Notes on Statistical and Machine Learning

Multivariate Regression

Chapter: *6* **Prepared by:** *Chenxi Zhou*

This note is prepared based on *Chapter 6, Multivariate Regression* in Izenman (2009).

I. Introduction

1. **Overview:** *Multivariate regression* is an extension of the multiple regression and has s output variables $Y = (Y_1, \cdots, Y_s)^\top \in \mathbb{R}^s$, each of whose behavior may be influenced by exactly the same set of predictors $X = (X_1, \cdots, X_p)^\top \in \mathbb{R}^p$.

2. **Goal:** We are interested in estimating the regression relationship between Y and X, taking into account the various *dependencies* between the p-dimensional vector X and the s-dimensional vector Y and the dependencies within X and within Y.

   Remark. In the multivariate regression setting,

   (a) the components of X are correlated with each other,

   (b) the components of Y are correlated with each other, and

   (c) the components of Y are correlated with components of X.

II. Random Design Case

1. **Setup:** Assume that the random vectors

$$X = (X_1, \cdots, X_p)^\top \qquad \text{and} \qquad Y = (Y_1, \cdots, Y_s)^\top$$

   are jointly distributed, where the mean vectors of X and Y are $\boldsymbol{\mu}_X$ and $\boldsymbol{\mu}_Y$, respectively, and the joint covariance matrix

$$\begin{pmatrix} \boldsymbol{\Sigma}_{XX} & \boldsymbol{\Sigma}_{XY} \\ \boldsymbol{\Sigma}_{YX} & \boldsymbol{\Sigma}_{YY} \end{pmatrix}.$$

   In addition, we assume that $s \leq p$.

2. **Classical Multivariate Regression Model:** Assume that Y is related to X by the following *multivariate linear model*

$$Y = \boldsymbol{\mu} + \boldsymbol{\Theta} X + \boldsymbol{\varepsilon}, \tag{1}$$

   where $\boldsymbol{\mu} \in \mathbb{R}^s$ and $\boldsymbol{\Theta} \in \mathbb{R}^{s \times p}$ are the unknown parameters to be estimated, and $\boldsymbol{\varepsilon} \in \mathbb{R}^s$ is the unobservable error component of the model with mean $\mathbf{0}_s$ and unknown error covariance matrix $\mathrm{Cov}[\boldsymbol{\varepsilon}] = \Sigma_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}$. We assume that X and $\boldsymbol{\varepsilon}$ are independent.

3. **Least Squares Estimate of Parameters:** We are interested in finding $\boldsymbol{\mu}$ and $\boldsymbol{\Theta}$ that minimize

$$L(\boldsymbol{\mu}, \boldsymbol{\Theta}) := \mathbb{E}\big[(Y - \boldsymbol{\mu} - \boldsymbol{\Theta} X)(Y - \boldsymbol{\mu} - \boldsymbol{\Theta} X)^\top\big], \tag{2}$$

where the expectation is taken over the joint distribution of X and Y.

(a) *First Derivation:* Let $Y_c := Y - \boldsymbol{\mu}_Y$ and $X_c := X - \boldsymbol{\mu}_X$ and assume $\boldsymbol{\Sigma}_{XX} \succ 0$. We have

$$\begin{aligned} L(\mu, \Theta) &= \mathbb{E}\big[(Y - \boldsymbol{\mu} - \boldsymbol{\Theta} X)(Y - \boldsymbol{\mu} - \boldsymbol{\Theta} X)^\top\big] \\ &= \mathbb{E}\big[Y_c Y_c^\top - Y_c X_c^\top \boldsymbol{\Theta}^\top - \boldsymbol{\Theta} X_c Y_c^\top + \boldsymbol{\Theta} X_c X_c^\top \boldsymbol{\Theta}^\top\big] \\ &\qquad + (\boldsymbol{\mu} - \boldsymbol{\mu}_Y + \boldsymbol{\Theta}\boldsymbol{\mu}_X)(\boldsymbol{\mu} - \boldsymbol{\mu}_Y + \boldsymbol{\Theta}\boldsymbol{\mu}_X)^\top \\ &= \boldsymbol{\Sigma}_{YY} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Theta}^\top - \boldsymbol{\Theta}\boldsymbol{\Sigma}_{XY} + \boldsymbol{\Theta}\boldsymbol{\Sigma}_{XX}\boldsymbol{\Theta}^\top \\ &\qquad + (\boldsymbol{\mu} - \boldsymbol{\mu}_Y + \boldsymbol{\Theta}\boldsymbol{\mu}_X)(\boldsymbol{\mu} - \boldsymbol{\mu}_Y + \boldsymbol{\Theta}\boldsymbol{\mu}_X)^\top. \end{aligned}$$

We note that

$$\begin{aligned} &- \boldsymbol{\Sigma}_{YX}\boldsymbol{\Theta}^\top - \boldsymbol{\Theta}\boldsymbol{\Sigma}_{XY} + \boldsymbol{\Theta}\boldsymbol{\Sigma}_{XX}\boldsymbol{\Theta}^\top \\ =\,& (\boldsymbol{\Theta}\boldsymbol{\Sigma}_{XX}^{1/2} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1/2})(\boldsymbol{\Theta}\boldsymbol{\Sigma}_{XX}^{1/2} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1/2})^\top - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}. \end{aligned}$$

Hence,

$$\begin{aligned} L(\boldsymbol{\mu}, \boldsymbol{\Theta}) &= \big(\boldsymbol{\Sigma}_{YY} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\big) + \big(\boldsymbol{\Theta}\boldsymbol{\Sigma}_{XX}^{1/2} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1/2}\big)\big(\boldsymbol{\Theta}\boldsymbol{\Sigma}_{XX}^{1/2} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1/2}\big)^\top \\ &\quad + \big(\boldsymbol{\mu} - \boldsymbol{\mu}_Y + \boldsymbol{\Theta}\boldsymbol{\mu}_X\big)\big(\boldsymbol{\mu} - \boldsymbol{\mu}_Y + \boldsymbol{\Theta}\boldsymbol{\mu}_X\big)^\top \\ &\geq \boldsymbol{\Sigma}_{YY} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}, \end{aligned}$$

with equality being held when

$$\boldsymbol{\Theta}^* := \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}, \qquad \boldsymbol{\mu}^* := \boldsymbol{\mu}_Y - \boldsymbol{\Theta}\boldsymbol{\mu}_X.$$

In other words,

$$(\boldsymbol{\mu}^*, \boldsymbol{\Theta}^*) = \operatorname*{arg\,min}_{\boldsymbol{\mu}, \boldsymbol{\Theta}} L(\boldsymbol{\mu}, \boldsymbol{\Theta}).$$

(b) *Second Derivation:* Assuming that we can interchange the derivative and the integral, we take the partial derivatives of $L(\boldsymbol{\mu}, \boldsymbol{\Theta})$ with respect to the two arguments and set the derivatives to zero:

$$\begin{aligned} \frac{\partial L(\boldsymbol{\mu}, \boldsymbol{\Theta})}{\partial \boldsymbol{\mu}} &= -2\boldsymbol{\mu}_Y^\top + 2\boldsymbol{\mu}^\top + 2\boldsymbol{\mu}_X\boldsymbol{\Theta} \overset{\text{set}}{=} \mathbf{0}_s, \\ \frac{\partial L(\boldsymbol{\mu}, \boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}} &= -2\,\mathbb{E}[(Y - \boldsymbol{\mu} - \boldsymbol{\Theta} X)X^\top] \overset{\text{set}}{=} \mathbf{0}_{s\times p}. \end{aligned}$$

It follows that $(\boldsymbol{\mu}^*, \boldsymbol{\Theta}^*) = \arg\min_{\boldsymbol{\mu}, \boldsymbol{\Theta}} L(\boldsymbol{\mu}, \boldsymbol{\Theta})$ must satisfy

$$\boldsymbol{\mu}^* = \boldsymbol{\mu}_Y - \boldsymbol{\Theta}^*\boldsymbol{\mu}_X, \qquad \text{and} \qquad \boldsymbol{\Theta}^* = \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}.$$

