A modified partial envelope tensor response regression

Wenxing Guo, School of Mathematics, Statistics and Actuarial Science, University of Essex

Narayanaswamy Balakrishnan, Department of Mathematics and Statistics, McMaster University, Canada

Shanshan Qin, School of Statistics, Tianjin University of Finance and Economics, China

ABSTRACT

The envelope model is a useful statistical technique that can be applied to multivariate linear regression problems. It aims to remove immaterial information via sufficient dimension reduction techniques while still gaining efficiency and providing accurate parameter estimates. Recently, envelope tensor versions have been developed to extend this technique to tensor data. In this work, a partial tensor envelope model is proposed that allows for a parsimonious version of tensor response regression when only certain predictors are of interest. The consistency and asymptotic normality of the regression coefficients estimator are also established theoretically, which provides a rigorous foundation for the proposed method. In numerical studies using both simulated and real-world data, the partial tensor envelope model is shown to outperform several existing methods in terms of the efficiency of the regression coefficients associated with the selected predictors.

KEYWORDS

dimension reduction, envelope model, sparsity principle, tensor regression

1 INTRODUCTION

Envelope models were proposed initially by Cook et al. (2010) as a method to simultaneously achieve dimension reduction on response and improved parameter estimation in multivariate linear regression. Serial developments and extensions of the envelope method have been made in the literature since then. Generally speaking, envelope regression models and associated inference are derived under different data regimes. In the context of the multivariate linear regression model, both response and predictors are vector-valued (one-way tensor). The developed methods include envelopes (Cook et al., 2010), partial envelopes (Su & Cook, 2011), envelopes and partial least-squares (Cook et al., 2013), and new envelope methods extending the linear model to a general multivariate context (Cook & Zhang, 2015), such as weighted least-squares, generalized linear model, and Cox regression. Moreover, Zhang et al. (2018) extended the envelope methodology beyond the usual multivariate regression setting to functional data analysis. In addition to vector-valued predictors, some envelope extensions allow the predictors to be tensor-valued (Zhang & Li, 2017). Ding and Cook (2018) studied complex structures in which the response is a random matrix variate, and predictors can be either scalar, vector, or matrix.

In many modern statistical applications, tensor-valued data, that is, multidimensional arrays, are commonly encountered in science, engineering, and medicine. Examples of tensor-valued data include electroencephalography (EEG, two-way tensor), anatomical magnetic resonance imaging (MRI, three-way tensor), and functional magnetic resonance imaging (fMRI, four-way tensor), among others (Zhang & Li, 2017). Therefore, methods that deal with multiple tensor-valued regressions have been developed. Kong et al. (2019) developed a low-rank linear regression model to correlate a matrix response with a high-dimensional vector of predictors when coefficient matrices have low-rank structures. Zhou et al. (2013) proposed a new family of tensor regression models that efficiently exploit the special structure of tensor covariates. Li and Zhang (2017) developed a parsimonious tensor response regression model with a multidimensional array (tensor) response and a vector predictor. Zhang and Chen (2020) discussed a principal envelope model, showing that any subset of principal components can preserve most of the information of the sample.

All the above-mentioned works tackle regression with vector-, matrix-, or tensor-valued predictors. However, little work exists on regression with a tensor-valued response. In this work, we study a class of envelope models with tensor-valued responses and vector-valued predictors. In

the case where some of the predictors are of particular interest, we propose a parsimonious tensor partial envelope by extending the multivariate linear regression of Su and Cook (2011) to a tensor version. Though the proposed method is an extension of their work, we can present many results formally under the framework of tensor-valued response. We show that the parsimonious models have milder restrictions and also can result in improved efficiency. Finally, we perform simulation studies and real data analysis to verify the developed theoretical results.

The rest of this paper is organized as follows. We first introduce tensor notations and review the existing envelope methods in Section 2. Section 3 presents the methodology of the proposed partial tensor envelope model, including model setup, parameter estimation, and theoretical properties of the estimator. Section 4 presents the simulations and real data analysis. Some concluding remarks are finally made in Section 5.

2 | NOTATIONS AND ENVELOPE MODELS REVISITED

2.1 | Notations

Throughout the paper, we use the following tensor notations and operations. More details can be found in Kolda and Bader (2009). A multidimensional array $\mathcal{A} \in \mathbb{R}^{r_1 \times \ldots \times r_m}$ is called an *m-order tensor*. In particular, vectors and matrices are tensors of orders one and two, respectively. The $vec(\mathcal{A})$ operator stacks the entries of a tensor $\mathcal{A}$ into a column vector, that is, an entry of $\mathcal{A}$, $a_{i_1 \ldots i_m}$ maps to the ℓ-th entry of $vec(\mathcal{A})$, where $\ell = 1 + \sum_{k=1}^{m} (i_k - 1) \Pi_{j=1}^{k-1} r_j$. Matricization, also known as unfolding or flattening, is the process of reordering the elements of an *m-way array* into a matrix. The *mode-k* matricization of a tensor $\mathcal{A}$ maps $\mathcal{A}$ into a matrix, denoted by $\boldsymbol{A}_{(\boldsymbol{k})} \in \mathbb{R}^{r_k \times \left(\prod_{j \neq k} r_j \right)}$. The *k-mode* (*matrix*) *product* of a tensor $\mathcal{A}$ and a matrix $\boldsymbol{M} \in \mathbb{R}^{l \times r_k}$ leads to an *m*-way tensor denoted by $\mathcal{A} \times_{\boldsymbol{k}} \boldsymbol{M} \in \mathbb{R}^{r_1 \times \ldots \times r_{k-1} \times l \times r_{k+1} \times \ldots \times r_m}$. Similarly, the *k*-mode (vector) product of a tensor $\mathcal{A}$ and a vector $\boldsymbol{X} \in \mathbb{R}^p$ results in an $(m-1)$-order tensor denoted by $\mathcal{A} \bar{\times}_{\boldsymbol{k}} \boldsymbol{X} \in \mathbb{R}^{r_1 \times \ldots \times r_{k-1} \times r_{k+1} \times \ldots \times r_m}$. The *Tucker decomposition* of a tensor is defined as

$$\mathcal{A} = \mathcal{M} \times_{\boldsymbol{1}} \boldsymbol{D_1} \times_{\boldsymbol{2}} \ldots \times_{\boldsymbol{m}} \boldsymbol{D_m} = [[\mathcal{M}; \boldsymbol{D_1}, \ldots, \boldsymbol{D_m}]],$$

where $\mathcal{M} \in \mathbb{R}^{t_1 \times \ldots \times t_m}$ is called the *core tensor* and its entries show the level of interaction between the different components, $\boldsymbol{D}_i \in \mathbb{R}^{r_i \times t_i}, i = 1, \ldots, m$, are the factor matrices; the second equality is the shorthand notation $[[\mathcal{M}; \boldsymbol{D_1}, \ldots, \boldsymbol{D_m}]]$ introduced in Kolda (2006). For a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{r \times r}$, the $vech(A) \in \mathbb{R}^{r(r+1)/2}$ stacks the unique entries of a symmetric matrix into a column vector.

