# Regression for matrix-valued data via Kronecker products factorization 

Yin-Jen Chen<br>Minh Tang<br>Department of Statistics, North Carolina State University

May 1, 2024


#### Abstract

We study the matrix-variate regression problem $Y_{i}=\sum_{k} \beta_{1 k} X_{i} \beta_{2 k}^{\top}+E_{i}$ for $i=1,2 \ldots, n$ in the high dimensional regime wherein the response $Y_{i}$ are matrices whose dimensions $p_{1} \times p_{2}$ outgrow both the sample size $n$ and the dimensions $q_{1} \times q_{2}$ of the predictor variables $X_{i}$ i.e., $q_{1}, q_{2} \ll n \ll p_{1}, p_{2}$. We propose an estimation algorithm, termed KRO-PRO-FAC, for estimating the parameters $\left\{\beta_{1 k}\right\} \subset \Re^{p_{1} \times q_{1}}$ and $\left\{\beta_{2 k}\right\} \subset \Re^{p_{2} \times q_{2}}$ that utilizes the Kronecker product factorization and rearrangement operations from Van Loan and Pitsianis (1993). The KRO-PRO-FAC algorithm is computationally efficient as it does not require estimating the covariance between the entries of the $\left\{Y_{i}\right\}$. We establish perturbation bounds between $\hat{\beta}_{1 k}-\beta_{1 k}$ and $\hat{\beta}_{2 k}-\beta_{2 k}$ in spectral norm for the setting where either the rows of $E_{i}$ or the columns of $E_{i}$ are independent sub-Gaussian random vectors. Numerical studies on simulated and real data indicate that our procedure is competitive, in terms of both estimation error and predictive accuracy, compared to other existing methods.


Keywords: matrix regression, Kronecker product, low-rank approximation, matrix perturbations

## 1 Introduction

Regression is one of the most important and widely studied inference tasks in statistics and machine learning. Traditional applications of regression mainly focus on settings where the response variables $Y$ are either scalars or, more generally, $Y \in \Re^{d}$ for some "small" $d$. With the recent advancements in computation and storage technology, it is now common to encounter scenarios where the responses are (large) matrices. Examples include data from multivariate bioassay study (Vølund, 1980), electroencephalography (Li and Zhang, 2017), images denoising (Kamm
and Nagy, 1998; Nagy, 1996; Cai et al., 2019), and factors models in econometrics (Chen et al., 2019; Wang et al., 2019).

These type of data naturally leads to the simple and intuitive notion of matrix-variate regression wherein, given a collection of predictor and response tuples $\left\{\left(X_{i}, Y_{i}\right)\right\}_{i=1}^{n}$ with $X_{i} \in \Re^{q_{1} \times q_{2}}$ and $Y_{i} \in \Re^{p_{1} \times p_{2}}$, one typically assumes that the $Y_{i}$ are related to the $X_{i}$ through the linear model

$$
\begin{equation*}
\operatorname{vec}\left(Y_{i}\right)=\nu \operatorname{vec}\left(X_{i}\right)+\operatorname{vec}\left(E_{i}\right), \quad i=1, \ldots, n \tag{1}
\end{equation*}
$$

where $\nu \in \Re^{p_{1} p_{2} \times q_{1} q_{2}}$ are the unknown regression coefficients, $E_{i} \in \Re^{p_{1} \times p_{2}}$ are the unobserved noise matrices, and vec denote the vectorization operator that concatenates the column vectors of the input matrix. We note that any linear model for $\left\{\left(X_{i}, Y_{i}\right)\right\}$ can be written in the form of Eq. (1) In the high-dimensional regime wherein the dimensions of response variables ( $p_{1} . p_{2}$ ) grow much faster than the sample size $n$, i.e., $p_{i} / n \rightarrow \infty$, the regression coefficient $\nu$ is overparameterized and consistent estimation of $\nu$ is generally unfeasible unless one impose some structural assumptions on Eq. (1) so as to reduce the effective number of parameters in $\nu$.

Two of the most widely studied and adopted regularity conditions for $\nu$ is that it is low-rank and/or sparse; see e.g., Yuan et al. (2007); Obozinski et al. (2011); Negahban and Wainwright (2011); Chen et al. (2012); Bunea et al. (2012); Bing and Wegkamp (2019); Zheng et al. (2019); Zou et al. (2020) and the references therein. In particular Negahban and Wainwright (2011) noted that low-rank constraints are analogous to imposing sparsity on the data without explicitly specifying any basis.

Despite the popularity of these sparse and/or low-rank assumptions, they do not lead to a significant reduction in complexity of $\nu$ when the feature vectors $\left\{X_{i}\right\}$ are low-dimensional but the response $\left\{Y_{i}\right\}$ are high-dimensional. More specifically, suppose $p_{1}=p_{2}=p, q_{1}=q_{2}=q, p \gg q$ and $p / n \rightarrow \infty$. If we only assume that $\nu$ is low-rank so that $\operatorname{rk}(\nu)=d \ll n$ then we still need to estimate on the order of $O\left(d\left(p^{2}+q^{2}\right)\right)$ parameters for any matrix factorization of $\nu$ (such as SVD) and is computationally infeasible as it is equivalent to the estimation of covariance matrices in high-dimensional univariate linear regression with $p$ predictor variables and $n$ scalar responses. In contrast if we assume sparsity on $\nu$ then, denoting the number of non-zero entries in $\nu$ by $s$, we will in general need $n=\omega(s)$ to estimate $\nu$ consistently. As $s / p^{2} \ll p^{-1}$, this implies that almost all of the entries in the responses $\left\{Y_{i}\right\}$ are ignorable. This is a rather strong assumption that should be justified on a case-by-case basis.

In this paper we consider a more refined variant of the low-rank assumption on $\nu$ by assuming that it admits a representation/ approximation in terms of a (sum of) Kronecker products of smaller matrices. More specifically, we shall assume that $\nu$ is of the form

$$
\nu=\sum_{k=1}^{d} \beta_{2 k} \otimes \beta_{1 k}
$$

for some collection of $p_{1} \times q_{1}$ matrices $\left\{\beta_{11}, \ldots, \beta_{1 d}\right\}$ and $p_{2} \times q_{2}$ matrices $\left\{\beta_{21}, \ldots, \beta_{2 d}\right\}$. The number of effective parameters in $\nu$ is then $d\left(p_{1} q_{1}+p_{2} q_{2}\right)$ which is substantially smaller than $O\left(d\left(p_{1} p_{2}+q_{1} q_{2}\right)\right)$ for $p_{1} \gg q_{1}, p_{2} \gg q_{2}$. See Beylkin and Mohlenkamp (2002); De Lathauwer et al. (2000); Tyrtyshnikov (2004) for futher discussion of Kronecker product factorization and its use in large-scale matrix approximations.

Finally the linear model in Eq. (1) with the Kronecker product structure for $\nu$ is equivalent to the bi-linear model

$$
\begin{equation*}
Y_{i}=\sum_{k=1}^{d} \beta_{1 k} X_{i} \beta_{2 k}^{\top}+E_{i} . \tag{2}
\end{equation*}
$$

Under this perspective the $\left\{\beta_{1 k}\right\}$ (resp. $\left\{\beta_{2 k}\right\}$ ) can be interpreted as the row effects (resp. column effects) of $X_{i}$ on the response $Y_{i}$. The special case of $d=1$ was considered previously in Ding and Cook (2016) wherein the authors studied estimation of $\beta_{11}$ and $\beta_{21}$ using two-step MLEs; see Section 2 for further discussions. In a related vein, Chen et al. (2019); Wang et al. (2019); Chen and Fan (2021) considered factor models for $Y_{i}$ of the form $Y_{i}=\beta_{1} X_{i} \beta_{2}^{\top}$ but, in contrast to the current paper, they assume that the $X_{i}$ are either unknown or unobserved. They then propose to estimate $\beta_{1}$ and $\beta_{2}$ via two-step PCA Finally, the bi-linear modeling of $\left\{Y_{i}\right\}$ also arise in the context of image recognition (Crainiceanu et al., 2011; Wang et al., 2016; Ye, 2005; Zhang, 2005). In particular Crainiceanu et al. (2011) proposed the notion of population value decomposition for summarizing images population $\left\{Y_{i}\right\}$ by assuming that $Y_{i} \approx P V_{i} D$ where $P$ and $D$ encode "population frame of reference" for all $\left\{Y_{i}\right\}$ while $V_{i}$ encode "subject-level" features specific to a given $Y_{i}$. Their $P$ and $D$ thus serve identical roles to that of $\left\{\beta_{11}, \beta_{21}\right\}$ in Eq. (2) (when $d=1$ ).

In this paper we study estimation of $\left\{\beta_{1 k}, \beta_{2 k}\right\}$ for the model in Eq. (2). Inspired by the work of Van Loan and Pitsianis (1993) on the nearest Kronecker product problem, we observe that $\nu$ exhibits a low-rank representation after reshaping and rearranging the entries of $\nu$. In other words, while $\nu=\sum_{k=1}^{d} \beta_{2 k} \otimes \beta_{1 k}$ itself need not be low-rank, its rearranged version still admits a low-rank representation or approximation. Leveraging this observation we propose an algorithm,
termed KRO-PRO-FAC, for estimating $\nu$ with computational complexity of $O\left(p_{1} p_{2} d q_{1} q_{2}\right)$ flops; see Section 2). We next studied the theoretical properties of the KRO-PRO-FAC algorithm and show that it yield, under reasonably mild conditions on the noise $E_{i}$ and the dimensions $p_{i}$ (compared with the sample size $n$ ), consistent estimates of the $\left\{\beta_{1 k}, \beta_{2 k}\right\}$; see Section 3. Numerical experiments on simulated and real data are presented in Section 4. In particular our procedure is shown to be competitive, in terms of both estimation error and predictive accuracy, to other existing methods.

## 2 Methodologies

We now introduce some basic notations used throughout this paper. For $p \in \mathbb{N}$, we denote the set $\{1, \ldots, p\}$ by $[p]$. Let $\mathcal{O}(\cdot), \mathcal{O}(\cdot)$ and $\Theta(\cdot)$ represent the standard big-O, little-o and big-theta relationships. For two arbitrary real sequences $\left(a_{n}\right)_{n \in \mathbb{N}}$ and $\left(b_{n}\right)_{n \in \mathbb{N}}$, we write $a_{n} \ll b_{n}$ if $a_{n} / b_{n}$ converges to 0 as $n \rightarrow \infty ; a_{n} \asymp b_{n}$ means $a_{n} / b_{n}$ has a finite and non-zero limit as $n \rightarrow \infty$. For an arbitrary matrix $M=\left(M_{i j}\right) \in \Re^{p \times q}$, the Frobenius norm, spectral norm and nuclear norm of $M$ are denoted by $\|M\|_{F},\|M\|_{2}$ and $\|M\|_{*}$, and if $M$ is square then $\operatorname{tr}(M)$ and $|M|$ denote its trace and determinant. The symbol ' $\otimes$ ' represents the Kronecker product between matrices while $\boldsymbol{I}_{p}$ denote the $p \times p$ identity matrix. The vectorization of a $p \times q$ matrix $M$ is defined as

$$
\operatorname{vec}(M)=\left[\begin{array}{lllllllllll}
a_{11} & \ldots & a_{p 1} & \ldots & a_{12} & \ldots & a_{p 2} & \ldots & a_{p 1} & \ldots & a_{p q}
\end{array}\right]^{T} \in \Re^{p q}
$$

and we denote its inverse by $\operatorname{vec}^{-1}(m, p, q)$ where $m$ is a vector in $\Re^{p q}$.

