Supplement: Proofs, Technical Details and additional results for “Matrix-variate regressions and envelope models”

Shanshan Ding and R. Dennis Cook

A Kronecker covariance matrix

Lemma 1. Let $\Delta_1 \in \mathbb{R}^{r \times r}$ and $\Delta_2 \in \mathbb{R}^{m \times m}$ be two positive definite matrices. If $\text{cov}[\text{vec}(Y)] = \Delta_2 \otimes \Delta_1$, then $\Delta_1 = \text{cov}_c(Y)/\text{tr}(\Delta_1)$ and $\Delta_2 = \text{cov}_r(Y)/\text{tr}(\Delta_2)$, where $\text{cov}_r(Y) = E[(Y - E(Y))(Y - E(Y))^T]$ and $\text{cov}_c(Y) = E[(Y - E(Y))(Y - E(Y))^T]$.

Proof. Let $Y - E(Y) = (Y_1, Y_2, \ldots, Y_m)$, where $Y_i$ denote the $i$-th column of $Y - E(Y)$, $i = 1, \ldots, m$. Then

$$\text{cov}[\text{vec}(Y)] = E \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_m \end{pmatrix} \begin{pmatrix} Y_1^T & Y_2^T & \cdots & Y_m^T \end{pmatrix} = \Delta_2 \otimes \Delta_1.$$ 

It follows that $E(Y_i Y_i^T) = \Delta_2 \delta_{2,ii}$, where $\delta_{2,ii}$ are the $i$-th diagonal elements of $\Delta_2$, $i = 1, \ldots, m$. Hence $\text{cov}_c(Y) = E[(Y - E(Y))(Y - E(Y))^T] = \sum_{i=1}^m E(Y_i Y_i^T) = \Delta_1 \text{tr}(\Delta_2)$ and $\Delta_1 = \text{cov}_c(Y)/\text{tr}(\Delta_2)$. Similarly, it can be shown that $\text{cov}_r(Y) = \Delta_2 \text{tr}(\Delta_1)$ and thus $\Delta_2 = \text{cov}_r(Y)/\text{tr}(\Delta_1)$. \qed

B Special formulations

Concomitant columns. When the column dimension of $X$ is the same as the corresponding dimension of $Y$, and in particular when the columns of $X$ and the columns of $Y$ represent repeated measures or similar characteristics, it might be appropriate to build (2.2) with $\beta_2 = I_m$,.

an $m \times m$ identity matrix, since each column of $Y$ is usually associated with the corresponding column of $X$. In this case, the matrix regression model can be simplified to:

$$ Y = \mu + \beta_1 X + \varepsilon, \quad (B.1) $$

so the regression of each column of $Y$ on the corresponding column of $X$ has the same coefficient matrix $\beta_1$. This model is the same as the matrix regression model proposed by Viroli (2012). In application, one can apply a likelihood ratio test or other model selection methods to select between (2.2) and (B.1). Similar comments apply to the row dimensions of $X$ and $Y$.

**Univariate predictor.** When the predictor $X \in \mathbb{R}^1$ is univariate, a matrix regression might allow a different linear regression for each element of $Y$ on $X$:

$$ Y = \mu + \beta X + \varepsilon, \quad (B.2) $$

where $\beta \in \mathbb{R}^{r \times m}$. In this case, (2.2) still has a multiplicative coefficient structure but now the column dimensions of $\beta_1$ and $\beta_2$ need not be restricted to match the dimension of $X$:

$$ Y = \mu + \beta_1 \beta_2^T X + \varepsilon, \quad (B.3) $$

where $\beta_1 \in \mathbb{R}^{r \times q}$ and $\beta_2 \in \mathbb{R}^{q \times m}$, for some $q \leq \min(r, m)$, are full rank matrices. The coefficient matrices $\beta_1$ and $\beta_2$ are not unique without imposing further constraints as for any nonsingular matrix $T \in \mathbb{R}^{q \times q}$, $\beta_1 \beta_2^T = \beta_1 T T^{-1} \beta_2^T$. Nevertheless, the column spans of $\beta_1$ and $\beta_2$ are identifiable, and $\beta_1$ and $\beta_2$ are also identifiable if we impose constraints like $\beta_1 \beta_1^T = I_d$ (or $\beta_2 \beta_2^T = I_d$). Thus, (B.3) is a reduced rank (Izenman, 1975) form of (B.2). Again, model selection tools can be applied to select the proper model between (B.2) and (B.3).

**Multivariate POD model.** The matrix-variate regression (2.2) can be seen as an extension of the partial one-dimensional regression model (POD) studied by Cook and Weisberg (2004). The POD was proposed to study data with one response variable, one grouping variable or factor, and several covariates. A typical POD model represents the response $Y \in \mathbb{R}^1$ at factor level $k$ as $Y_{(k)} = \eta_{0k} + \eta_{1k} X^T \beta + \varepsilon$, $k = 1, \ldots, g$, where $\eta_{0k}$ and $\eta_{1k}$ represent the $k$-th level factor effect, and $X^T \beta$ represents the one-dimensional covariate effect ($X, \beta \in \mathbb{R}^p$). If each subject contains measurements over all $g$ categories, then a multivariate response $Y = (Y_{(1)}, \ldots, Y_{(g)})^T$
will be observed for each subject. In this case, the POD model can be extended as

$$\mathbf{Y} = (\eta_0 + \eta_{11}\mathbf{X}^T\beta, \ldots, \eta_0g + \eta_{1g}\mathbf{X}^T\beta)^T + \varepsilon = \eta_0 + \eta_1\mathbf{X}^T\beta + \varepsilon,$$

(B.4)

where $\eta_0 = (\eta_{01}, \ldots, \eta_{0g})^T$ and $\eta_1 = (\eta_{11}, \ldots, \eta_{1g})^T$ are $g$ dimensional vectors. The mean function of this model has the exactly the same coefficient structure as shown in (2.2) except the response and the predictor are vector-valued. When $\mathbf{Y} \in \mathbb{R}^{g \times m}$ is matrix-valued, for example, $m$ response variables measured at each factor level, (B.4) can be extended to the exact matrix-variate regression setting.

C Maximum likelihood estimation

Matrix normal density function. If $\mathbf{Z}$ follows a matrix normal distribution with mean $\mathbf{M}$, and row and column covariance matrices $\Sigma_1$ and $\Sigma_2$, then the density function of $\mathbf{Z}$ is

$$f(\mathbf{Z}|\mathbf{M}, \Sigma_1, \Sigma_2) = (2\pi)^{-\frac{rm}{2}} |\Sigma_2|^{-\frac{r}{2}} |\Sigma_1|^{-\frac{m}{2}} \exp \left( -\frac{1}{2} \text{tr}\{\Sigma_2^{-1}(\mathbf{Z} - \mathbf{M})^T\Sigma_1^{-1}(\mathbf{Z} - \mathbf{M})\} \right).$$

MLE of (2.2). Based on the matrix normal density function, the log-likelihood function of (2.2) is

$$l(\mu, \beta_1, \beta_2, \Sigma_1, \Sigma_2) = c - \frac{nr}{2} \log |\Sigma_2| - \frac{nm}{2} \log |\Sigma_1|$$

$$- \frac{1}{2} \sum_{i=1}^{n} \text{tr}\{\Sigma_2^{-1}(\mathbf{Y}_i - \bar{\mathbf{Y}} - \mathbf{B}_{1|2}\mathbf{X}_i\beta_2^T)^T\Sigma_1^{-1}(\mathbf{Y}_i - \bar{\mathbf{Y}} - \mathbf{B}_{1|2}\mathbf{X}_i\beta_2^T)\},$$

(C.1)

where $c = -nrm \log 2\pi/2$. Since $\bar{\mathbf{X}} = 0$, the MLE $\hat{\mu} = \bar{\mathbf{Y}}$. Then given $\beta_2$ and $\Sigma_2$, by taking matrix derivatives, we have the MLE of $\beta_1$ as

$$\mathbf{B}_{1|2} = \left( \sum_{i=1}^{n}(\mathbf{Y}_i - \bar{\mathbf{Y}})\Sigma_2^{-1}\beta_2\mathbf{X}_i^T \right)\left( \sum_{i=1}^{n}\mathbf{X}_i\beta_2^T\Sigma_2^{-1}\beta_2\mathbf{X}_i^T \right)^{-1} := \mathbf{C}_2\mathbf{M}_2^{-1},$$

where $\mathbf{C}_2 = \sum_{i=1}^{n}(\mathbf{Y}_i - \bar{\mathbf{Y}})\Sigma_2^{-1}\beta_2\mathbf{X}_i^T$ and $\mathbf{M}_2 = \sum_{i=1}^{n}\mathbf{X}_i\beta_2^T\Sigma_2^{-1}\beta_2\mathbf{X}_i^T$, and the MLE of $\Sigma_1$ as

$$\mathbf{S}_{\text{res|2}} = (nm)^{-1} \sum_{i=1}^{n}(\mathbf{Y}_i - \bar{\mathbf{Y}} - \mathbf{B}_{1|2}\mathbf{X}_i\beta_2^T)\Sigma_2^{-1}(\mathbf{Y}_i - \bar{\mathbf{Y}} - \mathbf{B}_{1|2}\mathbf{X}_i\beta_2^T)^T.$$
Correspondingly, given $\beta_1$ and $\Sigma_1$, the MLE of $\beta_2$ and $\Sigma_2$ can be obtained as (2.4).

**MLE of (4.3).** Again, as $\bar{X} = 0$, the MLE $\hat{\mu} = \bar{Y}$. Then given $\beta_2$ and $\Sigma_2$, the log-likelihood function of (4.3) is $l(L, \eta_1, \Omega_1, \Omega_{10}) =$

$$c - \frac{nr}{2} \log |\Sigma_2| - \frac{nm}{2} \log |\Omega_{10}| - \frac{1}{2} \sum_{i=1}^{n} \text{tr}\{\Sigma_2^{-1}(Y_i - \bar{Y})^T L_0 \Omega_{10}^{-1} L_0^T (Y_i - \bar{Y})\}$$

$$- \frac{nm}{2} \log |\Omega_1| - \frac{1}{2} \sum_{i=1}^{n} \text{tr}\{\Sigma_2^{-1}(L^T Y_i - L^T \bar{Y} - \eta_1 X_i \beta_2^T) \Omega_1^{-1}(L^T Y_i - L^T \bar{Y} - \eta_1 X_i \beta_2^T)\}.$$ (C.2)

For fixed $L$, the MLE $\hat{\eta}_1 = L^T [\sum_{i=1}^{n} (Y_i - \bar{Y}) \Sigma_2^{-1} \beta_2 X_i^T] (\sum_{i=1}^{n} X_i \beta_2^T \Sigma_2^{-1} \beta_2 X_i^T)^{-1} := L^T B_{1|2}$. Substituting $\hat{\eta}_1$ back to (C.2), it can be found that $\hat{\Omega}_1 = L^T S_{Y|2} L$ and $\hat{\Omega}_{10} = L^T S_{Y|2} L$, where $S_{Y|2} = (nm)^{-1} \sum_{i=1}^{n} (Y_i - \bar{Y}) \Sigma_2^{-1} (Y_i - \bar{Y})^T$ is the second sample column moment of $Y$. These results lead to

$$\hat{l}(L) = c - \frac{nr}{2} \log |\Sigma_2| - \frac{nm}{2} \log |L_0^T S_{Y|2} L_0| - \frac{nm}{2} \log |L^T S_{Y|2} L| - \frac{nmr}{2}.$$ 

Therefore, given $\beta_2$ and $\Sigma_2$, the MLE of $L$ is

$$\hat{L} = \arg\min_{\text{span}(B) \in \mathcal{G}(u_1, r)} \log |B^T S_{Y|2} B| + \log |B^T S_{Y|2}^{-1} B|$$

by the fact that $|L_0^T S_{Y|2} L_0| = |S_{Y|2}| \cdot |L^T S_{Y|2}^{-1} L|$. Correspondingly, $\hat{\beta}_{1|2} = \hat{L} \hat{\eta}_1 = P_L B_{1|2}$ and $\hat{\Sigma}_{1|2} = P_L S_{Y|2} P_L + P_{L_0} S_{Y|2} P_{L_0}$. By taking the transpose of (4.3), the MLEs of $\beta_2$ and $\Sigma_2$ given $\beta_1$ and $\Sigma_1$ can be derived in the same way. We thus omit this part.

### D Matrix regression with group effects

Sometimes the goal of a study may be about the association between matrix-valued responses and certain group effects. Suppose that the data are collected from $g$ groups and no other predictors exist. Let $Y_{(i)j} \in \mathbb{R}^{r \times m}$ be the matrix-valued response for the $j$-th subject in the $i$-th group. Then an extended form of (B.2) for estimating the group effects is

$$Y_{(i)j} = \mu + \alpha_i + \varepsilon_{(i)j} \quad (D.1)$$
for \( i = 1, \ldots, g \) and \( j = 1, \ldots, n_i \), where \( \mu \) represents the overall mean, \( \alpha_i \in \mathbb{R}^{r \times m} \) represents the difference between the \( i \)-th group mean and the overall mean, \( \varepsilon_{(i)j} \) are i.i.d. matrix-variate errors as defined in (2.2), and \( n_i \) is the sample size of Group \( i \). The group effects satisfy

\[
\sum_{i=1}^g n_i \alpha_i = 0.
\]

**Envelope model of (D.1).** For enveloping the group effects in (D.1), let \( \text{span}(\alpha_1, \ldots, \alpha_g) = A \) and \( \text{span}\{(\alpha_1, \ldots, \alpha_g)^T\} = B \). Assume that there exist subspaces \( S_L \subseteq \mathbb{R}^r \) and \( S_R \subseteq \mathbb{R}^m \) such that:

\[
A \subseteq S_L, \quad \Sigma_1 = \Sigma_{S_L} + \Sigma_{S_L^\perp},
\]

\[
B \subseteq S_R, \quad \Sigma_2 = \Sigma_{S_R} + \Sigma_{S_R^\perp},
\]

where \( \Sigma_{S_L} = P_{S_L} \Sigma_1 P_{S_L} \), \( \Sigma_{S_L^\perp} = Q_{S_L} \Sigma_1 Q_{S_L} \), \( \Sigma_{S_R} = P_{S_R} \Sigma_2 P_{S_R} \) and \( \Sigma_{S_R^\perp} = Q_{S_R} \Sigma_2 Q_{S_R} \). Then the smallest subspaces that satisfy (D.2) and (D.3) are the \( \Sigma_1 \)-envelope of \( A \) and \( \Sigma_2 \)-envelope of \( B \), denoted as \( \mathcal{E}_{\Sigma_1}(A) \) and \( \mathcal{E}_{\Sigma_2}(B) \), or \( E_1 \) and \( E_2 \). Let \( L \in \mathbb{R}^{r \times u_1}(u_1 \leq r) \) and \( R \in \mathbb{R}^{m \times u_2}(u_2 \leq m) \) be semi-orthogonal bases of \( \mathcal{E}_{\Sigma_1}(A) \) and \( \mathcal{E}_{\Sigma_2}(B) \) respectively. Under (D.2) and (D.3), the envelope model of (D.1) is

\[
Y_{(i)j} = \mu + L \eta_i R^T + \varepsilon_{(i)j}, \quad i = 1, \ldots, g, \quad j = 1, \ldots, n_i,
\]

\[
\Sigma_1 = \Sigma_{\mathcal{E}_1} + \Sigma_{\mathcal{E}_1^\perp} = L \Omega_1 L^T + L_0 \Omega_{10} L_0^T
\]

\[
\Sigma_2 = \Sigma_{\mathcal{E}_2} + \Sigma_{\mathcal{E}_2^\perp} = R \Omega_2 R^T + R_0 \Omega_{20} R_0^T,
\]

where \( \eta_i \in \mathbb{R}^{u_1 \times u_2}, \ i = 1, \ldots, g, \) are the coordinate matrices of \( \alpha_i \) with respect to \( L \) and \( R \), and they satisfy \( \sum_{i=1}^g n_i \eta_i = 0 \).