2.2 | Response envelope model

Envelope method was developed by Cook et al. (2010) for the multivariate linear model,

$$\boldsymbol{Y} = \boldsymbol{\beta X} + \boldsymbol{\varepsilon}, \text{with } \operatorname{cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}, \tag{1}$$

where $\boldsymbol{Y} \in \mathbb{R}^r$ is the random response vector, $\boldsymbol{X} \in \mathbb{R}^p$ is a vector of predictors, and the random error vector $\boldsymbol{\varepsilon} \in \mathbb{R}^r$ is normally distributed with mean $\boldsymbol{0}$ and unknown $\boldsymbol{\Sigma}$. The envelope method is built on a key assumption that some linear combinations of $\boldsymbol{Y}$ are immaterial to the regression, while other linear combinations of $\boldsymbol{Y}$ depend on $\boldsymbol{X}$ and are thus important to the regression. In fact, envelopes separate the material and immaterial parts of $\boldsymbol{Y}$. More specifically, let $\boldsymbol{P_\xi Y}$ denote the projection of $\boldsymbol{Y}$ onto a subspace $\xi \subseteq \mathbb{R}^r$ with the following two properties: (i) the marginal distribution of $\boldsymbol{Q_\xi Y}$ does not depend on $\boldsymbol{X}$, and (ii) $\boldsymbol{P_\xi Y}$ is conditionally independent of $\boldsymbol{Q_\xi Y}$ given $\boldsymbol{X}$. The two conditions, when combined, imply that the distribution of $\boldsymbol{Q_\xi Y}$ is not affected marginally by $\boldsymbol{X}$ or through an association with $\boldsymbol{P_\xi Y}$. As a result, changes in $\boldsymbol{X}$ influence this distribution only through $\boldsymbol{P_\xi Y}$. The $\boldsymbol{\Sigma}$-envelope of span $(\boldsymbol{\beta})$, denoted by $\xi_{\boldsymbol{\Sigma}}(\mathcal{B})$, is defined as the intersection of all reducing subspaces of $\boldsymbol{\Sigma}$ containing $\mathcal{B}$. Let $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0) \in \mathbb{R}^{r \times r}$ be an orthogonal matrix with $\boldsymbol{\Gamma} \in \mathbb{R}^{r \times u}$ being a column orthogonal matrix and span $(\boldsymbol{\Gamma}) = \xi_{\boldsymbol{\Sigma}}(\mathcal{B})$, and u denotes the dimension of $\xi_{\boldsymbol{\Sigma}}(\mathcal{B})$. This then leads directly to the following envelope model,

$$\boldsymbol{Y} = \boldsymbol{\beta X} + \boldsymbol{\varepsilon}, \text{with } \boldsymbol{\Sigma} = \boldsymbol{\Gamma \Omega \Gamma}^{\mathsf{T}} + \boldsymbol{\Gamma}_0 \boldsymbol{\Omega}_0 {\boldsymbol{\Gamma}_0}^{\mathsf{T}}, \text{and } \boldsymbol{\beta} = \boldsymbol{\Gamma \eta}, \tag{2}$$

where $\boldsymbol{\eta} \in \mathbb{R}^{u \times p}$ represents the coordinates of $\boldsymbol{\beta}$ relative to the basis $\boldsymbol{\Gamma}$, $\boldsymbol{\Omega} \in \mathbb{S}^{u \times u}$ and $\boldsymbol{\Omega}_0 \in \mathbb{S}^{(r-u) \times (r-u)}$ are both positive definite matrices, and $\boldsymbol{\eta}$, $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ depend on the basis $\boldsymbol{\Gamma}$. It should be mentioned that the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ depend only on $\xi_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})$ instead of the basis. The envelope estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}} = \boldsymbol{P}_{\hat{\boldsymbol{\xi}}} \hat{\boldsymbol{\beta}}_{\boldsymbol{OLS}}$, is the projection of the ordinary least-squares (OLS) estimator $\hat{\boldsymbol{\beta}}_{\boldsymbol{OLS}}$ of $\boldsymbol{\beta}$ onto the estimated envelope space. A detailed review of envelope models can be found in Cook (2018) and Lee and Su (2020).

2.3 | Tensor response model

Now, we review the tensor response model. For an m-order tensor response variable $\mathcal{Y} \in \mathbb{R}^{r_1 \times \cdots \times r_m}$, and a vector of predictor variable $\boldsymbol{X} \in \mathbb{R}^p$, Li and Zhang (2017) studied a class of tensor response linear model,

$$\mathcal{Y} = \mathcal{B} \overline{\times}_{(m+1)} \boldsymbol{X} + \boldsymbol{\varepsilon}, \tag{3}$$

where the symbol $\overline{\times}_{(m+1)}$ represents the $(m+1)$-mode vector product, $\mathcal{B} \in \mathbb{R}^{r_1 \times \cdots \times r_m \times p}$ denotes an $(m+1)$-order tensor regression coefficient that is the parameter of interest, and $\boldsymbol{\varepsilon} \in \mathbb{R}^{r_1 \times \cdots \times r_m}$ is an m-order tensor denoting a random error that is independent of $\boldsymbol{X}$ and has mean $\boldsymbol{0}$. In this model, it is assumed that the covariance of $\boldsymbol{\varepsilon}$ has a separable Kronecker covariance such that $cov[vec(\boldsymbol{\varepsilon})] = \boldsymbol{\Sigma} = \boldsymbol{\Sigma_m} \otimes \ldots \otimes \boldsymbol{\Sigma_1}$, where $\boldsymbol{\Sigma_j}$, $j = 1, \ldots, m$, are positive-definite matrices. The assumption of separable structure is also used in some other works about tensors (Fosdick & Hoff, 2014; Hoff, 2011; Li & Zhang, 2017; Zhang & Li, 2017). It is useful to reduce the number of free parameters in tensor regression. An alternative form of the tensor response linear model (3) is

$$vec(\mathcal{Y}) = \boldsymbol{B}_{(m+1)}^{T} \boldsymbol{X} + vec(\boldsymbol{\varepsilon}), \tag{4}$$

where $\boldsymbol{B}_{(m+1)} \in \mathbb{R}^{p \times \left(\prod_{j=1}^{m} r_j \right)}$ is the coefficient matrix that can be regarded as mode-$(m+1)$ matricization of the tensor coefficient $\mathcal{B}$. The main difference between (3) and (4) is that there is no separable covariance structure restricted to $\boldsymbol{\varepsilon}$ in (4). Thus, for a given set of sample data with size n, the OLS estimator based on the model (3) is given by

$$\hat{\mathcal{B}}_{\boldsymbol{OLS}} = \mathcal{U} \times_{(m+1)} \left[\left(\boldsymbol{F}\boldsymbol{F}^{T} \right)^{-1} \boldsymbol{F} \right], \tag{5}$$

where $\mathcal{U} \in \mathbb{R}^{r_1 \times \ldots \times r_m \times n}$ and $\boldsymbol{F} \in \mathbb{R}^{p \times n}$ are the stacked response array of $\mathcal{Y}$ and predictor matrix of $\boldsymbol{X}$, respectively.

3 | METHODOLOGY

3.1 | Partial tensor envelope model

Li and Zhang (2017) and Zhang and Li (2017) gave a generalized sparsity principle similar to the vector situation for the tensor response linear model. Assume that we can find a series of subspaces, $S_k \subseteq \mathbb{R}^{r_k}$, $k = 1, \ldots, m$, such that

$$\mathcal{Y} \times_k \boldsymbol{Q}_k | \boldsymbol{X} \sim \mathcal{Y} \times_k \boldsymbol{Q}_k, \ \mathcal{Y} \times_k \boldsymbol{Q}_k \perp\!\!\!\perp \mathcal{Y} \times_k \boldsymbol{P}_k | \boldsymbol{X}, \tag{6}$$

where $\boldsymbol{P}_k \in \mathbb{R}^{r_k \times r_k}$ is the projection matrix onto S_k, $\boldsymbol{Q}_k = \boldsymbol{I}_{r_k} - \boldsymbol{P}_k$ is the projection matrix onto the complement space of S_k, and the symbol $\perp\!\!\!\perp$ denotes statistical independence. For any S_k with those properties, $\mathcal{Y} \times_k \boldsymbol{P}_k$ carries all of the material information and perhaps some immaterial information, while $\mathcal{Y} \times_k \boldsymbol{Q}_k$ carries only immaterial information. Therefore, we refer to $\mathcal{Y} \times_k \boldsymbol{P}_k$ informally as the material part and to $\mathcal{Y} \times_k \boldsymbol{Q}_k$ as the immaterial part of the regression $\mathcal{Y}$ on $\boldsymbol{X}$, respectively. Using a *Tucker decomposition*, (6) can be expressed as

$$\mathcal{Q}(\mathcal{Y}) | \boldsymbol{X} \sim \mathcal{Q}(\mathcal{Y}), \ \mathcal{Q}(\mathcal{Y}) \perp\!\!\!\perp \mathcal{P}(\mathcal{Y}) | \boldsymbol{X}. \tag{7}$$

where $\mathcal{Q}(\mathcal{Y}) = [[\mathcal{Y}; \boldsymbol{P_1}, \ldots, \boldsymbol{P_m}]] \in \mathcal{R}^{r_1 \times \ldots \times r_m}$ is a Tucker decomposition with $\mathcal{Y}$ as the core tensor and $\boldsymbol{P_1}, \ldots, \boldsymbol{P_m}$ as the factor matrices, and $\mathcal{Q}(\mathcal{Y}) = \mathcal{Y} - \mathcal{P}(\mathcal{Y})$. This results in $\mathcal{Y} = \mathcal{P}(\mathcal{Y}) + \mathcal{Q}(\mathcal{Y})$, where $\mathcal{P}(\mathcal{Y})$ is the material part and $\mathcal{Q}(\boldsymbol{Y})$ is the immatertial part.

