_1^{-1}) \mathbf{Y}_{i(k)}^\intercal,$$

where $\boldsymbol{\delta}_{i(k)}$ is the k-th matricization of the residual $\boldsymbol{\delta}_i$,

$$\begin{aligned} \boldsymbol{\delta}_i \quad &= \quad \mathbf{Y}_i - \mathbb{Y} \times_1 \mathbf{P}_{\mathbf{\Gamma}_1} \times_2 \cdots \times_{(k-1)} \mathbf{P}_{\mathbf{\Gamma}_{k-1}} \times_{(k+1)} \mathbf{P}_{\mathbf{\Gamma}_{k+1}} \times_{k+2} \cdots \times_m \mathbf{P}_{\mathbf{\Gamma}_m} \times_{(m+1)} \mathbf{X}_i^\intercal \{(\mathbb{X}\mathbb{X}^\intercal)^{-1}\mathbb{X}\} \\ &= \quad \mathbf{Y}_i - [\![\widehat{\mathbf{B}}_{\mathrm{OLS}}; \mathbf{P}_{\mathbf{\Gamma}_1}, \ldots, \mathbf{P}_{\mathbf{\Gamma}_{k-1}}, \mathbf{I}_{r_k}, \mathbf{P}_{\mathbf{\Gamma}_{k+1}}, \ldots, \mathbf{P}_{\mathbf{\Gamma}_m}, \mathbf{I}_p]\!] \bar{\times}_{(m+1)} \mathbf{X}_i. \end{aligned}$$

F.1 Estimation of other parameters given $\{\mathbf{\Gamma}_k\}_{k=1}^m$

We first decompose the $\log|\mathbf{\Sigma}|$ term as

$$\log|\mathbf{\Sigma}| \quad = \quad \log|\mathbf{\Sigma}_m \otimes \cdots \otimes \mathbf{\Sigma}_1| = \sum_{k=1}^{m} \{(\prod_{j\neq k} r_k) \log|\mathbf{\Sigma}_k|\} = \sum_{k=1}^{m} \{(\prod_{j\neq k} r_k)(\log|\mathbf{\Omega}_k| + \log|\mathbf{\Omega}_{0k}|)\},$$

which is essentially $2m$ additive terms of $\log|\mathbf{\Omega}_k|$ and $\log|\mathbf{\Omega}_{0k}|$. The conditional log-likelihood can be separated into two independent parts regarding $\mathbb{P}(\mathbf{Y})|\mathbf{X}$ and $\mathbb{Q}(\mathbf{Y}) \sim \mathbb{Q}(\mathbf{Y})|\mathbf{X}$. Studying the regression of $\mathbb{P}(\mathbf{Y})$ on $\mathbf{X}$ is essentially studying that of $\mathbf{Z} = [\![\mathbf{Y}; \mathbf{\Gamma}_1^\intercal, \ldots, \mathbf{\Gamma}_m^\intercal]\!]$ on $\mathbf{X}$. Following the discussion in the paper, we have the MLEs for $\mathbf{\Theta}$, $\mathbf{B}$ and $\{\mathbf{\Omega}_k\}_{k=1}^m$ given $\{\mathbf{\Gamma}_k\}_{k=1}^m$.

We next derive the MLE equations for $\{\mathbf{\Omega}_{0k}\}_{k=1}^m$ with given $\{\mathbf{\Gamma}_k\}_{k=1}^m$. Without loss of generality, we write down our derivations with respect to $\mathbf{\Omega}_{01}$. We decompose $\mathbf{\Sigma}^{-1} = \mathbf{\Delta}_1 + \mathbf{\Delta}_{01}$, according to $\mathbf{\Omega}_1$ and $\mathbf{\Omega}_{01}$ in the decomposition of $\mathbf{\Sigma}_1^{-1} = \mathbf{\Gamma}_1 \mathbf{\Omega}_1^{-1} \mathbf{\Gamma}_1^\intercal + \mathbf{\Gamma}_{01} \mathbf{\Omega}_{01}^{-1} \mathbf{\Gamma}_{01}^\intercal$.

$$\begin{aligned} \mathbf{\Delta}_1 \quad &= \quad \mathbf{\Sigma}_m^{-1} \otimes \cdots \otimes \mathbf{\Sigma}_2^{-1} \otimes (\mathbf{\Gamma}_1 \mathbf{\Omega}_1^{-1} \mathbf{\Gamma}_1^\intercal) \\ &= \quad (\mathbf{\Gamma}_m \mathbf{\Omega}_m^{-1} \mathbf{\Gamma}_m^\intercal + \mathbf{\Gamma}_{0m} \mathbf{\Omega}_{0m}^{-1} \mathbf{\Gamma}_{0m}^\intercal) \otimes \cdots \otimes (\mathbf{\Gamma}_2 \mathbf{\Omega}_2^{-1} \mathbf{\Gamma}_2^\intercal + \mathbf{\Gamma}_{02} \mathbf{\Omega}_{02}^{-1} \mathbf{\Gamma}_{02}^\intercal) \otimes (\mathbf{\Gamma}_1 \mathbf{\Omega}_1^{-1} \mathbf{\Gamma}_1^\intercal), \\ \mathbf{\Delta}_{01} \quad &= \quad \mathbf{\Sigma}_m^{-1} \otimes \cdots \otimes \mathbf{\Sigma}_2^{-1} \otimes (\mathbf{\Gamma}_{01} \mathbf{\Omega}_{01}^{-1} \mathbf{\Gamma}_{01}^\intercal) \\ &= \quad (\mathbf{\Gamma}_m \mathbf{\Omega}_m^{-1} \mathbf{\Gamma}_m^\intercal + \mathbf{\Gamma}_{0m} \mathbf{\Omega}_{0m}^{-1} \mathbf{\Gamma}_{0m}^\intercal) \otimes \cdots \otimes (\mathbf{\Gamma}_2 \mathbf{\Omega}_2^{-1} \mathbf{\Gamma}_2^\intercal + \mathbf{\Gamma}_{02} \mathbf{\Omega}_{02}^{-1} \mathbf{\Gamma}_{02}^\intercal) \otimes (\mathbf{\Gamma}_{01} \mathbf{\Omega}_{01}^{-1} \mathbf{\Gamma}_{01}^\intercal). \end{aligned}$$