The estimation procedure of (D.4) is similar to the two-step iterative algorithm given in Section 4.3. We thus omit it.

### E One-sided enveloping

Consider the envelope model for (B.1). Since only \( \beta_1 \) is of the primary interest, we use \( \Sigma_1 \)-envelope of \( \beta_1 \) to formulate the envelope model of (B.1) as:

\[
Y = \mu + L \eta_1 X + \varepsilon
\]

\[
\Sigma_1 = \Sigma_{\mathcal{E}_1} + \Sigma_{\mathcal{E}_1^\perp} = L \Omega_1 L^T + L_0 \Omega_{10} L_0^T
\]
where the parameters are the same defined as in (4.3).

To estimate the model parameters, one can apply the algorithm in Section 4.3 with some simplification: (1) as \( \beta_2 = I_m \), only one initial value \( \Sigma_{20} \) is needed in Step 1; (2) In Step 3, since \( \Sigma_2 \) is not enveloped, the estimator of \( \Sigma_2 \) is then simplified to 
\[
\hat{\Sigma}_2 = (nr)^{-1} \sum_{i=1}^{n} (Y_i - \bar{Y} - \hat{\beta}_1 X_i)^T \hat{\Sigma}_1 (Y_i - \bar{Y} - \hat{\beta}_1 X_i),
\]
given \( \hat{\beta}_1 \) and \( \hat{\Sigma}_1 \). In addition, since \( \beta_1 \) is uniquely identified, no rescaling is needed.

### F Simulations studies for goodness of fit in Section 2.4

We evaluate the goodness of fit of the matrix-variate regression (2.2) relative to the vector regression (2.1) based on LRT, AIC and BIC. Two simulation settings were considered: 1) we generated data from model (2.2) with \( Y, X \in \mathbb{R}^{5 \times 2} \), where the elements of \( \mu, \beta_1 \) and \( \beta_2 \) were obtained from independent standard normal variables, and the two covariance matrices are identity matrices multiplied by 0.5; 2) we generated data from model (2.1) with the same data dimensions, where the elements of \( \gamma \) and \( \nu \) were selected from independent standard normal variables, and the covariance matrix \( \Sigma \) was chosen to be an identity matrix multiplied by 0.25. Then we fitted both matrix-variate regression and vector regression for each setting and compared the two models via the the goodness of fit test and the information criteria, such as AIC and BIC. We randomly generated 100 replicates for each setting. The frequencies of the correct model selection over the 100 replicates are summarized in Table 1 for setting 1 and in Table 2 for setting 2. For the likelihood ratio test, the significance level was chosen to be 0.01.

Note that in the first simulation setting, the true model is a matrix-variate regression model. Then both the coefficient and covariance matrices are structured in the vectorized version, and thus the null hypothesis \( H_0 : \nu = \beta_2 \otimes \beta_1, \Sigma = \Sigma_2 \otimes \Sigma_1 \) holds for the goodness of fit test. A correct model selection will occur when we fail to reject \( H_0 \). Table 1 shows that the LRT performs fairly well for identifying the true model over different sample sizes. As \( n \) increases, the chi-squared approximation of the LRT test statistic become more accurate, and thus leading to more accurate model selection. In addition, both AIC and BIC were able to identify the true model over all 100 replicates.

In the second simulation setting, the true model does not have a matrix regression structure. Hence the null hypothesis \( H_0 : \nu = \beta_2 \otimes \beta_1, \Sigma = \Sigma_2 \otimes \Sigma_1 \) for the goodness of fit test does not hold. Correspondingly, the true model will be chosen when the null hypothesis is rejected. We
see from Table 2 that the LRT was able to successfully select the true model over all replicates and all sample sizes, so do AIC and BIC. The results imply that when the regression parameters do not possess the kronecker structure, the goodness fit test, and the information criteria can correctly suggest the less restricted model.

Table 2: Frequencies of correct model selection over 100 replicates for setting 2.

<table>
<thead>
<tr>
<th>method</th>
<th>n = 200</th>
<th>n = 300</th>
<th>n = 500</th>
<th>n = 800</th>
<th>n = 1000</th>
<th>n = 1500</th>
</tr>
</thead>
<tbody>
<tr>
<td>LRT</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>AIC</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>BIC</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

G Proofs of Propositions 1 and 2

Proof of Proposition 1. Let ‘vec’ and ‘vech’ denote the vectorization and half-vectorization operators: for any symmetric $B \in \mathbb{R}^{p \times q}$, $\text{vech}(B) = C_q \text{vec}(B)$ and $\text{vec}(B) = E_q \text{vech}(B)$, where $C_q$ and $E_q$ are the unique contraction matrix and the expansion matrix that connect the operators (Henderson and Searle, 1979).

Under normality, the Fisher information of $\gamma$ in (2.1) is

$$J = \begin{pmatrix} \Sigma_{\text{vec}(X)} \otimes \Sigma^{-1} & 0 \\ 0 & \frac{1}{2} E_{r m}^T (\Sigma^{-1} \otimes \Sigma^{-1}) E_{r m} \end{pmatrix},$$

where $\Sigma_{\text{vec}(X)} = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \text{vec}(X_i) \text{vec}(X_i)^T$ assuming that $X$ is centered. Correspondingly, $\sqrt{n}(\hat{\gamma}_s - \gamma)$ converges in distribution to a normal random vector with mean zero and covariance matrix $J^{-1}$. Now consider the gradient matrix $H = (\partial h_i / \partial \theta_j)_{i,j}$. It can be shown that

$$\partial h_1 / \partial \theta_1 = \partial \text{vec}(\beta_2 \otimes \beta_1) / \partial \text{vec}(\beta_1)^T = (I_q \otimes K_{pm} \otimes I_r)(\text{vec}(\beta_2) \otimes I_{rp}),$$

$$\partial h_1 / \partial \theta_2 = \partial \text{vec}(\beta_2 \otimes \beta_1) / \partial \text{vec}(\beta_2)^T = (I_q \otimes K_{pm} \otimes I_r)(I_{mq} \otimes \text{vec}(\beta_1)).$$
Hence according to Proposition 4.1 (Shapiro, 1986) along with the asymptotic distribution of the normal log likelihood function, it satisfies the four conditions listed in Section 3, Shapiro (1986). To apply Shapiro’s result, we first define the minimum discrepancy function and employ Proposition 4.1 in Shapiro (1986) to derive the asymptotic distribution. In order to consider avar$(\hat{\gamma})$, we have \( \sqrt{n}(\hat{\gamma} - \gamma) \) converges in distribution to a normal random vector with mean zero and covariance matrix \( \text{avar}(\sqrt{n}\hat{\gamma}) = H(H^THH)^T \).

To prove the asymptotic relative efficiency of \( \hat{\gamma}_m \) to \( \hat{\gamma}_s \), consider \( \text{avar}(\sqrt{n}\hat{\gamma}_s) - \text{avar}(\sqrt{n}\hat{\gamma}_m) = J^{-1} - H(H^THH)^T = J^{-1/2}Q_{J^{1/2}H}J^{-1/2} \), where \( Q_{J^{1/2}H} \) is the projection matrix onto the orthogonal compliment of span(\( J^{1/2}H \)). Therefore, \( \text{avar}(\sqrt{n}\hat{\gamma}_s) - \text{avar}(\sqrt{n}\hat{\gamma}_m) \geq 0 \). This completes the proof.

**Proof of Proposition 2**. This proof is similar to the proof of Proposition 1 as (4.3) is also an overparameterized model of (2.1). Let \( U = \partial \phi(\zeta)/\partial \zeta^T \). Using the chain rule, we have \( U = \partial h(\theta)/\partial \theta^T \cdot \partial g(\zeta)/\partial \zeta^T = HG_1 \). Then based on the same minimum discrepancy function \( F(\hat{\gamma}_s, \gamma) \) and Proposition 4.1 in Shapiro (1986), it follows that \( \sqrt{n}(\phi(\hat{\zeta}) - \phi(\zeta)) \) converges in distribution to a normal random vector with mean zero and covariance matrix \( \text{avar}(\sqrt{n}\phi(\hat{\zeta})) = HG_1(G_1^TH^THG_1)^TG_1^TH^T \).

Now we show the asymptotic relative efficiency of the envelope estimator \( \phi(\hat{\zeta}) \) to \( \hat{\gamma}_m \). Consider \( \text{avar}(\sqrt{n}\hat{\gamma}_s) - \text{avar}(\sqrt{n}\hat{\gamma}_m) = H(H^THH)^T - HG_1(G_1^TH^THG_1)^TG_1^TH^T = J^{-1/2}(P_{J^{1/2}H} - P_{J^{1/2}H})J^{-1/2} \), where \( P_{J^{1/2}H} \) is the projection matrix onto span(\( J^{1/2}H \)) and \( P_{J^{1/2}H} \) is the projection matrix onto span(\( J^{1/2}H \)). Since span(\( J^{1/2}H \)) \subseteq span(\( J^{1/2}H \)), \( P_{J^{1/2}H} - P_{J^{1/2}H} \) \geq 0. Thus, \( \text{avar}(\sqrt{n}\hat{\gamma}_m) - \text{avar}(\sqrt{n}\phi(\hat{\zeta})) \geq 0 \).
H Proofs of Propositions 3 and 4

Proof of Proposition 3. To develop the asymptotic distribution of the estimator of \( \theta \), the row and column parameter matrices \( \beta_1, \beta_2, \Sigma_1 \), and \( \Sigma_2 \) are required to be uniquely defined. According to the model setting, we require both \( \beta_2 \) and \( \Sigma_2 \) to have unit Frobenius norms and positive elements in their first rows and columns, that is, \( \beta_{11,2} > 0 \), and \( \Sigma_{11,2} > 0 \).

The MLE \( \hat{\theta} \) of \( \theta \) in (2.2) can be obtained by rescaling a non-conditioned MLE \( \hat{\theta}^* = (\text{vec}(B_1^*), \text{vec}(B_2^*), \text{vech}(S_1^*), \text{vech}(S_2^*))^T \) as

\[
\hat{\theta} = m(\hat{\theta}^*) = (c \text{vec}(B_1^*), \text{vec}(B_2^*)/c, d \text{vech}(S_1^*)^T, \text{vech}(S_2^*)^T/d)^T,
\]

where \( c = \text{sign}(B_{11,2}^*)||B_2^*||_F \), \( d = \text{sign}(S_{11,2}^*)||S_2^*||_F \), and \( B_{11,2}^* \) and \( S_{11,2}^* \) are the elements in the first rows and columns of \( B_j^* \) and \( S_j^* \), respectively. In addition, since \( B_j^* \) and \( S_j^* \) are proportional to \( B_j \) and \( S_j \), there exists a parameter vector

\[
\theta^* = (\text{vec}(\beta_1^*), \text{vec}(\beta_2^*), \text{vech}(\Sigma_1^*), \text{vech}(\Sigma_2^*))^T
\]

such that \( \hat{\theta}^* \) is \( \sqrt{n} \) consistent to \( \theta^* \) and is the MLE of \( \theta^* \). Moreover, \( \theta = m(\theta^*) = (c_1 \text{vec}(\beta_1^*), \text{vec}(\beta_2^*)/c_1, d_1 \text{vech}(\Sigma_1^*)^T, \text{vech}(\Sigma_2^*)^T/d_1)^T \), where \( c_1 = \text{sign}(\beta_{11,2}^*)||\beta_2^*||_F \), \( d_1 = \text{sign}(\Sigma_{11,2}^*)||\Sigma_2^*||_F \), and \( \beta_{11,2}^* \) and \( \Sigma_{11,2}^* \) are the elements in the first rows and columns of \( \beta_2^* \) and \( \Sigma_2^* \).

Let \( E_n \) denote the empirical mean. Denote \( J^* \) as the Fisher information of \( \theta^* \). Then \( \sqrt{n}(\theta^* - \theta) \) converges to a normal random vector with mean zero and covariance matrix \( J^{*-1} \). Thus, using the Delta method, we have \( \sqrt{n}(\hat{\theta} - \theta) \) converges to a normal random vector with mean zero and covariance matrix \( K^* J^{*-1} K^*^T \), where \( K^* = \partial \theta / \partial \theta^* \).

\[
\begin{pmatrix}
 c_1 I_{r p_1} & \frac{1}{c_1} \text{vec}(\beta_1) \text{vec}(\beta_2)^T & 0 & 0 \\
 0 & \frac{1}{c_1} (I_{mp_2} - \text{vec}(\beta_2) \text{vec}(\beta_2)^T) & 0 & 0 \\
 0 & 0 & d_1 I_{r(r+1)/2} & \frac{1}{d_1} \text{vech}(\Sigma_1) \text{vech}(\Sigma_2)^T E_m^T E_m \\
 0 & 0 & 0 & \frac{1}{d_1} (I_{m(m+1)/2} - \text{vech}(\Sigma_2) \text{vech}(\Sigma_2)^T E_m^T E_m)
\end{pmatrix}
\]