In practice, part of the predictors may be of special interest. In this case, we partition $\boldsymbol{X}$ into two sets of predictors $\boldsymbol{X_1} \in \mathbb{R}^{p_1}$ and $\boldsymbol{X_2} \in \mathbb{R}^{p_2}$, with $p_1 + p_2 = p$, and also partition the columns of $\mathcal{B}$ into $\mathcal{B}_1 \in \mathbb{R}^{r_1 \times \ldots \times r_m \times p_1}$ and $\mathcal{B}_2 \in \mathbb{R}^{r_1 \times \ldots \times r_m \times p_2}$. Then, the partial tensor response linear model can be expressed as

$$\mathcal{Y} = \mathcal{B}_1 \overline{\times}_{(m+1)} \boldsymbol{X_1} + \mathcal{B}_2 \overline{\times}_{(m+1)} \boldsymbol{X_2} + \boldsymbol{\varepsilon}, \tag{8}$$

where $\mathcal{B}_1$ is the coefficient associated with the predictors of interest. Accordingly, an alternative representation of (8) is given by

$$vec(\mathcal{Y}) = \boldsymbol{B}_{1(m+1)}^{T}\boldsymbol{X_1} + \boldsymbol{B}_{2(m+1)}^{T}\boldsymbol{X_2} + vec(\boldsymbol{\varepsilon}), \tag{9}$$

where $\boldsymbol{B}_{1(m+1)} \in \mathbb{R}^{p_1 \times \left(\prod_{j=1}^{m} r_j\right)}$ and $\boldsymbol{B}_{1(m+1)} \in \mathbb{R}^{p_2 \times \left(\prod_{j=1}^{m} r_j\right)}$ are the coefficient matrix that can be regarded as mode-$(m+1)$ matricization of the tensor coefficient $\mathcal{B}_1$ and $\mathcal{B}_2$, respectively.

Now, we consider the $\boldsymbol{\Sigma}$-envelope for $\mathscr{B}_1 = \text{span}(\mathcal{B}_1)$ and $\mathcal{B}_2$ as unrestricted parameter, which results in the parametric structure span $(\mathcal{B}_1) \subseteq \mathcal{T}_{\Sigma}(\mathscr{B}_1)$ and $\boldsymbol{\Sigma_k} = \boldsymbol{P_{1k}}\boldsymbol{\Sigma_k}\boldsymbol{P_{1k}} + \boldsymbol{Q_{1k}}\boldsymbol{\Sigma_k}\boldsymbol{Q_{1k}}$, where $\boldsymbol{P_{1k}}$ is the projection matrix onto $span(\mathscr{B}_1)$ and $\boldsymbol{Q_{1k}} = \boldsymbol{I_{1k}} - \boldsymbol{P_{1k}}$. Motivated by Su and Cook (2011), let $\boldsymbol{R_{1|2}}$ represent the population residuals from the regression $\boldsymbol{X_1}$ on $\boldsymbol{X_2}$. Then, the partial tensor linear model can be expressed as

$$\mathcal{Y} = \mathcal{B}_1 \overline{\times}_{(m+1)} \boldsymbol{R_{1|2}} + \mathcal{B}_2^{*} \overline{\times}_{(m+1)} \boldsymbol{X_2} + \boldsymbol{\varepsilon}, \tag{10}$$

or

$$vec(\mathcal{Y}) = \boldsymbol{B}_{1(m+1)}^{T}\boldsymbol{R_{1|2}} + \boldsymbol{B}_{2(m+1)}^{*T}\boldsymbol{X_2} + vec(\boldsymbol{\varepsilon}), \tag{11}$$

where $\mathcal{B}_2^{*}$ is a linear combination of $\mathcal{B}_1$ and $\mathcal{B}_2$. Furthermore, let $\mathcal{R}_{\boldsymbol{Y|2}} = \mathcal{Y} - \mathcal{B}_2^{*} \overline{\times}_{(m+1)} \boldsymbol{X_2}$, corresponding to the population residuals from the regression of $\mathcal{Y}$ on $\boldsymbol{X_2}$ alone. Thus, a linear model involving $\mathcal{B}_1$ alone can be parameterized as

$$\mathcal{R}_{\boldsymbol{Y|2}} = \mathcal{B}_1 \overline{\times}_{(m+1)} \boldsymbol{R_{1|2}} + \boldsymbol{\varepsilon}, \tag{12}$$

or alternatively as

$$vec(\mathcal{R}_{\boldsymbol{Y|2}}) = \boldsymbol{B_{1(m+1)}}^{T}\boldsymbol{R_{1|2}} + vec(\boldsymbol{\varepsilon}). \tag{13}$$

In practice, we use $\hat{\boldsymbol{R}}_{\boldsymbol{1|2}} = \boldsymbol{X_1} - Cov(\boldsymbol{X_1}, \boldsymbol{X_2})\left(\boldsymbol{\Sigma}_{\boldsymbol{X_2}}\right)^{-1}\boldsymbol{X_2}$ and $\hat{\mathcal{R}}_{\boldsymbol{Y|2}} = \mathcal{Y} - Cov(\mathcal{Y}, \boldsymbol{X_2}) \times_{(m+1)} \left(\boldsymbol{\Sigma}_{\boldsymbol{X_2}}\right)^{-1}\boldsymbol{X_2}$ as estimators of $\boldsymbol{R_{1|2}}$ and $\mathcal{R}_{\boldsymbol{Y|2}}$, respectively, where $\Sigma_{\boldsymbol{X_2}}$ denotes sample covariance matrix of predictors $\boldsymbol{X_2}$.

3.2 | Parameter estimation

Our goal is to estimate $\mathcal{B}_1$ by the tensor envelope $\mathcal{T}_{\Sigma}(\mathscr{B}_1)$, including the estimation of $\boldsymbol{\Sigma}$, and then estimate $\mathcal{B}_2$ via the ordinary least-squares (OLS) method to fit the residuals $\mathcal{Y} - \widehat{\mathcal{B}_1} \overline{\times}_{(m+1)} \boldsymbol{X_1}$ on $\boldsymbol{X_2}$. Given a set of data samples with size n, the objective function is given as follows:

$$l(\mathcal{B}_1, \boldsymbol{\Sigma}) = \log|\boldsymbol{\Sigma}| + \frac{1}{n}\sum_{i=1}^{n}\left\{vec\left(\mathcal{R}_{\boldsymbol{Y|2}}^{i}\right) - \boldsymbol{B_{1(m+1)}}^{T}\boldsymbol{R_{1|2}^{i}}\right\}^{T}\boldsymbol{\Sigma}^{-1}\left\{vec\left(\mathcal{R}_{\boldsymbol{Y|2}}^{i}\right) - \boldsymbol{B_{1(m+1)}}^{T}\boldsymbol{R_{1|2}^{i}}\right\}. \tag{14}$$

The procedure to estimate parameters is given by the following:

Step 1: Initialization, $\mathcal{B}_1^{(0)}$, $\boldsymbol{\Sigma}^{(0)} = \boldsymbol{\Sigma_m^{(0)}} \otimes \ldots \otimes \boldsymbol{\Sigma_1^{(0)}}$.