Hence, we can write the negative partial log-likelihood for solving $\boldsymbol{\Omega}_{01}$ as

$$\ell(\boldsymbol{\Omega}_{01}|\{\boldsymbol{\Gamma}_k\}_{k=1}^m) \simeq (\prod_{j>1} r_j)\log|\boldsymbol{\Omega}_{01}| + n^{-1}\sum_{i=1}^n \{\mathrm{vec}(\mathbf{Y}_i) - \mathbf{B}_{(m+1)}^\intercal \mathbf{X}_i\}^\intercal \boldsymbol{\Delta}_{01}\{\mathrm{vec}(\mathbf{Y}_i) - \mathbf{B}_{(m+1)}^\intercal \mathbf{X}_i\}.$$

Recall that $\mathbf{B}_{(m+1)} = \boldsymbol{\eta}_{(m+1)}(\boldsymbol{\Gamma}_m^\intercal \otimes \cdots \otimes \boldsymbol{\Gamma}_1^\intercal)$, hence $\boldsymbol{\Delta}_{01}\mathbf{B}_{(m+1)}^\intercal = 0$ as a result of $\boldsymbol{\Gamma}_{01}\boldsymbol{\Omega}_{01}^{-1}\boldsymbol{\Gamma}_{01}^\intercal \cdot \boldsymbol{\Gamma}_1 = 0$. Therefore,

$$\ell(\boldsymbol{\Omega}_{01}|\{\boldsymbol{\Gamma}_k\}_{k=1}^m) \simeq (\prod_{j>1} r_j)\log|\boldsymbol{\Omega}_{01}| + n^{-1}\sum_{i=1}^n \mathrm{vec}^\intercal(\mathbf{Y}_i)\boldsymbol{\Delta}_{01}\mathrm{vec}(\mathbf{Y}_i).$$

The quadratic form $\mathrm{vec}^\intercal(\mathbf{Y}_i)\boldsymbol{\Delta}_{01}\mathrm{vec}(\mathbf{Y}_i)$ equals the squared norm of $\mathrm{vec}(\mathbf{Y}_i \times_1 \boldsymbol{\Omega}_{01}^{-1/2}\boldsymbol{\Gamma}_{01}^\intercal \times_2 \boldsymbol{\Sigma}_2^{-1/2} \times \cdots \times_m \boldsymbol{\Sigma}_m^{-1/2}) \equiv \mathrm{vec}(\mathbf{V}_i)$. By definition of tensor norm, $\|\mathbf{V}_i\|^2 = \|\mathrm{vec}(\mathbf{V}_i)\|^2 = \|\mathbf{V}_{i(1)}\|_F^2 = \mathrm{tr}(\mathbf{V}_{i(1)}^\intercal \mathbf{V}_{i(1)})$, where $\mathbf{V}_{i(1)}$ is the mode-1 matricization of $\mathbf{V}_i$:

$$\mathbf{V}_{i(1)} = (\boldsymbol{\Omega}_{01}^{-1/2}\boldsymbol{\Gamma}_{01}^\intercal)\mathbf{Y}_{i(1)}\left(\boldsymbol{\Sigma}_m^{-1/2} \otimes \cdots \otimes \boldsymbol{\Sigma}_2^{-1/2}\right).$$

Hence

$$\mathrm{tr}(\mathbf{V}_{i(1)}^\intercal \mathbf{V}_{i(1)}) = \mathrm{tr}(\mathbf{V}_{i(1)}\mathbf{V}_{i(1)}^\intercal) = \mathrm{tr}\left\{\boldsymbol{\Omega}_{01}^{-1}\boldsymbol{\Gamma}_{01}^\intercal \mathbf{Y}_{i(1)}\left(\boldsymbol{\Sigma}_m^{-1} \otimes \cdots \otimes \boldsymbol{\Sigma}_2^{-1}\right)\mathbf{Y}_{i(1)}^\intercal \boldsymbol{\Gamma}_{01}\right\}.$$

The partial conditional log-likelihood becomes

$$\begin{aligned}\ell(\boldsymbol{\Omega}_{01}|\{\boldsymbol{\Gamma}_k\}_{k=1}^m) &\simeq (\prod_{j>1} r_j)\log|\boldsymbol{\Omega}_{01}| + n^{-1}\sum_{i=1}^n \mathrm{tr}(\mathbf{V}_{i(1)}^\intercal \mathbf{V}_{i(1)}) \\ &= (\prod_{j>1} r_j)\log|\boldsymbol{\Omega}_{01}| + n^{-1}\mathrm{tr}\left[\boldsymbol{\Omega}_{01}^{-1}\sum_{i=1}^n \{\boldsymbol{\Gamma}_{01}^\intercal \mathbf{Y}_{i(1)}\left(\boldsymbol{\Sigma}_m^{-1} \otimes \cdots \otimes \boldsymbol{\Sigma}_2^{-1}\right)\mathbf{Y}_{i(1)}^\intercal \boldsymbol{\Gamma}_{01}\}\right],\end{aligned}$$

which lead to the following equations for iteratively solving $\boldsymbol{\Omega}_{01}$.

$$\boldsymbol{\Omega}_{01} = (n\prod_{j>1} r_j)^{-1}\sum_{i=1}^n \boldsymbol{\Gamma}_{01}^\intercal \mathbf{Y}_{i(1)}(\boldsymbol{\Sigma}_m^{-1} \otimes \cdots \otimes \boldsymbol{\Sigma}_2^{-1})\mathbf{Y}_{i(1)}^\intercal \boldsymbol{\Gamma}_{01},$$

where $\boldsymbol{\Sigma}_k = \boldsymbol{\Gamma}_k\boldsymbol{\Omega}_k\boldsymbol{\Gamma}_k^\intercal + \boldsymbol{\Gamma}_{0k}\boldsymbol{\Omega}_{0k}\boldsymbol{\Gamma}_{0k}^\intercal$ for $k = 1, \ldots, m$. It is then easy to obtain the following result, for any k,

$$\boldsymbol{\Omega}_{0k} = (n\prod_{j\neq k} r_j)^{-1}\sum_{i=1}^n \boldsymbol{\Gamma}_{0k}^\intercal \mathbf{Y}_{i(k)}(\boldsymbol{\Sigma}_m^{-1} \otimes \cdots \otimes \boldsymbol{\Sigma}_{k+1}^{-1} \otimes \boldsymbol{\Sigma}_{k-1}^{-1} \otimes \cdots \otimes \boldsymbol{\Sigma}_1^{-1})\mathbf{Y}_{i(k)}^\intercal \boldsymbol{\Gamma}_{0k}.$$