### 2.1 Dual Kronecker products structure

We first discuss the special case of Eq. (2) with $d=1$, i.e., given a collection of matrix-variate predictors $\left\{X_{i}\right\}_{i=1}^{n} \subset \Re^{q_{1} \times q_{2}}$ and matrix-variate responses $\left\{Y_{i}\right\}_{i=1}^{n} \subset \Re^{p_{1} \times p_{2}}$, we consider the bilinear model of the form

$$
\begin{equation*}
Y_{i}=\beta_{1} X_{i} \beta_{2}^{\top}+E_{i}, \quad i \in[n] \tag{3}
\end{equation*}
$$

where $\beta_{1} \in \Re^{p_{1} \times q_{1}}$ and $\beta_{2} \in \Re^{p_{2} \times q_{2}}$ are the unknown regression coefficients and $E_{i}$ are unobserved noise matrices. Under this model each column (resp. row) of $Y_{i}$ is a noisy perturbation of some
linear combination of the columns of $\beta_{1}$ (resp. rows of $\beta_{2}^{\top}$ ). Recall that Eq. (3) can be rewritten in vectors form as

$$
\begin{equation*}
\operatorname{vec}^{\top}\left(Y_{i}\right)=\operatorname{vec}^{\top}\left(X_{i}\right)\left(\beta_{2}^{\top} \otimes \beta_{1}^{\top}\right)+\operatorname{vec}^{\top}\left(E_{i}\right), \quad i \in[n] \tag{4}
\end{equation*}
$$

and thus, by collecting all the $\operatorname{vec}^{\top}\left(Y_{i}\right)$ into a matrix and letting $\nu=\beta_{2} \otimes \beta_{1}$, leads to the linear regression model

$$
\begin{equation*}
\mathcal{Y}=\mathcal{X} \nu^{T}+\mathcal{E} \tag{5}
\end{equation*}
$$

where $\mathcal{Y}$ and $\mathcal{E}$ are $n \times p_{1} p_{2}$ matrices whose rows are the $\operatorname{vec}^{\top}\left(Y_{i}\right)$ and $\operatorname{vec}^{\top}\left(E_{i}\right)$ respectively while the design matrix $\mathcal{X} \in \Re^{n \times q_{1} q_{2}}$ has rows $\operatorname{vec}^{\top}\left(X_{i}\right)$. Several variants of formulation in Eq. (4) and Eq. (5) have been discussed in the literature; see e.g., Zhao and Leng (2014) for the case of $p_{1}=p_{2}=1$ and Kong et al. (2019) for the case of $q_{1}=q_{2}=1$ and $\nu$ being low-rank. Here Eq. (5) assumes that the mean coefficient $\nu$ admits a Kronecker product representation of $\beta_{2}$ and $\beta_{1}$. Without the Kronecker product factorization, $\nu$ can be easily overparameterized with $\mathcal{O}\left(p_{1} p_{2} q_{1} q_{2}\right)$ elements to be estimated. While $\nu$ is identifiable, is parametrization in terms of $\beta_{1}$ and $\beta_{2}$ is only identifiable up to a constant, i.e., $\beta_{2} \otimes \beta_{1}=c \beta_{2} \otimes c^{-1} \beta_{1}$ for any non-zero constant $c$.

As we allude to in the introduction, the bi-linear model in Eq. (3) had been studied previously in Ding and Cook (2016) and we now describe the pertinent details of this work in the context of the current paper. Denote the covariance matrix of $E_{i}$ as $\Sigma_{\mathrm{vec}(E)}$. Ding and Cook (2016) then assume that $\Sigma_{\mathrm{vec}(E)}$ can be decomposed as $\Sigma_{\mathrm{vec}(E)}=\Sigma_{2} \otimes \Sigma_{1}$; here $\Sigma_{1}$ and $\Sigma_{2}$ represent the covariance matrix for the rows and and columns of $Y_{i}$ respectively, i.e.,

$$
\begin{equation*}
\Sigma_{1}=\mathbb{E}\left\{\left(Y_{1}-\mathbb{E}\left(Y_{1}\right)\right)\left(Y_{1}-\mathbb{E}\left(Y_{1}\right)\right)^{T}\right\}, \quad \Sigma_{2}=\mathbb{E}\left\{\left(Y_{1}-\mathbb{E}\left(Y_{1}\right)\right)^{T}\left(Y_{1}-\mathbb{E}\left(Y_{1}\right)\right)\right\} \tag{6}
\end{equation*}
$$

The above structure for $\Sigma_{\mathrm{vec}(E)}$ is quite natural for longitudinal data wherein each subject is measured repeatedly over two different domains. For example the rows of $Y_{i}$ can record measurements over time while the columns of $Y_{i}$ record different covariates. Given Eq. (6), the number of parameters in $\Sigma_{\text {vec }(E)}$ is then reduced drasticically from $\mathcal{O}\left(p_{1}^{2} p_{2}^{2}\right)$ to $\mathcal{O}\left(p_{1}^{2}+p_{2}^{2}\right)$. This in turn allows the dimension $p_{1}$ and $p_{2}$ to possibly outgrow the sample size $n$, i.e., $n \ll \min \left\{p_{1}, p_{2}\right\}$.

Ding and Cook (2016) then consider MLE estimation of $\beta_{1}, \beta_{2}, \Sigma_{1}, \Sigma_{2}$ by further assuming that the $E_{i}$ follows the matrix normal distribution, i.e., $\operatorname{vec}\left(E_{i}\right) \sim \mathcal{N}\left(0, \Sigma_{\text {vec }(E)}\right)$; for more on the matrix normal distribution see De Waal (1985); Gupta and Nagar (1999) and the references therein. With
the above Kronecker product structure for the mean and covariance of $Y_{i}$, the log-likelihood of $\left\{Y_{i}\right\}$ given $\left\{X_{i}\right\}$ is (ignoring unimportant constants)

$$
\begin{equation*}
2 \ell\left(\left\{Y_{i}\right\} ; \theta\right)=-n p_{1} \ln \left|\Sigma_{1}\right|-n p_{2} \ln \left|\Sigma_{2}\right|-\sum_{i=1}^{n} \operatorname{tr}\left\{\Sigma_{2}^{-1}\left(Y_{i}-\beta_{1} X_{i} \beta_{2}^{\top}\right)^{T} \Sigma_{1}^{-1}\left(Y_{i}-\beta_{1} X_{i} \beta_{2}^{\top}\right)\right\} \tag{7}
\end{equation*}
$$

where $\theta=\left(\beta_{1}, \beta_{2}, \Sigma_{1}, \Sigma_{2}\right)$. Let $\hat{\theta}$ denote the MLE of $\theta$ from Eq. (7). As there are no closed-form expression $\hat{\theta}$, Ding and Cook (2016) proposed a two-stage iterative algorithm for finding $\hat{\theta}$ that is motivated by earlier work of Dutilleul (1999). More specifically the algorithm sequentially updates either the row parameters $\left(\beta_{1}, \Sigma_{1}\right)$ or the column parameters $\left(\beta_{2}, \Sigma_{2}\right)$, with the remaining parameters hold fixed. While the dual Kronecker product structure and the resulting MLE procedure provides a convenient way to model both the mean and covariance of the rows (and columns) simultaneously, there are nevertheless two major concerns regarding this approach. Firstly the MLE procedure is guaranteed to converge only to a stationary point but not necessarily a global optimum. Secondly, the update for $\left(\beta_{1}, \Sigma_{1}\right)$ (resp. $\left.\left(\beta_{2}, \Sigma_{2}\right)\right)$ require inverting $\Sigma_{2}$ (resp. $\Sigma_{1}$ ) and thus each updates involve possibly $\mathcal{O}\left(n\left(p_{1}^{2} p_{2}+p_{1} p_{2}^{2}\right)\right)$ flops, which is computationally prohibitive for moderate and/or large values of $p_{1}$ and $p_{2}$. In light of the above drawbacks we propose in Section 2.2 a more computationally efficient procedure which estimates only $\beta_{1}$ and $\beta_{2}$ but not $\Sigma_{1}, \Sigma_{2}$ or $\operatorname{Cov}\left[\operatorname{vec}\left(Y_{i}\right)\right]$.

### 2.2 Kronecker products factorization and low-rank approximation

If we assume a high-dimensional setting where the sample size $n$ is small or comparable to the dimensions $\min \left\{p_{1}, p_{2}\right\}$ of the response then it is generally the case that we can not estimate $\operatorname{Cov}\left(Y_{i}\right)$ to any reasonable degree of accuracy. One simple and intuitive remedy to this issue is to ignore the structure in $\operatorname{Cov}\left(Y_{i}\right)$ and instead focus our effort on estimating $\nu$.

Our starting point is the observation that although the OLS estimate $\tilde{\nu}=\left[\left(\mathcal{X}^{\top} \mathcal{X}\right)^{-1} \mathcal{X}^{\top} \mathcal{Y}\right]^{\top}$ is a simple and elegant estimate of $\nu$, it does not share the same Kronecker product structure as that for $\nu=\left(\beta_{2} \otimes \beta_{1}\right)$. It is thus natural to consider projecting $\tilde{\nu}$ onto the set formed by Kronecker products of matrices with appropriate dimensions. In particular let $p_{1}, q_{1}, p_{2}, q_{2}$ be positive integers and $M$ be a matrix of dimensions $p_{1} p_{2} \times q_{1} q_{2}$. The nearest Kronecker product approximation to
$M$ with respect to the dimensions $\left\{p_{i}, q_{i}\right\}$ is defined as

$$
\begin{equation*}
\underset{\beta_{1} \in \Re^{p_{1} \times q_{1}}, \beta_{2} \in \Re^{p_{2} \times q_{2}}}{\arg \min }\left\|M-\beta_{2} \otimes \beta_{1}\right\|_{F}^{2} . \tag{8}
\end{equation*}
$$

Van Loan and Pitsianis (1993) showed that Eq. (8) has a closed-form solution given by the truncated SVD of a rearranged version of $M$. More specifically first partition $M$ into smaller matrices $M_{i j} \in \Re^{p_{1} \times q_{1}}$ for $1 \leq i \leq p_{2}$ and $1 \leq j \leq q_{2}$, i.e.,

$$
M=\left[\begin{array}{cccc}
M_{11} & M_{12} & \cdots & M_{1 q_{2}}  \tag{9}\\
M_{21} & M_{22} & \cdots & M_{2 q_{2}} \\
\vdots & \vdots & \ddots & \vdots \\
M_{p_{2} 1} & M_{p_{2} 2} & \cdots & M_{p_{2} q_{2}}
\end{array}\right]
$$