Step 2: Estimate the envelope basis $\left\{\Gamma_{\boldsymbol{k}}^{(\boldsymbol{t+1})}\right\}_{k=1}^{m}$, given $\mathcal{B}_1^{(t)}$ and $\boldsymbol{\Sigma}^{(t)}$. This can be obtained by minimizing the objective function in (15), subject to $\Phi_{\boldsymbol{k}}^{T}\Phi_{\boldsymbol{k}} = \boldsymbol{I}_{u_k}$,

$$f_k^{(t)}(\Phi_{\boldsymbol{k}}) = \log\left|\Phi_{\boldsymbol{k}}^{T}\boldsymbol{\Sigma}_{\boldsymbol{k}}^{(t)}\Phi_{\boldsymbol{k}}\right| + \log\left|\Phi_{\boldsymbol{k}}^{T}\left(\boldsymbol{N}_{\boldsymbol{k}}^{(t)}\right)^{-1}\Phi_{\boldsymbol{k}}\right|, \tag{15}$$

where $\boldsymbol{N_k}^{(t)} = \left(n\Pi_{j \neq k} r_j\right)^{-1}\sum_{i=1}^{n}\mathcal{R}_{\boldsymbol{Y|2}}^{\boldsymbol{i(k)}}\left\{\left(\boldsymbol{\Sigma_m^{(t)}}\right)^{-1} \otimes \ldots \otimes \left(\boldsymbol{\Sigma_{k+1}^{(t)}}\right)^{-1} \otimes \left(\boldsymbol{\Sigma_{k-1}^{(t)}}\right)^{-1} \otimes \ldots \otimes \left(\boldsymbol{\Sigma_1^{(t)}}\right)^{-1}\right\}\left(\mathcal{R}_{\boldsymbol{Y|2}}^{\boldsymbol{i(k)}}\right)^{T}$.

Step 3: Estimate $\mathcal{B}_1^{(\boldsymbol{t+1})}$, $\boldsymbol{\Sigma}^{(\boldsymbol{t+1})} = \boldsymbol{\Sigma_m^{(t+1)}} \otimes \ldots \otimes \boldsymbol{\Sigma_1^{(t+1)}}$ and $\mathcal{B}_2^{(\boldsymbol{t+1})}$, given $\left\{\Gamma_{\boldsymbol{k}}^{(\boldsymbol{t+1})}\right\}_{k=1}^{m}$. First, $\mathcal{B}_1^{(t+1)}$ and $\boldsymbol{\Sigma}_{\boldsymbol{k}}^{(t+1)}$ are updated respectively by

$$\mathcal{B}_1^{(t+1)} = \mathcal{U} \times_1 \boldsymbol{\Gamma}_1^{(t+1)}\left(\boldsymbol{\Gamma}_1^{(t+1)}\right)^{T} \times_2 \ldots \times_m \boldsymbol{\Gamma}_m^{(t+1)}\left(\boldsymbol{\Gamma}_m^{(t+1)}\right)^{T} \times_{(m+1)} \left[\left(\boldsymbol{F}\boldsymbol{F}^{T}\right)^{-1}\boldsymbol{F}\right]$$

and

$$\boldsymbol{\Sigma}_k^{(t+1)} = \boldsymbol{\Gamma}_k^{(t+1)} \boldsymbol{\Omega}_k^{(t+1)} \left(\boldsymbol{\Gamma}_k^{(t+1)}\right)^\top + \boldsymbol{\Gamma}_{0k}^{(t+1)} \boldsymbol{\Omega}_{0k}^{(t+1)} \left(\boldsymbol{\Gamma}_{0k}^{(t+1)}\right)^\top.$$

After that, $\mathcal{B}_2^{(t+1)}$ is updated via OLS to fit the residuals $\mathcal{Y} - \widehat{\mathcal{B}_1}^{(t+1)} \times_{(m+1)} \boldsymbol{X_1}$ on $\boldsymbol{X_2}$.

Step 4: Repeat steps 2–3 until convergence by satisfying the termination condition, say, the objective function smaller than the desired level of tolerance.

We remark that the envelope dimension of u_k is not required to be given in step 1 since no envelope-based estimator is needed. The initial sample covariance matrix $\boldsymbol{\Sigma}_k^{(0)}$ is an OLS estimator for the full model $\boldsymbol{Y} = \boldsymbol{\beta X} + \boldsymbol{\varepsilon}$, which is $\sqrt{n}$ consistent. In step 2, the right hand of (15) is equivalent to $\log\left|\boldsymbol{\Gamma}_k \boldsymbol{\Sigma}_k^{(t)} \boldsymbol{\Gamma}_k\right| + \log\left|\boldsymbol{\Gamma}_k \left(\mathbf{N}_k^{(t)}\right)^{-1} \boldsymbol{\Gamma}_k\right|$. Then the envelope basis can be estimated via

$$\hat{\boldsymbol{\Gamma}}_k^{(t+1)} = \arg\min_{\boldsymbol{\Gamma}_k} \log\left|\boldsymbol{\Gamma}_k^T \boldsymbol{\Sigma}_k^{(t)} \boldsymbol{\Gamma}_k\right| + \log\left|\boldsymbol{\Gamma}_k^T \left(\mathbf{N}_k^{(t)}\right)^{-1} \boldsymbol{\Gamma}_k\right|.$$

3.3 | Asymptotic properties of the estimators

We show asymptotic properties of the estimators $\hat{\mathcal{B}}_1$ and $\hat{\mathcal{B}}_2$. First, we prove their $\sqrt{n}$-consistent under some weak conditions, where the error term in model (8) is not limited to be normally distributed.

Theorem 1. Suppose the error term $vec(\boldsymbol{\varepsilon}_i)$, $i = 1,\ldots,n$, in the model (8) are independent and identically distributed with the finite fourth moment and the starting value $\boldsymbol{\Sigma}_k^{(0)}$ of the covariance estimator is $\sqrt{n}$-consistent, $k = 1,\ldots,m$. Then, both $\hat{\mathcal{B}}_1$ and $\hat{\mathcal{B}}_2$ are $\sqrt{n}$-consistent.

Proof of Theorem 1 is given in Appendix A. We now show their asymptotic normalities. Denote

$$g = \begin{pmatrix} g_1 \\ g_2 \\ g_3 \end{pmatrix} = \begin{pmatrix} vec(\mathcal{B}_2) \\ vec(\mathcal{B}_1) \\ vech(\boldsymbol{\Sigma}) \end{pmatrix}, \psi = \begin{pmatrix} \psi_1 \\ \psi_2 \\ \psi_3 \\ \vdots \\ \psi_{m+2} \end{pmatrix} = \begin{pmatrix} vec(\mathcal{B}_2) \\ vec(\mathcal{B}_1) \\ vech(\boldsymbol{\Sigma_1}) \\ \vdots \\ vech(\boldsymbol{\Sigma_m}) \end{pmatrix}, \theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_{3m+2}L \end{pmatrix},$$

where $\theta_1 = vec(\mathcal{B}_2)$, $\theta_2 = vec(\boldsymbol{\Theta})$, $\{\theta_j\}_{j=3}^{m+2} = \{vec(\boldsymbol{\Gamma_k})\}_{k=1}^m$, $\{\theta_j\}_{j=m+3}^{2m+2} = \{vech(\boldsymbol{\Omega_k})\}_{k=1}^m$, $\{\theta_j\}_{j=2m+3}^{3m+2} = \{vech(\boldsymbol{\Omega_{0k}})\}_{k=1}^m$.

Theorem 2. Suppose the error term $vec(\boldsymbol{\varepsilon_i})$, $i = 1,\ldots,n$, in the model (8) are independent and identically distributed with a normal distribution. Then, $\sqrt{n}\left(vec\left(\hat{\mathcal{B}}_j\right) - vec(\mathcal{B}_j)\right), j = 1,2$, converge in distribution to normal random vectors.

Proof of Theorem 2 is given in Appendix A. For the proof of Theorem 2, we use Proposition 4.1 in Shapiro (1986) to derive the asymptotic distribution: $\sqrt{n}(\hat{g} - g) \sim N(\mathbf{0}, \boldsymbol{\Lambda_0})$, where the details of the $\boldsymbol{\Lambda_0}$ is given in the proof of Theorem 2. The asymptotic covariances for $vec\left(\hat{\mathcal{B}}_2\right)$ and $vec\left(\hat{\mathcal{B}}_1\right)$ are the first two diagonal blocks of $\boldsymbol{\Lambda}_0$. Due to the complexity of the asymptotic covariances, the closed form is not provided. Since Shapiro's result is built on the normal assumption, the normality of $vec(\boldsymbol{\varepsilon}_i)$ is thus required in Theorem 2. When $vec(\boldsymbol{\varepsilon}_i)$ is not normal, the asymptotical normality of $vec\left(\hat{\mathcal{B}}_j\right)$ may still hold. But the asymptotic covariance matrix will be even more complex than that in Theorem 2.