Here, $\mathbf{\Theta}^*$ is the *full-rank regression coefficient matrix* of Y on X, and

$$Y = \boldsymbol{\mu}_Y + \mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}(X - \boldsymbol{\mu}_X)$$

is the *full-rank linear regression function* of Y on X. Here, the "full-rank" refers to the rank of $\mathbf{\Theta}$. At the minimum of L, the error is

$$\boldsymbol{\varepsilon} = Y - \boldsymbol{\mu}_Y - \mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}(X - \boldsymbol{\mu}_X) = Y_c - \mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}X_c.$$

In particular, note that

$$\begin{aligned}\mathbb{E}[\boldsymbol{\varepsilon}] &= \mathbf{0}_s,\\ \mathrm{Var}[\boldsymbol{\varepsilon}] &= \mathbf{\Sigma}_{YY} - \mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}\mathbf{\Sigma}_{XY},\\ \mathbb{E}[\boldsymbol{\varepsilon}X_c^\top] &= \mathbf{0}_{s\times p}.\end{aligned}$$

4. **Multivariate Reduced-Rank Regression:**

   (a) *Model Specification:* Consider the multivariate linear regression model given by

   $$Y = \boldsymbol{\mu} + \mathbf{C}X + \boldsymbol{\varepsilon}, \tag{3}$$

   where $\boldsymbol{\mu} \in \mathbb{R}^s$ and $\mathbf{C} \in \mathbb{R}^{s\times p}$ are unknown regression parameters and $\boldsymbol{\varepsilon} \in \mathbb{R}^s$ is the unobservable error with mean $\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}_s$ and covariance matrix $\mathrm{Cov}[\boldsymbol{\varepsilon}] = \mathbf{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}$. We assume that $\boldsymbol{\varepsilon}$ and X are independent. Here, we allow the rank of the regression coefficient matrix $\mathbf{C}$ to be deficient, i.e.,

   $$\mathrm{rank}(C) = t \le \min s = \{s, p\}. \tag{4}$$

   Remark 1. The "reduced-rank" condition (4) on $\mathbf{C}$ brings a true multivariate feature into the model, implying that there may be a number of *linear constraints* on the set of regression coefficients in the model.

   Remark 2. The name *reduced-rank regression* is used to distinguish the case $1 \le t < s$ from the *full-rank regression* with $t = s$.

   (b) *Question of Interest:* When $\mathrm{rank}(\mathbf{C}) = t$, there exits two full-rank matrices, an $s \times t$ matrix $\mathbf{A}$ and a $t \times p$ matrix $\mathbf{B}$, such that $\mathbf{C} = \mathbf{AB}$. Note that this decomposition is *not* unique since we can always find a nonsingular $t \times t$ matrix T such that

   $$\mathbf{C} = \mathbf{AB} = (\mathbf{AT})(\mathbf{T}^{-1}\mathbf{B}) = \mathbf{DE}.$$

   With the decomposition $\mathbf{C} = \mathbf{AB}$, we can write (3) as

   $$Y = \boldsymbol{\mu} + \mathbf{AB}X + \boldsymbol{\varepsilon}.$$

   We wish to estimate the unknown parameters $\mathbf{A}$, $\mathbf{B}$ and $\boldsymbol{\mu}$.

   (c) *Effective Dimensionality:* The rank of the matrix $\mathbf{C}$ here, t, is a meta-parameter called the *effective dimensionality* of the multivariate regression.

(d) *Review of the Eckart-Young's Inequality:* Suppose both $\mathbf{A}$ and $\mathbf{B}$ are matrices of size $m \times n$. Assume that $\mathbf{B}$ has the reduced rank $\mathrm{rank}(\mathbf{B}) = t$ and $\mathbf{A}$ is of full rank, $\mathrm{rank}(\mathbf{A}) = \min\{m, n\}$. We use $\mathbf{B}$ to approximate $\mathbf{A}$. Then,

$$\lambda_j\big((\mathbf{A} - \mathbf{B})(\mathbf{A} - \mathbf{B})^\top\big) \geq \lambda_{j+t}(\mathbf{A}\mathbf{A}^\top),$$

with equality if

$$\mathbf{B} = \sum_{i=1}^{t} \lambda_i^{1/2} \mathbf{u}_i \mathbf{v}_i,$$

where $\lambda_i = \lambda_i(\mathbf{A}\mathbf{A}^\top)$, $\mathbf{u}_i$ is the i-th eigenvector of $\mathbf{A}\mathbf{A}^\top$, and $\mathbf{v}_i$ is the i-th eigenvector of $\mathbf{A}^\top\mathbf{A}$.

(e) *Minimizing a Weighted Least Squares Criterion:* We minimize the following weighted sum-of-squares criterion

$$L(t) := \mathbb{E}\big[(Y - \boldsymbol{\mu} - \mathbf{A}\mathbf{B}X)^\top \boldsymbol{\Gamma} (Y - \boldsymbol{\mu} - \mathbf{A}\mathbf{B}X)\big], \tag{5}$$

where $\boldsymbol{\Gamma} \in \mathbb{R}^{s \times s}$ is a positive definite symmetric matrix of weights and the expectation is taken over the joint distribution of (X, Y). By letting $Y_c := Y - \boldsymbol{\mu}_Y$ and $X_c := X - \boldsymbol{\mu}_X$, first note that

$$\begin{aligned} L(t) &\geq \mathbb{E}\big[(Y_c - \mathbf{C}X_c)^\top \boldsymbol{\Gamma} (Y_c - \mathbf{C}X_c)\big] \\ &= \mathrm{trace}\big(\boldsymbol{\Sigma}_{YY}^* - \mathbf{C}^*\boldsymbol{\Sigma}_{XY}^* - \boldsymbol{\Sigma}_{YX}^*\mathbf{C}^* + \mathbf{C}^*\boldsymbol{\Sigma}_{XX}^*\mathbf{C}^*\big) \\ &= \mathrm{trace}\big((\boldsymbol{\Sigma}_{YY}^* - \boldsymbol{\Sigma}_{YX}^*\boldsymbol{\Sigma}_{XX}^{*-1}\boldsymbol{\Sigma}_{XY}^*)\big) \\ &\qquad + \mathrm{trace}\big((\mathbf{C}^*\boldsymbol{\Sigma}_{XX}^{*1/2} - \boldsymbol{\Sigma}_{YX}^*\boldsymbol{\Sigma}_{XX}^{*-1/2})(\mathbf{C}^*\boldsymbol{\Sigma}_{XX}^{*1/2} - \boldsymbol{\Sigma}_{YX}^*\boldsymbol{\Sigma}_{XX}^{*-1/2})^\top\big), \end{aligned} \tag{6}$$

where the last equality follows by completing the perfect square, and $\boldsymbol{\Sigma}_{XX}^* := \boldsymbol{\Sigma}_{XX}$, $\boldsymbol{\Sigma}_{YY}^* := \boldsymbol{\Gamma}^{1/2}\boldsymbol{\Sigma}_{YY}\boldsymbol{\Gamma}^{1/2}$, $\boldsymbol{\Sigma}_{XY}^* := \boldsymbol{\Sigma}_{XY}\boldsymbol{\Gamma}^{1/2}$, and $\mathbf{C}^* := \boldsymbol{\Gamma}^{1/2}\mathbf{C}$.