F.2 Estimation of $\boldsymbol{\Gamma}_1$ given $\{\boldsymbol{\Gamma}_k, \boldsymbol{\Omega}_k, \boldsymbol{\Omega}_{0k}\}_{k=2}^m$

Treating the other parameters $\{\boldsymbol{\Gamma}_k, \boldsymbol{\Omega}_k, \boldsymbol{\Omega}_{0k}\}_{k=2}^m$ as fixed constants, we write $\mathbf{B} = \mathbf{B}(\boldsymbol{\Gamma}_1)$ and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\Gamma}_1, \boldsymbol{\Omega}_1(\boldsymbol{\Gamma}_1), \boldsymbol{\Omega}_{01}(\boldsymbol{\Gamma}_1))$ as functions of $\boldsymbol{\Gamma}_1$. We then plug them into the likelihood $\ell(\mathbf{B}, \boldsymbol{\Sigma})$ to partially optimize over $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_{01}$ analytically, and then the objective function for optimizing over $\boldsymbol{\Gamma}_1 \in \mathbb{R}^{r_1 \times u_1}$ as follows.

First, ignoring all the fixed constants, the log-determinant term in $\ell(\mathbf{B}, \boldsymbol{\Sigma})$ becomes

$$\log|\boldsymbol{\Sigma}| \simeq (\prod_{j=2}^m r_j)\{\log|\boldsymbol{\Omega}_1(\boldsymbol{\Gamma}_1)| + \log|\boldsymbol{\Omega}_{01}(\boldsymbol{\Gamma}_1)|\}.$$

Similar to the previous section, we can decompose $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Sigma}_m^{-1} \otimes \cdots \otimes \boldsymbol{\Sigma}_1^{-1}$ into two parts according to the decomposition $\boldsymbol{\Sigma}_1^{-1} = \boldsymbol{\Gamma}_1 \boldsymbol{\Omega}_1^{-1} \boldsymbol{\Gamma}_1^\intercal + \boldsymbol{\Gamma}_{01} \boldsymbol{\Omega}_{01}^{-1} \boldsymbol{\Gamma}_{01}^\intercal$. Then we can write

$$\begin{aligned}
&\sum_{i=1}^n \{\mathrm{vec}(\mathbf{Y}_i) - \mathbf{B}_{(m+1)}^\intercal \mathbf{X}_i\} \boldsymbol{\Sigma}^{-1} \{\mathrm{vec}(\mathbf{Y}_i) - \mathbf{B}_{(m+1)}^\intercal \mathbf{X}_i\} \\
= &\sum_{i=1}^n \mathrm{tr}\left\{\boldsymbol{\Omega}_{01}^{-1} \boldsymbol{\Gamma}_{01}^\intercal \mathbf{Y}_{i(1)} (\boldsymbol{\Sigma}_m^{-1} \otimes \cdots \otimes \boldsymbol{\Sigma}_2^{-1}) \mathbf{Y}_{i(1)}^\intercal \boldsymbol{\Gamma}_{01}\right\} + \sum_{i=1}^n \mathrm{tr}\left\{\boldsymbol{\Omega}_1^{-1} \boldsymbol{\Gamma}_1^\intercal \mathbf{e}_{i(1)} (\boldsymbol{\Sigma}_m^{-1} \otimes \cdots \otimes \boldsymbol{\Sigma}_2^{-1}) \mathbf{e}_{i(1)}^\intercal \boldsymbol{\Gamma}_1\right\},
\end{aligned}$$

where $\mathbf{e}_i \equiv \mathbf{e}_i(\boldsymbol{\Gamma}_1) = \mathbf{Y}_i - \mathbf{B}(\boldsymbol{\Gamma}_1) \times_{(m+1)} \mathbf{X}_i^\intercal$ and $\mathbf{B}(\boldsymbol{\Gamma}_1) = \mathbb{Y} \times_1 \mathbf{P}_{\boldsymbol{\Gamma}_1} \times_2 \cdots \times_m \mathbf{P}_{\boldsymbol{\Gamma}_m} \times_{(m+1)} \{(\mathbb{X}\mathbb{X}^\intercal)^{-1}\mathbb{X}\}$. Then,

$$\begin{aligned}
\boldsymbol{\Gamma}_1^\intercal \mathbf{e}_{i(1)} &= \left(\mathbf{Y}_i \times_1 \boldsymbol{\Gamma}_1^\intercal - \mathbb{Y} \times_1 \boldsymbol{\Gamma}_1^\intercal \mathbf{P}_{\boldsymbol{\Gamma}_1} \times_2 \cdots \times_m \mathbf{P}_{\boldsymbol{\Gamma}_m} \times_{(m+1)} \mathbf{X}_i^\intercal \{(\mathbb{X}\mathbb{X}^\intercal)^{-1}\mathbb{X}\}\right)_{(1)} \\
&= \left(\mathbf{Y}_i \times_1 \boldsymbol{\Gamma}_1^\intercal - \mathbb{Y} \times_1 \boldsymbol{\Gamma}_1^\intercal \times_2 \cdots \times_m \mathbf{P}_{\boldsymbol{\Gamma}_m} \times_{(m+1)} \mathbf{X}_i^\intercal \{(\mathbb{X}\mathbb{X}^\intercal)^{-1}\mathbb{X}\}\right)_{(1)} \\
&= \left\{\left(\mathbf{Y}_i - \mathbb{Y} \times_1 \mathbf{I}_{r_1} \times_2 \cdots \times_m \mathbf{P}_{\boldsymbol{\Gamma}_m} \times_{(m+1)} \mathbf{X}_i^\intercal \{(\mathbb{X}\mathbb{X}^\intercal)^{-1}\mathbb{X}\}\right) \times_1 \boldsymbol{\Gamma}_1^\intercal\right\}_{(1)} \\
&= \boldsymbol{\Gamma}_1^\intercal \boldsymbol{\delta}_{i(1)},
\end{aligned}$$