3.4 | Selection of *u*

In the above discussion, the dimension u of the envelope is assumed to be known. In practice, however, this is unknown. Common methods to select u include cross-validation (CV), likelihood ratio testing (LRT), or an information criterion like AIC, BIC, and so forth. Cook (2018) discussed the details of these methods. AIC tends to select a model that contains the true model, but it usually overestimates u. BIC tends to select the correct u with probability tending to one as $n \to \infty$, but it may be slow to respond in small samples. LRT tends to perform best with small samples, but asymptotically it makes an error with rate α. In any particular application, the factors that constitute a small or large sample depend on other characteristics of the regression model, including the strength of the signal. The cross-validation method tends to balance variance and bias in the selection of u and so may naturally result in choices that are different from those suggested by LRT or

information criterion. In the simulations, we have listed and discussed all possible values of envelope dimensions of u_k by setting $u_1 = \ldots = u_m$. In real data analysis, we select the envelope dimension $u_k, k = 1,\ldots,m$ respectively via BIC given by Li and Zhang (2017), an explicit formula of BIC that minimizes

$$\mathrm{BIC}_k(u_k) = -\frac{n}{2}\log\left|\boldsymbol{\Gamma}_k^\top \boldsymbol{\Sigma}_k^{(0)} \boldsymbol{\Gamma}_k\right| - \frac{n}{2}\log\left|\boldsymbol{\Gamma}_k^\top \left(\mathbf{N}_k^{(0)}\right)^{-1} \boldsymbol{\Gamma}_k\right| + \log(n) p u_k.$$

More discussion on the reasonability of the BIC criterion is given by Li and Zhang (2017).

4 | NUMERICAL STUDY

4.1 | Simulations

In this section, we consider a 3-order tensor dataset with error covariance matrices generated from $\boldsymbol{\Sigma}_k = \boldsymbol{\Gamma}_k \boldsymbol{\Omega}_k \boldsymbol{\Gamma}_k^T + \boldsymbol{\Gamma}_{0k} \boldsymbol{\Omega}_{0k} \boldsymbol{\Gamma}_{0k}^T$, $k = 1,2,3$, where $(\boldsymbol{\Gamma}_k, \boldsymbol{\Gamma}_{0k})$ are obtained by standardizing an $r_k \times r_k$ matrix of independent uniform(0, 1) variables, $\boldsymbol{\Omega}_k = \mathbf{I}_{u_k}$ and $\boldsymbol{\Omega}_0 = 0.01\mathbf{I}_{r_k - u_k}$. For $\mathcal{B} \in \mathbb{R}^{r_1 \times r_2 \times r_3 \times p}$, $\boldsymbol{B}_k = \boldsymbol{\Gamma}_k \boldsymbol{\eta}_k \in \mathbb{R}^{r_k \times p}$, where the entries in $\boldsymbol{\eta}_k$ are generated from uniform(0, 1) identically and independently, $k = 1,2,3$. $\boldsymbol{X}$ follows a multivariate normal distribution with mean $\boldsymbol{0}$ and covariance matrix $\mathbf{I}_{p_k}$. The error criterion is defined as

$$SMSE = E\left\|\hat{\mathcal{B}}_{\boldsymbol{1}} - \mathcal{B}_{\boldsymbol{1}}\right\|. \tag{16}$$

We then compare the proposed method with the envelope and OLS estimators in terms of the accuracy and stability of SMSE. The performance of the methods is assessed using simulation settings with $p = 10, \boldsymbol{u} = (1,1,1)^\top, \boldsymbol{r} = (10,10,10)^\top, c = 2$. We also consider different n and candidates of u. Table 1 presents the averages and standard deviations of SMSE that were obtained based on 500 simulations. As Table 1 shows, when envelope dimensions $\boldsymbol{u}$ -are correctly selected, the proposed P-Envelope achieves the smallest SMSE, and so -does its standard deviation. It shows that the P-Envelope method outperforms the Envelope and OLS uniformly in terms of prediction accuracy and robustness. Moreover, when the dimension $\boldsymbol{u}$ is misspecified, the proposed method still performs best and has the least variation for different dimensions.

4.2 | Real data analysis

In this section, we use data based on blood sugar concentration in rabbits after insulin injection. These data were analyzed earlier by Ding and Cook (2018) and Vølund (1980). The experiment used 36 rabbits and divided them equally into four groups, each with different treatments

TABLE 1 The average SMSE and standard deviations (in parenthesis) over 500 simulations.

		$\boldsymbol{u}$				
$\boldsymbol{n}$	**Method**	**(1,1,1)**	**(2,2,2)**	**(3,3,3)**	**(4,4,4)**	**(5,5,5)**
100	P-Envelope	0.4345(0.0725)	0.6409(1.1246)	1.0440(1.5327)	1.9131(2.8940)	2.8297(3.2181)
	Envelope	1.0615(3.3561)	1.9894(4.0962)	2.9078(4.5642)	3.5572(5.1294)	4.0327(5.8183)
	OLS	5.5664(8.7157)	5.5664(8.7157)	5.5664(8.7157)	5.5664(8.7157)	5.5664(8.7157)
200	P-Envelope	0.4490(0.2584)	0.5716(0.8046)	1.0661(2.3145)	1.7246(3.5551)	2.4647(4.0131)
	Envelope	1.8618(3.5259)	2.4586(6.2476)	2.8574(7.9752)	3.2291(10.1167)	3.4892(10.8583)
	OLS	4.5200(13.4513)	4.5200(13.4513)	4.5200(13.4513)	4.5200(13.4513)	4.5200(13.4513)
500	P-Envelope	0.4348(0.0997)	0.4967(0.3034)	0.7515(0.9606)	1.1596(1.5257)	1.6656(1.8179)
	Envelope	1.8099(2.4935)	1.9654(2.5297)	2.1077(2.6188)	2.2273(2.6786)	2.3300(2.7256)
	OLS	2.7710(3.0195)	2.7710(3.0195)	2.7710(3.0195)	2.7710(3.0195)	2.7710(3.0195)
1000	P-Envelope	0.4298(0.0731)	0.4938(0.2504)	0.6892(0.7150)	1.0686(1.3298)	1.5026(1.9128)
	Envelope	1.7919(2.3752)	1.8833(2.3987)	1.9565(2.4320)	2.0317(2.4896)	2.0990(2.5571)
	OLS	2.3804(2.8114)	2.3804(2.8114)	2.3804(2.8114)	2.3804(2.8114)	2.3804(2.8114)

and dose levels. Let S_1 and T_1 denote standard treatment and test treatment with low dose levels, 0.75 units per rabbit, while S_2 and T_2 denote standard treatment and test treatment with high dose levels, 1.50 units per rabbit. After administering the insulin dose every day, the blood sugar concentration levels of rabbits in each group were measured at 0, 1, 2, 3, 4, and 5 h. Our interest is to study the percentage decreases in blood sugar concentrations at 1, 2, 3, 4, and 5 h compared with the initial concentration at 0 h. We then write the measurements for each rabbit as a matrix $\boldsymbol{Y} \in \mathbb{R}^{5\times 2}$. The rows denote the percentage decreases in blood sugar concentration per hour per day under two different treatments. The