Now, we assume that $\mathrm{rank}(\mathbf{C}) = t$. According to the Eckart-Young's Inequality, the second trace is minimized when

$$\mathbf{C}^*\boldsymbol{\Sigma}_{XX}^{*1/2} = \sum_{i=1}^{t} \lambda_i^{1/2} \mathbf{v}_i \mathbf{u}_i,$$

where λ_i is the i-th eigenvalue of $\boldsymbol{\Sigma}_{YX}^*\boldsymbol{\Sigma}_{XX}^{*-1}\boldsymbol{\Sigma}_{XY}^* = \boldsymbol{\Gamma}^{1/2}\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\boldsymbol{\Gamma}^{1/2}$, and $\mathbf{v}_i$ is the eigenvector associated with λ_i, and

$$\mathbf{u}_i = \lambda_i^{-1/2}\boldsymbol{\Sigma}_{XX}^{*-1/2}\boldsymbol{\Sigma}_{XY}^*\mathbf{v}_i = \lambda_i^{-1/2}\boldsymbol{\Sigma}_{XX}^{-1/2}\boldsymbol{\Sigma}_{XY}\boldsymbol{\Gamma}^{1/2}\mathbf{v}_i.$$

It follows that the optimal $\mathbf{C}$ of reduced rank t that minimizes L is

$$\mathbf{C}^{(t)} = \boldsymbol{\Gamma}^{-1/2}\left(\sum_{j=1}^{t} \mathbf{v}_j\mathbf{v}_j^\top\right)\boldsymbol{\Gamma}^{1/2}\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}. \tag{7}$$

This $\mathbf{C}^{(t)}$ is called the *reduced-rank regression coefficient matrix* with rank t and weight matrix $\mathbf{\Gamma}$. It follows that L is minimized by letting $\boldsymbol{\mu}$, $\mathbf{A}$ and $\mathbf{B}$ be the following functions of t

$$\begin{aligned}
\boldsymbol{\mu}^{(t)} &= \boldsymbol{\mu}_Y - \mathbf{A}^{(t)}\mathbf{B}^{(t)}\boldsymbol{\mu}_X, && (8)\\
\mathbf{A}^{(t)} &= \mathbf{\Gamma}^{-1/2}\mathbf{V}_t, && (9)\\
\mathbf{B}^{(t)} &= \mathbf{V}_t^\top\mathbf{\Gamma}^{1/2}\mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}, && (10)
\end{aligned}$$

where $\mathbf{V}_t = (\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_t)$ is an $s \times t$ matrix and $\mathbf{v}_j$ is the eigenvector associated with the j-th largest eigenvalue of the matrix $\mathbf{\Gamma}^{1/2}\mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}\mathbf{\Sigma}_{XY}\mathbf{\Gamma}^{1/2}$.

(f) *Minimum of $W(t)$:* With $\mathbf{C}^{(t)}$ defined in (7), the minimum value of L is

$$\begin{aligned}
L_{\min}(t) &= \mathbb{E}\big[(Y - \boldsymbol{\mu} - \mathbf{C}^{(t)}X)^\top\mathbf{\Gamma}(Y - \boldsymbol{\mu} - \mathbf{C}^{(t)}X)\big]\\
&= \mathrm{trace}(\mathbf{\Sigma}_{YY}\mathbf{\Gamma}) - \sum_{j=1}^{t}\lambda_j.
\end{aligned}$$

(g) *Two Remarks:*

i. If we choose $t = s$, $\mathbf{C}^{(s)}$ reduces to the full-rank regression coefficient matrix $\mathbf{\Theta} = \mathbf{C}^{(s)}$;

ii. For any t and any positive-definite matrix $\mathbf{\Gamma}$, $\mathbf{C}^{(t)}$ and $\mathbf{\Theta}$ are related by

$$\mathbf{C}^{(t)} = \mathbf{P}_{\mathbf{\Gamma}}^{(t)}\mathbf{\Theta},$$

where

$$\mathbf{P}_{\mathbf{\Gamma}}^{(t)} = \mathbf{\Gamma}^{-1/2}\left(\sum_{j=1}^{t}\mathbf{v}_j\mathbf{v}_j^\top\right)\mathbf{\Gamma}^{1/2}.$$

(h) *Special Cases of Reduced-Rank Regression:*

i. If $X \equiv Y$ and $\mathbf{\Gamma} = \mathbf{I}_p$, we obtain Hotelling's principal component analysis;

ii. If $\mathbf{\Gamma} = \mathbf{\Sigma}_{YY}^{-1}$, we obtain the Hotelling's canonical correlation analysis;

iii. If $\mathbf{\Gamma} = \mathbf{\Sigma}_{YY}^{-1}$ and let Y be a vector of binary variables indicating the class belonging of observations, we obtain Fisher's linear discriminant analysis.

(i) *Sample Estimates:* Let $\{(\mathbf{x}_i^\top, \mathbf{y}_i^\top)^\top\}_{i=1}^n$ be n i.i.d observations from $(X^\top, Y^\top)^\top$. Then,

i. The mean vectors, $\boldsymbol{\mu}_X$ and $\boldsymbol{\mu}_Y$, can be estimated by

$$\hat{\boldsymbol{\mu}}_X = \bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i, \qquad \text{and} \qquad \hat{\boldsymbol{\mu}}_Y = \bar{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i,$$

respectively.

ii. For all $i = 1, \cdots, n$, let

$$\mathbf{x}_{c,i} = \mathbf{x}_i - \bar{\mathbf{x}}, \qquad \text{and} \qquad \mathbf{y}_{c,i} = \mathbf{y}_i - \bar{\mathbf{y}}$$

be the centered observations, and let

$$\mathbf{X}_c = \big[\mathbf{x}_{c,1}, \cdots, \mathbf{x}_{c,n}\big] \in \mathbb{R}^{p\times n}, \qquad \text{and} \qquad \mathbf{Y}_c = \big[\mathbf{y}_{c,1}, \cdots, \mathbf{y}_{c,n}\big] \in \mathbb{R}^{s\times n}$$

be the center data matrix.

iii. The covariance matrices, $\boldsymbol{\Sigma}_{XX}$, $\boldsymbol{\Sigma}_{XY}$, $\boldsymbol{\Sigma}_{YX}$ and $\boldsymbol{\Sigma}_{YY}$, can be estimated by

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}_{XX} &= \frac{1}{n}\mathbf{X}_c\mathbf{X}_c^\top, \\ \widehat{\boldsymbol{\Sigma}}_{YX} = \widehat{\boldsymbol{\Sigma}}_{XY}^\top &= \frac{1}{n}\mathbf{Y}_c\mathbf{X}_c^\top, \\ \widehat{\boldsymbol{\Sigma}}_{YY} &= \frac{1}{n}\mathbf{Y}_c\mathbf{Y}_c^\top.\end{aligned}$$

iv. Matrices $\mathbf{A}^{(t)}$ in (9) and $\mathbf{B}^{(t)}$ in (10) can be estimated by

$$\begin{aligned}\widehat{\mathbf{A}}^{(t)} &= \boldsymbol{\Gamma}^{-1/2}\widehat{\mathbf{V}}_t, \\ \widehat{\mathbf{B}}^{(t)} &= \widehat{\mathbf{V}}_t^\top\boldsymbol{\Gamma}^{1/2}\widehat{\boldsymbol{\Sigma}}_{YX}\widehat{\boldsymbol{\Sigma}}_{XX},\end{aligned}$$

where

$$\widehat{\mathbf{V}}_t = (\hat{\mathbf{v}}_1, \cdots, \hat{\mathbf{v}}_t)$$

is an $s \times t$-matrix with the j-th column, $\hat{\mathbf{v}}_j$, being the eigenvector associated with the j-th largest eigenvalue $\hat{\lambda}_j$ of the $s \times s$ symmetric matrix

$$\boldsymbol{\Gamma}^{1/2}\widehat{\boldsymbol{\Sigma}}_{YX}\widehat{\boldsymbol{\Sigma}}_{XX}^{-1}\widehat{\boldsymbol{\Sigma}}_{XY}\boldsymbol{\Gamma}^{1/2},$$

for $j = 1, 2, \cdots, p$. The reduced-rank regression coefficient matrix $\mathbf{C}^{(t)}$ in (7) can be estimated by

$$\widehat{\mathbf{C}}^{(t)} = \boldsymbol{\Gamma}^{-1/2}\left(\sum_{j=1}^{t}\hat{\mathbf{v}}_j\hat{\mathbf{v}}_j^\top\right)\boldsymbol{\Gamma}^{1/2}\widehat{\boldsymbol{\Sigma}}_{YX}\widehat{\boldsymbol{\Sigma}}_{XX}^{-1},$$

and the full-rank regression coefficient matrix $\boldsymbol{\Theta}$ can be estimated by

$$\widehat{\boldsymbol{\Theta}} = \widehat{\mathbf{C}}^{(s)} = \widehat{\boldsymbol{\Sigma}}_{YX}\widehat{\boldsymbol{\Sigma}}_{XX}^{-1}.$$

v. *Estimation in the Presence of Singularity:* In the case where $\widehat{\boldsymbol{\Sigma}}_{XX}$ and/or $\widehat{\boldsymbol{\Sigma}}_{YY}$ are singular, we replace them by a slight perturbation of their diagonal entries using the idea of the ridge regression