Next define the rearrangement operation $R(\cdot): \Re^{p_{1} p_{2} \times q_{1} q_{2}} \rightarrow \Re^{p_{2} q_{2} \times p_{1} q_{1}}$ by

$$
R(M)=\left[\begin{array}{c}
A_{1}  \tag{10}\\
A_{2} \\
\ldots \\
A_{q_{2}}
\end{array}\right], \quad A_{j}=\left[\begin{array}{c}
\operatorname{vec}\left(M_{1 j}\right)^{\top} \\
\operatorname{vec}\left(M_{2 j}\right)^{\top} \\
\ldots \\
\operatorname{vec}\left(M_{p_{2} j}\right)^{\top}
\end{array}\right]
$$

We emphasize that $R(M)$ and $M$ generally have different dimensions. In particular, if $p_{1} \asymp p_{2}$, $q_{1} \asymp q_{2}$ and $q_{i} \ll p_{i}$ then $M$ is a tall matrix but the dimensions of $R(M)$ are comparable. The solution of Eq. (8) is then equivalent to finding the closest rank-1 representation of $R(M)$, i.e.,

$$
\begin{equation*}
\min _{\beta_{1}, \beta_{2}}\left\|M-\beta_{2} \otimes \beta_{1}\right\|_{F}=\min _{\beta_{1}, \beta_{2}}\left\|R(M)-\operatorname{vec}\left(\beta_{2}\right) \operatorname{vec}\left(\beta_{1}\right)^{\top}\right\|_{F} \tag{11}
\end{equation*}
$$

and thus, by the Eckart-Young-Mirsky theorem (Eckart and Young, 1936), we can take $\beta_{1}=$ $\sigma_{1}^{1 / 2} \operatorname{vec}^{-1}\left(\mathcal{V}_{1}, p_{1}, q_{1}\right)$ and $\beta_{2}=\sigma_{1}^{1 / 2} \mathrm{vec}^{-1}\left(\mathcal{U}_{1}, p_{2}, q_{2}\right)$ where $\sigma_{1}, \mathcal{U}_{1}$ and $\mathcal{V}_{1}$ are the leading singular values and (left and right) singular vectors of $R(M)$, respectively. See Van Loan and Pitsianis (1993) for more details. In summary if we assume a Kronecker product structure for the regression coefficient $\nu$ in Eq. (5) then our estimate for $\nu, \beta_{1}$ and $\beta_{2}$ is given by

Step 1: Let $\tilde{\nu} \leftarrow\left[\left(\mathcal{X}^{\top} \mathcal{X}\right)^{-1} \mathcal{X}^{\top} \mathcal{Y}\right]^{\top} \in \Re^{p_{1} p_{2} \times q_{1} q_{2}}$ be the OLS estimate of $\nu$ and let $R(\tilde{\nu})$ be its Pitsianis-Van Loan rearrangement; see Eq. (10).

Step 2: Compute the $\operatorname{SVD} R(\tilde{\nu})=\sum_{k=1}^{r} \hat{\sigma}_{k} \hat{\mathcal{U}}_{k} \hat{\mathcal{V}}_{k}^{\top}$ with $\hat{\sigma}_{1} \geq \cdots \geq \hat{\sigma}_{r}$ and $r \leq \min \left\{p_{1} q_{1}, p_{2} q_{2}\right\}$.

Step 3: Let $\operatorname{vec}\left(\hat{\beta}_{2}\right)=\hat{\sigma}_{1}^{1 / 2} \hat{\mathcal{U}}_{1}$ and $\operatorname{vec}\left(\hat{\beta}_{1}\right)=\hat{\sigma}_{1}^{1 / 2} \hat{\mathcal{V}}_{1}$
Step 4: Output the estimate $\hat{\nu} \leftarrow \hat{\beta}_{2} \otimes \hat{\beta}_{1}$ for $\nu$.
We emphasize that, despite the close connection between Kronecker products approximation and low-rank approximations described in Eq. (11), the assumption of a Kronecker factorization for $\nu$ is quite different from the assumption that $\nu$ is low-rank. Indeed, the rank of $\nu$ can be as large as $q_{1} q_{2}$ (assuming $p_{1} p_{2} \geq q_{1} q_{2}$ ) even when $R(\nu)$ is a rank- 1 matrix. This difference distinguishes our work from those which introduce penalty terms to induce low-rank structure on $\nu$ directly; see e.g., Kong et al. (2019); Wang et al. (2021); Feng et al. (2021) for recent examples of this latter approach. We now consider a simple simulation study to further illustrate this distinction.

Example. We set the dimensions of $Y_{i}$ and $X_{i}$ as $p_{1}=p_{2}=q_{1}=q_{2}=10$. We then generate $n=3000$ samples of the $\left\{\left(X_{i}, Y_{i}\right)\right\}$ pair according to the model $Y_{i}=\beta_{1} X_{i} \beta_{2}^{\top}+E_{i}$ where the $\left\{X_{i}\right\}$ are iid random vectors with $\operatorname{vec}\left(X_{i}\right) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and the $\operatorname{vec}\left(E_{i}\right)$ are also iid random vectors with $\operatorname{vec}\left(E_{i}\right) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Given the $\left\{\left(X_{i}, Y_{i}\right)\right\}$ we first compute the OLS estimate $\tilde{\nu}$ and its rearranged version $R(\tilde{\nu})$. Next define, for a matrix $M$ and an integer $k \geq 1$, the function $f_{k}(M)=$ $\left(\sum_{i=1}^{k} \sigma_{i}(M)\right) /\|M\|_{*}$ corresponding to the (normalized) sum of the $k$ largest singular values of $M$; here $\|M\|_{*}$ denotes the nuclear norm of $M$. We then compute, for each $k \in[100]$, the quantity $f_{k}(\tilde{\nu})$ and $f_{k}(R(\tilde{\nu}))$. Finally we repeat the above steps for 1000 Monte Carlo replicates. We note that the $\beta_{1}$ and $\beta_{2}$ are fixed constants and do not vary with the Monte Carlo replicates.

Figure 1 plots the (normalized) cumulative sum of the first $k$ singular values of $\tilde{\nu}$ and $R(\tilde{\nu})$ for $k$ varying in $\{1,2, \ldots, 100\}$; note that $f_{k}(\tilde{\nu})=f_{k}(R(\tilde{\nu}))=1$ when $k=100$. From Figure 1 we see that the largest singular value of $R(\tilde{\nu})$ accounts for, on average, roughly $87 \%$ of $\|R(\tilde{\nu})\|_{*}$ and thus a rank-1 approximation of $R(\tilde{\nu})$ is expected to preserve most of the information in $R(\tilde{\nu})$ while also removing the noise from the small singular values in $R(\tilde{\nu})$. In contrast the largest singular value of $\tilde{\nu}$ only explains $5 \%$ of $\|\tilde{\nu}\|_{*}$ and thus computing $\hat{\nu}$ using low-rank approximations to $\tilde{\nu}$ is possibly problematic.

### 2.3 KRO-PRO-FAC algorithm

A natural extension of the optimization problem in Eq. (8) is to approximate a matrix $\nu$ using a sum of $d$ Kronecker products which, by the above discussions, can be related to the sum of $d$


Figure 1: cumulative singular value (averaged over 1000 replications) for (a) the OLS estimate $\tilde{\nu}$ and (b) the rearranged estimate $R(\tilde{\nu})$
rank-1 matrices via

$$
\begin{equation*}
\underset{\left\{\left(\beta_{1 k}, \beta_{2 k}\right)\right\}}{\arg \min }\left\|\nu-\sum_{k=1}^{d} \beta_{2 k} \otimes \beta_{1 k}\right\|_{F}=\underset{\left\{\left(\beta_{1 k}, \beta_{2 k}\right)\right\}}{\arg \min }\left\|R(\nu)-\sum_{k=1}^{d} \operatorname{vec}\left(\beta_{2 k}\right) \operatorname{vec}\left(\beta_{1 k}\right)^{\top}\right\|_{F} . \tag{12}
\end{equation*}
$$

A solution of Eq. (12) is then once again given by the truncated SVD of $R(\nu)$. Eq. (12) furthermore suggests a more general version of the regression problem in Eq. (3), namely that

$$
\begin{equation*}
Y_{i}=\sum_{k=1}^{d} \beta_{1 k} X_{i} \beta_{2 k}^{\top}+E_{i}, \quad i \in[n] \tag{13}
\end{equation*}
$$

with $d \ll n \ll \min \left\{p_{1} q_{1}, p_{2} q_{2}\right\}$. Eq. (13) can be rewritten as

$$
\begin{equation*}
\mathcal{Y}=\mathcal{X} \nu^{T}+\mathcal{E}, \quad \nu=\sum_{k=1}^{d} \beta_{2 k} \otimes \beta_{1 k} \tag{14}
\end{equation*}
$$

Here we refer to $d$ in Eq. (14), as the Kronecker product rank of $\nu$. For ease of exposition (and without loss of generality) we shall assume that the $\left\{\beta_{1 k}, \beta_{2 k}\right\}$ are orthogonal, i.e., $\operatorname{vec}\left(\beta_{1 s}\right)^{\top} \operatorname{vec}\left(\beta_{1 t}\right)=$ $\operatorname{vec}\left(\beta_{2 s}\right)^{\top} \operatorname{vec}\left(\beta_{2 t}\right)=0$ for all $s \neq t$ and $\left\|\beta_{1 s}\right\|_{F}=\left\|\beta_{2 s}\right\|_{F}$.

Our estimate for $\nu$ and $\left\{\beta_{1 k}, \beta_{2 k}\right\}$ proceeds in an analogous manner to that described in Section 2.2. In particular we first compute the OLS estimate $\tilde{\nu}=\left[\left(\mathcal{X}^{\top} \mathcal{X}\right)^{-1} \mathcal{X}^{\top} \mathcal{Y}\right]^{\top}$, then rearrange $\tilde{\nu}$ to obtain $R(\tilde{\nu})$, and finally compute the truncated SVD of $R(\tilde{\nu})$ to keep only the $d$ largest singular values and singular vectors. We termed this procedure as the KRO-PRO-FAC (Kronecker product factorization) estimate of $\nu$. See Algorithm 1 for a more formal descriptions. As the Kronecker product rank $d$ of $\nu$ is generally unknown, we estimate it using the ratio of singular values as described in Lam and Yao (2012) and Ahn and Horenstein (2013), i.e., we estimate $d$ by

$$
\begin{equation*}
\hat{d}=\underset{j \in\{1, \ldots, \bar{d}\}}{\arg \max } \hat{\sigma}_{j} / \hat{\sigma}_{j+1} \tag{15}
\end{equation*}
$$

where $\bar{d}$ is a pre-specified constant and $\hat{\sigma}_{k}$ 's are the singular values of $R(\tilde{\nu})$ in a descending order.
The computational complexity of the KRO-PRO-FAC algorithm is $\mathcal{O}\left(p_{1} p_{2} q_{1} q_{2} \min \left\{p_{1} q_{1}, p_{2} q_{2}\right\}\right)$ with the main computational bottleneck being the SVD of $R(\tilde{\nu})$. If $d$ is either known or is estimated to be much smaller than the dimensions of $R(\tilde{\nu})$ then the cost of the SVD step reduces to $\mathcal{O}\left(p_{1} p_{2} q_{1} q_{2} d\right)$ flops by using either Lanczos bidiagonalization and/or randomized SVD, see e.g., Halko et al. (2011); Musco and Musco (2015); Tropp et al. (2017) and the references therein. Hence the complexity for the full algorithm itself drops to $\mathcal{O}\left(n p_{1} p_{2} q_{1} q_{2}\right)$. In contrast, any algorithm that involves estimating the covariance matrices for the rows and/or columns will requires at least $\mathcal{O}\left(n\left(p_{1}^{2} p_{2}+p_{1} p_{2}^{2}\right)\right)$ flops which is an enormous computational burden for large values of $p_{1}$ and/or $p_{2}$.