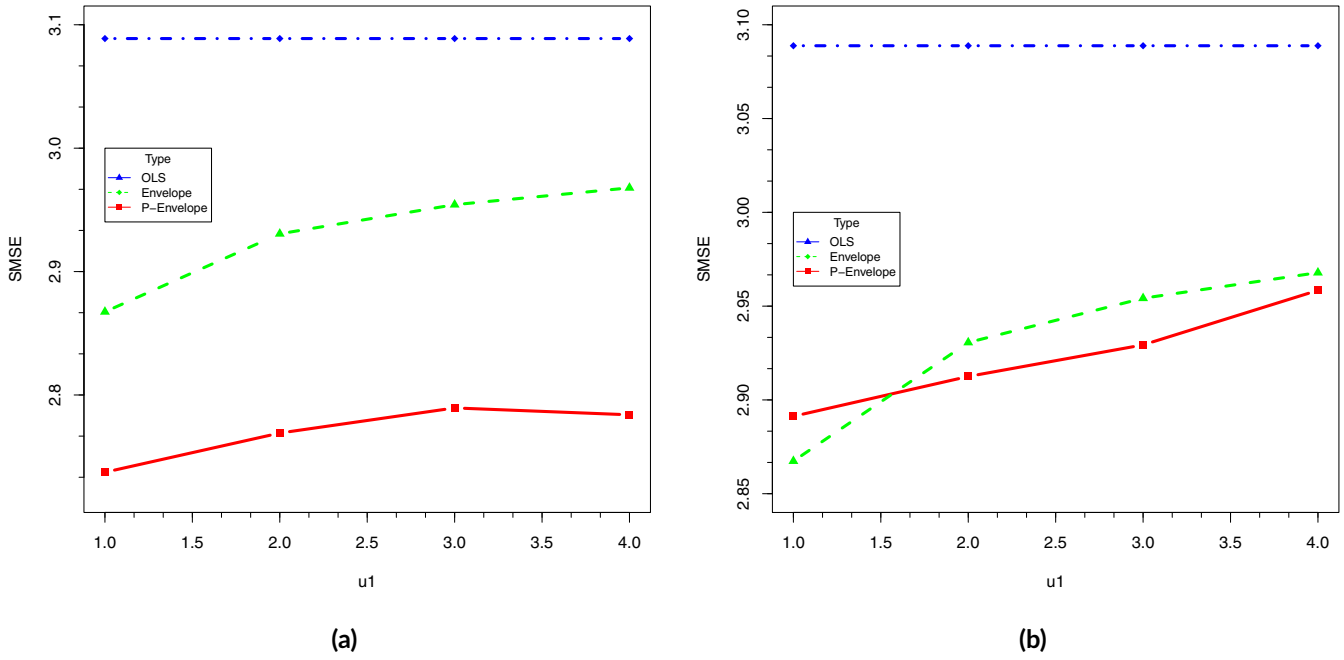

FIGURE 1 SMSE of three methods over different values of u_1, and P-Envelope denotes the envelope technology used for (a) X_1 and X_3, (b) X_1 and X_2.

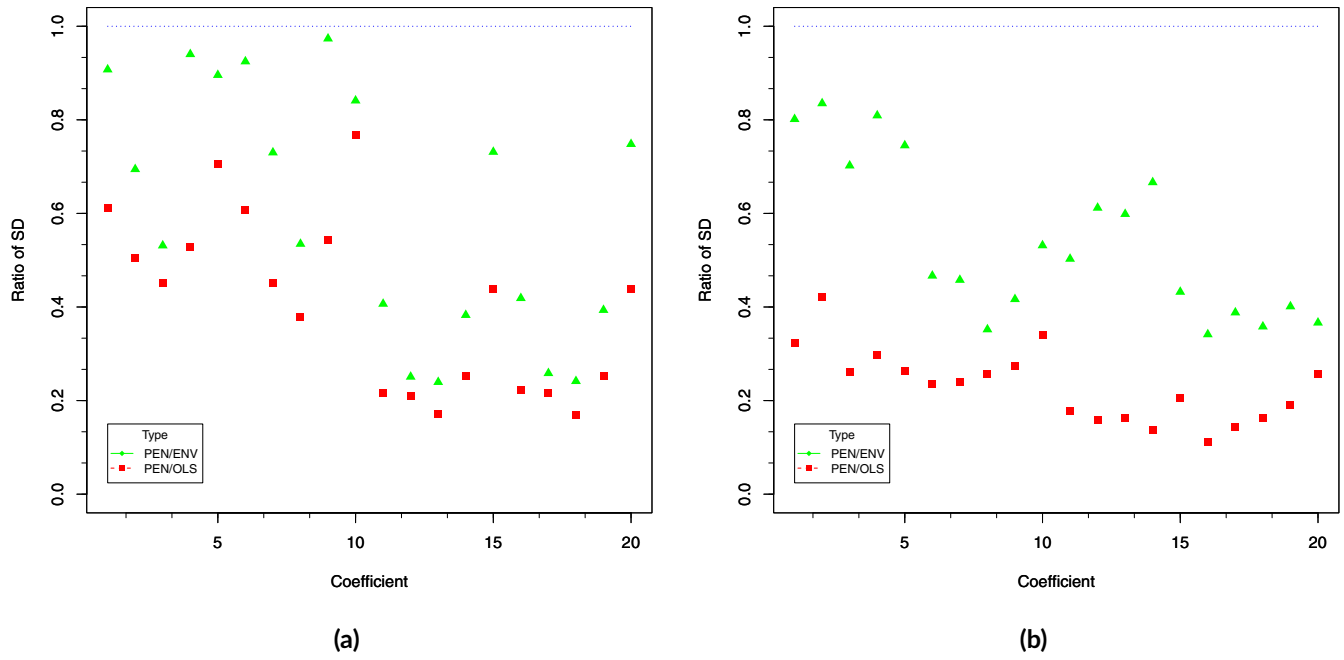

FIGURE 2 Ratio of the standard deviation of the estimated value of the corresponding coefficients. P-Envelope denotes the envelope technology used for (a) X_1 and X_3, (b) X_1 and X_2. P-Envelope/Envelope denotes the ratio of the proposed partial envelope method to the standard envelope method, and P-Envelope/OLS denotes the ratio of the proposed partial envelope method to the ordinary least squares method.

columns are the percentage reductions under two different treatments assigned on day one and day two. Two treatments and two different days form a predictor vector $\boldsymbol{X} \in \mathbb{R}^{4\times 1}$. For groups 1–4, the predictor vector $\boldsymbol{X} = (X_1, X_2, X_3, X_4)^T$ can be designated as $(0.75,0,0,1.5)^T$, $(1.5,0,0,0.75)^T$, $(0,1.5,0.75,0)^T$, and $(0,0.75,1.5,0)^T$, respectively. This study aims to find the relationship between these predictors and the percentage decreases in blood sugar concentration in terms of the model (10).

For these real data, we consider the proposed partial envelope method under two scenarios. In the first scenario, we use envelope technology for the effects of the two treatments on the first day, and the ordinary least-squares method for the effects on the second day. For the second scenario, we are interested in the standard treatment effects in two different days. Using BIC, we select the envelope dimension u_1 to be one. From Figure 1, we see that, for both scenarios, the smallest prediction errors occur at $u_1 = 1$. Moreover, as seen in Figure 1, the proposed partial envelope method and the standard envelope method improve the OLS method. However, the P-Envelope method performs the best over different choices of u_1. Figure 1b shows that both the partial envelope and standard envelope methods have similar performance in terms of prediction error. Figure 2 shows the standard deviations of the regression coefficient estimates. The partial envelope method achieves the smallest deviations, revealing that the proposed method outperforms other methods in terms of stability.

5 | CONCLUDING REMARKS

This paper proposes a partial envelope regression model that allows the response to be tensor-valued and the predictors to be vector-valued. It leads to a parsimonious version of tensor response while some predictors are of interest. Although the proposed method is an extension of a series of envelope methods, we formally report many results under the tensor-valued response framework. Theoretically, we prove that the regression coefficient estimators are consistent and asymptotic normality under mild conditions, resulting in improved efficiency. Experimentally, we show that the proposed method achieves significant gains in accuracy compared to some other methods.

ACKNOWLEDGMENTS

Shanshan Qin is supported by the National Natural Science Foundation of China [grant number 12201454] and the National Natural Science Foundation of China – Mathematical Tianyuan Fund [grant number 12226333].

REFERENCES

Cook, R. D. (2018). *An introduction to envelopes: Dimension reduction for efficient estimation in multivariate statistics*: John Wiley & Sons.

Cook, R. D., Helland, I. S., & Su, Z. (2013). Envelopes and partial least squares regression. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, *75*(5), 851–877.

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

Cook, R. D., & Zhang, X. (2015). Foundations for envelope models and methods. *Journal of the American Statistical Association*, *110*(510), 599–611.

Ding, S., & Cook, R. D. (2018). Matrix variate regressions and envelope models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *80*(2), 387–408.

Fosdick, B. K., & Hoff, P. D. (2014). Separable factor analysis with applications to mortality data. *The Annals of Applied Statistics*, *8*(1), 120.

Hoff, P. D. (2011). Separable covariance arrays via the tucker product, with applications to multivariate relational data. *Bayesian Analysis*, *8*, 179–196. Kolda, T. G., & Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review*, *51*(3), 455–500.