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}_{XX}^{(\eta)} &= \frac{1}{n}\big(\mathbf{X}_c\mathbf{X}_c^\top + \eta\cdot\mathbf{I}_d\big), \\ \widehat{\boldsymbol{\Sigma}}_{YY}^{(\eta)} &= \frac{1}{n}\big(\mathbf{Y}_c\mathbf{Y}_c^\top + \eta\cdot\mathbf{I}_d\big),\end{aligned}$$

respectively, where $\eta > 0$ is a tuning parameter. Everything else, such obtaining estimates of $\mathbf{C}^{(t)}$, $\mathbf{A}^{(k)}$ and $\mathbf{B}^{(t)}$, can be proceeded as before and will depend on the choice of $\eta > 0$.

(j) *Assessing the Effective Dimensionality:* In order to choose the value of $t \in \{1, 2, \cdots, s\}$, we choose the smallest integer such that the reduced-rank regression of Y on X with that integer as rank is as close (to be specified) as possible to the corresponding full-rank regression.

   i. Method 1. Let $L_{\min}(t)$ denote the minimum value of $L(t)$ for a fixed value of t. The reduction in $L_{\min}(t)$ by increasing the rank from $t = t_0$ to $t = t_1$ with $t_0 < t_1$ is

$$L_{\min}(t_0) - L_{\min}(t_1) = \sum_{j=t_0+1}^{t_1} \lambda_j,$$

   which only depends on the eigenvalues of $\mathbf{\Gamma}^{1/2}\mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}\mathbf{\Sigma}_{XY}\mathbf{\Gamma}^{1/2}$. Therefore, the rank of $\mathbf{C}$ can be assessed by some monotone function of the sequence of ordered sample eigenvalues $\{\hat{\lambda}_j\}_{j=1}^{s}$, in which $\hat{\lambda}_j$ is compared with suitable reference values for each j, or by the sum of some monotone function of the smallest $s - t_0$ sample eigenvalues.

   ii. Method 2 — Rank Trace. Suppose the true rank of $\mathbf{C}$ is t^*. The main idea behind *rank trace* is the following:

      A. for $1 \leq t < t^*$, the entries in both the estimated regression coefficient matrix and the residual covariance matrix change significantly each time we increase the rank;

      B. as soon as the true rank t^* is achieved, these two matrices stabilize.

   The algorithm is provided in Algorithm 1.

Algorithm 1 Using Rank Trace to Assess the Effective Dimensionality of a Multivariate Regression

1: Define $\widehat{\mathbf{C}}^{(0)} = \mathbf{0}_{s\times p}$ and $\widehat{\mathbf{\Sigma}}_{\varepsilon\varepsilon}^{(0)} = \widehat{\mathbf{\Sigma}}_{YY}$.
2: For $t = 1$ to s:

(a) compute $\widehat{\mathbf{C}}^{(t)}$ and $\widehat{\mathbf{\Sigma}}_{\varepsilon\varepsilon}^{(t)}$, and set $\widehat{\mathbf{C}}^{(s)} = \widehat{\mathbf{\Theta}}$ and $\widehat{\mathbf{\Sigma}}_{\varepsilon\varepsilon}^{(s)} = \widehat{\mathbf{\Sigma}}_{\varepsilon\varepsilon}$.

(b) compute

$$\Delta\widehat{\mathbf{C}}^{(t)} = \frac{\|\widehat{\mathbf{\Theta}} - \widehat{\mathbf{C}}^{(t)}\|}{\|\widehat{\mathbf{\Theta}}\|}, \qquad \Delta\widehat{\mathbf{\Sigma}}_{\varepsilon\varepsilon}^{(t)} = \frac{\|\widehat{\mathbf{\Sigma}}_{\varepsilon\varepsilon} - \widehat{\mathbf{\Sigma}}_{\varepsilon\varepsilon}^{(t)}\|}{\|\widehat{\mathbf{\Sigma}}_{\varepsilon\varepsilon} - \widehat{\mathbf{\Sigma}}_{YY}\|},$$

where $\|\mathbf{A}\| = \sqrt{\mathrm{trace}(\mathbf{A}\mathbf{A}^\top)} = \big(\sum_i \sum_j a_{ij}^2\big)^{1/2}$.

3: Make a scatterplot of the s points

$$\big(\Delta\widehat{\mathbf{C}}^{(t)}, \Delta\widehat{\mathbf{\Sigma}}_{\varepsilon\varepsilon}^{(t)}\big),$$

for $t = 0, 1, \cdots, s$, and join up successive points on the plot. This is called the *rank trace* for the multivariate reduce-rank regression of Y onto X.
4: Assess the rank of $\mathbf{C}$ as the *smallest* rank for which both coordinates from Step 3 are approximately zero.

Note the following:

- The first point in the rank trace, corresponding to $t = 0$, is always plotted at $(1, 1)$ and the last point, corresponding to $t = s$, is always plotted at $(0, 0)$;
- The horizontal coordinate, $\Delta\widehat{\mathbf{C}}^{(t)}$, gives a quantitative representation of the difference between a reduced-rank regression coefficient matrix and its full-rank analog;
- The vertical coordinate, $\Delta\widehat{\mathbf{\Sigma}}_{\varepsilon\varepsilon}^{(t)}$, shows the proportionate reduction in the residual variance matrix in using a simple full-rank model rather than the reduced-rank model.

iii. Method 3 — Cross Validation. For each rank t, compute a sequence of estimates of prediction error using the cross validation. Identify the smallest rank such that, for larger ranks, the prediction error has stabilized and does *not* decrease significantly; this is similar to saying that at $\hat{t}$, there is an elbow in the plot of prediction error against the rank.

III. Fixed Design Case

1. **Assumptions:** Let $Y = (Y_1, \cdots, Y_s)^\top$ be an s-dimensional random vector with mean $\boldsymbol{\mu}_Y \in \mathbb{R}^s$ and the covariance matrix $\boldsymbol{\Sigma}_{YY} \in \mathbb{R}^{s\times s}$, and let $X = (X_1, \cdots, X_p)^\top$ be a p-dimensional *nonstochastic* (i.e., fixed) vector.

2. **Data:** Let $\{(\mathbf{x}_i^\top, \mathbf{y}_i^\top)^\top\}_{i=1}^n$ be n i.i.d observations from $(X^\top, Y^\top)^\top$. Define the matrices $\mathbf{X}$ and $\mathbf{Y}$ by

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_n \end{bmatrix} \in \mathbb{R}^{p\times n}, \qquad \text{and} \qquad \mathbf{Y} = \begin{bmatrix} \mathbf{y}_1 & \cdots & \mathbf{y}_n \end{bmatrix} \in \mathbb{R}^{s\times n},$$

respectively.