Remark 1. We note that even if $\nu$ does not have the form as specified in Eq. (13) it can nevertheless be well-approximated by a sum of Kronecker products. Kronecker products provide a computational efficient building block for approximating large matrices in numerical linear algebra application. See Beylkin and Mohlenkamp (2002); De Lathauwer et al. (2000); Tyrtyshnikov (2004) for some general theory and see (Kamm and Nagy, 1998; Nagy, 1996; Werner et al., 2008; Greenewald and Hero, 2015) for specific examples in image restoration and covariance estimation. We emphasize that if $M$ is a $p_{1} p_{2} \times q_{1} q_{2}$ matrix with $p_{1} p_{2} \gg q_{1} q_{2}$ then a rank $d$ SVD of $M$ will require computing left singular vectors of length $p_{1} p_{2}$ while its Kronecker product factorization only require computing factors of dimensions $p_{1} \times q_{1}$ and $p_{2} \times q_{2}$.

Remark 2. We note that Kronecker products factorization also featured prominently in the work of Cai et al. (2019) but their research question is susbtantially different from that considered in the
current paper. In particular our setting is that of linear regression where the goal is to estimate the factorization $\left(\beta_{1}, \beta_{2}\right)$ of $\nu$ given both the responses $\left\{Y_{i}\right\}$ and feature vectors $\left\{X_{i}\right\}$, i.e., our estimation of $\left(\beta_{1}, \beta_{2}\right)$ is a supervised learning problem. In contrast Cai et al. (2019) uses Kronecker product approximation to perform dimension reduction of the $\left\{Y_{i}\right\}$ without observing any $\left\{X_{i}\right\}$, i.e., they are considering an unsupervised learning problem.

```
Algorithm 1: KRO-PRO-FAC algorithm
    Input: \(\mathcal{Y}, \mathcal{X}\) and \(\left(p_{2}, q_{2}\right),\left(p_{1}, q_{1}\right), \bar{d}\)
    Output: \(\left(\hat{\beta}_{1 k}, \hat{\beta}_{2 k}\right)\)
    1 Compute the OLS estimate \(\tilde{\nu} \leftarrow\left[\left(\mathcal{X}^{\top} \mathcal{X}\right)^{-1} \mathcal{X}^{\top} \mathcal{Y}\right]^{\top}\).
    2 Rearrange \(\tilde{\nu}\) to get \(R(\tilde{\nu})\) by Eq. (10).
    \({ }_{3}\) Perform SVD on \(R(\tilde{\nu})\), i.e., \(R(\tilde{\nu})=\sum_{k=1}^{\bar{d}} \hat{\sigma}_{k} \hat{\mathcal{U}}_{k} \hat{\mathcal{V}}_{k}^{\top}\) with \(\hat{\sigma}_{1} \geq \hat{\sigma}_{2} \geq \cdots \geq \hat{\sigma}_{\bar{d}}\).
    4 Estimate \(d\) by \(\hat{d}=\arg \max _{j \in\{1, \ldots, \bar{d}\}} \hat{\sigma}_{j} / \hat{\sigma}_{j+1}\)
    5 Set \(\operatorname{vec}\left(\hat{\beta}_{2 k}\right)=\hat{\sigma}_{k}^{1 / 2} \hat{\mathcal{U}}_{k}\) and \(\operatorname{vec}\left(\hat{\beta}_{1 k}\right)=\hat{\sigma}_{k}^{1 / 2} \hat{\mathcal{V}}_{k}\).
    \({ }_{6}\) Output \(\hat{\nu} \leftarrow \sum_{k=1}^{\hat{d}} \hat{\beta}_{2 k} \otimes \hat{\beta}_{1 k}\)
```


## 3 Theoretical Results

We now study large-sample and/or asymptotic results for the estimates of $\left\{\beta_{1 k}, \beta_{2 k}\right\}$ obtained by the KRO-PRO-FAC algorithm. Recall that, from our earlier discussions in Section 2, the rearranged OLS estimate $R(\tilde{\nu})$ can be viewed as a sum of rank- 1 matrices $R(\nu)=\sum_{k=1}^{d} \operatorname{vec}\left(\beta_{2 k}\right) \operatorname{vec}\left(\beta_{1 k}\right)^{\top}$ additively perturbed by the noise matrix $R(\tilde{\mathcal{E}})$ where $\tilde{\mathcal{E}}=\left[\left(\mathcal{X}^{\top} \mathcal{X}\right)^{-1} \mathcal{X}^{\top} \mathcal{E}\right]^{\top}$. Therefore, if $\|\tilde{\mathcal{E}}\|$ is sufficiently small compared to $\|R(\nu)\|$, then we can apply classical results in matrix perturbation theory such as the sin- $\Theta$ theorem (Wedin, 1972) to show that the leading singular vectors of $R(\tilde{\nu})$ are "close" to the $\operatorname{vec}\left(\beta_{1 k}\right)$ and $\operatorname{vec}\left(\beta_{2 k}\right)$.

We now make the above description precise. Let $R(\nu)$ be a rank $d$ matrix for some fixed constant $d$ not depending on $p_{1}, p_{2}$ and $n$. Denote the SVD of $R(\nu)$ by $R(\nu)=\mathcal{U D}^{\top}$ where $\mathcal{D}=\operatorname{diag}\left(\sigma_{k}\right)$ is a $d \times d$ diagonal matrix of singular values, $\mathcal{V}=\left(\mathcal{V}_{1}, \ldots, \mathcal{V}_{d}\right)$ is a $p_{1} q_{1} \times d$ orthonormal matrix of right singular vectors and $\mathcal{U}=\left(\mathcal{U}_{1}, \ldots, \mathcal{U}_{d}\right)$ is a $p_{2} q_{2} \times d$ orthonormal matrix of left singular vectors. Next let $\hat{\mathcal{U}} \hat{\mathcal{D}} \hat{\mathcal{V}}^{\top}$ denote the truncated SVD corresponding to the $d$ largest
singular values and singular vectors of $R(\tilde{\nu})$. We first make an assumption on the relationship between the matrix dimensions $p_{1}, p_{2}, q_{1}, q_{2}$ and the sample size $n$ as well as the growth rate for the singular values of $R(\nu)$.

Condition 1. Let $p_{1}, p_{2}, q_{1}, q_{2}$ and $n$ satisfy

$$
\frac{q_{1}}{q_{2}}=\Theta(1), \quad \frac{p_{1}}{p_{2}}=\Theta(1), \quad q_{1} q_{2} \ll n, \quad \ln p_{i}=o(n), i=1,2 .
$$

Furthermore, for sufficiently large $p_{1}, p_{2}$, assume that the singular values of $R(\nu)$ satisfy

$$
\sigma_{k}=\mathcal{O}\left(p_{1}\right), i=1,2, \ldots, d
$$

Condition 1 implies that $R(\nu)$ have bounded condition number.
We next recall the notion of a sub-Gaussian random vector
Definition 1. Let $Z$ be a mean zero random variable. Then $Z$ is said to be sub-Gaussian with variance proxy $\sigma^{2}$ if, for all $t>0$ we have

$$
\begin{equation*}
\mathbb{P}(|Z|>t) \leq 2 \exp \left(-\frac{t^{2}}{2 \sigma^{2}}\right) \tag{16}
\end{equation*}
$$

In other words, the tail probability of $Z$ behaves similarly to that of a Gaussian distribution with variance $\sigma^{2}$. A mean zero random vector $\boldsymbol{Z} \in \Re^{p}$ is then said to be a sub-Gaussian random vector with covariance proxy $\Sigma$ if $w^{\top} \boldsymbol{Z}$ is sub-Gaussian with variance proxy $w^{\top} \Sigma w$ for all $w \in \Re^{p}$. See Section 2.5 and Section 3.4 of Vershynin (2018) for further discussion and characterizations of sub-Gaussian random vectors.

Now let $\left\{\xi_{1}, \ldots, \xi_{n}\right\}$ be iid mean zero sub-Gaussian random vectors in $\Re^{p_{1} p_{2}}$ with covariance proxy $\mathcal{I}$ where $\mathcal{I}$ is the $p_{1} p_{2} \times p_{1} p_{2}$ identity matrix. We shall assume that the noise matrices $E_{i}$ are of the form

$$
\begin{equation*}
\operatorname{vec}\left(E_{i}\right)=\Sigma_{\operatorname{vec}(E)}^{1 / 2} \xi_{i}, \quad i \in[n] \tag{17}
\end{equation*}
$$

for some $p_{1} p_{2} \times p_{1} p_{2}$ positive definite matrix $\Sigma_{\operatorname{vec}(E)}^{1 / 2}$ satisfying the following condition.
Condition 2. $\Sigma_{\mathrm{vec}(E)}$ is a block diagonal matrix, i.e., $\Sigma_{\mathrm{vec}(E)}=\operatorname{diag}\left(\boldsymbol{\Sigma}_{1}, \boldsymbol{\Sigma}_{2}, \ldots, \boldsymbol{\Sigma}_{p_{2}}\right)$ where each diagonal block is of size $p_{1} \times p_{1}$. Furthermore there exists a positive constant $\mathcal{C}$ independent of $p_{1}$, $p_{2}$ and $n$ such that

$$
\begin{equation*}
\max _{k \in\left[p_{2}\right]} \max _{s \in\left[p_{1}\right]} \boldsymbol{\Sigma}_{k}(s, s) \leq \mathcal{C} . \tag{18}
\end{equation*}
$$

where $\boldsymbol{\Sigma}_{k}(s, t)$ is the $(s, t)$ entry of $\boldsymbol{\Sigma}_{k}$.

Remark 3. We note that the block diagonal structure posited in Assumption 2 is different from and arguably more flexible than assuming a Kronecker product structure for $\Sigma_{\mathrm{vec}(E)}$. More specifically an arbitrary $\Sigma_{\mathrm{vec}(E)}$ has $O\left(p_{1}^{2} p_{2}^{2}\right)$ parameters. If $\Sigma_{\mathrm{vec}(E)}$ can be factored into the Kronecker product of a $p_{1} \times p_{1}$ matrix and a $p_{2} \times p_{2}$ matrix then the number of parameters is reduced drastically to $O\left(p_{1}^{2}+p_{2}^{2}\right)$ parameters. It was noted in Barratt (2018) that $O\left(p_{1}^{2}+p_{2}^{2}\right)$ parameters is potentially too few as it preclude the use of some common matrix-variate Gaussian distribution to model $\Sigma_{\text {vec }(E)}$. In contrast, under Assumption (2), $\Sigma_{\text {vec }(E)}$ has $O\left(p_{2} p_{1}^{2}\right)$ parameters. If $p_{1} \asymp p_{2} \asymp p$ then the above three scenarios correspond to $O\left(p^{4}\right), O\left(p^{2}\right)$ and $O\left(p^{3}\right)$ parameters, respectively. Finally we note that Assumption 2 is satisfied whenever the columns of $E_{i}$ are uncorrelated. A similar condition can be formulated for the case when the rows of $E_{i}$ are uncorrelated. These conditions are milder than assuming that the entries of $E_{i}$ are mutually independent as is done in Obozinski et al. (2011); Negahban and Wainwright (2011); Bunea et al. (2012); Bing and Wegkamp (2019); Zou et al. (2020).