where $\boldsymbol{\delta}_i = \left(\mathbf{Y}_i - \mathbb{Y} \times_1 \mathbf{I}_{r_1} \times_2 \cdots \times_m \mathbf{P}_{\boldsymbol{\Gamma}_m} \times_{(m+1)} \mathbf{X}_i^\intercal \{(\mathbb{X}\mathbb{X}^\intercal)^{-1}\mathbb{X}\}\right)$ does not involve $\boldsymbol{\Gamma}_1$. The partially maximized negative log-likelihood now becomes

$$
\begin{aligned}
\ell(\mathbf{B},\boldsymbol{\Sigma}) &= \log|\boldsymbol{\Sigma}| + n^{-1}\sum_{i=1}^{n}\{\mathrm{vec}(\mathbf{Y}_i)-\mathbf{B}_{(m+1)}^{\intercal}\mathbf{X}_i\}\boldsymbol{\Sigma}^{-1}\{\mathrm{vec}(\mathbf{Y}_i)-\mathbf{B}_{(m+1)}^{\intercal}\mathbf{X}_i\} \\
&\simeq (\prod_{j=2}^{m} r_j)\{\log|\boldsymbol{\Omega}_1(\boldsymbol{\Gamma}_1)| + \log|\boldsymbol{\Omega}_{01}(\boldsymbol{\Gamma}_1)|\} \\
&+ n^{-1}\sum_{i=1}^{n}\mathrm{tr}\left\{\boldsymbol{\Omega}_{01}^{-1}\boldsymbol{\Gamma}_{01}^{\intercal}\mathbf{Y}_{i(1)}(\boldsymbol{\Sigma}_m^{-1}\otimes\cdots\otimes\boldsymbol{\Sigma}_2^{-1})\mathbf{Y}_{i(1)}^{\intercal}\boldsymbol{\Gamma}_{01}\right\} \\
&+ n^{-1}\sum_{i=1}^{n}\mathrm{tr}\left\{\boldsymbol{\Omega}_{1}^{-1}\boldsymbol{\Gamma}_{1}^{\intercal}\boldsymbol{\delta}_{i(1)}(\boldsymbol{\Sigma}_m^{-1}\otimes\cdots\otimes\boldsymbol{\Sigma}_2^{-1})\boldsymbol{\delta}_{i(1)}^{\intercal}\boldsymbol{\Gamma}_{1}\right\},
\end{aligned}
$$

which leads to partial MLE of $\boldsymbol{\Omega}_1(\boldsymbol{\Gamma}_1)$ and $\boldsymbol{\Omega}_{01}(\boldsymbol{\Gamma}_1)$ as

$$
\begin{aligned}
\boldsymbol{\Omega}_1(\boldsymbol{\Gamma}_1) &= (n\prod_{j=2}^{m} r_j)^{-1}\sum_{i=1}^{n}\boldsymbol{\Gamma}_1^{\intercal}\boldsymbol{\delta}_{i(1)}(\boldsymbol{\Sigma}_m^{-1}\otimes\cdots\otimes\boldsymbol{\Sigma}_2^{-1})\boldsymbol{\delta}_{i(1)}^{\intercal}\boldsymbol{\Gamma}_1 \\
\boldsymbol{\Omega}_{01}(\boldsymbol{\Gamma}_1) &= (n\prod_{j=2}^{m} r_j)^{-1}\sum_{i=1}^{n}\boldsymbol{\Gamma}_{01}^{\intercal}\mathbf{Y}_{i(1)}(\boldsymbol{\Sigma}_m^{-1}\otimes\cdots\otimes\boldsymbol{\Sigma}_2^{-1})\mathbf{Y}_{i(1)}^{\intercal}\boldsymbol{\Gamma}_{01}.
\end{aligned}
$$

Then, substitute these back to $\ell(\mathbf{B},\boldsymbol{\Sigma})$ to get the likelihood-based objective function for $\boldsymbol{\Gamma}_1$:

$$
\begin{aligned}
\mathrm{F}_n(\boldsymbol{\Gamma}_1) &= \log|(n\prod_{j=2}^{m} r_j)^{-1}\sum_{i=1}^{n}\boldsymbol{\Gamma}_1^{\intercal}\boldsymbol{\delta}_{i(1)}(\boldsymbol{\Sigma}_m^{-1}\otimes\cdots\otimes\boldsymbol{\Sigma}_2^{-1})\boldsymbol{\delta}_{i(1)}^{\intercal}\boldsymbol{\Gamma}_1| \\
&+ \log|(n\prod_{j=2}^{m} r_j)^{-1}\sum_{i=1}^{n}\boldsymbol{\Gamma}_{01}^{\intercal}\mathbf{Y}_{i(1)}(\boldsymbol{\Sigma}_m^{-1}\otimes\cdots\otimes\boldsymbol{\Sigma}_2^{-1})\mathbf{Y}_{i(1)}^{\intercal}\boldsymbol{\Gamma}_{01}| \\
&\simeq \log|\boldsymbol{\Gamma}_1^{\intercal}\mathbf{M}_1\boldsymbol{\Gamma}_1| + \log|\boldsymbol{\Gamma}_1^{\intercal}\mathbf{N}_1^{-1}\boldsymbol{\Gamma}_1|,
\end{aligned}
$$

where $\mathbf{M}_1 = (n\prod_{j=2}^{m} r_j)^{-1}\sum_{i=1}^{n}\boldsymbol{\delta}_{i(1)}(\boldsymbol{\Sigma}_m^{-1}\otimes\cdots\otimes\boldsymbol{\Sigma}_2^{-1})\boldsymbol{\delta}_{i(1)}^{\intercal}$ and $\mathbf{N}_1 = (n\prod_{j=2}^{m} r_j)^{-1}\sum_{i=1}^{n}\mathbf{Y}_{i(1)}(\boldsymbol{\Sigma}_m^{-1}\otimes\cdots\otimes\boldsymbol{\Sigma}_2^{-1})\mathbf{Y}_{i(1)}^{\intercal}$. The last step of the above equations come from the fact that $\log|\boldsymbol{\Gamma}_{01}^{\intercal}\mathbf{A}\boldsymbol{\Gamma}_{01}| \simeq \log|\boldsymbol{\Gamma}_1\mathbf{A}^{-1}\boldsymbol{\Gamma}_1|$ for any positive definite symmetric matrix $\mathbf{A}$.