We next derive the Fisher information \( J^* \). For convenience, we first develop the Fisher information of \( \delta = (\delta_1^T, \delta_2^T, \delta_3^T, \delta_4^T)^T = (\text{vec}(\beta_1^T), \text{vec}(\beta_2^T), \text{vech}(\Sigma_1^{-1})^T, \text{vech}(\Sigma_2^{-1})^T)^T \), denoted as \( I_\delta \). Hence \( I_\delta = (I_{\delta,\delta})_{i,j} \), where \( I_{\delta,\delta} = -\lim_{n \to \infty} E_n [\partial^2 l / \partial \delta_i \partial \delta_j] \) and \( l \) is the log-likelihood...
function (C.1). After some algebra, it can be shown that

$$\frac{\partial^2 l}{\partial \delta_1 \partial \delta_1^T} = -\sum_{i=1}^{n} X_i \beta_2^T \Sigma_2^{-1} \beta_2 X_i^T \otimes \Sigma_1^{-1}$$

$$\frac{\partial^2 l}{\partial \delta_1 \partial \delta_2^T} = \sum_{i=1}^{n} X_i \otimes \Sigma_1^{-1} Y_i \Sigma_2^{-1} - \left( \sum_{i=1}^{n} X_i \otimes \Sigma_1^{-1} \beta_1^T X_i \right) \left[ I_{p_2} \otimes \beta_2^T \Sigma_2^{-1} + (\beta_2^T \Sigma_2^{-1} \otimes I_m) K_{mp_2} \right]$$

$$= \sum_{i=1}^{n} X_i \otimes \Sigma_1^{-1} \epsilon_i \Sigma_2^{-1} - \left( \sum_{i=1}^{n} X_i \beta_2^T \Sigma_2^{-1} \otimes \Sigma_1^{-1} \beta_1^T X_i \right) K_{mp_2}$$

$$\frac{\partial^2 l}{\partial \delta_1 \partial \delta_3^T} = \partial \text{vec} \left( \sum_{i=1}^{n} \Sigma_1^{-1} \epsilon_i \Sigma_2^{-1} \beta_2^T X_i^T \right) / \partial \text{vech} \left( \Sigma_1^{-1} \right)^T$$

$$= \left( \sum_{i=1}^{n} X_i \beta_2^T \Sigma_2^{-1} \epsilon_i \otimes I_r \right) E_r$$

$$\frac{\partial^2 l}{\partial \delta_1 \partial \delta_4^T} = \partial \text{vec} \left( \sum_{i=1}^{n} \Sigma_1^{-1} \epsilon_i \Sigma_2^{-1} \beta_2^T X_i^T \right) / \partial \text{vech} \left( \Sigma_2^{-1} \right)^T = \left( \sum_{i=1}^{n} X_i \beta_2^T \otimes \Sigma_1^{-1} \epsilon_i \right) E_m.$$  

As $X_i$'s are nonstochastic, and in addition $E(\epsilon_i) = 0$, $i = 1, \ldots, n$, it follows that $I_{\delta_1 \delta_1} = N_1 \otimes \Sigma_1^{-1}$, $I_{\delta_1 \delta_2} = N_2$ and both $I_{\delta_1 \delta_3}$, and $I_{\delta_1 \delta_4}$ are zero matrices, where $N_1 = \lim_{n \to \infty} E_n(X \beta_2^T \Sigma_2^{-1} \beta_2 X^T)$ and $N_2 = \lim_{n \to \infty} E_n(X \beta_2^T \Sigma_2^{-1} \otimes \Sigma_1^{-1} \beta_1^T X) K_{mp_2}$. Similarly, it can be shown that $I_{\delta_2 \delta_2} = N_3 \otimes \Sigma_2^{-1}$, and $I_{\delta_2 \delta_3}$ and $I_{\delta_2 \delta_4}$ are zero matrices, where $N_3 = \lim_{n \to \infty} E_n(X^T \beta_1^T \Sigma_1^{-1} \beta_1 X)$. By taking the second derivatives of $l$ respected to $\delta_3$ and $\delta_4$, we have

$$\frac{\partial^2 l}{\partial \delta_3 \partial \delta_3^T} = -\frac{nm}{2} E_r^T (\Sigma_1^* \otimes \Sigma_1^*) E_r$$

$$\frac{\partial^2 l}{\partial \delta_4 \partial \delta_4^T} = -\frac{nr}{2} E_m^T (\Sigma_2^* \otimes \Sigma_2^*) E_m$$

$$\frac{\partial^2 l}{\partial \delta_3 \partial \delta_4^T} = -\frac{1}{2} \sum_{i=1}^{n} E_i^T \left[ (Y_i - \mu - \beta_1^T X_i ^T \beta_2^T) \otimes (Y_i - \mu - \beta_1^T X_i ^T \beta_2^T) \right] E_m.$$  

Hence $I_{\delta_3 \delta_3} = \frac{m}{2} E_r^T (\Sigma_1^* \otimes \Sigma_1^*) E_r$, $I_{\delta_3 \delta_4} = \frac{r}{2} E_m^T (\Sigma_2^* \otimes \Sigma_2^*) E_m$, and $I_{\delta_3 \delta_4} = \frac{1}{2} E_r^T E_r \text{vech}(\Sigma_1^*) \text{vech}(\Sigma_2^*)^T E_m^T E_m$ (von Rosen, 1988; Pan and Fang, 2000).

Because $\theta^* = (\text{vec}(\beta_1^T), \text{vec}(\beta_2^T), \text{vech}(\Sigma_1^*), \text{vech}(\Sigma_2^*))^T = f(\delta)$ is a function of $\delta$, by the Delta method, the Fisher information of $\theta^*$ is then $J^* = \{ f'(\delta) \}^{-1}$, where $f'(\delta) = \partial f / \partial \delta = \text{blockdiag}[L_{p_1}, L_{mp_2}, -E_r^T (\Sigma_1^* \otimes \Sigma_1^*) C_r, -E_m^T (\Sigma_2^* \otimes \Sigma_2^*) C_m]$ is a block diagonal matrix.
After some matrix multiplications, this gives

\[
\begin{pmatrix}
N_1 \otimes \Sigma_1^{*-1} & N_2 & 0 & 0 \\
N_2^T & N_3 \otimes \Sigma_2^{*-1} & 0 & 0 \\
0 & 0 & \frac{1}{2} E_r^T \Pi_1 E_r & \frac{1}{2} E_r^T \Pi_1 \Psi \Pi_2 E_m \\
0 & 0 & \frac{1}{2} E_m^T \Pi_2 \Psi^T \Pi_1 E_r & \frac{1}{2} E_m^T \Pi_2 E_m
\end{pmatrix},
\]

where \( \Psi = \text{vec}(\Sigma_1^*)\text{vec}(\Sigma_2^*)^T \), \( \Pi_1 = \Sigma_1^{*-1} \otimes \Sigma_1^{*-1} \) and \( \Pi_2 = \Sigma_2^{*-1} \otimes \Sigma_2^{*-1} \).

The last step is to simplify the asymptotic variance of \( \hat{\theta} \). Consider the decomposition of \( K^* \) as \( K^* = KC \), where

\[
K = \begin{pmatrix}
I_{rp_1} & \text{vec}(\beta_1)^T \text{vec}(\beta_2)^T & 0 & 0 \\
0 & I_{mp_2} - \text{vec}(\beta_2)^T \text{vec}(\beta_2)^T & 0 & 0 \\
0 & 0 & I_{r(r+1)/2} & \text{vech}(\Sigma_1) \text{vech}(\Sigma_2)^T E_m^T E_m \\
0 & 0 & 0 & I_{m(m+1)/2} - \text{vech}(\Sigma_2) \text{vech}(\Sigma_2)^T E_m^T E_m
\end{pmatrix},
\]

and \( C = \text{blockdiag}(c_1 I_{rp_1}, I_{mp_2}/c_1, d_1 I_{r(r+1)/2}, I_{m(m+1)/2}/d_1) \). Based on the relationship between \( \theta \) and \( \theta^* \), it is easy to check that \( K^* J^{*-1} K^*^T = K J_1^{-1} K^T \), where \( J_1 \) is a similar matrix as \( J^* \) but with \( \beta_j^* \) and \( \Sigma_j^* \) in \( J^* \) replaced by the corresponding \( \beta_j \) and \( \Sigma_j \).

**Proof of Proposition 4.** For notation convenience, let \( \hat{\theta}_e = (\text{vec}(\hat{\beta}_1)^T, \text{vec}(\hat{\beta}_2)^T, \text{vech}(\hat{\Sigma}_1)^T, \text{vech}(\hat{\Sigma}_2)^T)^T = g(\hat{\zeta}) \) be the envelope MLE of \( \theta \). Similarly, \( \hat{\theta}_e \) can be obtained by normalizing and rescaling a non-constrained envelope estimator \( \hat{\theta}_e^* = g(\hat{\zeta}^*) = (\text{vec}(\hat{\beta}_1)^T, \text{vec}(\hat{\beta}_2)^T, \text{vech}(\hat{\Sigma}_1)^T, \text{vech}(\hat{\Sigma}_2)^T)^T \) as \( \hat{\theta}_e = m(\hat{\theta}_e^*) = (c \text{vec}(\hat{\beta}_1)^T, \text{vec}(\hat{\beta}_2)^T/c, d \text{vech}(\hat{\Sigma}_1)^T, \text{vech}(\hat{\Sigma}_2)^T/d)^T \), where \( \hat{\beta}_j^* \) and \( \hat{\Sigma}_j^* \) are without normalization constraints and are over-parameterized in terms of

\[
\hat{\zeta}^* = (\text{vec}(\hat{\eta}_1)^T, \text{vec}(\hat{\mathbf{L}})^T, \text{vec}(\hat{\eta}_2)^T, \text{vec}(\hat{\mathbf{R}})^T, \text{vech}(\hat{\Omega}_1)^T, \text{vech}(\hat{\Omega}_1^*)^T, \text{vech}(\hat{\Omega}_2)^T, \text{vech}(\hat{\Omega}_2^*)^T, \text{vech}(\hat{\Omega}_{20})^T)^T,
\]

where \( c = \text{sign}(\hat{\beta}_{11,2})\|\hat{\beta}_2\|_F \), \( d = \text{sign}(\hat{\Sigma}_{11,2})\|\hat{\Sigma}_2\|_F \), and \( \hat{\beta}_{11,2} \) and \( \hat{\Sigma}_{11,2} \) are the elements in the first rows and columns of \( \hat{\beta}_2^* \) and \( \hat{\Sigma}_2^* \).

In addition, there must exist an unconstrained parameter vector

\[
\theta^{**} = (\text{vec}(\beta_1^*)^T, \text{vec}(\beta_2^*)^T, \text{vech}(\Sigma_1^{**})^T, \text{vech}(\Sigma_2^{**})^T)^T,
\]
such that $\hat{\theta}_e^*$ is the envelope MLE of $\theta^{**}$ and

$$
\theta = m(\theta^{**}) = (c_2\text{vec}(\beta_1^{**})^T, \text{vec}(\beta_2^{**})^T/c_2, d_2\text{vech}(\Sigma_1^{**})^T, \text{vech}(\Sigma_2^{**})^T/d_2)^T,
$$

where $c_2 = \text{sign}(\beta_{11}^{**})\|\beta_2^{**}\|_F$, $d_2 = \text{sign}(\Sigma_{11}^{**})\|\Sigma_2^{**}\|_F$ with $\beta_{11}^{**}$ and $\Sigma_{11}^{**}$ being the elements in the first rows and columns of $\beta_2^{**}$ and $\Sigma_2^{**}$, respectively. Moreover, $\theta^{**} = g(\zeta^*)$ is overparameterized in terms of the corresponding envelope parameters

$$
\zeta^* = (\text{vec}(\eta_1^*)^T, \text{vec}(L)^T, \text{vec}(\eta_2^*)^T, \text{vec}(R)^T, \text{vech}(\Omega_1^*)^T, \text{vech}(\Omega_1^{10})^T, \text{vech}(\Omega_2^*)^T)^T.
$$

Let $\hat{\theta}^{**}$ be the MLE of $\theta^{**}$ without enveloping, then $\sqrt{n}(\hat{\theta}^{**} - \theta^{**})$ converges to a normal random vector with mean zero and covariance matrix $J^{**-1}$, where $J^{**}$ is the Fisher information of $\theta^{**}$, which has the same formulation as $J^*$ except that $\beta_j^*$ and $\Sigma_j^*$ in $J^*$ are replaced by the corresponding $\beta_j^{**}$ and $\Sigma_j^{**}$. As $\hat{\theta}_e^*$ is an overparameterized estimator of $\theta^{**}$, we can again apply Proposition 4.1 in Shapiro (1986). It follows that $\sqrt{n}(\hat{\theta}_e^* - \theta^{**})$ converges to a normal random vector with mean zero and covariance matrix $G_1^*(G_1^{**T}J^{**}G_1^*)^T$, where $G_1^* = \partial g(\zeta^*)/\partial \zeta^{**T}$.

Finally, since $\hat{\theta}_e = m(\hat{\theta}_e^*)$, using the Delta method, we have $\sqrt{n}(\hat{\theta}_e - \theta) = \sqrt{n}(g(\hat{\zeta}) - g(\zeta))$ converges to a normal random vector with mean zero and covariance matrix $K^{**}G_1^*(G_1^{**T}J^{**}G_1^*)^T G_1^{**T}K^{**T}$, where $K^{**} = KC_1$ with

$$
C_1 = \text{blockdiag}(c_2 I_{p_1}, I_{mp_2}/c_2, d_2 I_{r(r+1)/2}, I_{m(m+1)/2}/d_2)
$$

and $K$ given in the proof of Proposition 3. To simplify the asymptotic variance of $g(\hat{\zeta})$, it is easy to check that $J^{**} = C_1 J_1 C_1$ and $C_1 G_1^* = G_1 C_1$, where $J_1$ is defined in the proof of Proposition 3 and $G_1 = \partial g(\zeta)/\partial \zeta^T =$

$$
\begin{pmatrix}
I_{p_1} \otimes L & \eta_1^T \otimes I_r & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & I_{p_2} \otimes R & \eta_2^T \otimes I_m & 0 & 0 & 0 & 0 \\
0 & G_{1,32} & 0 & 0 & C_r(L \otimes L)E_{u_1} & C_r(L_0 \otimes L_0)E_{r-u_1} & 0 & 0 \\
0 & 0 & G_{1,44} & 0 & 0 & G_{1,47} & G_{1,48} & \\
\end{pmatrix},
$$

with $G_{1,32} = 2C_r(L\Omega_1 \otimes I_r - L \otimes L_0\Omega_{10}^L)I_r$, $G_{1,44} = 2C_m(R\Omega_2 \otimes I_m - R \otimes R_0\Omega_{20}^R)$, $G_{1,47} = C_m(R \otimes R)E_{u_2}$ and $G_{1,48} = C_m(R_0 \otimes R_0)E_{m-u_2}$. Therefore, we have $\text{avar}(\sqrt{n}g(\hat{\zeta})) = KG_1(G_1^{**T}J_1G_1)^T G_1^{**T}K^T$. 

To show the asymptotic relative efficiency of the envelope estimator \( g(\hat{\zeta}) \) to \( \hat{\theta} \), consider 
\[
\text{avar}(\sqrt{n}\hat{\theta}) - \text{avar}(\sqrt{n}g(\hat{\zeta})) = KJ^{-1}K^T - KG_1(G_1^T J_1 G_1)^T G_1^T K^T = K(J_1^{-1} - G_1(G_1^T J_1 G_1)^T G_1^T)K^T = KJ_1^{-1/2}Q_{J_1^1/2}^G J_1^{-1/2}K^T,
\]
where \( Q_{J_1^1/2}^G \) is the projection matrix onto the orthogonal compliment of \( \text{span}(J_1^{1/2} G_1) \). Thus, \( \text{avar}(\sqrt{n}\hat{\theta}) - \text{avar}(\sqrt{n}g(\hat{\zeta})) \geq 0 \).