With the above assumptions in place, we now state our theoretical results for bounding the estimation error between $\hat{\mathcal{U}}$ (resp. $\hat{\mathcal{V}}$ ) and $\mathcal{U}$ (resp. $\mathcal{V}$ ). These errors are stated in terms of the $\sin -\Theta$ distance between linear subspaces, i.e., given two orthonormal matrices $\mathcal{W}_{1}$ and $\mathcal{W}_{2}$ the $\sin -\Theta$ distance between the linear subspaces spanned by $\mathcal{W}_{1}$ and $\mathcal{W}_{2}$ is defined as

$$
\begin{equation*}
\left\|\sin \Theta\left(\mathcal{W}_{1}, \mathcal{W}_{2}\right)\right\|=\sqrt{1-\sigma_{\min }^{2}\left(\mathcal{W}_{1}, \mathcal{W}_{2}\right)} \tag{19}
\end{equation*}
$$

where $\sigma_{\min }\left(\mathcal{W}_{1}, \mathcal{W}_{2}\right)$ is the minimum singular value of $\mathcal{W}_{1}^{\top} \mathcal{W}_{2}$.

Theorem 1. Let $\left\{\left(X_{i}, Y_{i}\right)\right\}$ satisfy the linear model in Eq. (13) for some fixed $d \geq 1$ not depending on $n$ and suppose that Condition 1 and 2 holds. Then there exists a constant $C>0$ such that, with probability at least $1-n^{-3}$, the following holds simultaneously,

$$
\begin{array}{r}
\left|\hat{\sigma}_{k}-\sigma_{k}\right| \leq C q_{1} q_{2} \max _{k \in p_{2}} \max _{s \in\left[p_{1}\right]} \boldsymbol{\Sigma}_{k}(s, s) \frac{\sqrt{p_{1}}+\sqrt{p_{2}}+\ln p_{1}+\ln p_{2}}{\sqrt{n}}, \\
\max \{\|\sin \Theta(\mathcal{U}, \hat{\mathcal{U}})\|,\|\sin \Theta(\mathcal{V}, \hat{\mathcal{V}})\|\} \leq C q_{1} q_{2} \max _{k \in p_{2}} \max _{s \in\left[p_{1}\right]} \boldsymbol{\Sigma}_{k}(s, s) \frac{\sqrt{p_{1}}+\sqrt{p_{2}}+\ln p_{1}+\ln p_{2}}{\sqrt{n p_{1} p_{2}}} \tag{21}
\end{array}
$$

Theorem 1 implies the following upper bound for the error of $\hat{\nu}=\sum_{k=1}^{d} \hat{\beta}_{2 k} \otimes \hat{\beta}_{1 k}$ as an estimate for $\nu=\sum_{k=1}^{d} \beta_{2 k} \otimes \beta_{1 k}$. In particular the relative error of $\hat{\nu}-\nu$ converges to 0 as $p_{1}, p_{2}$ and $n$ diverge and thus $\hat{\nu}$ is a consistent estimate for $\nu$.

Corollary 1. Suppose $p_{1}=p_{2}=p$ and consider the setting in Theorem 1. Then there exists a constant $C>0$ such that with probability at least $1-n^{-3}$,

$$
\begin{equation*}
\frac{\|\hat{\nu}-\nu\|_{F}}{\|\nu\|_{F}} \leq \frac{C}{\sqrt{n p}} \tag{22}
\end{equation*}
$$

## 4 Numerical experiments

We evaluate the numerical performance of the KRO-PRO-FAC algorithm through a few simulation studies and real data analysis.

### 4.1 Simulation studies

For the simulation experiments we set the dimensions of $Y_{i}$ and $X_{i}$ to be $p_{1}=p_{2}=500$ and $q_{1}=q_{2}=2$ while the sample size $n$ is chosen in $\{200,400,1000,2000,3000\}$. For ease of exposition we only consider the special case of Eq. (13) with $d=1$, and thus $\nu=\beta_{2} \otimes \beta_{1}$ where $\beta_{1}$ and $\beta_{2}$ are $500 \times 2$ matrices. We first generate $\operatorname{vec}\left(\beta_{1}\right)$ from the standard multivariate normal distribution on $\Re^{1000}$ and similarly for $\operatorname{vec}\left(\beta_{2}\right)$; note that neither $\beta_{1}$ nor $\beta_{2}$ are expected to be sparse and furthermore the estimation of these $\beta_{k}$ when $n=200$ or $n=400$ falls within the setting of regression with high-dimensional responses. We then generate $X_{1}, X_{2}, \ldots, X_{n}$ where the vec $\left(X_{i}\right)$ are iid standard multivariate normals in $\Re^{4}$.

Given the $\left\{X_{i}\right\}$ we then consider the following 4 different models for the random noises $\left\{E_{i}\right\}$. The first three models corresponds to $\operatorname{vec}\left(E_{i}\right) \in \Re^{25000}$ that are multivariate normals while the last model corresponds to $\operatorname{vec}\left(E_{i}\right)$ with entries independently sampled from Student's $t$ distribution with 5 degrees of freedom. The entries of $E_{i}$ for Model 4 have heavier tails compared to that for Models 1-3.

Model 1: Identity covariance: $\Sigma_{\mathrm{vec}(E)}=\mathbf{I}_{p_{1} p_{2} \times p_{1} p_{2}}$ and vec $\left(E_{i}\right)$ 's are generated independently from the standard multivariate normal distribution.

Model 2: Banded covariance: $\Sigma_{\text {vec }(E)}=\mathbf{L L}^{\top}$ where $\mathbf{L}$ is a lower triangular banded matrix in $\Re^{p_{1} p_{2} \times p_{2} p_{2}}$ with $\mathbf{L}_{i j}=0$ for $i<j$ or $i-j>b$. The bandwidth $b$ is set to 5 and the diagonal elements are generated from $\mathcal{N}(3,1)$ and the non-zero off-diagonal elements are generated from $\mathcal{N}(0,1)$. $\mathbf{L}$ is fixed over the 100 replications.

Model 3: $A R(1): \Sigma_{\operatorname{vec}(E)}=\left(\rho^{|i-j|}\right)_{p_{1} p_{2} \times p_{1} p_{2}}$ with $\rho=0.9$. Here we generate vec $\left(E_{1}\right), \ldots, \operatorname{vec}\left(E_{n}\right)$ based on Matlab codes from arima.

Model 4: Heavy-tailed: $\Sigma_{\mathrm{vec}(E)}$ is proportional to $\mathbf{I}_{p_{1} p_{2} \times p_{1} p_{2}}$ and the entries of vec $\left(E_{i}\right)$ 's are random samples from the Student's t-distribution with 5 degrees of freedom.

For each choice of the noise model for $E_{i}$ we then generate $\left\{Y_{i}\right\}$ according to Eq. (3) and then estimate $\hat{\nu}$ based on the $\left\{X_{i}, Y_{i}\right\}$ using the KRO-PRO-FAC algorithm. For illustrative comparisons we considered, in addition to the default described in Algorithm 1, two other variants which performs rank regularization of either the responses or the OLS estimate. More specifically the first variant uses, instead of the observed $Y_{i}$, its truncated rank- $\alpha \operatorname{SVD} Y_{i}^{(\alpha)}$ for estimating $\nu$. We termed this variant as KRO-PRO-FAC $(\alpha)$ and note that it is motivated by the fact that while $Y_{i}$ is, with probability 1 full rank, $\mathbb{E}\left[Y_{i}\right]=\beta_{1} X_{i} \beta_{2}^{\top}$ is low-rank for all $i$ and thus a rank-regularized version of the $\left\{Y_{i}\right\}$ might lead to better estimate of $\nu$. The second variant also performs rank regularization, but on the OLS estimate $\tilde{\nu}$ as opposed to the responses $\left\{Y_{i}\right\}$. Letting $\tilde{\nu}^{(\gamma)}$ be the truncated rank $-\gamma$ SVD of $\tilde{\nu}$ we then perform the remaining steps of Algorithm 1 with $\tilde{\nu}^{(\gamma)}$ in place of $\tilde{\nu}$. We termed this variant as rdu-rank-KRO $(\gamma)$ and note that it is motivated by the notion of reduced-rank-regression in Izenman (1975). For this simulation we chose $\alpha=\gamma=2$.

Finally we also estimate $\nu$ using the MLE based procedure described in Ding and Cook (2016). Recall that this MLE based approach posits both a Kronecer product structure for both the regression coefficient $\nu$ and the covariance matrix of $\operatorname{vec}\left(E_{i}\right)$. We use the implementation from is based on R codes from MatrixEnv and denote the resulting estimates as dual-KRO-MLE. Table 1 summarizes some key differences between the 4 methods described above. For numerical comparisons we evaluate the relative errors $\|\hat{\nu}-\nu\|_{F} /\|\nu\|_{F}$ for each methods and averaged these over 100 independent Monte Carlo replicates. The results are presented in Table 2 through Table 5 for the four noise models described above.

For Model 1 we see from Table 2 that both the KRO-PRO-FAC and KRO-PRO-FAC $(\alpha)$ method have the smallest estimation error. The dual-KRO-MLE estimate is substantially less accurate compared to that of KRO-PRO-FAC and KRO-PRO-FAC ( $\alpha$ ) especially when the sample size is small, e.g., $n=200$ or $n=400$. This is in a sense expected as the entries of $E_{i}$ are iid and thus there are few if any benefits in estimating and/or incorporating the covariance structure of $\left\{Y_{i}\right\}$. Finally, the rdu-rank-KRO $(\gamma)$ method has the highest estimation error and this observation

Table 1: Method Comparison

| method | data | $\nu$ estimation | Kronecker structure on |
| :--- | :---: | :---: | :---: |
| KRO-PRO-FAC | $Y_{i}$ | OLS | mean |
| KRO-PRO-FAC $(\alpha)$ | rank- $Y_{i}^{(\alpha)}$ | OLS | mean |
| rdu-rank-KRO $(\gamma)$ | $Y_{i}$ | rank- $\gamma$ OLS $\tilde{\nu}^{(\gamma)}$ | mean |
| dual-KRO-MLE | $Y_{i}$ | column \& row separate estimates | mean \& covariance |

Table 2: Average relative estimation error (\%) under the identity covariance (Model 1)

| $\\|\hat{\nu}-\nu\\|_{F} /\\|\nu\\|_{F}$ | sample size (n) |  |  |  |  |
| :--- | ---: | ---: | ---: | ---: | ---: |
|  | 200 | 400 | 1000 | 2000 | 3000 |
| KRO-PRO-FAC | 0.339 | 0.237 | 0.151 | 0.106 | 0.087 |
| KRO-PRO-FAC $(\alpha=2)$ | 0.339 | 0.238 | 0.151 | 0.106 | 0.087 |
| rdu-rank-KRO $(\gamma=2)$ | 63.998 | 63.998 | 63.997 | 63.997 | 63.997 |
| dual-KRO-MLE | 76.938 | 52.702 | 20.721 | 8.377 | 6.563 |

also extends to the results for Model 2 through 4 as presented in Table 3 through Table 5 This is once again expected as, recalling the earlier discussions in Example 2.2, the Kronecker structure in the regression coefficient $\nu$ is fundamentally different from assuming $\nu$ to be low-rank. In other words imposing rank constraints on $\tilde{\nu}$ only leads to information loss due to model misspecification.