G Proof for Theorem 3

From the proof of Theorem 2, we have seen that $\widehat{\mathbf{B}}_{\boldsymbol{\Gamma}} = [\![\widehat{\mathbf{B}}_{\mathrm{OLS}};\mathbf{P}_{\boldsymbol{\Gamma}_1},\ldots,\mathbf{P}_{\boldsymbol{\Gamma}_m},\mathbf{I}_p]\!]$, where $\mathbf{P}_{\boldsymbol{\Gamma}_k}$, $k=1,\ldots,m$, are all true projections onto the envelopes (i.e. population values). Then $\mathrm{vec}(\widehat{\mathbf{B}}_{\boldsymbol{\Gamma}}) = (\mathbf{I}_p\otimes\mathbf{P}_{\boldsymbol{\Gamma}_m}\otimes\cdots\otimes\mathbf{P}_{\boldsymbol{\Gamma}_1})\mathrm{vec}(\widehat{\mathbf{B}}_{\mathrm{OLS}}) := \mathbf{P}_t\cdot\mathrm{vec}(\widehat{\mathbf{B}}_{\mathrm{OLS}})$. The OLS estimator $\mathrm{vec}(\widehat{\mathbf{B}}_{\mathrm{OLS}})$ is $\sqrt{n}$-consistent and asymptotically normal with mean zero covariance equals to $\mathbf{U}_{\mathrm{OLS}} =$

$\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}$. Since $\mathbf{P}_t = \mathbf{I}_p \otimes \mathbf{P}_{\boldsymbol{\Gamma}_m} \otimes \cdots \otimes \mathbf{P}_{\boldsymbol{\Gamma}_1}$ is fixed as population truth, we have proven that the envelope estimator $\widehat{\mathbf{B}}_{\boldsymbol{\Gamma}}$ is $\sqrt{n}$-consistent and asymptotically normal with mean zero covariance $\mathbf{U}_{\boldsymbol{\Gamma}} = \mathbf{P}_t \mathbf{U}_{\mathrm{OLS}} \mathbf{P}_t = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \mathbf{P}_{\boldsymbol{\Gamma}_m} \boldsymbol{\Sigma}_m \mathbf{P}_{\boldsymbol{\Gamma}_m} \otimes \cdots \otimes \mathbf{P}_{\boldsymbol{\Gamma}_1} \boldsymbol{\Sigma}_1 \mathbf{P}_{\boldsymbol{\Gamma}_1}$.

H Proof for Theorem 4

Recall that the parameter vectors involved in this Theorem are:

$$\mathbf{h} = \begin{pmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \end{pmatrix} = \begin{pmatrix} \mathrm{vec}(\mathbf{B}) \\ \mathrm{vech}(\boldsymbol{\Sigma}) \end{pmatrix}, \quad \boldsymbol{\phi} = \begin{pmatrix} \boldsymbol{\phi}_1 \\ \boldsymbol{\phi}_2 \\ \vdots \\ \boldsymbol{\phi}_{m+1} \end{pmatrix} = \begin{pmatrix} \mathrm{vec}(\mathbf{B}) \\ \mathrm{vech}(\boldsymbol{\Sigma}_1) \\ \vdots \\ \mathrm{vech}(\boldsymbol{\Sigma}_m) \end{pmatrix}, \quad \boldsymbol{\xi} = \begin{pmatrix} \boldsymbol{\xi}_1 \\ \vdots \\ \boldsymbol{\xi}_{3m+1} \end{pmatrix},$$

where $\boldsymbol{\xi}_1 = \mathrm{vec}(\boldsymbol{\Theta})$, $\{\boldsymbol{\xi}_j\}_{j=2}^{m+1} = \{\mathrm{vec}(\boldsymbol{\Gamma}_k)\}_{k=1}^{m}$, $\{\boldsymbol{\xi}_j\}_{j=m+2}^{2m+1} = \{\mathrm{vech}(\boldsymbol{\Omega}_k)\}_{k=1}^{m}$, $\{\boldsymbol{\xi}_j\}_{j=2m+2}^{3m+1} = \{\mathrm{vech}(\boldsymbol{\Omega}_{0k})\}_{k=1}^{m}$. The length of vectors are monotonically decreasing from $\mathbf{h}$ to $\boldsymbol{\phi}$ and then to $\boldsymbol{\xi}$, because they corresponding to three nested model assumptions: (M1) $\mathbf{h}$ corresponding to the unrestricted, vectorized linear regression model, or the OLS model; (M2) $\boldsymbol{\phi}$ is from the Kronecker covariance assumption, so letting $\mathbf{h} = \mathbf{h}(\boldsymbol{\phi})$ is essentially imposing the Kronecker matrix structure that $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_m \otimes \cdots \otimes \boldsymbol{\Sigma}_1$; (M3) $\boldsymbol{\xi}$ is based on envelope assumptions of $\mathcal{E}_{\boldsymbol{\Sigma}_k}(\mathbf{B}_{(k)})$, $k = 1, \ldots, m$.

From Theorem 2, we know that $\mathbf{h}^{(\infty)} = \mathbf{h}(\boldsymbol{\phi}^{(\infty)}) = \mathbf{h}(\boldsymbol{\xi}^{(\infty)})$ is the MLE under the envelope assumption, i.e. model assumption (M3). Also, $\mathbf{h}^{(0)} = \mathbf{h}(\boldsymbol{\phi}^{(0)})$ contains the OLS estimator and the Kronecker covariance estimator from Manceur and Dutilleul (2013). Hence it is the MLE under tensor normal assumption, i.e. model assumption (M2). Since $\mathbf{h} = \mathbf{h}(\boldsymbol{\phi}) = \mathbf{h}(\boldsymbol{\xi})$ is overparameterized, from Shapiro's (1986) we see the following results: $\sqrt{n}(\mathbf{h}^{(0)} - \mathbf{h}_{\mathrm{True}}) \to N(0, \mathbf{V}_{\mathbf{h}0})$ and $\sqrt{n}(\mathbf{h}^{(\infty)} - \mathbf{h}_{\mathrm{True}}) \to N(0, \mathbf{V}_{\mathbf{h}\infty})$, where $\mathbf{V}_{\mathbf{h}0} = \mathbf{H}(\mathbf{H}^{\intercal}\mathbf{J}_{\mathbf{h}}\mathbf{H})^{\dagger}\mathbf{H}^{\intercal}$, $\mathbf{V}_{\mathbf{h}\infty} = \mathbf{K}(\mathbf{K}^{\intercal}\mathbf{J}_{\mathbf{h}}\mathbf{K})^{\dagger}\mathbf{K}^{\intercal}$ and the gradient matrices $\mathbf{H} = \partial\mathbf{h}(\boldsymbol{\phi})/\partial\boldsymbol{\phi}$ and $\mathbf{K} = \partial\mathbf{h}(\boldsymbol{\xi})/\partial\boldsymbol{\xi}$. Moreover, the Fisher information matrix $\mathbf{J}_{\mathbf{h}}$ has the same form as in the usual vector-response linear regression model,