Define the following quantities

$$\begin{aligned} \bar{\mathbf{x}} &:= \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i, \\ \bar{\mathbf{y}} &:= \frac{1}{n}\sum_{i=1}^n \mathbf{y}_i, \\ \bar{\mathbf{X}} &:= \begin{bmatrix} \bar{\mathbf{x}} & \cdots & \bar{\mathbf{x}} \end{bmatrix} \in \mathbb{R}^{p\times n}, \\ \bar{\mathbf{Y}} &:= \begin{bmatrix} \bar{\mathbf{y}} & \cdots & \bar{\mathbf{y}} \end{bmatrix} \in \mathbb{R}^{s\times n}. \end{aligned}$$

Defined the centered versions of $\mathbf{X}$ and $\mathbf{Y}$ to be

$$\mathbf{X}_c := \mathbf{X} - \bar{\mathbf{X}}, \qquad \text{and} \qquad \mathbf{Y}_c := \mathbf{Y} - \bar{\mathbf{Y}}.$$

3. **Classical Multivariate Regression Model:**

   (a) *Model specification:* Suppose each component of Y depends on the same set of predictors $X_1, \cdots, X_p$ in the following way

$$Y_j = \mu_j + \theta_{\theta,1}X_1 + \theta_{\theta,2}X_2 + \cdots + \theta_{\theta,p}X_p + \varepsilon_j, \qquad \text{for all } j = 1, \cdots, s, \tag{11}$$

   where μ_j is the intercept term and ε_j is the error variable with zero mean.
   With the data, we can write (11) collectively as

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\Theta}\mathbf{X} + \mathbf{E}, \tag{12}$$

   where $\boldsymbol{\mu} \in \mathbb{R}^{s\times n}$ is the matrix of the intercept terms, $\boldsymbol{\Theta} \in \mathbb{R}^{s\times p}$ is the matrix of the regression coefficients, and $\mathbf{E} = [\mathbf{E}_1, \mathbf{E}_2, \cdots, \mathbf{E}_n]$ is the $s\times n$ error matrix. We assume that each column of $\mathbf{E}$ has mean $\mathbf{0}_s$ and the common unknown nonsingular $s \times s$ covariance matrix $\boldsymbol{\Sigma}_{\mathbf{EE}}$.

   (b) *Centered version:* We remove $\boldsymbol{\mu}$ from the equation by centering $\mathbf{X}$ and $\mathbf{Y}$ and then estimate $\boldsymbol{\Theta}$ directly. To this end, we let

$$\boldsymbol{\mu} = \bar{\mathbf{Y}} - \boldsymbol{\Theta}\bar{\mathbf{X}}. \tag{13}$$

   Then, the model (12) becomes

$$\mathbf{Y}_c = \boldsymbol{\Theta}\mathbf{X}_c + \mathbf{E}. \tag{14}$$

(c) *Least squares estimation:* Apply the "vec" operation to both sides of (14), we obtain

$$\mathrm{vec}(\mathbf{Y}_c) = (\mathbf{I}_s \otimes \mathbf{X}_c^\top)\mathrm{vec}(\mathbf{\Theta}) + \mathrm{vec}(\mathbf{E}), \tag{15}$$

where we have $\mathrm{vec}(\mathbf{Y}_c) \in \mathbb{R}^{sn\times 1}$, $\mathbf{I}_s \otimes \mathbf{X}_c^\top \in \mathbb{R}^{sn\times sp}$, $\mathrm{vec}(\mathbf{\Theta}) \in \mathbb{R}^{sp\times 1}$ and $\mathrm{vec}(\mathbf{E}) \in \mathbb{R}^{sn\times 1}$.

Inspecting (15), we see it is a multiple linear regression problem. The error $\mathrm{vec}(\mathbf{E})$ has the mean vector $\mathbf{0}_{sn}$ and $sn \times sn$ block-diagonal covariance matrix

$$\mathrm{Var}[\mathrm{vec}(\mathbf{E})] = \mathbb{E}\big[\mathrm{vec}(\mathbf{E})\mathrm{vec}(\mathbf{E})^\top\big] = \mathbf{\Sigma}_{\mathbf{EE}} \otimes \mathbf{I}_n.$$

Assuming $\mathbf{X}_c\mathbf{X}_c^\top$ is nonsingular, we estimate $\mathrm{vec}(\mathbf{\Theta})$ using the generalized least-squares method and solve the following minimization problem

$$\underset{\mathbf{\Theta}}{\text{minimize}} \left\{ \big(\mathrm{vec}(\mathbf{Y}_c) - (\mathbf{I}_s \otimes \mathbf{X}_c^\top)\mathrm{vec}(\mathbf{\Theta})\big)[\mathrm{Var}[\mathrm{vec}(\mathbf{E})]]^{-1}\big(\mathrm{vec}(\mathbf{Y}_c) - (\mathbf{I}_s \otimes \mathbf{X}_c^\top)\mathrm{vec}(\mathbf{\Theta})\big) \right\}.$$

Then,

$$\begin{aligned}\mathrm{vec}(\widehat{\mathbf{\Theta}}) &= ((\mathbf{I}_s \otimes \mathbf{X}_c)(\mathbf{\Sigma}_{\mathbf{EE}} \otimes \mathbf{I}_n)^{-1}(\mathbf{I}_s \otimes \mathbf{X}_c^\top))^{-1}(\mathbf{I}_s \otimes \mathbf{X}_c)(\mathbf{\Sigma}_{\mathbf{EE}} \otimes \mathbf{I}_n)^{-1}\mathrm{vec}(\mathbf{Y}_c) \\ &= (\mathbf{I}_s \otimes (\mathbf{X}_c\mathbf{X}_c^\top)^{-1}\mathbf{X}_c)\mathrm{vec}(\mathbf{Y}_c),\end{aligned}$$

using the properties of the Kronecker product.

If we "un-vec" everything, we obtain

$$\widehat{\mathbf{\Theta}} = \mathbf{Y}_c\mathbf{X}_c^\top(\mathbf{X}_c\mathbf{X}_c^\top)^{-1}, \tag{16}$$

$$\widehat{\boldsymbol{\mu}} = \bar{\mathbf{Y}} - \widehat{\mathbf{\Theta}}\bar{\mathbf{X}}. \tag{17}$$

(d) *Minimum-variance linear unbiased estimator of* $\mathrm{trace}(\mathbf{A\Theta})$*:* Let $\mathbf{A}$ be a fixed matrix. Then, under the conditions above and if $\mathbf{X}_c\mathbf{X}_c^\top$ is nonsingular, then the minimum-variance linear unbiased estimator of $\mathrm{trace}(\mathbf{A\Theta})$ is given by $\mathrm{trace}(\mathbf{A}\widehat{\mathbf{\Theta}})$.

(e) *Interpretation of* $\widehat{\mathbf{\Theta}}$*:* Suppose we transpose both sides of (14) so that

$$\mathbf{Y}_c^\top = \mathbf{X}_c^\top\mathbf{\Theta}^\top + \mathbf{E}^\top.$$

Let $\tilde{\mathbf{y}}_j \in \mathbb{R}^n$ be the j-th column vector of $\mathbf{Y}_c^\top$, which represents all the n mean-centered observations on the j-th output variable, for $j = 1, \cdots, s$. Then, $\tilde{\mathbf{y}}_j \in \mathbb{R}^n$ can be modeled by the multiple regression equation

$$\tilde{\mathbf{y}}_j = \mathbf{X}_c^\top\boldsymbol{\theta}_j + \mathbf{e}_j,$$

where $\boldsymbol{\theta}_j$ is the j-th column of $\mathbf{\Theta}^\top$ and $\mathbf{e}_j$ is the j-th column of $\mathbf{E}^\top$. The ordinary least-square estimator of $\boldsymbol{\theta}_j$ is

$$\hat{\boldsymbol{\theta}}_j = (\mathbf{X}_c\mathbf{X}_c^\top)^{-1}\mathbf{X}_c\tilde{\mathbf{y}}_j,$$

which is exactly the j-th row of (16).

Thus, simultaneous (unrestricted) least-squares estimation applied to all the s equations of the multivariate regression model yields the *same* results as does equation-by-equation least-squares. As a result, nothing is gained by estimating the equations jointly, even though the output variables Y may be correlated.

In other words, even though the variables in Y may be correlated, the LS estimator, $\widehat{\mathbf{\Theta}}$, of $\mathbf{\Theta}$ does *not* contain any reference to that correlation.