For Model 2 we see from Table 3 that the KRO-PRO-FAC algorithm has the smallest estimation error with the KRO-PRO-FAC $(\alpha)$ variant being slightly worse. The estimate obtained from the

Table 3: Average relative estimation error (\%) under the banded covariance with a bandwidth $b=5($ Model 2$)$

| $\\|\hat{\nu}-\nu\\|_{F} /\\|\nu\\|_{F}$ | sample size (n) |  |  |  |  |  |
| :--- | ---: | ---: | ---: | ---: | ---: | :---: |
|  | 200 | 400 | 1000 | 2000 | 3000 |  |
| KRO-PRO-FAC | 1.508 | 1.059 | 0.666 | 0.472 | 0.385 |  |
| KRO-PRO-FAC $(\alpha=2)$ | 1.552 | 1.116 | 0.746 | 0.576 | 0.506 |  |
| rdu-rank-KRO $(\gamma=2)$ | 63.963 | 63.997 | 63.999 | 63.998 | 63.998 |  |
| dual-KRO-MLE | 38.934 | 19.116 | 7.211 | 2.292 | 2.418 |  |

Table 4: Average relative estimation error (\%) under the $\operatorname{AR}(1)$ setting with $\rho=0.9$ (Model 3)

| $\\|\hat{\nu}-\nu\\|_{F} /\\|\nu\\|_{F}$ | sample size (n) |  |  |  |  |
| :--- | ---: | ---: | ---: | ---: | ---: |
|  | 200 | 400 | 1000 | 2000 | 3000 |
| KRO-PRO-FAC | 0.351 | 0.247 | 0.157 | 0.110 | 0.090 |
| KRO-PRO-FAC $(\alpha=2)$ | 0.512 | 0.440 | 0.391 | 0.373 | 0.371 |
| rdu-rank-KRO $(\gamma=2)$ | 63.998 | 63.998 | 63.997 | 63.997 | 63.997 |
| dual-KRO-MLE | 0.250 | 0.177 | 0.113 | 0.079 | 0.064 |

Table 5: Average relative estimation error (\%) under the Student's t-distribution with 5 degrees of freedom (Model 4)

| $-\nu\left\\|_{F} /\right\\| \nu \\|_{F}$ | sample size (n) |  |  |  |  |
| :--- | ---: | ---: | ---: | ---: | ---: |
|  | 200 | 400 | 1000 | 2000 | 3000 |
| KRO-PRO-FAC | 0.439 | 0.308 | 0.194 | 0.137 | 0.112 |
| KRO-PRO-FAC $(\alpha=2)$ | 0.440 | 0.308 | 0.195 | 0.137 | 0.113 |
| rdu-rank-KRO $(\gamma=2)$ | 63.998 | 63.998 | 63.997 | 63.997 | 63.997 |
| dual-KRO-MLE | 69.049 | 39.734 | 12.677 | 7.141 | 5.010 |

dual-KRO-MLE algorithm is noticably worse compared to both the KRO-PRO-FAC and KRO-PRO-FAC $(\alpha)$ and furthermore appeared to be sensitive to the sample size $n$, i.e., its estimation error is much larger than its competitors when $n=200$ or $n=400$. We note that for this Model $2,90 \%$ of the non-zero correlations in $\Sigma_{\mathrm{vec}(E)}$ have absolute value less than 0.5 , which suggests either weak or mild dependence among rows and columns in $\operatorname{vec}\left(Y_{i}\right)$.

For Model 3 we see from Table 4 that the dual-KRO-MLE algorithm yields the most accurate estimates with errors that are slightly smaller than that of KRO-PRO-FAC and KRO-PRO-FAC $(\alpha)$ methods. There is thus value in joint modeling of the mean $\nu$ and the covariance structure for the $\left\{Y_{i}\right\}$. Note, however, that the KRO-PRO-FAC algorithm is much less computationally demanding compared to dual-KRO-MLE.

Finally, for Model 4 we see from Table 5 that the KRO-PRO-FAC algorithm outperforms all of its competitors. In particular it is slightly better than $\operatorname{KRO}-\operatorname{PRO}-\operatorname{FAC}(\alpha)$ and is much better than dual-KRO-MLE. These results are similar to that in Table 2 and one possible explanation for this similarity is that both models induce the same covariance structure for $\left\{Y_{i}\right\}$.

### 4.2 Real data analysis

We now apply the KRO-PRO-FAC algorithm to the electroencephacology (EEG) dataset from the UC Irvine Machine Learning Repository. The data arises from a study of EEG measurements related to alcoholoism in which there are 122 subjects from either the alcoholic group ( 77 subjects) or the control group ( 45 subjects). For each subject a series of voltage measurements is made at 256 different time points from 64 different regions of the scalp, i.e., the EEG response for the $i$ th subject in the $j$ th group (with $j=1$ and $j=2$ denoting the alcoholic and control) can be viewed as a matrix $Y_{i j}$ with 256 rows and 64 columns. A key research question for this dataset is to identify which of the 64 brain channel accounts for most of the differences in voltages measurements between the two groups.

To answer the above inquiry we partition the data according to the subject grouping and fit a bi-linear model of the form Eq (13) to each group. As the EEG dataset contains no other covariates, this lead to a model of the form

$$
\begin{equation*}
\operatorname{vec}\left(Y_{i j}\right)=\left(\sum_{k=1}^{d^{(j)}} \beta_{2 k}^{(j)} \otimes \beta_{1 k}^{(j)}\right)+\operatorname{vec}\left(E_{i j}\right), \quad i \in\left[n_{j}\right], j=1,2 \tag{23}
\end{equation*}
$$

where $\beta_{2 k}^{(j)} \in \Re^{64 \times 1}$ and $\beta_{1 k}^{(j)} \in \Re^{256 \times 1}$. In other words, the mean response $\nu^{(j)}=\mathbb{E}\left[\operatorname{vec}\left(Y_{i j}\right)\right]$ for the $j$ th group is a sum of $d^{(j)}$ Kronecker products and thus $\nu^{(1)}-\nu^{(2)}$ is the effect of alcoholism (when compared to the control group) on the voltage measurements. We emphasize that the number of Kronecker factors $d^{(j)}$ are possibly different between the two groups. We apply the KRO-PRO-FAC algorithm to these $\left\{Y_{i j}\right\}$ with $d^{(1)}=2$ and $d^{(2)}=3$ chosen via the singular value ratio criterion as described in Eq. (15). Let $\hat{\nu}^{(1)}=\sum_{k=1}^{2} \hat{\beta}_{2 k}^{(1)} \otimes \hat{\beta}_{1 k}^{(1)}$ and $\hat{\nu}^{(2)}=\sum_{k=1}^{2} \hat{\beta}_{2 k}^{(2)} \otimes \hat{\beta}_{1 k}^{(2)}$ be the resulting estimates of $\nu^{(1)}$ and $\nu^{(2)}$.

Given these $\hat{\nu}^{(1)}$ and $\hat{\nu}^{(2)}$ we then follow the same post-processing steps described in Ding and Cook (2016) for multiple testing among the brain locations. Firsly, we isolate the alcoholism effects of each channel by averaging out the time effects $\hat{\nu}^{(1)}-\hat{\nu}^{(2)}$, ie., we take the column means of the $\operatorname{vec}^{-1}\left(\hat{\nu}^{(1)}-\hat{\nu}^{(2)}, 256,64\right)$ where $\operatorname{vec}^{-1}(\cdot, 256,64)$ yields a matrix of dimensions $256 \times 64$. This yields in a vector $\hat{\theta} \in \Re^{64}$ which we then conduct multiple t-test for the null hypothesis that $\mathbb{H}_{0}: \theta_{i}=0$ and compute the resulting p-values. Finally we apply the Benjamini-Yekutieli procedure (Benjamini and Yekutieli, 2001) to adjust these $p$-values.

The left panel of Figure 2 reports these adjusted p-values (on a $\log _{10}$ scale). For comparisons we also repeat the same post-processing analysis but replaced the estimates $\hat{\nu}^{(1)}$ and $\hat{\nu}^{(2)}$ by the the OLS estimate $\bar{Y}^{(1)}=n_{1}^{-1} \sum_{i \in n_{1}} Y_{i 1}$ and $\bar{Y}^{(2)}=n_{2}^{-1} \sum_{i \in n_{2}} Y_{i 2}$ and present the adjusted $p$-values for these OLS estimates in the right panel of Figure 2. Figure 2 indicates that, for a significant level of 0.05 , the KRO-PRO-FAC estimates lead to the detectation of 20 possibly relevant channels while the OLS estimates detect only 3 possibly relevant channels. We note that Li and Zhang (2017); Ding and Cook (2016) also analyzed the same data set and their estimates detect 24 and 26 possibly relevant channels, respectively. Our detections using the KRO-PRO-FAC estimates are thus comparable with those from Li and Zhang (2017); Ding and Cook (2016); indeed they all detected the regions from 21 to 25 , from 44 to 52 and from 57 and 62 . The main benefit of using the KRO-PRO-FAC estimates is that they can be commputed efficiently and do not depend on knowing or estimating $\operatorname{Cov}\left[\operatorname{vec}\left(Y_{i j}\right)\right]$.

## 5 Conclusion

In this paper we consider matrix regression $Y_{i}=\sum_{k} \beta_{1 k} X_{i} \beta_{2 k}^{\top}+E_{i}$ where the responses $Y_{i}$ are highdimensional matrices and propose a computationally efficient procedure for estimating $\left\{\beta_{1 k}, \beta_{2 k}\right\}$


Figure 2: Benjamini-Yekutieli adjusted p-values (on a scale of $-\log _{10}$ ) for 64 brain channels obtained from the (a) Kronecker products estimates in Eq. (23) with $d=2$ and $d=3$ for the alcoholic and control group (b) OLS estimates
based on the nearest Kronecker products approximation to the OLS estimate $\hat{\nu}$ of $\nu=\sum_{k} \beta_{2 k} \otimes$ $\beta_{1 k}$. We now mention three potential directions for future research. The empirical results in Section 4.1 show that the KRO-PRO-FAC procedure has smallest estimation error when the noise entries for $E_{i}$ are independent and is slightly worse than the dual-KRO-MLE procedure of Ding and Cook (2016) when the noise of $E_{i}$ are highly correlated. As the dual-KRO-MLE is somewhat computationally demanding, it will be valuable to refine our KRO-PRO-FAC procedure for handling highly dependent rows and columns without compromising its computation efficiency. Secondly, the performance of the low-rank variant KRO-PRO-FAC $(\alpha)$ is also quite competitive but its theoretical property is currently unaddressed. Finally, for many type of matrix data such as those arising in image analysis, the ordering of the rows and columns for these matrices are based on latent but important features. For example, pixels' intensities in an image usually exhibit some continuity in both vertical and horizontal directions. How to meaningfully extract these latent features and incorporate them into the matrix regression problem is an open and interesting research question.