$$\mathbf{J}_{\mathbf{h}} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{X}} \otimes \boldsymbol{\Sigma}^{-1} & 0 \\ 0 & \frac{1}{2}\mathbf{E}_d^T(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{E}_d \end{pmatrix},$$

where the expansion matrix $\mathbf{E}_d \in \mathbb{R}^{d^2 \times d(d+1)/2}$ has the corresponding dimension $d = \prod_{k=1}^{m} r_k$. Then,

$$\mathbf{H}(\mathbf{H}^\intercal \mathbf{J}_\mathbf{h} \mathbf{H})^\dagger \mathbf{H}^\intercal = \mathbf{J}_\mathbf{h}^{-1/2} \cdot \mathbf{J}_\mathbf{h}^{1/2} \mathbf{H} (\mathbf{H}^\intercal \mathbf{J}_\mathbf{h}^{1/2} \mathbf{J}_\mathbf{h}^{1/2} \mathbf{H})^\dagger \mathbf{H}^\intercal \mathbf{J}_\mathbf{h}^{1/2} \cdot \mathbf{J}_\mathbf{h}^{-1/2} = \mathbf{J}_\mathbf{h}^{-1/2} \mathbf{P}_{\mathbf{J}_\mathbf{h}^{1/2}\mathbf{H}} \mathbf{J}_\mathbf{h}^{-1/2},$$

and similarly, $\mathbf{K}(\mathbf{K}^\intercal \mathbf{J}_\mathbf{h} \mathbf{K})^\dagger \mathbf{K}^\intercal = \mathbf{J}_\mathbf{h}^{-1/2} \mathbf{P}_{\mathbf{J}_\mathbf{h}^{1/2}\mathbf{K}} \mathbf{J}_\mathbf{h}^{-1/2}$. By chain rule, we can write

$$\mathbf{K} = \partial \mathbf{h}(\boldsymbol{\xi}) / \partial \boldsymbol{\xi} = \partial \mathbf{h}(\boldsymbol{\phi}) / \partial \boldsymbol{\phi} \cdot \partial \boldsymbol{\phi}(\boldsymbol{\xi}) / \partial \boldsymbol{\xi} = \mathbf{H} \cdot \partial \boldsymbol{\phi}(\boldsymbol{\xi}) / \partial \boldsymbol{\xi}.$$

Therefore $\mathrm{span}(\mathbf{K}) \subseteq \mathrm{span}(\mathbf{H})$ and $\mathrm{span}(\mathbf{J}_\mathbf{h}^{1/2}\mathbf{K}) \subseteq \mathrm{span}(\mathbf{J}_\mathbf{h}^{1/2}\mathbf{H})$, which implies that $\mathbf{P}_{\mathbf{J}_\mathbf{h}^{1/2}\mathbf{K}} = \mathbf{P}_{\mathbf{J}_\mathbf{h}^{1/2}\mathbf{H}} \mathbf{P}_{\mathbf{J}_\mathbf{h}^{1/2}\mathbf{K}} = \mathbf{P}_{\mathbf{J}_\mathbf{h}^{1/2}\mathbf{K}} \mathbf{P}_{\mathbf{J}_\mathbf{h}^{1/2}\mathbf{H}}$. Finally, we have arrived at

$$\begin{aligned}
\mathbf{V}_{\mathbf{h}0} - \mathbf{V}_{\mathbf{h}\infty} &= \mathbf{H}(\mathbf{H}^\intercal \mathbf{J}_\mathbf{h} \mathbf{H})^\dagger \mathbf{H}^\intercal - \mathbf{K}(\mathbf{K}^\intercal \mathbf{J}_\mathbf{h} \mathbf{K})^\dagger \mathbf{K}^\intercal \\
&= \mathbf{J}_\mathbf{h}^{-1/2} \mathbf{P}_{\mathbf{J}_\mathbf{h}^{1/2}\mathbf{H}} \mathbf{J}_\mathbf{h}^{-1/2} - \mathbf{J}_\mathbf{h}^{-1/2} \mathbf{P}_{\mathbf{J}_\mathbf{h}^{1/2}\mathbf{K}} \mathbf{J}_\mathbf{h}^{-1/2} \\
&= \mathbf{J}_\mathbf{h}^{-1/2} \left(\mathbf{P}_{\mathbf{J}_\mathbf{h}^{1/2}\mathbf{H}} - \mathbf{P}_{\mathbf{J}_\mathbf{h}^{1/2}\mathbf{K}} \right) \mathbf{J}_\mathbf{h}^{-1/2} \\
&= \mathbf{J}_\mathbf{h}^{-1/2} \left(\mathbf{P}_{\mathbf{J}_\mathbf{h}^{1/2}\mathbf{H}} - \mathbf{P}_{\mathbf{J}_\mathbf{h}^{1/2}\mathbf{H}} \mathbf{P}_{\mathbf{J}_\mathbf{h}^{1/2}\mathbf{K}} \right) \mathbf{J}_\mathbf{h}^{-1/2} \\
&= \mathbf{J}_\mathbf{h}^{-1/2} \mathbf{P}_{\mathbf{J}_\mathbf{h}^{1/2}\mathbf{H}} \mathbf{Q}_{\mathbf{J}_\mathbf{h}^{1/2}\mathbf{K}} \mathbf{J}_\mathbf{h}^{-1/2} \geq 0.
\end{aligned}$$