(f) *Covariance Matrix of* $\widehat{\mathbf{\Theta}}$*:* We derive the covariance matrix of $\widehat{\mathbf{\Theta}}$. By the relationship $\mathbf{Y}_c = \mathbf{\Theta}\mathbf{X}_c + \mathbf{E}$, we have

$$\begin{aligned}\widehat{\mathbf{\Theta}} &= \mathbf{Y}_c\mathbf{X}_c^\top(\mathbf{X}_c\mathbf{X}_c^\top)^{-1} \\ &= (\mathbf{\Theta}\mathbf{X}_c + \mathbf{E})\mathbf{X}_c^\top(\mathbf{X}_c\mathbf{X}_c^\top)^{-1} \\ &= \mathbf{\Theta} + \mathbf{E}\mathbf{X}_c^\top(\mathbf{X}_c\mathbf{X}_c^\top)^{-1}.\end{aligned}$$

It follows that

$$\mathrm{vec}(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}) = \mathrm{vec}(\mathbf{E}\mathbf{X}_c^\top(\mathbf{X}_c\mathbf{X}_c^\top)^{-1}) = (\mathbf{I}_s \otimes (\mathbf{X}_c\mathbf{X}_c^\top)^{-1}\mathbf{X}_c)\mathrm{vec}(\mathbf{E}),$$

and, hence,

$$\begin{aligned}\mathrm{Var}[\mathrm{vec}(\widehat{\mathbf{\Theta}})] &= \mathbb{E}[\mathrm{vec}(\widehat{\mathbf{\Theta}} - \mathbf{\Theta})\mathrm{vec}(\widehat{\mathbf{\Theta}} - \mathbf{\Theta})^\top] \\ &= (\mathbf{I}_s \otimes (\mathbf{X}_c\mathbf{X}_c^\top)^{-1}\mathbf{X}_c)(\mathbf{\Sigma}_{\mathbf{EE}} \otimes \mathbf{I}_n)(\mathbf{I}_s \otimes \mathbf{X}_c^\top(\mathbf{X}_c\mathbf{X}_c^\top)^{-1}) \\ &= \mathbf{\Sigma}_{\mathbf{EE}} \otimes (\mathbf{X}_c\mathbf{X}_c^\top)^{-1}.\end{aligned}$$

(g) *Distribution of* $\widehat{\mathbf{\Theta}}$*:* If we assume that the errors in the model (12) are distributed as i.i.d Gaussian random vectors,

$$\mathbf{E}_i \sim \mathrm{Normal}(\mathbf{0}_s, \mathbf{\Sigma}_{\mathbf{EE}}), \qquad \text{for all } i = 1, \cdots, n,$$

where $\mathbf{E}_i$ denotes the i-th column of $\mathbf{E}$, then,

$$\mathrm{vec}(\widehat{\mathbf{\Theta}}) \sim \mathrm{Normal}(\mathrm{vec}(\mathbf{\Theta}), \mathbf{\Sigma}_{\mathbf{EE}} \otimes (\mathbf{X}_c\mathbf{X}_c^\top)^{-1}).$$

(h) *Fitted Values:* The $s \times n$ matrix $\widehat{\mathbf{Y}}$ of fitted values is given by

$$\widehat{\mathbf{Y}} = \widehat{\boldsymbol{\mu}} + \widehat{\mathbf{\Theta}}\mathbf{X} = \bar{\mathbf{Y}} + \widehat{\mathbf{\Theta}}(\mathbf{X} - \bar{\mathbf{X}}).$$

Also, we have

$$\widehat{\mathbf{Y}}_c = \widehat{\mathbf{\Theta}}\mathbf{X}_c = \mathbf{Y}_c\mathbf{X}_c^\top(\mathbf{X}_c\mathbf{X}_c^\top)^{-1}\mathbf{X}_c = \mathbf{Y}_c\mathbf{H},$$

where the $n \times n$ matrix $\mathbf{H} = \mathbf{X}_c^\top(\mathbf{X}_c\mathbf{X}_c^\top)^{-1}\mathbf{X}_c$ is the *hat matrix*.

(i) *Residual Matrix:* The $s \times n$ residual matrix $\widehat{\mathbf{E}}$ is the difference between the observed and fitted values of $\mathbf{Y}$, i.e.,

$$\widehat{\mathbf{E}} := \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y}_c - \widehat{\mathbf{\Theta}}\mathbf{X}_c = \mathbf{Y}_c(\mathbf{I}_n - \mathbf{H}),$$

In addition, we also have

$$\widehat{\mathbf{E}} = (\boldsymbol{\Theta}\mathbf{X}_c + \mathbf{E}) - (\boldsymbol{\Theta} + \mathbf{E}\mathbf{X}_c^\top(\mathbf{X}_c\mathbf{X}_c^\top)^{-1})\mathbf{X}_c = \mathbf{E}(\mathbf{I}_n - \mathbf{H}).$$

It follows that

$$\begin{aligned}\mathbb{E}[\mathrm{vec}(\widehat{\mathbf{E}})] &= \mathbf{0}_{ns},\\ \mathrm{Var}[\mathrm{vec}(\widehat{\mathbf{E}})] &= \boldsymbol{\Sigma}_{\mathbf{EE}} \otimes (\mathbf{I}_n - \mathbf{H}).\end{aligned}$$

(j) *Estimation of* $\boldsymbol{\Sigma}_{\mathbf{EE}}$*:* The $s \times s$ matrix version of the residual sum of squares is

$$\mathbf{S}_{\mathbf{E}} := \widehat{\mathbf{E}}\widehat{\mathbf{E}}^\top = (\mathbf{Y}_c(\mathbf{I}_n - \mathbf{H}))(\mathbf{Y}_c(\mathbf{I}_n - \mathbf{H}))^\top = \mathbf{Y}_c(\mathbf{I}_n - \mathbf{H})\mathbf{Y}_c^\top.$$

Also, it is easy to show

$$\mathbf{S}_{\mathbf{E}} = \mathbf{E}(\mathbf{I}_n - \mathbf{H})\mathbf{E}^\top.$$

Let $\mathbf{E}_j$ be the j-th row of $\mathbf{E}$. Then, the (j,k)-th element of $\mathbf{S}_{\mathbf{E}}$ can be written as

$$[\mathbf{S}_{\mathbf{E}}]_{(j,k)} = \mathbf{E}_j(\mathbf{I}_n - \mathbf{H})\mathbf{E}_k^\top,$$

whence,

$$\begin{aligned}\mathbb{E}\big[[\mathbf{S}_{\mathbf{E}}]_{(j,k)}\big] &= \mathbb{E}[\mathrm{trace}(\mathbf{I}_n - \mathbf{H})\mathbf{E}_k^\top\mathbf{E}_j]\\ &= \mathrm{trace}(\mathbf{I}_n - \mathbf{H})[\boldsymbol{\Sigma}_{\mathbf{EE}}]_{(j,k)}\\ &= (n-p)[\boldsymbol{\Sigma}_{\mathbf{EE}}]_{(j,k)}.\end{aligned}$$

The matrix

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{EE}} := \frac{1}{n-p}\mathbf{S}_{\mathbf{E}}$$

is called the *residual covariance matrix.*

(k) *Properties of* $\widehat{\boldsymbol{\Sigma}}_{\mathbf{EE}}$*:*

i. The matrix $\widehat{\boldsymbol{\Sigma}}_{\mathbf{EE}}$ is is statistically independent of $\widehat{\boldsymbol{\Theta}}$;

ii. The matrix $\widehat{\boldsymbol{\Sigma}}_{\mathbf{EE}}$ has a Wishart distribution with $n-p$ degrees of freedom and expectation $\boldsymbol{\Sigma}_{\mathbf{EE}}$.