I Asymptotic variances of vec(\( \hat{\beta}_1 \)) and vec(\( \hat{\beta}_2 \))

Since the row and column coefficient matrices, vec(\( \beta_1 \)) and vec(\( \beta_2 \)), are often of interest in regression analysis, we give the asymptotic variances of the envelope coefficient estimators vec(\( \hat{\beta}_1 \)) and vec(\( \hat{\beta}_2 \)) in this section. Let

\[
G = \begin{pmatrix}
I_{p_1} \otimes L & \eta_1^T \otimes L_0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & I_{p_2} \otimes R & \eta_2^T \otimes R_0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & G_{32} & 0 & 0 & C_r(L \otimes L)E_{n_1} & C_r(L_0 \otimes L_0)E_{r-n_1} & 0 & 0 \\
0 & 0 & 0 & G_{44} & 0 & 0 & G_{1,47} & G_{1,48} & \\
\end{pmatrix},
\]

where \( G_{32} = 2C_r(L \Omega_1 \otimes L_0 - L \otimes L_0 \Omega_{10}) \), \( G_{44} = 2C_m(R \Omega_2 \otimes R_0 - R \otimes R_0 \Omega_{20}) \). Since \( G \) is a non-singular column transformation matrix of \( G_1 \), that is, \( \text{span}(G) = \text{span}(G_1) \), and \( \text{avar}(\sqrt{n}g(\hat{\zeta})) \) depends on \( G_1 \) only through its column space \( \text{span}(G_1) \), we have 
\[
\text{avar}(\sqrt{n}g(\hat{\zeta})) = KG(G^T J_1 G)^T G^T K.
\]

Now let \( G_{,1} = (I_{p_1} \otimes L^T, 0^T, 0^T, 0^T)^T \), \( G_{,2} = (\eta_1 \otimes L_0^T, 0^T, G_{32}^T, 0^T)^T \), \( G_{,3} = (0^T, I_{p_2} \otimes R^T, 0^T, 0^T)^T \) and \( G_{,4} = (0^T, \eta_2 \otimes R_0^T, 0^T, G_{44}^T)^T \) be the first four block columns of \( G \), and \( G^* = (G_{,5}, G_{,6}, G_{,7}, G_{,8}) \) be the matrix containing the rest block columns of \( G \). After some matrix operations, it can be shown that \( G^T J_1 G = \text{blockdiag}(W, G^* T J_1 G^*) \), where

\[
W = (G_{,1}, G_{,2}, G_{,3}, G_{,4})^T J_1 (G_{,1}, G_{,2}, G_{,3}, G_{,4}) = \begin{pmatrix} A_1 & A_2 \\ A_2^T & A_3 \end{pmatrix},
\]

with \( A_1 = \text{blockdiag}(G_{,1}^T J_1 G_{,1}, G_{,1}^T J_1 G_{,2}) \), \( A_3 = \text{blockdiag}(G_{,3}^T J_1 G_{,3}, G_{,4}^T J_1 G_{,4}) \), and

\[
A_2 = \begin{pmatrix} G_{,1}^T J_1 G_{,3} & G_{,1}^T J_1 G_{,4} \\ G_{,2}^T J_1 G_{,3} & G_{,2}^T J_1 G_{,4} \end{pmatrix}.
\]
The expressions of $A_1$, $A_2$ and $A_3$ are given by

$$G_1^T J_1 G_1 = N_1 \otimes \Omega_1^{-1}$$

$$G_2^T J_1 G_2 = \eta_1 N_1 \eta_1^T \otimes \Omega_{10}^{-1} + m(\Omega_1 \otimes \Omega_{10}^{-1} + \Omega_1^{-1} \otimes \Omega_{10} - 2I_{u_1} \otimes I_{r-u_1})$$

$$G_3^T J_1 G_3 = N_3 \otimes \Omega_2^{-1}$$

$$G_4^T J_1 G_4 = \eta_2 N_3 \eta_2^T \otimes \Omega_{20}^{-1} + r(\Omega_2 \otimes \Omega_{20}^{-1} + \Omega_2^{-1} \otimes \Omega_{20} - 2I_{u_2} \otimes I_{m-u_2})$$

$$G_1^T J_1 G_3 = (I_{p_1} \otimes L^T)N_2(I_{p_2} \otimes R)$$

$$G_1^T J_1 G_4 = (I_{p_1} \otimes L^T)N_2(\eta_2^T \otimes R_0)$$

$$G_2^T J_1 G_3 = (\eta_1 \otimes L_0^T)N_2(I_{p_2} \otimes R)$$

$$G_2^T J_1 G_4 = (\eta_1 \otimes L_0^T)N_2(\eta_2^T \otimes R_0) + 2(L^T \otimes \Omega_{10}^{-1}L_0^T - \Omega_1^{-1}L^T \otimes L_0^T)\Psi$$

$$\times (R \otimes R_0 \Omega_{20}^{-1} - R \Omega_2^{-1} \otimes R_0)K_{u_2(m-u_2)}.$$

**Proposition 7.** Under the same conditions of Proposition \ref{prop}, \(\sqrt{n}[\text{vec}(\beta_1) - \text{vec}(\beta_1)] \) and \(\sqrt{n}[\text{vec}(\beta_2) - \text{vec}(\beta_2)] \) converge in distribution to normal random vectors with mean zero and covariance matrices \(\text{avar}[^\sqrt{n} \text{vec}(\hat{\beta}_1)] = D_1 W^T D_1^T \) and \(\text{avar}[^\sqrt{n} \text{vec}(\hat{\beta}_2)] = D_2(A_3 - A_2^T A_1 \hat{A}_2)^T \), where \(D_1 = [I_{p_1} \otimes L, \eta_1^T \otimes L_0, (I_{p_1} \otimes L)\text{vec}(\eta_1)^T, 0_{r p_1 u_2(m-u_2)}] \) and \(D_2 = [(I_{p_2} \otimes L)(I_{p_2 u_2 - \text{vec}(\eta_2)^T}, \eta_2^T \otimes R_0) \]

Proof. Denote \(K_1 = (I_{r p_1}, \text{vec}(\beta_1)^T, 0, 0) \) and \(K_2 = (0, I_{m u_2 - \text{vec}(\beta_2)^T}, 0, 0) \) to be the first two block rows of \(K \). Let \(s_1 \) be the column dimension of \(G^* \) and let \(s_2 \) be the sum of the column dimensions of \(G_1 \) and \(G_2 \). Then we have \(K_1 G = (D_1, 0_{r p_1 s_1}) \) and \(K_2 G = (0_{m u_2 s_2}, D_2, 0_{m u_2 s_1}) \). Therefore, \(\text{avar}[^\sqrt{n} \text{vec}(\hat{\beta}_1)] = K_1 G (G^T J_1 G)^T G^T K_1^T = D_1 W^T D_1^T \) and \(\text{avar}[^\sqrt{n} \text{vec}(\hat{\beta}_2)] = K_2 G (G^T J_1 G)^T G^T K_2^T = (0_{m u_2 s_2}, D_2) W^T (0_{m u_2 s_2}, D_2)^T = D_2(A_3 - A_2^T A_1 \hat{A}_2)^T \).

**J Proofs of Propositions 5 and 6**

**Proof of Proposition 5** Without loss of generality, we assume that \(\mu = 0 \). Under \(\eqref{model} \) with a non-normal error \(\varepsilon \sim (0, \Sigma_1, \Sigma_2) \), it is still reasonable to estimate the parameters \(\beta_1, \beta_2, \Sigma_1, \Sigma_2 \) by minimizing the objective function \(E_n[M_{\hat{\theta}}(X, Y)] = \frac{1}{n} \sum_{i=1}^{n} M_{\hat{\theta}}(X_i, Y_i) = \frac{r}{2} \log|\tilde{\Sigma}_2| + \frac{m}{2} \log|\tilde{\Sigma}_1| + E_n\{\text{tr}[\tilde{\Sigma}_2^{-1}(Y - \tilde{\beta}_1 X \tilde{\beta}_2^T) \tilde{\Sigma}_1^{-1}(Y - \tilde{\beta}_1 X \tilde{\beta}_2^T)]\} \), \(J.1\)
over the parameter space \( \theta^* = (\text{vec}(\vec{\beta_1})^T, \text{vec}(\vec{\beta_2})^T, \text{vech}(\vec{\Sigma_1})^T, \text{vech}(\vec{\Sigma_2})^T)^T \), where

\[
M_\theta(X_i, Y_i) = \frac{r}{2} \log |\Sigma| + \frac{m}{2} \log |\Sigma_1| + \text{tr}[\Sigma^{-1}(Y_i - \vec{\beta_1}X_i\vec{\beta_2})^T\Sigma^{-1}(Y_i - \vec{\beta_1}X_i\vec{\beta_2})],
\]

and \( \vec{\beta_2} \) and \( \vec{\Sigma_2} \) are constrained to have unit Frobenius norm and to have positive elements in their first rows and columns. The objective function is reasonable because (i) at the population level, it gives exactly \( \Sigma_1 = E(\epsilon\Sigma_2^{-1}\epsilon^T)/m = E(\epsilon\epsilon^T)/\text{tr}(\Sigma_2) \) and \( \Sigma_2 = E(\epsilon^T\Sigma_1^{-1}\epsilon)/r = E(\epsilon^T\epsilon)/\text{tr}(\Sigma_1) \) for given \( \beta_1 \) and \( \beta_2 \); and (ii) the coefficient matrices \( \beta_1 \) and \( \beta_2 \) are estimated based on a generalized least square method, which is a common approach for estimating model parameters in linear regression. Since the objective function is the same as in the normal setting, the iterative algorithm proposed in Section 4.1 can still be applied for the parameter estimation although now (J.1) is not related to likelihood anymore. Correspondingly, the obtained estimators \( \hat{\theta} \) are not MLEs but instead \( \hat{M} \)-estimators.

The following proof shares the similar logic as shown in the proof of Proposition 3. For simplicity, we keep the same notations as used under the normality setting to represent \( M \)-estimators under the non-normal setting. Let \( \theta^* = (\text{vec}(B_1^*)^T, \text{vec}(B_2^*)^T, \text{vech}(S_1^*)^T, \text{vech}(S_2^*)^T)^T \) be an unconstrained \( M \)-estimator from (J.1) over the unconstrained parameter space \( \hat{\theta}^* = (\text{vec}(\hat{\mu}^*)^T, \text{vec}(\hat{\beta_1}^*)^T, \text{vec}(\hat{\beta_2}^*)^T, \text{vech}(\hat{\Sigma_1}^*)^T, \text{vech}(\hat{\Sigma_2}^*)^T)^T \). Then \( \hat{\theta} \) can be obtained by

\[
\hat{\theta} = m(\hat{\theta}^*) = (c\text{vec}(B_1^*)^T, \text{vec}(B_2^*)^T/c, d\text{vech}(S_1^*)^T, \text{vech}(S_2^*)^T/d)^T,
\]

where \( c = \text{sign}(B_{11,2})\|B_2^*\|_F, d = \text{sign}(S_{11,2})\|S_2^*\|_F \), and \( B_{11,2}^* \) and \( S_{11,2}^* \) are the elements in the first rows and columns of \( B_2^* \) and \( S_2^* \). Hence the asymptotic distribution of \( \hat{\theta} \) can be derived based on the asymptotic distribution of \( \theta^* \) and the Delta method.

Since \( \hat{\theta}^* \) is an \( M \)-estimator obtained from (J.1), based on the standard results regarding \( M \)-estimators in van der Vaart (1998), \( \hat{\theta}^* \) is \( \sqrt{n} \) consistent to a non-constrained parameter vector \( \theta^* = (\text{vec}(\beta_1^*)^T, \text{vec}(\beta_2^*)^T, \text{vech}(\Sigma_1^*)^T, \text{vech}(\Sigma_2^*)^T)^T \), where \( \theta^* \) is associated with \( \theta \) such that \( \theta = m(\theta^*) = (c_1\text{vec}(\beta_1^*)^T, \text{vec}(\beta_2^*)^T/c_1, d_1\text{vech}(\Sigma_1^*)^T, \text{vech}(\Sigma_2^*)^T/d_1)^T \), with \( c_1 = \text{sign}(\beta_{11,2})\|\beta_2^*\|_F, d_1 = \text{sign}(\Sigma_{11,2})\|\Sigma_2^*\|_F \), and \( \beta_{11,2}^* \) and \( \Sigma_{11,2}^* \) are the elements in the first rows and columns of the corresponding \( \beta_2^* \) and \( \Sigma_2^* \).

Let \( \phi_{\theta^*}(X_i, Y_i) = \partial M_{\theta^*}/\partial \theta^* = (\partial M_{\theta^*}/\partial \text{vec}(\vec{\beta_1})^T, \partial M_{\theta^*}/\partial \text{vec}(\vec{\beta_2})^T, \partial M_{\theta^*}/\partial \text{vech}(\vec{\Sigma_1})^T, \partial M_{\theta^*}/\partial \text{vech}(\vec{\Sigma_2})^T)^T \), and \( \phi_n(\theta^*) = 1/n \sum_{i=1}^n \phi_{\theta^*}(X_i, Y_i) \). Then the limiting derivative of \( \phi_n(\theta^*) \) at the true \( \theta^* \) is \( \lim_{n \to \infty} \hat{\phi}_n(\theta^*) = \lim_{n \to \infty} \partial \phi_n(\theta^*)/\partial \theta^{*T} = J^* \). The expression of \( J^* \) is the same
as the $J^*$ given in Section 4 but now $J^*$ does not stand for Fisher information. Let $\Phi_{\theta^*} = \lim_{n \to \infty} \phi_n(\theta^*)\phi_n(\theta^*)^T$ at the true $\theta^*$. According to (5.20) in van der Vaart (1998), when $J^*$ is invertible at the true $\theta^*$ and $\|\Phi_n(\theta^*)\|_F = \|\partial [J^*(\theta^*)] / \partial \theta^T\|_F$ is $O_p(1)$ at a point $\theta^*$ between the estimator $\hat{\theta}$ and the true $\theta^*$, the $M$-estimator $\hat{\theta}$ obtained from (4.1) is asymptotically normal. More specifically, $\sqrt{n}(\hat{\theta} - \theta^*)$ converges to a normal random vector with mean zero and asymptotic covariance matrix $J^*^{-1} \Phi_{\theta^*} J^*^{-1}$.

Finally, let $K^* = \partial \theta / \partial \theta^T$. Using the Delta method, we see that $\sqrt{n}(\hat{\theta} - \theta)$ converges to a normal random vector with mean zero and asymptotic covariance matrix $K^* J^*^{-1} \Phi_{\theta^*} J^*^{-1} K^T$. To obtain the final expression of the asymptotic covariance matrix of $\hat{\theta}$, it is not hard to check that $K^* = KC$, $J^* = CJ_1C$, and $\Phi_{\theta^*} = C \Phi_{\theta} C$, where $K$, $C$, and $J_1$ are given in Section 4 and $\Phi_{\theta^*} = \lim_{n \to \infty} \phi_n(\theta)\phi_n(\theta)^T$. Hence we have $\text{aov}(\sqrt{n}\hat{\theta}) = KJ_1^{-1} \Phi_{\theta^*} J_1^{-1} K^T$.