So far, we have proved the main part of the Proposition 6. We next provide details about the gradient matrices $\mathbf{H}$ and $\mathbf{K}$. First of all, since $\mathbf{h}_1 = \boldsymbol{\phi}_1 = \mathrm{vec}(\mathbf{B})$, we have

$$\mathbf{H} = \begin{pmatrix} \mathbf{I}_{p \prod_{k=1}^{m} r_k} & 0 & \cdots & 0 \\ 0 & \frac{\partial \mathrm{vech}(\boldsymbol{\Sigma})}{\partial \mathrm{vech}(\boldsymbol{\Sigma}_1)} & \cdots & \frac{\partial \mathrm{vech}(\boldsymbol{\Sigma})}{\partial \mathrm{vech}(\boldsymbol{\Sigma}_m)} \end{pmatrix}. \tag{H1}$$

For a symmetric matrix $\mathbf{A} \in \mathbb{R}^{a \times a}$, $\mathrm{vech}(\mathbf{A}) = \mathbf{C}_a \mathrm{vec}(\mathbf{A})$ and $\mathrm{vec}(\mathbf{A}) = \mathbf{E}_a \mathrm{vech}(\mathbf{A})$, where $\mathbf{C}_a \in \mathbb{R}^{a(a+1)/2 \times a^2}$ is the contraction matrix and $\mathbf{E}_a \in \mathbb{R}^{a^2 \times a(a+1)/2}$ is the extraction matrix. Therefore,

$$\frac{\partial \mathrm{vech}(\boldsymbol{\Sigma})}{\partial \mathrm{vech}(\boldsymbol{\Sigma}_k)} = \mathbf{C}_{\prod_{j=1}^{m} r_j} \frac{\partial \mathrm{vec}(\boldsymbol{\Sigma})}{\partial \mathrm{vec}(\boldsymbol{\Sigma}_k)} \mathbf{E}_{r_k}, \quad k = 1, \ldots, m.$$

We then use the Kronecker structure $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_m \otimes \cdots \otimes \boldsymbol{\Sigma}_1$ to calculate $\frac{\partial \mathrm{vec}(\boldsymbol{\Sigma})}{\partial \mathrm{vec}(\boldsymbol{\Sigma}_k)}$. For $k = 1$ and $k = m$, we can use the formulas from Fackler (2005) for the derivatives of $\mathrm{vec}(\mathbf{A} \otimes \mathbf{B})$ over $\mathrm{vec}(\mathbf{A})$ and over $\mathrm{vec}(\mathbf{B})$. For $2 \leq k \leq m - 1$, there is no compact way of writing down the matrix form but elementwise derivatives, which is straightforward but non-trivial.

Next we write $\mathbf{K}$ as

$$\mathbf{K} = \mathbf{H} \cdot \frac{\partial \phi(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}} = \mathbf{HR}, \quad \mathbf{R} = \Big(\Big(\frac{\partial \mathrm{vec}(\mathbf{B})}{\partial \boldsymbol{\xi}}\Big)^\intercal \quad \Big(\frac{\partial \mathrm{vech}(\boldsymbol{\Sigma}_1)}{\partial \boldsymbol{\xi}}\Big)^\intercal \quad \cdots \quad \Big(\frac{\partial \mathrm{vech}(\boldsymbol{\Sigma}_m)}{\partial \boldsymbol{\xi}}\Big)^\intercal \Big)^\intercal, \tag{H2}$$

where we calculate each blocks of $\mathbf{R}$ in the following. Recall that $\boldsymbol{\xi} = (\boldsymbol{\xi}_1^\intercal, \ldots, \boldsymbol{\xi}_{3m+1}^\intercal)^\intercal$ and $\boldsymbol{\xi}_1 = \mathrm{vec}(\boldsymbol{\Theta})$, $\{\boldsymbol{\xi}_j\}_{j=2}^{m+1} = \{\mathrm{vec}(\boldsymbol{\Gamma}_k)\}_{k=1}^m$, $\{\boldsymbol{\xi}_j\}_{j=m+2}^{2m+1} = \{\mathrm{vech}(\boldsymbol{\Omega}_k)\}_{k=1}^m$, $\{\boldsymbol{\xi}_j\}_{j=2m+2}^{3m+1} = \{\mathrm{vech}(\boldsymbol{\Omega}_{0k})\}_{k=1}^m$.

The first block in $\mathbf{R}$ is

$$\frac{\partial \mathrm{vec}(\mathbf{B})}{\partial \boldsymbol{\xi}} = \Big(\frac{\partial \mathrm{vec}(\mathbf{B})}{\partial \mathrm{vec}(\boldsymbol{\Theta})} \quad \frac{\partial \mathrm{vec}(\mathbf{B})}{\partial \mathrm{vec}(\boldsymbol{\Gamma}_1)} \quad \cdots \quad \frac{\partial \mathrm{vec}(\mathbf{B})}{\partial \mathrm{vec}(\boldsymbol{\Gamma}_m)} \quad 0 \quad \cdots \quad 0 \Big), \tag{H3}$$

where the zeros are because of $\mathbf{B}$ does not depend on $\boldsymbol{\Omega}_k$ or $\boldsymbol{\Omega}_{0k}$. We re-write $\mathrm{vec}(\mathbf{B})$ as

$$\begin{aligned} \mathrm{vec}(\mathbf{B}) &= \mathrm{vec}(\mathbf{B}_{(m+1)}^\intercal) = \mathrm{vec}\left((\boldsymbol{\Gamma}_m \otimes \cdots \otimes \boldsymbol{\Gamma}_1)\boldsymbol{\Theta}_{(m+1)}^\intercal\right) \\ &= (\mathbf{I}_p \otimes \boldsymbol{\Gamma}_m \otimes \cdots \otimes \boldsymbol{\Gamma}_1)\mathrm{vec}(\boldsymbol{\Theta}_{(m+1)}^\intercal) \\ &= (\mathbf{I}_p \otimes \boldsymbol{\Gamma}_m \otimes \cdots \otimes \boldsymbol{\Gamma}_1)\mathrm{vec}(\boldsymbol{\Theta}). \end{aligned}$$

Thus,

$$\frac{\partial \mathrm{vec}(\mathbf{B})}{\partial \mathrm{vec}(\boldsymbol{\Theta})} = (\mathbf{I}_p \otimes \boldsymbol{\Gamma}_m \otimes \cdots \otimes \boldsymbol{\Gamma}_1). \tag{H4}$$