Kolda, T. G. (2006). Multilinear operators for higher-order decompositions. (*Tech. Rep.*): Sandia National Laboratories (SNL), Albuquerque, NM, and Livermore, CA.

Kong, D., An, B., Zhang, J., & Zhu, H. (2019). L2RM: Low-rank linear regression models for high-dimensional matrix responses. *Journal of the American Statistical Association*, *115*, 403–424.

Lee, M., & Su, Z. (2020). A review of envelope models. *International Statistical Review*, *88*(3), 658–676.

Li, L., & Zhang, X. (2017). Parsimonious tensor response regression. *Journal of the American Statistical Association*, *112*(519), 1131–1146.

Shapiro, A. (1986). Asymptotic theory of overparameterized structural models. *Journal of the American Statistical Association*, *81*(393), 142–149.

Su, Z., & Cook, R. D. (2011). Partial envelopes for efficient estimation in multivariate linear regression. *Biometrika*, *98*(1), 133–146.

Vølund, A. (1980). Multivariate bioassay. *Biometrics*, *36*(2), 225–236.

Zhang, J., & Chen, X. (2020). Principal envelope model. *Journal of Statistical Planning and Inference*, *206*, 249–262.

Zhang, X., & Li, L. (2017). Tensor envelope partial least-squares regression. *Technometrics*, *59*(4), 426–436.

Zhang, X., Wang, C., & Wu, Y. (2018). Functional envelope for model-free sufficient dimension reduction. *Journal of Multivariate Analysis*, *163*, 37–50.

Zhou, H., Li, L., & Zhu, H. (2013). Tensor regression with applications in neuroimaging data analysis. Journal of the American Statistical Association, 108(502), 540–552.

APPENDIX A: APPENDIX PROOFS OF THEOREMS

A.1 | Proof of Theorem 1

Proof. For $\hat{\mathcal{B}}_1$, the proof is similar to that of Theorem 1 in Li and Zhang (2017) and is therefore omitted.

Next, we prove the coefficient $\mathcal{B}_2$ is also $\sqrt{n}$-consistent. Let $\mathcal{Y}^*$ denote $\mathcal{Y} - \hat{\mathcal{B}}_1 \overline{\times}_{(m+1)} \boldsymbol{X}_1$. Then, Equation (8) can be expressed as

$$\mathcal{Y}^* = \mathcal{B}_2 \overline{\times}_{(m+1)} \boldsymbol{X}_2 + \boldsymbol{\varepsilon}. \tag{A1}$$

This results in the least-squares estimator of $\mathcal{B}_2$ based on model (A1), given the data $\{(\boldsymbol{X}_{1i}, \boldsymbol{X}_{2i}, \boldsymbol{Y}_i)\}_{i=1}^n$, as

$$\begin{aligned} \hat{\mathcal{B}}_{2OLS} &= \mathcal{U}^* \times_{(m+1)} \left[\left(\boldsymbol{F}_2 \boldsymbol{F}_2^T \right)^{-1} \boldsymbol{F}_2 \right] \\ &= \left[\mathcal{U} - \hat{\mathcal{B}}_1 \overline{\times}_{(m+1)} \boldsymbol{F}_1 \right] \times_{(m+1)} \left[\left(\boldsymbol{F}_2 \boldsymbol{F}_2^T \right)^{-1} \boldsymbol{F}_2 \right]. \end{aligned} \tag{A2}$$

Recall that $\mathcal{B}_1 - \hat{\mathcal{B}}_1 = O_p\left(n^{-1/2}\right)$, $\boldsymbol{\varepsilon} = O_p\left(n^{-1/2}\right)$ and that $\boldsymbol{F}$ is bounded. Then, by (A2), we have

$$\begin{aligned} \hat{\mathcal{B}}_{2OLS} &= \left[\mathcal{U} - \hat{\mathcal{B}}_1 \overline{\times}_{(m+1)} \boldsymbol{F}_1 \right] \times_{(m+1)} \left[\left(\boldsymbol{F}_2 \boldsymbol{F}_2^T \right)^{-1} \boldsymbol{F}_2 \right] \\ &= \left[\left(\mathcal{B}_1 - \hat{\mathcal{B}}_1 \right) \times_{(m+1)} \boldsymbol{F}_1 + \mathcal{B}_2 \times_{(m+1)} \boldsymbol{F}_2 + \boldsymbol{\varepsilon} \right] \times_{(m+1)} \left[\left(\boldsymbol{F}_2 \boldsymbol{F}_2^T \right)^{-1} \boldsymbol{F}_2 \right] \\ &= \left(\mathcal{B}_1 - \hat{\mathcal{B}}_1 \right) \times_{(m+1)} \left[\boldsymbol{F}_1{}^T \left(\boldsymbol{F}_2 \boldsymbol{F}_2^T \right)^{-1} \boldsymbol{F}_2 \right] + \mathcal{B}_2 \times_{(m+1)} \left[\boldsymbol{F}_2{}^T \left(\boldsymbol{F}_2 \boldsymbol{F}_2^T \right)^{-1} \boldsymbol{F}_2 \right] \\ &\quad + \boldsymbol{\varepsilon} \times_{(m+1)} \left[\left(\boldsymbol{F}_2 \boldsymbol{F}_2^T \right)^{-1} \boldsymbol{F}_2 \right]. \end{aligned}$$

This leads to

$$\begin{aligned} \hat{\mathcal{B}}_{2OLS} \times_{(m+1)} \left[\boldsymbol{F}_2{}^T \left(\boldsymbol{F}_2 \boldsymbol{F}_2^T \right)^{-1} \boldsymbol{F}_2 \right] &= \left(\mathcal{B}_1 - \hat{\mathcal{B}}_1 \right) \times_{(m+1)} \left[\boldsymbol{F}_1{}^T \left(\boldsymbol{F}_2 \boldsymbol{F}_2^T \right)^{-1} \boldsymbol{F}_2 \right] \\ &\quad + \mathcal{B}_2 \times_{(m+1)} \left[\boldsymbol{F}_2{}^T \left(\boldsymbol{F}_2 \boldsymbol{F}_2^T \right)^{-1} \boldsymbol{F}_2 \right] + \boldsymbol{\varepsilon} \times_{(m+1)} \left[\left(\boldsymbol{F}_2 \boldsymbol{F}_2^T \right)^{-1} \boldsymbol{F}_2 \right]; \end{aligned}$$

that is,

$$\begin{aligned} \left[\hat{\mathcal{B}}_{2OLS} - \mathcal{B}_2 \right] \times_{(m+1)} \left[\boldsymbol{F}_2{}^T \left(\boldsymbol{F}_2 \boldsymbol{F}_2^T \right)^{-1} \boldsymbol{F}_2 \right] &= \left(\mathcal{B}_1 - \hat{\mathcal{B}}_1 \right) \times_{(m+1)} \left[\boldsymbol{F}_1{}^T \left(\boldsymbol{F}_2 \boldsymbol{F}_2^T \right)^{-1} \boldsymbol{F}_2 \right] \\ &\quad + \boldsymbol{\varepsilon} \times_{(m+1)} \left[\left(\boldsymbol{F}_2 \boldsymbol{F}_2^T \right)^{-1} \boldsymbol{F}_2 \right]. \end{aligned}$$

Thus, we have

$$\hat{\mathcal{B}}_{2OLS} - \mathcal{B}_2 = O_p\left(n^{-1/2}\right) + O_p\left(n^{-1/2}\right) = O_p\left(n^{-1/2}\right),$$

which completes the proof of Theorem 1.