## References

Ahn, S. C. and Horenstein, A. R. (2013) Eigenvalue ratio test for the number of factors. Econometrica, 81, 1203-1227.

Barratt, S. (2018) A matrix Gaussian distribution. arXiv preprint arXiv:1804.11010.
Benjamini, Y. and Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. Annals of Statistics, 1165-1188.

Beylkin, G. and Mohlenkamp, M. J. (2002) Numerical operator calculus in higher dimensions. Proceedings of the National Academy of Sciences, 99, 10246-10251.

Bhatia, R. (2013) Matrix analysis. Springer.

Bing, X. and Wegkamp, M. H. (2019) Adaptive estimation of the rank of the coefficient matrix in high-dimensional multivariate response regression models. The Annals of Statistics, 47, 31573184.

Bunea, F., She, Y. and Wegkamp, M. H. (2012) Joint variable and rank selection for parsimonious estimation of high-dimensional matrices. The Annals of Statistics, 40, 2359-2388.

Cai, C., Chen, R. and Xiao, H. (2019) Kopa: Automated Kronecker product approximation. arXiv preprint arXiv:1912.02392.

Chen, E. Y. and Fan, J. (2021) Statistical inference for high-dimensional matrix-variate factor models. Journal of the American Statistical Association, 1-18.

Chen, E. Y., Tsay, R. S. and Chen, R. (2019) Constrained factor models for high-dimensional matrix-variate time series. Journal of the American Statistical Association.

Chen, K., Chan, K.-S. and Stenseth, N. C. (2012) Reduced rank stochastic regression with a sparse singular value decomposition. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 74, 203-221.

Crainiceanu, C. M., Caffo, B. S., Luo, S., Zipunnikov, V. M. and Punjabi, N. M. (2011) Population value decomposition, a framework for the analysis of image populations. Journal of the American Statistical Association, 106, 775-790.

De Lathauwer, L., De Moor, B. and Vandewalle, J. (2000) A multilinear singular value decomposition. SIAM Journal on Matrix Analysis and Applications, 21, 1253-1278.

De Waal, D. (1985) Matrix-valued distributions. In Encyclopedia of statistical sciences, 326-333. Wiley Online Library.

Ding, S. and Cook, R. (2016) Matrix-variate regressions and envelope models. Journal of the Royal Statistical Society: Series B, $\mathbf{8 0}$.

Dutilleul, P. (1999) The MLE algorithm for the matrix normal distribution. Journal of Statistical Computation and Simulation, 64, 105-123.

Eckart, C. and Young, G. (1936) The approximation of one matrix by another of lower rank. Psychometrika, 1, 211-218.

Feng, Y., Xiao, L. and Chi, E. C. (2021) Sparse single index models for multivariate responses. Journal of Computational and Graphical Statistics, 30, 115-124.

Greenewald, K. and Hero, A. O. (2015) Robust Kronecker product PCA for spatio-temporal covariance estimation. IEEE Transactions on Signal Processing, 63, 6368-6378.

Gupta, A. K. and Nagar, D. K. (1999) Matrix variate distributions. Chapman and Hall/CRC.
Halko, N., Martinsson, P.-G. and Tropp, J. A. (2011) Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. SIAM review, 53, 217-288.

Izenman, A. J. (1975) Reduced-rank regression for the multivariate linear model. Journal of Multivariate Analysis, 5, 248-264.

Kamm, J. and Nagy, J. G. (1998) Kronecker product and SVD approximations in image restoration. Linear Algebra and its Applications, 284, 177-192.

Kong, D., An, B., Zhang, J. and Zhu, H. (2019) L2rm: Low-rank linear regression models for high-dimensional matrix responses. Journal of the American Statistical Association.

Lam, C. and Yao, Q. (2012) Factor modeling for high-dimensional time series: inference for the number of factors. Annals of Statistics, 694-726.

Li, L. and Zhang, X. (2017) Parsimonious tensor response regression. Journal of the American Statistical Association, 112, 1131-1146.

Musco, C. and Musco, C. (2015) Randomized block krylov methods for stronger and faster approximate singular value decomposition. Advances in neural information processing systems, 28.

Nagy, J. G. (1996) Decomposition of block Toeplitz matrices into a sum of Kronecker products with applications in image restoration. Tech. rep., Southern Methodist University.

Negahban, S. and Wainwright, M. J. (2011) Estimation of (near) low-rank matrices with noise and high-dimensional scaling. The Annals of Statistics, 39, 1069-1097.

Obozinski, G., Wainwright, M. J. and Jordan, M. I. (2011) Support union recovery in highdimensional multivariate regression. The Annals of Statistics, 39, 1-47.

Tropp, J. A., Yurtsever, A., Udell, M. and Cevher, V. (2017) Practical sketching algorithms for low-rank matrix approximation. SIAM Journal on Matrix Analysis and Applications, 38, 1454-1485.

Tyrtyshnikov, E. (2004) Kronecker-product approximations for some function-related matrices. Linear Algebra and its Applications, 379, 423-437.

Van Loan, C. F. and Pitsianis, N. (1993) Approximation with Kronecker products. In Linear algebra for large scale and real-time applications, 293-314. Springer.

Vershynin, R. (2018) High-Dimensional Probability: An Introduction with Applications in Data Science. Cambridge University Press.

Vølund, A. (1980) Multivariate bioassay. Biometrics, 225-236.
Wang, D., Liu, X. and Chen, R. (2019) Factor models for matrix-valued high-dimensional time series. Journal of Econometrics, 208, 231-248.

Wang, D., Shen, H. and Truong, Y. (2016) Efficient dimension reduction for high-dimensional matrix-valued data. Neurocomputing, 190, 25-34.

Wang, D., Zheng, Y., Lian, H. and Li, G. (2021) High-dimensional vector autoregressive time series modeling via tensor decomposition. Journal of the American Statistical Association, 1-19.

Wedin, P.- $\AA$. (1972) Perturbation bounds in connection with singular value decomposition. BIT Numerical Mathematics, 12, 99-111.

Werner, K., Jansson, M. and Stoica, P. (2008) On estimation of covariance matrices with kronecker product structure. IEEE Transactions on Signal Processing, 56, 478-491.

Ye, J. (2005) Generalized low rank approximations of matrices. Machine Learning, 61, 167-191.

Yuan, M., Ekici, A., Lu, Z. and Monteiro, R. (2007) Dimension reduction and coefficient estimation in multivariate linear regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69, 329-346.

Zhang, D. (2005) (2d)^ 2 pca: Two-directional two-dimensional pca for efficient face representation and recognition. Neurocomputing, 69, 224-231.

Zhao, J. and Leng, C. (2014) Structured Lasso for regression with matrix covariates. Statistica Sinica, 799-814.

Zheng, Z., Bahadori, M. T., Liu, Y. and Lv, J. (2019) Scalable interpretable multi-response regression via seed. J. Mach. Learn. Res., 20, 1-34.

Zou, C., Ke, Y. and Zhang, W. (2020) Estimation of low rank high-dimensional multivariate linear models for multi-response data. Journal of the American Statistical Association, 1-11.

## Appendix A: Proofs of Stated Results

## A. 1 Proof of Theorem 1

For convenience of notation, we take

$$
\begin{align*}
& \underbrace{\left(\mathcal{X}^{\top} \mathcal{X}\right)^{-1} \mathcal{X}^{\top}}_{q_{1} q_{2} \times n}=\left[\mathcal{C}_{1}, \mathcal{C}_{2}, \ldots, \mathcal{C}_{q_{1} q_{2}}\right]^{\top}, \mathcal{C}_{1}, \mathcal{C}_{2}, \ldots, \mathcal{C}_{q_{1} q_{2}} \in \Re^{n}  \tag{24}\\
& \underbrace{\mathcal{E}}_{n \times p_{1} p_{2}}=\left[\begin{array}{c}
\operatorname{vec}\left(E_{1}\right)^{\top} \\
\vdots \\
\operatorname{vec}\left(E_{n}\right)^{\top}
\end{array}\right]=\left[\mathcal{D}_{1}, \mathcal{D}_{2}, \ldots, \mathcal{D}_{p_{1} p_{2}}\right]=\left[\mathcal{F}_{1}, \mathcal{F}_{2}, \ldots, \mathcal{F}_{p_{1} p_{2}}\right] \Sigma_{\operatorname{vec}(E)}^{1 / 2} \tag{25}
\end{align*}
$$

where $\mathcal{D}_{1}, \mathcal{D}_{2} \cdots \mathcal{D}_{p_{1} p_{2}} \in \Re^{n}$ and $\mathcal{F}_{1}, \mathcal{F}_{2} \cdots \mathcal{F}_{p_{1} p_{2}} \in \Re^{n}$ with $\mathcal{F}_{s}=\left(\xi_{1 s}, \ldots, \xi_{n s}\right)^{\top}$ being independent with independent entries.

Set $\left.\Delta=\left[\left(\mathcal{X}^{\top} \mathcal{X}\right)^{-1} \mathcal{X}^{\top} \mathcal{E}\right)\right]^{\top}$ and recall that $R(\nu)=\sum_{k=1}^{d} \operatorname{vec}\left(\beta_{2 k}\right) \operatorname{vec}\left(\beta_{1 k}\right)^{\top}=\sum_{k=1}^{d} \sigma_{k} \mathcal{U}_{k} \mathcal{V}_{k}^{\top}$. We then have

$$
R(\hat{\nu})=\sum_{k=1}^{d} \sigma_{k} \mathcal{U}_{k} \mathcal{V}_{k}^{\top}+R(\Delta)
$$

By Weyl's inequality (Problem III.6.13 in Bhatia (2013)) and Wedin sin- $\Theta$ (Wedin, 1972) theorem we have

$$
\begin{align*}
\left|\hat{\sigma}_{k}-\sigma_{k}\right| & \leq\|R(\Delta)\|  \tag{26}\\
\left\|\sin \Theta\left(\mathcal{U}_{1}, \hat{\mathcal{U}}_{1}\right)\right\|_{2},\left\|\sin \Theta\left(\mathcal{V}_{1}, \hat{\mathcal{V}}_{1}\right)\right\|_{2} & \leq \frac{\min \left\{\left\|\hat{\mathcal{U}}_{1}^{\top} R(\Delta)\right\|_{2},\left\|R(\Delta) \hat{\mathcal{V}}_{1}\right\|_{2}\right\}}{\hat{\sigma}_{d}} \leq \frac{\|R(\Delta)\|_{2}}{\hat{\sigma}_{d}} \tag{27}
\end{align*}
$$