(l) *Estimator of* $\mathrm{Var}[\widehat{\boldsymbol{\Theta}}]$*:* Using the results above, we can estimate $\mathrm{Var}[\widehat{\boldsymbol{\Theta}}] = \boldsymbol{\Sigma}_{\mathbf{EE}} \otimes (\mathbf{X}_c\mathbf{X}_c^\top)^{-1}$ by

$$\widehat{\mathrm{Var}}[\mathrm{vec}(\widehat{\boldsymbol{\Theta}})] = \widehat{\boldsymbol{\Sigma}}_{\mathbf{EE}} \otimes (\mathbf{X}_c\mathbf{X}_c^\top)^{-1}. \tag{18}$$

(m) *Confidence interval:* Let $\boldsymbol{\gamma} \in \mathbb{R}^{sp}$ be an arbitrary vector and consider to construct a confidence interval of $\boldsymbol{\gamma}^\top\mathrm{vec}(\boldsymbol{\Theta})$. Assuming the error vectors are distributed as

$$\mathbf{E}_i \sim \mathrm{Normal}(\mathbf{0}_s, \boldsymbol{\Sigma}_{\mathbf{EE}}), \qquad \text{for all } i = 1, \cdots, n,$$

we have the quantity

$$t = \frac{\boldsymbol{\gamma}^\top \mathrm{vec}(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})}{\sqrt{\boldsymbol{\gamma}^\top (\widehat{\boldsymbol{\Sigma}}_{\mathbf{EE}} \otimes (\mathbf{X}_c \mathbf{X}_c^\top)^{-1}) \boldsymbol{\gamma}}} \tag{19}$$

has the Student's t-distribution with $n-p$ degrees of freedom. Thus, a $(1-\alpha) \times 100\%$ confidence interval for $\boldsymbol{\gamma}^\top \mathrm{vec}(\boldsymbol{\Theta})$ can be given by

$$\boldsymbol{\gamma}^\top \mathrm{vec}(\widehat{\boldsymbol{\Theta}}) \pm t_{n-p,1-\frac{\alpha}{2}} \sqrt{\boldsymbol{\gamma}^\top (\widehat{\boldsymbol{\Sigma}}_{\mathbf{EE}} \otimes (\mathbf{X}_c \mathbf{X}_c^\top)^{-1}) \boldsymbol{\gamma}},$$

where $t_{n-p,1-\frac{\alpha}{2}}$ is the $(1-\alpha/2) \times 100\%$ percentile of the t-distribution with degrees of freedom $n-p$.

4. **Linear Constrained Estimation:** Consider the following model with centered data matrices $\mathbf{X}_c$ and $\mathbf{Y}_c$

$$\mathbf{Y}_c = \boldsymbol{\Theta} \mathbf{X}_c + \mathbf{E}.$$

We require $\boldsymbol{\Theta}$ to satisfy a set of known linear constraints of the form

$$\mathbf{K} \boldsymbol{\Theta} \mathbf{L} = \boldsymbol{\Gamma},$$

where the matrix $\mathbf{K} \in \mathbb{R}^{m \times s}$ and the matrix $\mathbf{L} \in \mathbb{R}^{p \times u}$ are full-rank matrices of known constants, and $\boldsymbol{\Gamma} \in \mathbb{R}^{m \times u}$ is a matrix of parameters (known or unknown). We often take $\boldsymbol{\Gamma} = \mathbf{0}_{m \times u}$. We require $m \leq s$ and $u \leq p$.

(a) *Example — variable selection:* Suppose we wish to study whether a specific subset of the p input variables has little or no effect on the behavior of the output variables. Suppose we arrange the rows of $\mathbf{X}_c$ so that

$$\mathbf{X}_c = \begin{pmatrix} \mathbf{X}_{c,1} \\ \mathbf{X}_{c,2} \end{pmatrix},$$

where $\mathbf{X}_{c,1} \in \mathbb{R}^{p_1 \times n}$ and $\mathbf{X}_{c,2} \in \mathbb{R}^{p_2 \times n}$ with $p_1 + p_2 = p$. Suppose we believe that the variables included in $\mathbf{X}_{c,2}$ do *not* belong in the regression. Corresponding to the partition of $\mathbf{X}_c$, we set $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)$, so that

$$\mathbf{Y}_c = \boldsymbol{\Theta}_1 \mathbf{X}_{c,1} + \boldsymbol{\Theta}_2 \mathbf{X}_{c,2} + \mathbf{E},$$

where $\boldsymbol{\Theta}_1 \in \mathbb{R}^{s \times p_1}$ and $\boldsymbol{\Theta}_2 \in \mathbb{R}^{s \times p_2}$.

To study whether the input variables included in $\mathbf{X}_{c,2}$ can be eliminated from the model, we set

$$\mathbf{K} = \mathbf{I}_s, \qquad \mathbf{L} = \begin{pmatrix} \mathbf{0}_{p_1 \times p_2} \\ \mathbf{I}_{p_2 \times p_2} \end{pmatrix},$$

so that $\mathbf{K} \boldsymbol{\Theta} \mathbf{L} = \boldsymbol{\Theta}_2 = \mathbf{0}_{s \times p_2}$.

(b) *Constrained least-squares estimation:* To estimate $\boldsymbol{\Theta}$ under the linear constraint $\mathbf{K}\boldsymbol{\Theta}\mathbf{L} = \boldsymbol{\Gamma}$, we consider the following optimization problem

$$\begin{aligned} &\underset{\boldsymbol{\Theta}}{\text{minimize}} \left\{ \text{trace}\Big((\mathbf{Y}_c - \boldsymbol{\Theta}\mathbf{X}_c)(\mathbf{Y}_c - \boldsymbol{\Theta}\mathbf{X}_c)^\top \Big) \right\} \\ &\text{subject to } \mathbf{K}\boldsymbol{\Theta}\mathbf{L} = \boldsymbol{\Gamma}. \end{aligned} \tag{20}$$

Let $\widehat{\boldsymbol{\Theta}}^*$ be the minimizer of (20) and $\boldsymbol{\Lambda}$ be a matrix of Lagrangian coefficients. The normal equations are

$$\begin{aligned} \widehat{\boldsymbol{\Theta}}^* \mathbf{X}_c \mathbf{X}_c^\top + \mathbf{K}^\top \boldsymbol{\Lambda} \mathbf{L}^\top &= \mathbf{Y}_c \mathbf{X}_c^\top, \\ \mathbf{K}\widehat{\boldsymbol{\Theta}}^* \mathbf{L} &= \boldsymbol{\Gamma}. \end{aligned} \tag{21}$$

It follows that

$$\widehat{\boldsymbol{\Theta}}^* = \widehat{\boldsymbol{\Theta}} - \mathbf{K}^\top \boldsymbol{\Lambda} \mathbf{L}^\top (\mathbf{X}_c \mathbf{X}_c^\top)^{-1},$$

where $\widehat{\boldsymbol{\Theta}} = \mathbf{Y}_c \mathbf{X}_c^\top (\mathbf{X}_c \mathbf{X}_c^\top)^{-1}$ as we have shown before. Then,

$$\mathbf{K}\big(\widehat{\boldsymbol{\Theta}} - \mathbf{K}^\top \boldsymbol{\Lambda} \mathbf{L}^\top (\mathbf{X}_c \mathbf{X}_c^\top)^{-1}\big)\mathbf{L} = \boldsymbol{\Gamma},$$

from which we obtain

$$\boldsymbol{\Lambda} = (\mathbf{K}\mathbf{K}^\top)^{-1}(\mathbf{K}\widehat{\boldsymbol{\Theta}}\mathbf{L} - \boldsymbol{\Gamma})(\mathbf{L}^\top (\mathbf{X}_c \mathbf{X}_c^\top)^{-1}\mathbf{L})^{-1},$$

assuming all inverses exist. Finally, we obtain

$$\widehat{\boldsymbol{\Theta}}^* = \widehat{\boldsymbol{\Theta}} - \mathbf{K}^\top (\mathbf{K}\mathbf{K}^\top)^{-1}(\mathbf{K}\widehat{\boldsymbol{\Theta}}\mathbf{L} - \boldsymbol{\Gamma})(\mathbf{L}^\top (\mathbf{X}_c \mathbf{X}_c^\top)^{-1}\mathbf{L})^{-1}\mathbf{L}^\top (\mathbf{X}_c \mathbf{X}_c^\top)^{-1}.$$