Regarding the conditions, we find that $J^*$ is invertible if and only if $J_1$ is invertible. In addition, it can be shown that $\Phi_n(\theta^*) = [(C_1 \otimes C_1) \Phi_n(\theta_1) K_1 C_1]$, where $K_1 = K(\theta)$ and $C_1 = C(\theta_1)$ are matrices $K$ and $C$ at $\theta_1$, a point between $\hat{\theta}$ and the true $\theta$, and $\Phi_n(\theta_1) = \partial \text{vec}[J_1(\theta)] / \partial \theta_1^T$ is defined in Proposition 5. Thus,

$$\|\Phi_n(\theta_1)\|_F \leq \|\Phi_n(\theta_1)\|_F : \|C_1 \otimes C_1\|_F : \|K_1 C_1\|_F = a \|\Phi_n(\theta_1)\|_F,$$

where $a$ is a constant, meaning that $\|\Phi_n(\theta_1)\|_F$ is $O_p(1)$ when $\|\Phi_n(\theta_1)\|_F$ is $O_p(1)$. This completes the proof.

**Proof of Proposition 6.** For (4.3) with a non-normal error $\varepsilon \sim (0, \Sigma_1, \Sigma_2)$, a good way to estimate the envelope parameters is to minimize (4.1) under the envelope re-parameterization. Then one can obtain the envelope estimator $g(\hat{\zeta})$ using the same algorithm proposed in Section 4.3. The rest of the proof is similar to the proof of Proposition 4. For convenience, we keep the same notations as used in Proposition 4 but now they are not associated to MLEs. Let $\hat{\theta}_e = (\text{vec}(\hat{\beta}_1)^T, \text{vec}(\hat{\beta}_2)^T, \text{vech}(\hat{\Sigma}_1)^T, \text{vech}(\hat{\Sigma}_2)^T)^T = g(\hat{\zeta})$ be the envelope $M$-estimator of $\theta$. Let $\hat{\theta}_e^* = (\text{vec}(\hat{\beta}_1)^T, \text{vec}(\hat{\beta}_2)^T, \text{vech}(\hat{\Sigma}_1)^T, \text{vech}(\hat{\Sigma}_2)^T)^T = g(\hat{\zeta}_e^*)$ be a unconstrained envelope $M$-estimator such that $\hat{\theta}_e^* = m(\hat{\theta}_e)^* = (\text{cvec}(\hat{\beta}_1)^T, \text{vec}(\hat{\beta}_2)^T/c, \text{dvec}(\hat{\Sigma}_1)^T, \text{vech}(\hat{\Sigma}_2)^T/d)^T$, where $\hat{\beta}_j$ and $\hat{\Sigma}_j$ are over-parameterized in terms of

$$\hat{\zeta}^* = (\text{vec}(\hat{\eta}_1)^T, \text{vec}(\hat{L})^T, \text{vec}(\hat{\eta}_2)^T, \text{vec}(\hat{R})^T, \text{vech}(\hat{\Omega}_1)^T, \text{vech}(\hat{\Omega}_2)^T, \text{vech}(\hat{\Omega}_{10})^T, \text{vech}(\hat{\Omega}_2)^T, \text{vech}(\hat{\Omega}_{20})^T)^T,$$

c = \text{sign}(\hat{\beta}_{11,2}^* \|\hat{\beta}_2^*\|_F, d = \text{sign}(\hat{\Sigma}_{11,2}^* \|\hat{\Sigma}_2^*\|_F$, and $\hat{\beta}_{11,2}^*$ and $\hat{\Sigma}_{11,2}^*$ are the elements in the first
rows and columns of $\hat{\beta}_2^*$ and $\hat{\Sigma}_2^*$, respectively. Then there exists a non-constrained parameter vector

$$\theta^{**} = (\text{vec}(\beta_1^{**})^T, \text{vec}(\beta_2^{**})^T, \text{vech}(\Sigma_1^{**})^T, \text{vech}(\Sigma_2^{**})^T)^T,$$

such that $\hat{\theta}_e^*$ is $\sqrt{n}$ consistent to $\theta^{**}$ and in addition

$$\theta = m(\theta^{**}) = (c_2 \text{vec}(\beta_1^{**})^T, \text{vec}(\beta_2^{**})^T / c_2, d_2 \text{vech}(\Sigma_1^{**})^T, \text{vech}(\Sigma_2^{**})^T / d_2)^T,$$

where $c_2 = \text{sign}(\beta_{11,2}^{**})\|\beta_2^{**}\|_F$, $d_2 = \text{sign}(\Sigma_{11,2}^{**})\|\Sigma_2^{**}\|_F$ with $\beta_{11,2}^{**}$ and $\Sigma_{11,2}^{**}$ being the elements in the first rows and columns of $\beta_2^{**}$ and $\Sigma_2^{**}$. Moreover, $\theta^{**} = g(\zeta^*)$ is overparameterized in terms of the corresponding envelope parameters

$$\zeta^* = (\text{vec}(\eta_1^*)^T, \text{vec}(L)^T, \text{vec}(\eta_2^*)^T, \text{vec}(R)^T, \text{vech}(\Omega_1^*)^T, \text{vech}(\Omega_{10}^*)^T, \text{vech}(\Omega_2^*)^T)^T.$$

Let $\hat{\theta}_1^*$ be the $M$-estimator of $\theta^{**}$ without enveloping. Based on the result from Proposition 5, $\sqrt{n}(\hat{\theta}_1^* - \theta^{**})$ converges to a normal random vector with mean zero and covariance matrix $J^{**-1}\Phi_{\theta^{**}} J^{**-1}$, where $J^{**}$ and $\Phi_{\theta^{**}}$ are same defined as $J^*$ and $\Phi_{\theta^*}$ in Proposition 5 except replacing $\theta^*$ with $\theta^{**}$. Since $\hat{\theta}_e^*$ is an over-parameterized estimator of $\theta^{**}$, applying Proposition 4.1 in Shapiro (1986), we have $\sqrt{n}(\hat{\theta}_e^* - \theta^{**})$ converges to a normal random vector with mean zero and covariance matrix $\Xi^* = G_1^*(G_1^{**T} J^{**} G_1^*) G_1^{**T} \Phi_{\theta^{**}} G_1^*(G_1^{**T} J^{**} G_1^*) G_1^{**T}$, where $G_1^* = \partial g(\zeta^*) / \partial \zeta^{**T}$.

Finally, let $K^{**} = \partial \theta / \partial \theta^{**}$. The Delta method indicates that $\sqrt{n}(\hat{\theta}_e - \theta) = \sqrt{n}(g(\hat{\zeta}) - g(\zeta))$ converges to a normal random vector with mean zero and covariance matrix $K^{***\Xi* K^{**T}}$. Furthermore, it can shown that $K^{**} = KC_1$, $J^{**} = C_1 J_1 C_1$, $G_1^* = C_1^{-1} G_1 C_1$, and $\Phi_{\theta^{**}} = C_1 \Phi_{\theta} C_1$, where $C_1$ = blockdiag$(c_2 I_{r_1}, I_{mp_2}/c_2, d_2 I_{(r+1)/2}, I_{m(m+1)/2}/d_2)$ is a block diagonal matrix, $K$, $J_1$ and $G_1$ are defined in Section H, and $\Phi_{\theta}$ is given in the proof of Proposition 5. Hence

$$\text{avarp}(\sqrt{n}g(\hat{\zeta})) = KG_1(G_1^{**T} J_1 G_1)^{\dagger} G_1^{**T} \Phi_{\theta} G_1(G_1^{**T} J_1 G_1)^{\dagger} G_1^{**T} K^{T}.$$

To show the the asymptotic efficiency of $g(\hat{\zeta})$ relative to $\hat{\theta}$, consider $\text{avarp}(\sqrt{n}\hat{\theta}) - \text{avarp}(\sqrt{n}g(\hat{\zeta})) = K\Gamma K^{T}$, where

$$\Gamma = J_1^{-1} \Phi_{\theta} J_1^{-1} - G_1(G_1^{**T} J_1 G_1)^{\dagger} G_1^{**T} \Phi_{\theta} G_1(G_1^{**T} J_1 G_1)^{\dagger} G_1^{**T}$$

$$= J_1^{1/2} J_1^{1/2} \Phi_{\theta} J_1^{-1/2} - P J_1^{1/2} G_1 J_1^{1/2} \Phi_{\theta} J_1^{1/2} P J_1^{1/2} G_1 J_1^{1/2}.$$
and $P_{J_1^{1/2}G_1}$ is the projection matrix onto $\text{span}(J_1^{1/2}G_1)$. When $\text{span}(J_1^{1/2}G_1)$ is a reducing subspace of $J_1^{1/2} \Phi \theta J_1^{1/2}$, $\Gamma \geq 0$. Thus, $\text{avar}(\sqrt{n} \hat{\theta}) - \text{avar}(\sqrt{n} \hat{g}(\hat{\zeta})) \geq 0$.

K Additional simulation results in Section 7

We assess the performance of the two matrix regression models (2.2) and (4.3) in terms of the individual row and column parameter estimations, particularly the estimation of $\beta_1$ and $\beta_2$. The data were generated in the same setting of Section 7 but to make the row and column parameter matrices identifiable, we normalized $\beta_2$ and $\Sigma_2$ such that each of them has Frobenius norm 1 and positive element in its first row and column. We then evaluated the estimation accuracy of $\beta_1$ and $\beta_2$ according to the criterion $|| \hat{\beta}_j - \beta_j ||_F$, $j = 1, 2$, for both models. The results are shown in Figure 5, which demonstrates similar patterns as the right panel of Figure 2 wherein the envelope model outperformed the matrix regression (2.2) for both $\beta_1$ and $\beta_2$ at all sample sizes.

![Figure 5: The average estimation errors of $\beta_1$ and $\beta_2$ from the envelope matrix regression (4.3) and the matrix regression (2.2).](image)

To demonstrate the substantial efficiency gains of the envelope model, we further compared the standard errors of the first elements in vec($\hat{\beta}_1$) and vec($\hat{\beta}_2$) obtained from (4.3) and (2.2). The asymptotic standard errors were estimated based on the results from Propositions 3 and 4. The results in Figure 6 show similar patterns as the right panel of Figure 3. For other elements in $\hat{\beta}_1$ and $\hat{\beta}_2$, the comparison results hold similarly. This again supports the accuracy of the asymptotic covariance matrices given in Section 5. In addition, the envelope model shows
asymptotic efficiency over all selected sample sizes. The ratios of the standard errors between the estimators from (2.2) and (4.3) over all elements in $\beta_1$ range from 2.7 to 11.8. The ratios of the standard errors between the corresponding estimators of $\beta_2$ range from 2.5 to 27.2.

Figure 6: The standard errors of the first elements in $\text{vec}(\hat{\beta}_1)$ and $\text{vec}(\hat{\beta}_2)$ obtained from (4.3) and (2.2). In both panels, the top three curves represent the standard errors from (2.2), and the bottom three curves indicate the standard errors from the envelope model (4.3).

L Simulation studies for matrix-variate regressions in high dimension without sparsity

To evaluate the numerical performance of (2.2) and (4.3) in high dimension settings, we varied the simulation setting in Section 7 with $Y \in \mathbb{R}^{50 \times 50}$, $X \in \mathbb{R}^{50 \times 50}$, $n = 100$, and kept the rest unchanged. In this case, since the dimensions of $\text{vec}(Y)$ and $\text{vec}(X)$ are 2500, and sample size is 100 and is much smaller, the conventional multivariate regression model (2.1) and Li and Zhang’s method are not estimable. We thus only applied the matrix-variate regression (2.2) and the envelope model (4.3) for the estimation. Table 3 shows the average estimation errors for the corresponding parameters measured in terms of the Frobenius norm of the estimates minus the truth over 100 random samples. The last two columns of the table show the average angles between the estimated envelopes and true envelopes.

It can be seen that in the high dimension setting, both the matrix-variate regression and envelope matrix-variate regression can still estimate the parameters accurately, while the envelope model improves the estimation dramatically.
### Table 3: Comparison of average estimation errors over 100 runs.

<table>
<thead>
<tr>
<th>Method</th>
<th>$\beta_2 \otimes \beta_1$</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$P_L$</th>
<th>$P_R$</th>
<th>angle $L$</th>
<th>angle $R$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matrix reg. (2.2)</td>
<td>13.98</td>
<td>11.53</td>
<td>0.07</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Env. matrix reg. (4.3)</td>
<td>1.71</td>
<td>1.25</td>
<td>0.01</td>
<td>0.02</td>
<td>0.02</td>
<td>0.67</td>
<td>0.67</td>
</tr>
<tr>
<td>Sparse env. est.</td>
<td>0.86</td>
<td>0.40</td>
<td>0.005</td>
<td>0.03</td>
<td>0.008</td>
<td>0.66</td>
<td>0.27</td>
</tr>
</tbody>
</table>

### M Simulation studies for sparse envelope matrix-variate regression in high dimension

The following simulation studies evaluate the performance of the sparse envelope matrix regression method. We considered two simulation settings:

Case 1: $Y \in \mathbb{R}^{200 \times 100}$, $X \in \mathbb{R}^{15 \times 10}$, $r_1 = 20$, $m_1 = 10$, $u_1 = 5$, $u_2 = 2$, and $n = 100$;

Case 2: $Y \in \mathbb{R}^{100 \times 100}$, $X \in \mathbb{R}^{50 \times 50}$, $r_1 = 10$, $m_1 = 5$, $u_1 = u_2 = 2$, and $n = 100$.

For each case, we generated data based on (4.3). Specifically, the semi-orthogonal matrices $L$ and $R$ each were generated from two parts with $L_{A_1}$ and $R_{A_2}$ obtained by orthogonalizing matrices of independent uniform (0,1) random variables, and with $L_{A_1}$ and $R_{A_2}$ set to be zero matrices. The rest of the parameters were generated in the same way as in the previous simulation setting. Tables 4 and 5 summarize the comparison results, each averaged over 100 random samples, based on models (2.2), (4.3) and the sparse envelope estimation method. Note that as discussed in Section 2.2, (2.2) and (4.3) are still applicable in the high dimensional setting. We see from Tables 4 and 5 that for both cases, the sparse envelope approach provides the smallest estimation errors among the three methods, while the nonsparse envelope matrix regression (4.3) outperforms the matrix regression (2.2) without enveloping.

Table 4: Comparison of average estimation errors for Case 1 (over 100 runs).

<table>
<thead>
<tr>
<th>Method</th>
<th>$\beta_2 \otimes \beta_1$</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$P_L$</th>
<th>$P_R$</th>
<th>angle $L$</th>
<th>angle $R$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matrix reg. (2.2)</td>
<td>8.46</td>
<td>7.43</td>
<td>0.10</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Env. matrix reg. (4.3)</td>
<td>1.28</td>
<td>1.15</td>
<td>0.01</td>
<td>0.09</td>
<td>0.03</td>
<td>1.82</td>
<td>0.87</td>
</tr>
<tr>
<td>Sparse env. est.</td>
<td>0.86</td>
<td>0.40</td>
<td>0.005</td>
<td>0.03</td>
<td>0.008</td>
<td>0.66</td>
<td>0.27</td>
</tr>
</tbody>
</table>

Table 6 reports the variable selection results for Case 1 and Case 2 over 100 repeated samples. We use $\bar{C}_L$ and $\bar{T}_L$ to denote the average numbers of correctly selected rows and incorrectly
Table 5: Comparison of average estimation errors for Case 2 (over 100 runs).