We next introduce the notation of re-arranging the vectorizations of a mode-N tensor $\mathbf{T} \in \mathbb{R}^{d_1 \times \cdots \times d_N}$: squared constant matrix$\boldsymbol{\Pi}_n^{\mathbf{T}}$ satisfies: $\mathrm{vec}(\mathbf{T}) = \boldsymbol{\Pi}_n^{\mathbf{T}}\mathrm{vec}(\mathbf{T}_{(n)})$. Therefore, $\mathrm{vec}(\mathbf{B}) = \boldsymbol{\Pi}_k^{\mathbf{B}}\mathrm{vec}(\mathbf{B}_{(k)})$ and hence

$$\begin{aligned} \mathrm{vec}(\mathbf{B}) &= \boldsymbol{\Pi}_k^{\mathbf{B}}\mathrm{vec}\left(\boldsymbol{\Gamma}_k\boldsymbol{\Theta}_{(k)}(\boldsymbol{\Gamma}_m^\intercal \otimes \cdots \otimes \boldsymbol{\Gamma}_{k+1}^\intercal \otimes \boldsymbol{\Gamma}_{k-1}^\intercal \otimes \cdots \otimes \boldsymbol{\Gamma}_1^\intercal)\right) \\ &= \boldsymbol{\Pi}_k^{\mathbf{B}}\left((\boldsymbol{\Gamma}_m \otimes \cdots \otimes \boldsymbol{\Gamma}_{k+1} \otimes \boldsymbol{\Gamma}_{k-1} \otimes \cdots \otimes \boldsymbol{\Gamma}_1)\boldsymbol{\Theta}_{(k)}^\intercal \otimes \mathbf{I}_{r_k}\right)\mathrm{vec}(\boldsymbol{\Gamma}_k), \end{aligned}$$

$$\frac{\partial \mathrm{vec}(\mathbf{B})}{\partial \mathrm{vec}(\boldsymbol{\Gamma}_k)} = \boldsymbol{\Pi}_k^{\mathbf{B}}\left((\boldsymbol{\Gamma}_m \otimes \cdots \otimes \boldsymbol{\Gamma}_{k+1} \otimes \boldsymbol{\Gamma}_{k-1} \otimes \cdots \otimes \boldsymbol{\Gamma}_1)\boldsymbol{\Theta}_{(k)}^\intercal \otimes \mathbf{I}_{r_k}\right). \tag{H5}$$

Finally for $\boldsymbol{\Sigma}_k = \boldsymbol{\Gamma}_k\boldsymbol{\Omega}_k\boldsymbol{\Gamma}_k^\intercal + \boldsymbol{\Gamma}_{0k}\boldsymbol{\Omega}_{0k}\boldsymbol{\Gamma}_{0k}^\intercal$, we have

$$\frac{\partial \mathrm{vech}(\boldsymbol{\Sigma}_k)}{\partial \boldsymbol{\xi}} = \Big(0 \quad \cdots \quad 0 \quad \frac{\partial \mathrm{vech}(\boldsymbol{\Sigma}_k)}{\partial \mathrm{vec}(\boldsymbol{\Gamma}_k)} \quad 0 \quad \cdots \quad 0 \quad \frac{\partial \mathrm{vech}(\boldsymbol{\Sigma}_k)}{\partial \mathrm{vech}(\boldsymbol{\Omega}_k)} \quad 0 \quad \cdots \quad 0 \quad \frac{\partial \mathrm{vech}(\boldsymbol{\Sigma}_k)}{\partial \mathrm{vech}(\boldsymbol{\Omega}_{0k})} \quad 0 \quad \cdots \quad 0 \Big), \tag{H6}$$

where the three nonzero elements are, (analogous to Cook et al. (2010)),

$$\begin{aligned}
\frac{\partial \mathrm{vech}(\boldsymbol{\Sigma}_k)}{\partial \mathrm{vec}(\boldsymbol{\Gamma}_k)} &= 2\mathbf{C}_{r_k}(\boldsymbol{\Gamma}_k\boldsymbol{\Omega}_k \otimes \mathbf{I}_{r_k} - \boldsymbol{\Gamma}_{r_k} \otimes \boldsymbol{\Gamma}_{0k}\boldsymbol{\Omega}_{0k}\boldsymbol{\Gamma}_{0k}^{\mathsf{T}}), \\
\frac{\partial \mathrm{vech}(\boldsymbol{\Sigma}_k)}{\partial \mathrm{vech}(\boldsymbol{\Omega}_k)} &= \mathbf{C}_{r_k}(\boldsymbol{\Gamma}_k \otimes \boldsymbol{\Gamma}_k)\mathbf{E}_{u_k}, \\
\frac{\partial \mathrm{vech}(\boldsymbol{\Sigma}_k)}{\partial \mathrm{vech}(\boldsymbol{\Omega}_{0k})} &= \mathbf{C}_{r_k}(\boldsymbol{\Gamma}_{0k} \otimes \boldsymbol{\Gamma}_{0k})\mathbf{E}_{r_k-u_k}.
\end{aligned}$$

Finally the explicit gradient matrices $\mathbf{H}$ is obtained by plugging each blocks (H3)–(H6) into (H1); similarly, the explicit gradient matrices $\mathbf{K}$ is obtained by plugging each blocks (H3)–(H6) into (H2). We have thus completed the proof of this theorem.

Additional References

Cook, R. D., Li, B. and Chiaromonte, F. (2010), Envelope models for parsimonious and efficient multivariate linear regression, *Statistica Sinica*, **20(3)**, 927–960.

Cook, R. D. and Zhang, X. (2015), Simultaneous envelope for multivariate linear regression, *Technometrics*, **57(1)**, 11–25.

Cook, R. D. and Zhang, X. (2016), Algorithms for envelope estimation, *Journal of Computational and Graphical Statistics*, **In press**.

Fackler, P. L. (2005). Notes on matrix calculus. `http://www4.ncsu.edu/~pfackler/MatCalc.pdf`, Technical Report.

Gupta, A. and Nagar, D. (1999), *Matrix variate distributions*, CRC Press.

Kolda, T. G. (2006). Multilinear operators for higher-order decompositions. United States, Department of Energy, Technical Report.

Lehmann, E. L. and Casella, G. (1998). *Theory of point estimation*, Springer Science & Business Media.

Manceur, A. M. and Dutilleul, P. (2013), Maximum likelihood estimation for the tensor normal distribution: Algorithm, minimum sample size, and empirical bias and dispersion, *Journal of Computational and Applied Mathematics*, **239**, 37–49.

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