It thus suffices to bound the spectral norm of $R(\Delta)$. First note that $R(\Delta)$ can be written as a block matrix of the form

$$
R(\Delta)=\left[\begin{array}{cccc}
\tilde{\Delta}_{11} & \tilde{\Delta}_{12} & \ldots & \tilde{\Delta}_{1 q_{1}} \\
\tilde{\Delta}_{21} & \tilde{\Delta}_{22} & \ldots & \tilde{\Delta}_{2 q_{1}} \\
\vdots & \vdots & \ddots & \vdots \\
\tilde{\Delta}_{q_{2} 1} & \tilde{\Delta}_{q_{2} 2} & \ldots & \tilde{\Delta}_{q_{2} q_{1}}
\end{array}\right], \quad \tilde{\Delta}_{k \ell}=\left[\begin{array}{ccc}
\mathcal{D}_{1}^{\top} \mathcal{C}_{(k-1) q_{1}+\ell} & \ldots & \mathcal{D}_{p_{1}}^{\top} \mathcal{C}_{(k-1) q_{1}+\ell} \\
\mathcal{D}_{p_{1}+1}^{\top} \mathcal{C}_{(k-1) q_{1}+\ell} & \ldots & \mathcal{D}_{2 p_{1}}^{\top} \mathcal{C}_{(k-1) q_{1}+\ell} \\
\vdots & \ddots & \vdots \\
\mathcal{D}_{\left(p_{2}-1\right) p_{1}+1}^{\top} \mathcal{C}_{(k-1) q_{2}+\ell} & \ldots & \mathcal{D}_{p_{1} p_{2}}^{\top} \mathcal{C}_{(k-1) q_{1}+\ell}
\end{array}\right]
$$

The matrix $\tilde{\Delta}_{k \ell}$ can be further expressed as

$$
\tilde{\Delta}_{k \ell}=\left[\begin{array}{llll}
\boldsymbol{\Sigma}_{1}^{1 / 2} \zeta_{1} & \boldsymbol{\Sigma}_{2}^{1 / 2} \zeta_{2} & \ldots & \boldsymbol{\Sigma}_{p_{2}}^{1 / 2} \zeta_{p_{1}} \tag{28}
\end{array}\right], \quad \zeta_{s}=\left[\mathcal{F}_{s}, \mathcal{F}_{p_{1}+s}, \mathcal{F}_{2 p_{1}+s}, \ldots, \mathcal{F}_{\left(p_{2}-1\right) p_{1}+s}\right]^{\top} \mathcal{C}_{(k-1) q_{1}+\ell}
$$

Note that for ease of exposition we had suppressed the dependency on $k$ and $\ell$ in the notation for $\zeta_{s}$. This should cause minimal confusion as we can fix some arbitrary $k$ and $\ell$ before proceeding with the subsequent derivations.

We now derive a concentration inequality for $\left\|\tilde{\Delta}_{k \ell}\right\|$ using a standard $\epsilon$-net argument.
Step 1: $\epsilon$ net Let $\epsilon=1 / 4$ and choose an $\epsilon$ net $\mathcal{M}$ of the sphere $\mathcal{S}^{p_{2}-1}$ and an $\epsilon$ net $\mathcal{R}$ of the sphere $\mathcal{S}^{p_{1}-1}$. We have

$$
|\mathcal{M}| \leq 9^{p_{2}}, \quad|\mathcal{R}| \leq 9^{p_{1}}
$$

The spectral norm of $\tilde{\Delta}_{k \ell}$ can then be bounded as

$$
\left\|\tilde{\Delta}_{k \ell}\right\|_{2} \leq 2 \max _{x \in \mathcal{M}, y \in \mathcal{R}}\left\langle\tilde{\Delta}_{k \ell} x, y\right\rangle
$$

Step 2: Concentration Fix $x \in \mathcal{M}$ and $y \in \mathcal{R}$. We then have

$$
\left\langle\tilde{\Delta}_{k \ell} x, y\right\rangle=\sum_{i=1}^{p_{2}} x_{i} \zeta_{i}^{\top} \Sigma_{i}^{1 / 2} y=\sum_{i=1}^{p_{2}} x_{i}\left[\sum_{s=1}^{p_{1}} \sum_{k=1}^{p_{1}} \boldsymbol{\Sigma}_{i}^{1 / 2}(k, s) \zeta_{i k} y_{s}\right]
$$

Using properties of the Orlicz $\Psi_{2}$-norm (see e.g., Proposition 2.6.1 in Vershynin (2018)) we have

$$
\begin{aligned}
\left\|\left\langle\tilde{\Delta}_{k \ell} x, y\right\rangle\right\|_{\psi_{2}}^{2} & \leq \mathcal{C} \sum_{i=1}^{p_{2}} x_{i}^{2}\left\|\sum_{s=1}^{p_{1}} \sum_{k=1}^{p_{1}} \Sigma_{i}^{1 / 2}(k, s) \zeta_{i k} y_{s}\right\|_{\psi_{2}}^{2} \\
& \leq \mathcal{C} \sum_{i=1}^{p_{2}} x_{i}^{2} \sum_{s=1}^{p_{1}} y_{s}^{2} \sum_{k=1}^{p_{1}}\left\|\boldsymbol{\Sigma}_{i}^{1 / 2}(k, s) \zeta_{i k}\right\|_{\psi_{2}}^{2} \\
& \leq \mathcal{C} \max _{i, k}\left\|\zeta_{i k}\right\|_{\psi_{2}}^{2} \sum_{i=1}^{p_{2}} x_{i}^{2} \sum_{s=1}^{p_{1}} y_{s}^{2} \sum_{k=1}^{p_{1}}\left|\Sigma_{i}^{1 / 2}(k, s)\right|^{2} \\
& \leq \mathcal{C} \max _{i, k}\left\|\zeta_{i k}\right\|_{\psi_{2}}^{2} \max _{i \in\left[p_{2}\right]} \max _{s \in\left[p_{1}\right]} \boldsymbol{\Sigma}_{i}(s, s) \\
& \leq \mathcal{C}\left\|\left(\mathcal{X}^{\top} \mathcal{X}\right)^{-1} \mathcal{X}\right\|_{2} \max _{i \in[m]} \max _{s \in[r]} \boldsymbol{\Sigma}_{i}(s, s)
\end{aligned}
$$

where the second to last inequality is because

$$
\sum_{k=1}^{p_{1}}\left|\boldsymbol{\Sigma}_{i}^{1 / 2}(k, s)\right|^{2}=\sum_{k=1}^{p_{1}} \boldsymbol{\Sigma}_{i}^{1 / 2}(s, k) \boldsymbol{\Sigma}_{i}^{1 / 2}(k, s)=\boldsymbol{\Sigma}_{i}(s, s)
$$

Let $K=\left\|\left(\mathcal{X}^{\top} \mathcal{X}\right)^{-1} \mathcal{X}\right\| \times \max _{i \in[m]} \max _{s \in[r]} \boldsymbol{\Sigma}_{i}(s, s)$. We therefore have, for all $u \geq 0$, that

$$
\begin{equation*}
P\left(\left\langle\tilde{\Delta}_{k \ell} x, y\right\rangle \geq u\right) \leq 2 \exp \left(-c u^{2} / K^{2}\right) \tag{29}
\end{equation*}
$$

Step 3: Union bound By union over the $\mathcal{M}$ and $\mathcal{R}$, then with probability $1-2 \exp \left(-u^{2}\right)$, we have for any $u>0$

$$
\begin{equation*}
\left\|\tilde{\Delta}_{k \ell}\right\| \leq \mathcal{C}\left\|\left(\mathcal{X}^{\top} \mathcal{X}\right)^{-1} \mathcal{X}\right\|_{2} \max _{i \in\left[p_{2}\right]} \max _{s \in\left[p_{1}\right]} \boldsymbol{\Sigma}_{i}(s, s)\left(\sqrt{p_{2}}+\sqrt{p_{1}}+u\right) \tag{30}
\end{equation*}
$$

This upper bound is independent of $\{k, \ell\}$ and since $\|R(\Delta)\| \leq \sum_{k=1}^{q_{1}} \sum_{\ell=1}^{q_{2}}\left\|\tilde{\Delta}_{k \ell}\right\|$, we obtain the desired results in Theorem 1.

## A. 2 Proof of Corollary 1

We will continue to use the same notations as that in the proof of Theorem 1. Let $\mathcal{P}_{0}=\mathcal{U} \mathcal{U}^{\top}$ and $\mathcal{P}_{1}=\mathcal{V} \mathcal{V}^{\top}$. Similarly, let $\hat{\mathcal{P}}_{0}=\hat{\mathcal{U}} \hat{\mathcal{U}}^{\top}$ and $\hat{\mathcal{P}}_{1}=\hat{\mathcal{V}} \hat{\mathcal{V}}^{\top}$. Note that these matrices are all of rank at most $d$. As $R(\nu)=\mathcal{P}_{0} R(\nu) \mathcal{P}_{1}$, we have

$$
\hat{\mathcal{P}}_{0} R(\tilde{\nu}) \hat{\mathcal{P}}_{1}-R(\nu)=\left(\hat{\mathcal{P}}_{0}-\mathcal{P}_{0}\right) R(\tilde{\nu}) \hat{\mathcal{P}}_{1}+\mathcal{P}_{0} R(\tilde{\nu})\left(\hat{\mathcal{P}}_{1}-\mathcal{P}_{1}\right)+\mathcal{P}_{0} R(\Delta) \mathcal{P}_{1}
$$

We therefore have

$$
\left\|\hat{\mathcal{P}}_{0} R(\tilde{\nu}) \hat{\mathcal{P}}_{1}-R(\nu)\right\|_{F} \leq \sqrt{d}\left\|\hat{\mathcal{P}}_{0}-\mathcal{P}_{0}\right\|_{2} \times\|R(\tilde{\nu})\|_{2}+\sqrt{d}\left\|\hat{\mathcal{P}}_{1}-\mathcal{P}_{1}\right\|_{2} \times\|R(\tilde{\nu})\|_{2}+\sqrt{d}\|R(\Delta)\|_{2}
$$

Now $\left\|\hat{\mathcal{P}}_{0}-\mathcal{P}_{0}\right\|_{2} \leq 2\|\sin \Theta(\hat{\mathcal{U}}, \mathcal{U})\|_{2}$ and similarly for $\left\|\hat{\mathcal{P}}_{1}-\mathcal{P}_{1}\right\|_{2}$. Then from the conditions in Assumption 1, we have

$$
\left\|\hat{\mathcal{P}}_{0} R(\tilde{\nu}) \hat{\mathcal{P}}_{1}-R(\nu)\right\|_{F} \leq 2 \sqrt{d}\|\sin \Theta(\hat{\mathcal{U}}, \mathcal{U})\|_{2} \times\left(\|R(\nu)\|_{2}+\|R(\Delta)\|_{2}\right)+\|R(\Delta)\|_{2}=\mathcal{O}\left(n^{-1 / 2} p^{1 / 2}\right)
$$

Finally, as $\hat{\nu}$ and $\nu$ are the inverse rearrangement of $\hat{\mathcal{P}}_{0} R(\tilde{\nu}) \hat{\mathcal{P}}_{1}$ and $R(\nu)$, respectively, we have

$$
\frac{\|\hat{\nu}-\nu\|_{F}}{\|\nu\|_{F}}=\frac{\left\|\hat{\mathcal{P}}_{0} R(\tilde{\nu}) \hat{\mathcal{P}}_{1}-R(\nu)\right\|_{F}}{\|\nu\|_{F}}=\mathcal{O}\left((n p)^{-1 / 2}\right)
$$

as desired.