<table>
<thead>
<tr>
<th>Method</th>
<th>$\beta_2 \otimes \beta_1$</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$P_L$</th>
<th>$P_R$</th>
<th>angle L</th>
<th>angle R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matrix reg. (2.2)</td>
<td>17.35</td>
<td>13.55</td>
<td>0.12</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Env. matrix reg. (4.3)</td>
<td>1.98</td>
<td>1.40</td>
<td>0.02</td>
<td>0.03</td>
<td>0.03</td>
<td>0.94</td>
<td>0.88</td>
</tr>
<tr>
<td>Sparse env. est.</td>
<td>0.86</td>
<td>0.70</td>
<td>0.005</td>
<td>0.01</td>
<td>0.005</td>
<td>0.34</td>
<td>0.19</td>
</tr>
</tbody>
</table>

selected rows in $\beta_1$, and use $C_R$ and $TC_R$ to denote the average numbers of correctly selected rows and incorrectly selected rows in $\beta_2$. The proposed sparse envelope estimation procedure correctly identified all the active sets and obtained no false discovery for both cases.

Table 6: Variable selection results for Case 1 and Case 2 (over 100 runs).

<table>
<thead>
<tr>
<th>Case</th>
<th>$C_L$</th>
<th>$C_R$</th>
<th>$TC_L$</th>
<th>$TC_R$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1</td>
<td>20</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Case 2</td>
<td>10</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

N Envelope dimension selection

In this section, we evaluate the performance of AIC, BIC, and cross validation for envelope dimension selection and propose stepwise selection methods for envelope dimension selection in high dimension. The idea of stepwise selection is similar to that in linear regression. Here we mainly focus on forward stepwise selection. Specifically, (1) we start from the smallest model with the envelope dimensions $u_1^{(0)} = u_2^{(0)} = 0$, and compute the value of the model selection criterion, for example, AIC, or BIC, or mean squared prediction error (MSPE) for cross valuation (CV). Set $t = 0$. (2) Update $u_1^{(t+1)}$ and $u_2^{(t+1)}$. Increase the envelope dimensions $u_1^{(t)}$ and $u_2^{(t)}$ by one respectively. That is, consider two cases: i) $u_1^{(t+1)} = u_1^{(t)} + 1$ and $u_2^{(t+1)} = u_2^{(t)}$, and ii) $u_1^{(t+1)} = u_1^{(t)}$ and $u_2^{(t+1)} = u_2^{(t)} + 1$. Compute the AIC (or BIC, or MSPE) for i) and ii), and then take the case with the smaller AIC (or BIC, or MSPE) for round $t + 1$. (3) Compare round $t + 1$ with round $t$, if there is no decrease in term of AIC (or BIC, or MSPE) after adding dimension, stop and set the final envelope dimensions to be $u_1^{(t+1)}$ and $u_2^{(t+1)}$; otherwise, set $t = t + 1$ and repeat (2) until no further decrease in model selection criterion.
We have completed simulation studies for model selection in both low and high dimension settings. The low dimension setting, referred as Case A, is the same as the simulation setting given in Section 7 of the manuscript, where the response is a $5 \times 5$ random matrix, the predictor is a $5 \times 5$ fixed matrix, and the envelope dimensions $u_1 = u_2 = 2$. In addition to likelihood ratio test (LRT), we used both full subset selection ($u_1$ and $u_2$ in the range [0,5]) and forward stepwise selection via AIC, BIC, and 5 fold-CV for this setting. For the high dimension setting, referred as Case B, we varied the response to be $50 \times 20$, the predictor to be $10 \times 10$, and $u_1 = u_2 = 5$. The rest of the setting remains the same except that we confine the range of $u_1$ and $u_2$ to be [0, 50] and [0,20]. We mainly used stepwise selection for evaluation in this setting. The CV was 5 fold-CV. The frequencies of the correct envelope dimension selection are summarized in Tables 7 and 9 which are based 100 random repetitions. The frequencies of selecting dimensions greater than or equal to the true dimensions are reported in Tables 8 and 10.

Table 7: Frequencies of correct envelope dimension selection for Case A ($Y \in \mathbb{R}^{5\times5}$ and $u_1 = u_2 = 2$) over 100 runs.

<table>
<thead>
<tr>
<th>Method</th>
<th>sample size n</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>200</td>
</tr>
<tr>
<td>AIC</td>
<td>49</td>
</tr>
<tr>
<td>BIC</td>
<td>100</td>
</tr>
<tr>
<td>LRT</td>
<td>100</td>
</tr>
<tr>
<td>CV</td>
<td>100</td>
</tr>
<tr>
<td>stepwise AIC</td>
<td>54</td>
</tr>
<tr>
<td>stepwise BIC</td>
<td>100</td>
</tr>
<tr>
<td>stepwise CV</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 7 indicates that the stepwise selection procedures are comparable to the corresponding full subset selection procedures. All the procedures can correctly identify the true envelope dimensions at very high percentages, except for AIC that tends to overfit and to select higher envelope dimensions than the truth, as shown in Table 8. In the case of overfitting, we might incur extra variation but no asymptotic bias. There is still some efficiency gains but the gains might not be maximum, so the selection error needs not be worrisome.

Table 9 further shows that in the high dimensional setting, the proposed stepwise selection method performs fairly well in terms of BIC and CV, while stepwise AIC tends to overfit. In fact, as shown in Table 10, both stepwise AIC and BIC suggest higher envelope dimensions
Table 8: Frequencies of selecting envelope dimensions greater than or equal to the true dimensions for Case A over 100 runs.

<table>
<thead>
<tr>
<th>Method</th>
<th>sample size n</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>200</td>
</tr>
<tr>
<td>AIC</td>
<td>100</td>
</tr>
<tr>
<td>BIC</td>
<td>100</td>
</tr>
<tr>
<td>LRT</td>
<td>100</td>
</tr>
<tr>
<td>CV</td>
<td>100</td>
</tr>
<tr>
<td>stepwise AIC</td>
<td>100</td>
</tr>
<tr>
<td>stepwise BIC</td>
<td>100</td>
</tr>
<tr>
<td>stepwise CV</td>
<td>100</td>
</tr>
</tbody>
</table>

when they are not able to identify the truth. So the selection error causes nothing to lose compared to the model without enveloping.

Table 9: Frequencies of correct envelope dimension selection for Case B ($Y \in \mathbb{R}^{50 \times 20}$ and $u_1 = u_2 = 5$) over 100 runs.

<table>
<thead>
<tr>
<th>Method</th>
<th>sample size n</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>200</td>
</tr>
<tr>
<td>stepwise AIC</td>
<td>2</td>
</tr>
<tr>
<td>stepwise BIC</td>
<td>91</td>
</tr>
<tr>
<td>stepwise CV</td>
<td>100</td>
</tr>
</tbody>
</table>

For non-normal errors, both information criteria and LRT will fail. We propose to use cross valuation to select the optimal envelope dimensions, and to use stepwise CV for the high dimensional setting. We completed the simulation study by varying the matrix normal error to matrix t distributed error (df=20) in the above high dimensional setting and retaining the rest the same. We refer to this setting as Case C. The simulation results are summaries in Tables 11 and 12. For comparison, we also reported the results from stepwise AIC and BIC. Again, stepwise CV correctly identifies the true envelope dimensions at high percentages over all selected sample sizes. The empirical performance of stepwise BIC is robust to the non-normal error case, especially when sample size gets large, while the AIC procedure still tends to overfit. Similar to the results in Case B, Table 12 shows that both stepwise AIC and BIC select greater envelope dimensions when they are not able to identify the truth.
Table 10: Frequencies of selecting envelope dimensions greater than or equal to the true dimensions for Case B over 100 runs.

<table>
<thead>
<tr>
<th>Method</th>
<th>sample size n</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>200</td>
</tr>
<tr>
<td>stepwise AIC</td>
<td>100</td>
</tr>
<tr>
<td>stepwise BIC</td>
<td>100</td>
</tr>
<tr>
<td>stepwise CV</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 11: Frequencies of correct envelope dimension selection for Case C ($Y \in \mathbb{R}^{50 \times 20}$, $u_1 = u_2 = 5$ and matrix t distributed error) over 100 runs.

<table>
<thead>
<tr>
<th>Method</th>
<th>sample size n</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>200</td>
</tr>
<tr>
<td>stepwise AIC</td>
<td>1</td>
</tr>
<tr>
<td>stepwise BIC</td>
<td>76</td>
</tr>
<tr>
<td>stepwise CV</td>
<td>98</td>
</tr>
</tbody>
</table>

Table 12: Frequencies of selecting envelope dimensions greater than or equal to the true dimensions for Case C over 100 runs.

<table>
<thead>
<tr>
<th>Method</th>
<th>sample size n</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>200</td>
</tr>
<tr>
<td>stepwise AIC</td>
<td>100</td>
</tr>
<tr>
<td>stepwise BIC</td>
<td>100</td>
</tr>
<tr>
<td>stepwise CV</td>
<td>98</td>
</tr>
</tbody>
</table>

We report the computation time of a single envelope model fitting and of the model selection of each method for Cases A and B in Tables 13-15. The computing time for Case C is similar to that for Case B. As shown in these tables, the envelope matrix regression method is computationally fast.
Table 13: Average computing time (in seconds) of fitting a single envelope model after model selection for Cases A and B (over 100 runs).

<table>
<thead>
<tr>
<th>sample size n</th>
<th>Setting</th>
<th>200</th>
<th>300</th>
<th>500</th>
<th>800</th>
<th>1000</th>
<th>1500</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case A</td>
<td>0.29</td>
<td>0.40</td>
<td>0.56</td>
<td>0.84</td>
<td>0.98</td>
<td>1.42</td>
<td></td>
</tr>
<tr>
<td>Case B</td>
<td>1.22</td>
<td>1.54</td>
<td>2.22</td>
<td>3.03</td>
<td>3.49</td>
<td>5.01</td>
<td></td>
</tr>
</tbody>
</table>

Table 14: Average computing time (in seconds) of model selection for Case A (over 100 runs).

<table>
<thead>
<tr>
<th>sample size n</th>
<th>Setting</th>
<th>200</th>
<th>300</th>
<th>500</th>
<th>800</th>
<th>1000</th>
<th>1500</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>8.49</td>
<td>12.10</td>
<td>16.03</td>
<td>23.68</td>
<td>28.64</td>
<td>41.10</td>
<td></td>
</tr>
<tr>
<td>BIC</td>
<td>8.49</td>
<td>12.10</td>
<td>16.03</td>
<td>23.68</td>
<td>28.64</td>
<td>41.10</td>
<td></td>
</tr>
<tr>
<td>LRT</td>
<td>3.80</td>
<td>5.46</td>
<td>7.18</td>
<td>10.74</td>
<td>12.79</td>
<td>18.80</td>
<td></td>
</tr>
<tr>
<td>CV</td>
<td>27.91</td>
<td>32.57</td>
<td>33.86</td>
<td>43.88</td>
<td>48.41</td>
<td>65.62</td>
<td></td>
</tr>
<tr>
<td>stepwise AIC</td>
<td>3.49</td>
<td>5.10</td>
<td>6.63</td>
<td>9.84</td>
<td>11.59</td>
<td>17.38</td>
<td></td>
</tr>
<tr>
<td>stepwise BIC</td>
<td>3.08</td>
<td>4.41</td>
<td>5.84</td>
<td>8.81</td>
<td>10.50</td>
<td>15.32</td>
<td></td>
</tr>
</tbody>
</table>

Table 15: Average computing time (in seconds) of model selection for Case B (over 100 runs).

<table>
<thead>
<tr>
<th>sample size n</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>200</td>
</tr>
<tr>
<td>stepwise AIC</td>
<td>136.08</td>
</tr>
<tr>
<td>stepwise BIC</td>
<td>105.65</td>
</tr>
<tr>
<td>stepwise CV</td>
<td>401.88</td>
</tr>
</tbody>
</table>

While the stepwise selection method can be useful for the high dimensional setting, it might inherit some issue from general stepwise selection procedure. Theoretical justification will be pursued in future work, but we think these simulation results are a good indicator of the efficacy of the methods.
O Envelope matrix-variate regression under misspecified envelope dimensions

We further investigate the performance of the envelope matrix regression under misspecified envelope dimensions. In the following simulation study, we generated data from the same setting in Section 7 with sample size \( n = 200 \), and computed the envelope estimators of \( \beta_1 \) and \( \beta_2 \) under over- and under-specified envelope dimensions. Then the bootstrap estimation variance and estimation mean squared error (MSE) were calculated for each envelope estimator. The bootstrap variances were computed based on 100 bootstraps. Table 16 summarizes the results for a randomly chosen element in \( \beta_1 \) and \( \beta_2 \). Note that the true envelope dimension in the simulation study is \((u_1, u_2) = (2, 2)\). When \((u_1, u_2)\) is over-specified to be \((5, 5)\), which is the full data dimension as \( Y \in \mathbb{R}^{5 \times 5} \), there is no immaterial information eliminated from the model fitting. Thus the envelope model (4.3) coincides with the matrix regression model (2.2).

We see from Table 16 that compared to the estimator of (2.2) at the full dimension \((5, 5)\), the envelope estimations in general lead to smaller estimation variances even when the envelope dimension is misspecified, because enveloping eliminates immaterial information or partial immaterial information from estimation. In addition, the envelope estimators under over-specified dimensions generally have smaller estimation bias in comparison to the under-specified estimators. This because overestimation leads to asymptotically unbiased estimators while underestimation leads to inconsistency.

Based on the bias-variance tradeoff, the envelope estimators can still have smaller MSEs under over- or under-specified dimensions than the estimator of (2.2), as shown in Table 16. Under over-specified dimensions, there is no loss using envelope estimation as the estimators have both smaller bias and variances. Under under-specified dimensions, even if the envelope estimators might have larger bias, they generally have smaller estimation variances. Then if the reduction of the estimation variances is substantial, the envelope estimators will give smaller MSEs. In fact, as shown in Section N the proposed dimension selection methods can correctly identify the true envelope dimensions at very high percentages. Even when they fail, they tend to over-estimate the dimensions. Hence it is worthwhile to employ envelope estimation.

In addition, Eck and Cook (2017) proposed a weighted envelope estimation approach to circumvent selection volatility in multivariate model estimation. The idea is to first obtain envelope estimators under different envelope dimensions and then to combine estimators by
Table 16: Bootstrap estimation variances and MSEs for envelope estimators under different \( u_1 \) and \( u_2 \). All the numbers \( \times 10^{-3} \) are the observed values.