A.1.1 | Proof of Theorem 2

Proof. As $g = g(\psi) = g(\theta)$ is overparameterized, by Proposition 4.1 in Shapiro (1986), we have $\sqrt{n}(\hat{g} - g)$ converging in distribution to $N(\mathbf{0}, \boldsymbol{\Lambda_0})$, where $\boldsymbol{\Lambda_0} = \boldsymbol{\Psi}\left(\boldsymbol{\Psi^T J \Psi}\right)^{+}\boldsymbol{\Psi^T}$ and $\boldsymbol{\Psi} = \partial g(\theta)/\partial\theta$ is the gradient matrix. Let $\boldsymbol{\Delta}$ represent the limit of the sample covariance matrix of $\boldsymbol{X}$ as the sample size n tends to infinity, and $\boldsymbol{\Delta}_{ij}(i,j = 1,2)$ represent the covariance matrices corresponding to the partition $\boldsymbol{X}_{ij}(i,j = 1,2)$ of $\boldsymbol{X}$. The Fisher information matrix $\boldsymbol{J}$ of $\left[vec(\mathcal{B}_2)^T, vec(\mathcal{B}_1)^T, vech(\boldsymbol{\Sigma})^T\right]^T$ is given by

$$\boldsymbol{J} = \begin{pmatrix} \boldsymbol{\Delta_{22}} \otimes \boldsymbol{\Sigma^{-1}} & \boldsymbol{\Delta_{21}} \otimes \boldsymbol{\Sigma^{-1}} & \mathbf{0} \\ \boldsymbol{\Delta_{12}} \otimes \boldsymbol{\Sigma^{-1}} & \boldsymbol{\Delta_{11}} \otimes \boldsymbol{\Sigma^{-1}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \frac{1}{2}\boldsymbol{E_h^T}\left(\boldsymbol{\Sigma^{-1}} \otimes \boldsymbol{\Sigma^{-1}}\right)\boldsymbol{E_h} \end{pmatrix},$$

where $\boldsymbol{E_h} \in \mathbb{R}^{h^2 \times h(h+1)/2}$ is the expansion matrix and $h = \Pi_{k=1}^{m} r_k$.

Moreover, let $\boldsymbol{M} = \partial g(\psi)/\partial\psi$ and $\boldsymbol{N} = \partial\psi(\theta)/\partial\theta$. By chain rule, we then obtain $\boldsymbol{\Psi} = \partial g(\theta)/\partial\theta = \partial g(\psi)/\partial\psi \cdot \partial\psi(\theta)/\partial\theta = \boldsymbol{MN}$. Further, we have

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{I}_{p_2\Pi_{k=1}^{m} r_k} & \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{p_1\Pi_{k=1}^{m} r_k} & \boldsymbol{0} & \dots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \frac{\partial vech(\boldsymbol{\Sigma})}{\partial vech(\Sigma_1)} & \dots & \frac{\partial vech(\boldsymbol{\Sigma})}{\partial vech(\Sigma_m)} \end{pmatrix}$$

and

$$\boldsymbol{N} = \left[\left(\frac{\partial vec(\mathcal{B}_2)}{\partial\theta}\right)^T, \left(\frac{\partial vec(\mathcal{B}_1)}{\partial\theta}\right)^T, \left(\frac{\partial vech(\boldsymbol{\Sigma_1})}{\partial\theta}\right)^T, \dots, \left(\frac{\partial vech(\boldsymbol{\Sigma_m})}{\partial\theta}\right)^T\right]^T,$$

where

$$\frac{\partial vec(\mathcal{B}_2)}{\partial\theta} = \left(\boldsymbol{I}_{p_2\Pi_{k=1}^{m} r_k}, \boldsymbol{0}, \boldsymbol{0}, \dots, \boldsymbol{0}\right).$$

As in Li and Zhang (2017), we then obtain

$$\frac{\partial vec(\mathcal{B}_1)}{\partial\theta} = \left[\frac{\partial vec(\mathcal{B}_1)}{\partial vec(\boldsymbol{\Theta})}, \frac{\partial vec(\mathcal{B}_1)}{\partial vec(\boldsymbol{\Gamma}_1)}, \dots, \frac{\partial vec(\mathcal{B}_1)}{\partial vec(\boldsymbol{\Gamma}_m)}, \mathbf{0}, \dots, \mathbf{0}\right],$$

where

$$\frac{\partial vec(\mathcal{B}_1)}{\partial\boldsymbol{\Theta}} = \left(\boldsymbol{I}_{\boldsymbol{p_1}} \otimes \boldsymbol{\Gamma_m} \otimes \dots \otimes \boldsymbol{\Gamma_1}\right),$$

$$\frac{\partial vec(\mathcal{B}_1)}{\partial vec(\boldsymbol{\Gamma}_k)} = \boldsymbol{\Pi}_k^{\mathcal{B}_1}\left[\left(\boldsymbol{\Gamma_m} \otimes \dots \otimes \boldsymbol{\Gamma_{k+1}} \otimes \boldsymbol{\Gamma_{k-1}} \otimes \dots \otimes \boldsymbol{\Gamma_1}\right)\Theta_{(k)}{}^T \otimes \boldsymbol{I}_{\boldsymbol{r_k}}\right]$$

and

$$\frac{\partial vech(\boldsymbol{\Sigma_k})}{\partial \theta}=\left[\mathbf{0},...,\mathbf{0},\frac{\partial vech(\boldsymbol{\Sigma_k})}{\partial vec(\boldsymbol{\Gamma_k})},\mathbf{0},...,\mathbf{0},\frac{\partial vech(\boldsymbol{\Sigma_k})}{\partial vech(\boldsymbol{\Omega_k})},\mathbf{0},...,\mathbf{0},\frac{\partial vech(\boldsymbol{\Sigma_k})}{\partial vech(\boldsymbol{\Omega_{0k}})},\mathbf{0},...,\mathbf{0}\right],$$

with

$$\frac{\partial vech(\boldsymbol{\Sigma_k})}{\partial vec(\boldsymbol{\Gamma_k})}=2\boldsymbol{C}_{r_k}\left(\boldsymbol{\Gamma_k}\boldsymbol{\Omega_k}\otimes\boldsymbol{I}_{r_k}-\boldsymbol{\Gamma}_{r_k}\otimes\boldsymbol{\Gamma_{0k}}\boldsymbol{\Omega_{0k}}\boldsymbol{\Gamma}_{\mathbf{0k}}^{T}\right),$$

$$\frac{\partial vech(\boldsymbol{\Sigma_k})}{\partial vec(\boldsymbol{\Omega_k})}=\boldsymbol{C}_{r_k}(\boldsymbol{\Gamma_k}\otimes\boldsymbol{\Gamma_k})\boldsymbol{E}_{u_k},$$

$$\frac{\partial vech(\boldsymbol{\Sigma_k})}{\partial vec(\boldsymbol{\Omega_{0k}})}=\boldsymbol{C}_{r_k}(\boldsymbol{\Gamma_{0k}}\otimes\boldsymbol{\Gamma_{0k}})\boldsymbol{E}_{r_k-u_k},$$

and $\boldsymbol{C}_{r_k}\in\mathbb{R}^{r_k(r_k+1)/2\times r_k^2}$ is the contraction matrix.

For $\frac{\partial vech(\boldsymbol{\Sigma})}{\partial vech(\boldsymbol{\Sigma_k})}$, we obtain

$$\frac{\partial vech(\boldsymbol{\Sigma})}{\partial vech(\boldsymbol{\Sigma_k})}=\boldsymbol{C}_{\Pi_{i=1}^{m}r_i}\frac{\partial vec(\boldsymbol{\Sigma})}{\partial vec(\boldsymbol{\Sigma_k})}\boldsymbol{E}_{r_k}.$$

Next, we calculate $\frac{\partial vec(\boldsymbol{\Sigma})}{\partial vec(\boldsymbol{\Sigma_k})}$. When $k=1$ and $k=m$, we have

$$\frac{\partial vec(\boldsymbol{\Sigma})}{\partial vec(\boldsymbol{\Sigma_1})}=\left(\boldsymbol{K}_{r^2}\otimes\boldsymbol{I}_{(\Pi_{i=2}^{m}r_i)^2}\right)(\boldsymbol{I}_{r_1}\otimes vec(\boldsymbol{\Sigma}_m\otimes...\otimes\boldsymbol{\Sigma}_2)\otimes\boldsymbol{I}_{r_1})$$

and

$$\frac{\partial vec(\boldsymbol{\Sigma})}{\partial vec(\boldsymbol{\Sigma_m})}=\left(\boldsymbol{I}_{\Pi_{i=1}^{m}r_i}\otimes\boldsymbol{K}_{\Pi_{i=1}^{m}r_i}\right)(\boldsymbol{I}_{r_m}\otimes vec(\boldsymbol{\Sigma}_{m-1}\otimes...\otimes\boldsymbol{\Sigma}_1)\otimes\boldsymbol{I}_{r_m}).$$

When $2\le k\le m-1$, these derivatives cannot be written in matrix form, but they are indeed unique. This completes the proof of Theorem 2. □