<table>
<thead>
<tr>
<th>((u_1, u_2))</th>
<th>(1,1)</th>
<th>(1,2)</th>
<th>(2,1)</th>
<th>(2,2)</th>
<th>(2,3)</th>
<th>(3,3)</th>
<th>(5,5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE ((\beta_1))</td>
<td>10.953</td>
<td>11.494</td>
<td>6.151</td>
<td>1.657</td>
<td>1.738</td>
<td>1.657</td>
<td>1.708</td>
</tr>
<tr>
<td>Boot. Var. ((\beta_2))</td>
<td>0.038</td>
<td>0.036</td>
<td>0.019</td>
<td>0.015</td>
<td>0.015</td>
<td>0.045</td>
<td>0.046</td>
</tr>
<tr>
<td>MSE ((\beta_2))</td>
<td>0.422</td>
<td>0.062</td>
<td>0.079</td>
<td>0.016</td>
<td>0.046</td>
<td>0.016</td>
<td>0.046</td>
</tr>
</tbody>
</table>

weighting with BIC weights. With this weighting scheme, Eck and Cook (2017) showed that the weighted estimator is \( \sqrt{n} \) consistent, and the residual bootstrap with respect to this estimator is asymptotically equivalent to the residual bootstrap with respect to the envelope estimator under the true dimension. Therefore, the weighted estimator can useful as it requires no dimension selection and the asymptotic efficiency is guaranteed. In the following, we extend the weighting approach to the matrix-variate setting and evaluate its numerical performance. We define the weighted estimators in the form

\[
\hat{\beta}_{w,l} = \sum_{i=1}^{r} \sum_{j=1}^{m} w_{ij} \beta_{l}^{(i,j)}, \ldots, l = 1, 2,
\]

where \( \sum_{i=1}^{r} \sum_{j=1}^{m} w_{ij} = 1, w_{ij} \geq 0 \) for \( i = 1, \ldots, r \) and \( j = 1, \ldots, m \), and \( \beta_{l}^{(i,j)} \) is the envelope estimator under the envelope dimension \((i, j)\). The weights \( w_{ij} \) are the BIC weights of the form

\[
w_{ij} = \frac{\exp(-B_{ij})}{\sum_{k=1}^{r} \sum_{g=1}^{m} \exp(-B_{kg})},
\]

where \( B_{ij} \) is the value of BIC for the envelope model under dimension \((i, j)\). Table 17 reports the comparison results of the weighted estimator, the envelope estimator under true dimension, and the estimator without enveloping for two randomly chosen elements in \( \beta_1 \) and \( \beta_2 \). The data are the same as used for generating Table 16.

It can be seen that the weighted envelope estimators are close to the envelope estimators at the true dimension in terms of estimation variance and mean squared error. Both the methods outperform the estimator from model \((2.2)\) without enveloping.
Table 17: Bootstrap estimation variances and MSEs for the weighted envelope estimator, and the envelope estimators at the true dimension \((u_1, u_2) = (2, 2)\) and at the full dimension \((u_1, u_2) = (5, 5)\). All the numbers \(\times 10^{-3}\) are the observed values.

<table>
<thead>
<tr>
<th></th>
<th>Est. at (2,2)</th>
<th>Weighted Est.</th>
<th>Est. at (5,5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boot.Var. ((\beta_1))</td>
<td>1.200</td>
<td>0.638</td>
<td>14.478</td>
</tr>
<tr>
<td>MSE ((\beta_1))</td>
<td>1.657</td>
<td>0.748</td>
<td>20.178</td>
</tr>
<tr>
<td>Boot.Var. ((\beta_2))</td>
<td>0.015</td>
<td>0.015</td>
<td>0.021</td>
</tr>
<tr>
<td>MSE ((\beta_2))</td>
<td>0.016</td>
<td>0.025</td>
<td>0.022</td>
</tr>
</tbody>
</table>

P Multivariate bioassay data

We applied the proposed matrix-variate regression models to a cross-over assay of insulin based on rabbit blood sugar concentration (Vølund, 1980). The design partitioned the animals into four groups of nine rabbits each, with a different treatment assigned to each group. Hence \(n = 36\). Let \(K_1\) and \(K_2\) denote the low and high dose levels, 0.75 units and 1.5 units of the standard treatment, and let \(T_1\) and \(T_2\) denote the same two dose levels of the test treatment. The treatment assignment is shown in Table 18.

<table>
<thead>
<tr>
<th>Group</th>
<th>1st day</th>
<th>2nd day</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(K_1)</td>
<td>(T_2)</td>
</tr>
<tr>
<td>2</td>
<td>(K_2)</td>
<td>(T_1)</td>
</tr>
<tr>
<td>3</td>
<td>(T_1)</td>
<td>(K_2)</td>
</tr>
<tr>
<td>4</td>
<td>(T_2)</td>
<td>(K_1)</td>
</tr>
</tbody>
</table>

After injection of the insulin dose each day, the blood sugar concentration of each rabbit was measured at 0, 1, 2, 3, 4, and 5 hours. We consider the percentage decreases of the blood sugar concentrations at 1, 2, 3, 4, and 5 hours relative to the initial concentration at 0 hours. The measurements on each rabbit then form a matrix \(Y \in \mathbb{R}^{5 \times 2}\), where the two columns represent the percentage decreases under two different treatments assigned on day 1 and day 2, and the rows indicate the hourly percentage decreases each day.

We modeled the insulin assay data with treatment effects and dose levels as covariates to explore how these factors affect the percentage decreases of the blood sugar concentrations. We defined the covariates of each rabbit as a matrix \(X \in \mathbb{R}^{2 \times 2}\), where the two rows of \(X\)
represent standard and the test treatments, and the two columns of $X$ represent day 1 and day 2. For example, for the rabbits in group 1, the covariate matrix $X = \text{diag}(0.75, 1.5)$, a diagonal matrix with elements 0.75 and 1.5. That is because the rabbits in group 1 received the standard treatment at the low dose level 0.75 in day 1, and received the test treatment at the high dose level 1.5 in day 2. Similarly, for the rabbits in groups 2, 3, and 4, the covariate matrices can be formed as $\text{diag}(1.5, 0.75)$, anti-$\text{diag}(1.5, 0.75)$ and anti-$\text{diag}(0.75, 1.5)$. We then fitted the relationship between the response $Y \in \mathbb{R}^{5 \times 2}$ and the predictor $X \in \mathbb{R}^{2 \times 2}$ based on (2.2), (B.1) and (2.1) as:

\[
Y = \mu + \beta_1 X^T \beta_2 + \epsilon, \quad \text{(P.1)}
\]

\[
Y = \mu + \beta X + \epsilon, \quad \text{(P.2)}
\]

\[
\text{vec}(Y) = \text{vec}(\mu) + B \text{vec}(X) + \text{vec}(\epsilon), \quad \text{(P.3)}
\]

where $\mu \in \mathbb{R}^{5 \times 2}$ is the intercept matrix, $\beta_1 \in \mathbb{R}^{5 \times 2}$, $\beta_2 \in \mathbb{R}^{2 \times 2}$, $\beta \in \mathbb{R}^{5 \times 2}$, and $B \in \mathbb{R}^{10 \times 4}$ are the corresponding coefficient matrices in the three models, and $\epsilon$ is the random error that is assumed to be matrix normal in the matrix-variate regression models (P.1) and (P.2), and $\text{vec}(\epsilon)$ is assumed to be multivariate normal under the multivariate regression (P.3). Since the blood sugar concentration levels observed in the first day cannot be affected by the treatment received in the second day, the elements in the upper-right corner of $\beta_2$ in (P.1) and of $B$ in (P.3) should be zeros. We estimated the two models under this setting.

Because (P.3), (P.1) and (P.2) are nested models, we applied likelihood ratio tests to select the best model among the three and concluded that (P.2) was sufficient. We then performed envelope fitting for (P.2) to achieve efficiency gains for the parameter estimation. Using BIC, the envelope dimension was selected to be $u_1 = 1$. The estimates (EST) and the asymptotic standard errors (SE) for each element in vec($\hat{\beta}$) are summarized in Table 19 with subscripts ‘e’ and ‘s’ indicating the results from the envelope fit and the fit from (P.2) without enveloping. In comparison to the fit from (P.2), the envelope model shows smaller standard errors for all the elements in vec($\hat{\beta}$). All the parameters in vec($\beta$) are significant based on the two models. Yet the magnitudes of the ratios EST/SE from the envelope fitting are mostly higher than those from the fitting without enveloping.

We further evaluated the sum of squared prediction error (SSPE) using 10-fold cross validation. The results based on the average testing set prediction error $\|\hat{Y}_{test} - Y\|_F$ are summarized in Figure 7, which shows that the envelope regression dominates for all choices of the envelope dimension $u_1$. The smallest prediction error occurs at the optimal dimension $u_1 = 1$. When
Table 19: Comparison of vec(\(\hat{\beta}\)) and the standard errors of vec(\(\hat{\beta}\)) based on the envelope fit and the fit from (P.2) without enveloping.

<table>
<thead>
<tr>
<th>vec((\hat{\beta}))</th>
<th>Envelope model</th>
<th>Model without enveloping</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EST(_e)</td>
<td>SE(_e)</td>
<td>EST(_e)/SE(_e)</td>
</tr>
<tr>
<td>7.74</td>
<td>1.78</td>
<td>4.35</td>
<td>6.34</td>
</tr>
<tr>
<td>13.07</td>
<td>2.48</td>
<td>5.28</td>
<td>14.42</td>
</tr>
<tr>
<td>18.98</td>
<td>3.22</td>
<td>5.9</td>
<td>21.12</td>
</tr>
<tr>
<td>17.41</td>
<td>3.11</td>
<td>5.6</td>
<td>16.67</td>
</tr>
<tr>
<td>9.63</td>
<td>2.17</td>
<td>4.43</td>
<td>6.03</td>
</tr>
<tr>
<td>7.33</td>
<td>1.73</td>
<td>4.23</td>
<td>6.33</td>
</tr>
<tr>
<td>12.37</td>
<td>2.44</td>
<td>5.06</td>
<td>12.45</td>
</tr>
<tr>
<td>17.96</td>
<td>3.21</td>
<td>5.6</td>
<td>20.1</td>
</tr>
<tr>
<td>16.47</td>
<td>3.09</td>
<td>5.34</td>
<td>16.42</td>
</tr>
<tr>
<td>9.11</td>
<td>2.12</td>
<td>4.3</td>
<td>5.68</td>
</tr>
</tbody>
</table>

\(u_1 = 5\), the envelope model coincides with the non-enveloped model and so the prediction errors of the two models are identical.

![Figure 7: The SSPEs of (P.2) with and without enveloping over different choices of \(u_1\).](image)

In addition, we also modeled the data with group effects using the method given in the Supplement Sections [D]. Let \(Y_{(ij)} \in \mathbb{R}^{2 \times 5}\) denote the matrix-variate measurements of the \(j\)-th rabbit in the \(i\)-th group. We modeled the group effects according to the matrix model (D.1). BIC suggested that this model is adequate in comparison to the corresponding vector group model on vec(\(Y_{(ij)}\)). In (D.1), the number of groups \(g\) is equal to 4. The sample sizes among groups are the same with \(n_i = 9\), \(i = 1, \ldots, 4\). To gain more efficiency, we also fitted the envelope model
and applied BIC to select the optimal envelope dimensions as $u_1 = 1$ and $u_2 = 2$. This means that only one material direction in the row space $\mathbb{R}^2$ and two material directions in the column space $\mathbb{R}^5$ are useful for the group effect estimation. We then estimated the group effects $\alpha_i$, $i = 1, 2, 3$, by both the envelope model (D.4) and the matrix model (D.1). Here only $g - 1$ group effects are free parameters as $\sum_{i=1}^{n} n_i \alpha_i = 0$. The corresponding asymptotic standard errors of the elements in vec($\hat{\alpha}_i$) were computed. The estimate (EST) and the asymptotic standard error (SE) for each element in vec($\hat{\alpha}_i$) are summarized in Table 20 with subscripts ‘e’ and ‘s’ indicate the results from the envelope and non-envelope fits respectively.

Table 20 shows that the envelope model (D.4) provides more efficient estimators than the matrix model (D.1), as the relative ratios of the standard errors of the envelope model to those of the matrix model range from 2.81 to 4.66, shown in the last column of Table 20. In addition, the envelope model is more likely to detect the treatment effects as the ratios EST_e/SE_e of the elements in vec($\hat{\alpha}_i$) are observed to be mostly significant in the envelope fitting. For instance, the rabbits in Group 1 were given treatment $K_1$, the standard treatment at the low dose level in Day 1. The envelope model shows that the percentage decreases of the blood sugar concentration of this treatment group are significantly less than the overall percentage decreases among all the four groups in Day 1. This implies less efficacy of the low dose standard treatment compared to the other treatments. Yet the matrix model (D.1) without enveloping fails to detect this effect.

Q Estimation and inference of Model (8.2)

MLE. To make the model more general, let $Y \in \mathbb{R}^{r \times m}$. For an i.i.d. random sample \{\(Y_i, X_i\), $i = 1, \ldots, n$, where $X_j = 1$ for $j = 1, \ldots, n_1$ and $X_j = 0$ for $j = n_1 + 1, \ldots, n$, let $\bar{Y}$ and $\bar{X}$ be the sample mean of $Y$ and $X$, let $Y_1$ be sample mean of the group with $X$ being 1, and let $Y_2$ be sample mean of the group with $X$ being 0. Based on the matrix normal density function, the log-likelihood function of (8.2) is

$$l(\mu, \eta, L, R, \Omega_1, \Omega_{10}, \Omega_2, \Omega_{20}) = c - \frac{nr}{2} \log | \Sigma_2 | - \frac{nm}{2} \log | \Sigma_1 |$$

$$- \frac{1}{2} \sum_{i=1}^{n} \text{tr}\{\Sigma_2^{-1}(Y_i - \mu - L\eta R^TX_i)^T \Sigma_1^{-1}(Y_i - \mu - L\eta R^TX_i)\},$$

(Q.1)
Table 20: Comparison of $\text{vec}(\hat{\alpha}_i)$ and the standard errors of $\text{vec}(\hat{\alpha}_i)$ from the envelope model and the model without enveloping.

<table>
<thead>
<tr>
<th></th>
<th>Envelope model</th>
<th></th>
<th>Model without enveloping</th>
<th>Ratio</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EST$_e$</td>
<td>SE$_e$</td>
<td>EST$_e$/SE$_e$</td>
<td>EST$_s$</td>
<td>SE$_s$</td>
</tr>
<tr>
<td>$\text{vec}(\hat{\alpha}_1)$</td>
<td>-2.71</td>
<td>0.39</td>
<td>-4.57</td>
<td>0.09</td>
<td>2.45</td>
</tr>
<tr>
<td></td>
<td>-4.52</td>
<td>0.57</td>
<td>-7.95</td>
<td>-0.7</td>
<td>2.42</td>
</tr>
<tr>
<td></td>
<td>-6.44</td>
<td>0.77</td>
<td>-8.33</td>
<td>-2.18</td>
<td>2.96</td>
</tr>
<tr>
<td></td>
<td>-5.99</td>
<td>0.76</td>
<td>-7.87</td>
<td>-3.9</td>
<td>2.92</td>
</tr>
<tr>
<td></td>
<td>-3.5</td>
<td>0.89</td>
<td>-3.91</td>
<td>-1.64</td>
<td>3.52</td>
</tr>
<tr>
<td></td>
<td>2.48</td>
<td>0.91</td>
<td>2.73</td>
<td>5.2</td>
<td>3.48</td>
</tr>
<tr>
<td></td>
<td>4.14</td>
<td>0.93</td>
<td>4.45</td>
<td>7.79</td>
<td>3.88</td>
</tr>
<tr>
<td></td>
<td>5.9</td>
<td>0.93</td>
<td>6.36</td>
<td>9.88</td>
<td>3.84</td>
</tr>
<tr>
<td></td>
<td>5.49</td>
<td>0.69</td>
<td>7.94</td>
<td>8.95</td>
<td>3.16</td>
</tr>
<tr>
<td></td>
<td>3.2</td>
<td>0.67</td>
<td>4.78</td>
<td>5.39</td>
<td>3.12</td>
</tr>
<tr>
<td></td>
<td>3.29</td>
<td>0.7</td>
<td>4.69</td>
<td>0.13</td>
<td>2.45</td>
</tr>
<tr>
<td></td>
<td>5.62</td>
<td>0.67</td>
<td>8.33</td>
<td>5.39</td>
<td>2.42</td>
</tr>
<tr>
<td></td>
<td>8.16</td>
<td>0.92</td>
<td>8.91</td>
<td>8.09</td>
<td>2.96</td>
</tr>
<tr>
<td></td>
<td>7.53</td>
<td>0.91</td>
<td>8.29</td>
<td>9.57</td>
<td>2.92</td>
</tr>
<tr>
<td></td>
<td>4.25</td>
<td>1.05</td>
<td>4.06</td>
<td>6.82</td>
<td>3.52</td>
</tr>
<tr>
<td></td>
<td>-3.02</td>
<td>1.08</td>
<td>-2.79</td>
<td>-4.66</td>
<td>3.48</td>
</tr>
<tr>
<td></td>
<td>-5.15</td>
<td>1.1</td>
<td>-4.68</td>
<td>-8.27</td>
<td>3.88</td>
</tr>
<tr>
<td></td>
<td>-7.47</td>
<td>1.11</td>
<td>-6.74</td>
<td>-7.06</td>
<td>3.84</td>
</tr>
<tr>
<td></td>
<td>-6.9</td>
<td>0.81</td>
<td>-8.52</td>
<td>-3.87</td>
<td>3.16</td>
</tr>
<tr>
<td></td>
<td>-3.9</td>
<td>0.79</td>
<td>-4.95</td>
<td>-1.07</td>
<td>3.12</td>
</tr>
<tr>
<td>$\text{vec}(\hat{\alpha}_2)$</td>
<td>-1.36</td>
<td>0.71</td>
<td>-1.92</td>
<td>1.23</td>
<td>2.45</td>
</tr>
<tr>
<td></td>
<td>-4.58</td>
<td>0.65</td>
<td>-7</td>
<td>-4.9</td>
<td>2.42</td>
</tr>
<tr>
<td></td>
<td>-9.37</td>
<td>0.88</td>
<td>-10.68</td>
<td>-10.78</td>
<td>2.96</td>
</tr>
<tr>
<td></td>
<td>-7.54</td>
<td>0.85</td>
<td>-8.85</td>
<td>-7.91</td>
<td>2.92</td>
</tr>
<tr>
<td></td>
<td>-1.83</td>
<td>1.2</td>
<td>-1.53</td>
<td>-2.9</td>
<td>3.52</td>
</tr>
<tr>
<td></td>
<td>1.25</td>
<td>1.24</td>
<td>1.01</td>
<td>5.25</td>
<td>3.48</td>
</tr>
<tr>
<td></td>
<td>4.2</td>
<td>1.17</td>
<td>3.57</td>
<td>6.49</td>
<td>3.88</td>
</tr>
<tr>
<td></td>
<td>8.59</td>
<td>1.17</td>
<td>7.33</td>
<td>7.92</td>
<td>3.84</td>
</tr>
<tr>
<td></td>
<td>6.91</td>
<td>0.77</td>
<td>9.01</td>
<td>4.03</td>
<td>3.16</td>
</tr>
<tr>
<td></td>
<td>1.68</td>
<td>0.71</td>
<td>2.36</td>
<td>-1.14</td>
<td>3.12</td>
</tr>
</tbody>
</table>

32
where $c = -nmr \log 2\pi /2$, $\Sigma_1 = L\Omega_1 L^T + L_0\Omega_{10} L_0^T$, and $\Sigma_2 = R\Omega_2 R^T + R_0\Omega_{20} R_0^T$. It can be shown that given $L$, $R$, and $\eta$, $\hat{\mu}_c = \tilde{Y} - L\eta R^T \bar{X}$, and given $L$ and $R$, $\tilde{\eta} = L^T (\tilde{Y}_1 - \tilde{Y}_2) R$ by substituting $\hat{\mu}_c$ back. Let $X_i^* = X_i - \bar{X}$. Given $R$, $\Omega_2$, and $\Omega_{20}$, by substituting $\tilde{\eta}_1$ back to (Q.1), we have $l(L, \Omega_1, \Omega_{10}) = \sum_{i=1}^n \text{tr} \{ \Sigma_2^{-1} (Y_i - \tilde{Y})^T L_0 \Omega_{10}^{-1} L_0^T (Y_i - \tilde{Y}) \} - \frac{nm}{2} \log | \Omega_1 | - \frac{1}{2} \sum_{i=1}^n \text{tr} \{ \Sigma_2^{-1} [L^T (Y_i - \tilde{Y} - (\tilde{Y}_1 - \tilde{Y}_2) P_R X_i^*)] \} \Omega_1^{-1} [L^T (Y_i - \tilde{Y} - (\tilde{Y}_1 - \tilde{Y}_2) P_R X_i^*)] \} \right]$. (Q.2)

For fixed $L$, it can be shown that $\hat{\Omega}_1 = L^T C_{\text{res}} L$ and $\hat{\Omega}_{10} = L_0^T C_Y L_0$, where $C_{\text{res}} = (nm)^{-1} \sum_{i=1}^n (Y_i - \tilde{Y} - (\tilde{Y}_1 - \tilde{Y}_2) P_R X_i^*) \Sigma_2^{-1} [Y_i - \tilde{Y} - (\tilde{Y}_1 - \tilde{Y}_2) P_R X_i^*)]^T$ and $C_Y = (nm)^{-1} \sum_{i=1}^n (Y_i - \tilde{Y}) \Sigma_2^{-1} (Y_i - \tilde{Y})^T$. Substituting $\hat{\Omega}_1$ and $\hat{\Omega}_{10}$ back to (Q.1), we have $\hat{\Omega}_1 = L^T C_{\text{res}} L$, $\hat{\Omega}_{10} = L_0^T C_Y L_0$. The MLEs of $R$, $\Omega_2$, and $\Omega_{20}$ given $L$, $\Omega_1$, and $\Omega_{10}$ can be derived similarly. We thus omit the details.

The estimation of (8.2) is implemented with a two-step iterative algorithm: 1) initialize $R$ and $\Sigma_2$; 2) given the (updated) estimates of $R$ and $\Sigma_2$, estimate $L$, $\Omega_1$, $\Omega_{10}$, and thus $\Sigma_1$; 3) given the (updated) estimates of $L$ and $\Sigma_1$, estimate $R$, $\Omega_2$, $\Omega_{20}$, and $\Sigma_2$; 4) repeat 2)-3) until the log likelihood function converges. Then $\hat{\beta} = P_L (\tilde{Y}_1 - \tilde{Y}_2) P_R$, $\hat{\Sigma}_1 = P_L \hat{C}_{\text{res}} P_L + P_{L_0} \hat{C}_Y P_{L_0}$, and $\hat{\Sigma}_2 = P_R \hat{R}_{\text{res}} P_R + P_{R_0} \hat{R}_Y P_{R_0}$, where $\hat{C}_{\text{res}}$ and $\hat{C}_Y$ are the estimates of $C_{\text{res}}$ and $C_Y$, and $\hat{R}_{\text{res}} = (nr)^{-1} \sum_{i=1}^n [Y_i - \tilde{Y} - P_L (\tilde{Y}_1 - \tilde{Y}_2) X_i^*]^T \hat{\Sigma}_1^{-1} [Y_i - \tilde{Y} - P_L (\tilde{Y}_1 - \tilde{Y}_2) X_i^*]$ and $\hat{R}_Y = (nr)^{-1} \sum_{i=1}^n (Y_i - \tilde{Y})^T \hat{\Sigma}_1^{-1} (Y_i - \tilde{Y})$.

Asymptotic normality. We next derive the asymptotic distribution of the envelope estimators of (8.2).
Lemma 2. Under \([8,1]\), the Fisher information of the parameter vector \(\theta = (\text{vec}(\mu_c)^T, \text{vec}(\beta)^T, \text{vech}(\Sigma_1)^T, \text{vech}(\Sigma_2)^T)^T\) is

\[
J = \begin{pmatrix}
\Delta & E_n(X)\Delta & 0 & 0 \\
E_n(X)\Delta & E_n(X^2)\Delta & 0 & 0 \\
0 & 0 & \frac{m}{2}E_r^T\Delta_1E_r & \frac{1}{2}E_r^T\Delta_1\Psi \Delta_2 E_m \\
0 & 0 & \frac{1}{2}E_m^T\Delta_2 \Psi^T \Delta_1E_r & \frac{1}{2}E_m^T \Psi \Delta_2 E_m
\end{pmatrix},
\]

where \(\Delta = \Sigma_2^{-1} \otimes \Sigma_1^{-1}, \Delta_1 = \Sigma_1^{-1} \otimes \Sigma_1^{-1}, \Delta_2 = \Sigma_2^{-1} \otimes \Sigma_2^{-1}, \Psi = \text{vec}(\Sigma_1)\text{vec}(\Sigma_2)^T\), and \(E_r\) and \(E_m\) are expansion matrices defined at the beginning of Section \(\text{C}\).

For simplicity, let \(\text{vec}(\mu_c), \text{vec}(\eta), \text{vec}(L), \text{vec}(R), \text{vech}(\Omega_1), \text{vech}(\Omega_{10}), \text{vech}(\Omega_2)\), and \(\text{vech}(\Omega_{20})\) in \([8.2]\) be denoted as \(\phi_1, \phi_2, \ldots, \phi_8\), respectively. Let \(\phi\) be the combined parameter vector \(\phi = (\phi_1^T, \phi_2^T, \ldots, \phi_8^T)^T\). Therefore, \(\text{vec}(\mu_c), \text{vec}(\beta), \text{vech}(\Sigma_1)\) and \(\text{vech}(\Sigma_2)\) can be expressed as functions of \(\phi\):

\[
h(\phi) = \begin{pmatrix}
\text{vec}(\mu_c) \\
\text{vec}(\beta) \\
\text{vech}(\Sigma_1) \\
\text{vech}(\Sigma_2)
\end{pmatrix} = \begin{pmatrix}
\text{vec}(\mu_c) \\
\text{vec}(L\eta R^T) \\
\text{vech}(L\Omega_1L^T + L_0\Omega_{10}L_0^T) \\
\text{vech}(R\Omega_2R^T + R_0\Omega_{20}R_0^T)
\end{pmatrix} = \begin{pmatrix}
h_1(\phi) \\
h_2(\phi) \\
h_3(\phi) \\
h_4(\phi)
\end{pmatrix}. \tag{Q.3}
\]

Because of the overparameterization of \([8.2]\), we apply Proposition 4.1 in Shapiro (1986) to derive the asymptotic distribution of the envelop estimator of \(\theta\), which is \(h(\hat{\phi})\).

Proposition 8. Under \([8.2]\), \(\sqrt{n}(h(\hat{\phi}) - h(\phi))\) converges in distribution to a normal random vector with mean zero and covariance matrix \(\Lambda = H(H^TJH)^{-1}H^T\), where \(H = \partial h(\phi)/\partial \phi^T\) and is given by

\[
\begin{pmatrix}
I_{rm} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & R \otimes L & R\eta^T \otimes L_0 & R_0 \otimes L\eta & 0 & 0 & 0 & 0 \\
0 & 0 & H_{33} & C_r(L \otimes L)E_{u_1} & C_r(L_0 \otimes L_0)E_{u_{-u_1}} & 0 & 0 \\
0 & 0 & 0 & H_{44} & 0 & 0 & H_{47} & H_{48}
\end{pmatrix},
\]

with \(H_{33} = 2C_r(L\Omega_1 \otimes L_0 - L \otimes L_0\Omega_{10}), H_{44} = 2C_m(R\Omega_2 \otimes R_0 - R \otimes R_0\Omega_{20})K_{u_2(m-u_2)}, H_{47} = C_m(R \otimes R)E_{u_2}\) and \(H_{48} = C_m(R_0 \otimes R_0)E_{m-u_2}\). Let \(V = J^{-1}\) be the asymptotic variance of \(\theta\) under \([8.1]\). Then the envelop estimators are asymptotically more efficient than
the MLEs of \([8.1]\), that is, \(V - \Lambda \geq 0\).

The asymptotic distribution of the envelope estimator \(h(\hat{\phi})\) can be directly obtained by applying Proposition 4.1 in Shapiro (1986). The asymptotic efficiency holds because \(V - \Lambda = J^{-1} - H(H^T J H)^{\dagger} H^T = J^{-\frac{1}{2}} (I - P_{J^2 H}) J^{-\frac{1}{2}} = J^{-\frac{1}{2}} Q_{J^2 H} J^{-\frac{1}{2}} \geq 0\).

In addition, it can be shown that the asymptotic variance of the envelope MLEs of \((\text{vec}(\beta)^T, \text{vech}(\Sigma_1)^T, \text{vech}(\Sigma_2)^T)^T\), denoted as \(h^*(\hat{\phi})\), is \(\text{avar}(\sqrt{n} h^*(\hat{\phi})) = H_1 (H_1^T J_1 H_1)^{\dagger} H_1^T\), where

\[
J_1 = \begin{pmatrix}
E_n[X - E_n(X)]^2 \Delta & 0 & 0 \\
0 & \frac{m}{2} E_r^T \Delta_1 E_r & \frac{1}{2} E_r^T \Delta_1 \Psi \Delta_2 E_m \\
0 & \frac{1}{2} E_m^T \Delta_2 \Psi^T \Delta_1 E_r & \frac{1}{2} E_m^T \Delta_2 E_m
\end{pmatrix},
\]

and

\[
H_1 = \begin{pmatrix}
R \otimes L & R \eta^T \otimes L_0 & R_0 \otimes L \eta & 0 & 0 & 0 & 0 \\
0 & H_{33} & 0 & C_r(L \otimes L) E_{u_1} & C_r(L_0 \otimes L_0) E_{r-u_1} & 0 & 0 \\
0 & 0 & H_{44} & 0 & 0 & H_{47} & H_{48}
\end{pmatrix}.
\]

REFERENCES