(c) *Multivariate Analysis of Variance (MANOVA):*

- Residual sum of squares: The residual sum of squares under the constrained model is given by

$$\begin{aligned} \mathbf{S}_{\mathbf{E}}^* &:= (\mathbf{Y}_c - \widehat{\boldsymbol{\Theta}}^* \mathbf{X}_c)(\mathbf{Y}_c - \widehat{\boldsymbol{\Theta}}^* \mathbf{X}_c)^\top \\ &= \big(\mathbf{Y}_c - \widehat{\boldsymbol{\Theta}}\mathbf{X}_c + (\widehat{\boldsymbol{\Theta}} - \widehat{\boldsymbol{\Theta}}^*)\mathbf{X}_c\big)\big(\mathbf{Y}_c - \widehat{\boldsymbol{\Theta}}\mathbf{X}_c + (\widehat{\boldsymbol{\Theta}} - \widehat{\boldsymbol{\Theta}}^*)\mathbf{X}_c\big)^\top \\ &= (\mathbf{Y}_c - \widehat{\boldsymbol{\Theta}}\mathbf{X}_c)\big(\mathbf{Y}_c - \widehat{\boldsymbol{\Theta}}\mathbf{X}_c\big)^\top + (\widehat{\boldsymbol{\Theta}} - \widehat{\boldsymbol{\Theta}}^*)\mathbf{X}_c \mathbf{X}_c^\top (\widehat{\boldsymbol{\Theta}} - \widehat{\boldsymbol{\Theta}}^*)^\top, \end{aligned} \tag{22}$$

  where

  - the first term on the RHS of (22) is the matrix version of the residual sum of squares, $\mathbf{S}_{\mathbf{E}}$, for the *unconstrained* model, and
  - the second term is the additional source of variation, $\mathbf{S}_h := \mathbf{S}_{\mathbf{E}} - \mathbf{S}_{\mathbf{E}}^*$, due to dropping the constraints.

- Regression sum of squares: The *regression sum of squares*, $\mathbf{S}_{\text{reg}}$, for the unconstrained model is given by

$$\begin{aligned}\mathbf{S}_{\text{reg}} &:= \widehat{\mathbf{\Theta}}\mathbf{X}_c\mathbf{X}_c^\top\widehat{\mathbf{\Theta}}^\top \\ &= \big(\widehat{\mathbf{\Theta}}^* + (\widehat{\mathbf{\Theta}} - \widehat{\mathbf{\Theta}}^*)\big)\mathbf{X}_c\mathbf{X}_c^\top\big(\widehat{\mathbf{\Theta}}^* + (\widehat{\mathbf{\Theta}} - \widehat{\mathbf{\Theta}}^*)\big)^\top \\ &= \widehat{\mathbf{\Theta}}^*\mathbf{X}_c\mathbf{X}_c^\top\widehat{\mathbf{\Theta}}^* + (\widehat{\mathbf{\Theta}} - \widehat{\mathbf{\Theta}}^*)\mathbf{X}_c\mathbf{X}_c^\top(\widehat{\mathbf{\Theta}} - \widehat{\mathbf{\Theta}}^*)^\top, \end{aligned} \tag{23}$$

  where
  - the first term on the RHS of (23) is $\mathbf{S}_{\text{reg}}^*$, the matrix version of the regression sum of squares for the constrained model, and
  - the second term is, again, $\mathbf{S}_h$.
- MANOVA Table: Let $k = \text{rank}(\mathbf{K})$. We have the following MANOVA table.

Source of Variation	df	Sum of Squares
Constrained Model	$p-k$	$\mathbf{S}_{\text{reg}}^* = \widehat{\mathbf{\Theta}}^*\mathbf{X}_c\mathbf{X}_c^\top\widehat{\mathbf{\Theta}}^*$
Due to dropping constraints	k	$\mathbf{S}_h = (\widehat{\mathbf{\Theta}} - \widehat{\mathbf{\Theta}}^*)\mathbf{X}_c\mathbf{X}_c^\top(\widehat{\mathbf{\Theta}} - \widehat{\mathbf{\Theta}}^*)^\top$
Unconstrained model	p	$\mathbf{S}_{\text{reg}} = \widehat{\mathbf{\Theta}}\mathbf{X}_c\mathbf{X}_c^\top\widehat{\mathbf{\Theta}}$
Residual	$n-p-1$	$\mathbf{S}_{\mathbf{E}} = (\mathbf{Y}_c - \widehat{\mathbf{\Theta}}\mathbf{X}_c)(\mathbf{Y}_c - \widehat{\mathbf{\Theta}}\mathbf{X}_c)^\top$
Total	$n-1$	$\mathbf{Y}_c\mathbf{Y}_c^\top$

(d) *Expectation of* $\mathbf{S}_{\text{reg}}$*:* With the previous results, we have

$$\mathbb{E}[\mathbf{S}_h] = \mathbf{D}(\mathbf{K}\widehat{\mathbf{\Theta}}\mathbf{L} - \mathbf{\Gamma})(\mathbf{L}^\top(\mathbf{X}_c\mathbf{X}_c^\top)^{-1}\mathbf{L})^{-1}(\mathbf{K}\widehat{\mathbf{\Theta}}\mathbf{L} - \mathbf{\Gamma})^\top\mathbf{D}^\top + \mathbf{F}\,\mathbb{E}[\mathbf{E}\mathbf{G}\mathbf{E}^\top]\mathbf{F}^\top,$$

where

$$\begin{aligned}\mathbf{D} &= \mathbf{K}^\top(\mathbf{K}\mathbf{K}^\top)^{-1}, \\ \mathbf{F} &= \mathbf{D}\mathbf{K}, \\ \mathbf{G} &= \mathbf{X}_c^\top(\mathbf{X}_c\mathbf{X}_c^\top)^{-1}\mathbf{L}(\mathbf{L}^\top(\mathbf{X}_c\mathbf{X}_c^\top)^{-1}\mathbf{L})^{-1}\mathbf{L}^\top(\mathbf{X}_c\mathbf{X}_c^\top)^{-1}\mathbf{X}_c^\top.\end{aligned}$$

It is easy to show

$$\mathbf{F}^2 = \mathbf{F} = \mathbf{F}^\top, \qquad \text{and} \qquad \mathbf{G}^2 = \mathbf{G} = \mathbf{G}^\top,$$

and

$$\mathbb{E}[\mathbf{E}\mathbf{G}\mathbf{E}^\top] = u\mathbf{\Sigma}_{\mathbf{E}\mathbf{E}},$$

where $u = \text{trace}(\mathbf{G}) = \text{trace}(\mathbf{I}_u)$ is the rank of $\mathbf{L}$.

(e) *Hypothesis testing:* We test

$$H_0 : \mathbf{K}\mathbf{\Theta}\mathbf{L} = \mathbf{\Gamma} \qquad \text{vs.} \qquad H_1 : \mathbf{K}\mathbf{\Theta}\mathbf{L} \neq \mathbf{\Gamma}.$$

Under H_0,

$$\mathbb{E}\left[\frac{1}{u}\mathbf{S}_h\right] = \mathbf{F}\mathbf{\Sigma}_{\mathbf{EE}}\mathbf{F}^\top,$$
$$\mathbb{E}\left[\frac{1}{n-p-1}\mathbf{S}_{\mathbf{E}}\right] = \mathbf{\Sigma}_{\mathbf{EE}}.$$

A formal significance test of H_0 vs. H_1 can be realized through a function (e.g., determinant, trace, or largest eigenvalue) of the quantity $\mathbf{F}\mathbf{S}_h\mathbf{F}^{-1}(\mathbf{F}\mathbf{S}_{\mathbf{E}}\mathbf{F}^\top)^{-1}$. Examples include

- Hotelling-Lawley trace statistic: $\text{trace}(\mathbf{S}_h\mathbf{S}_{\mathbf{E}}^{-1})$,
- Roy's largest root: $\lambda_{\max}(\mathbf{S}_h\mathbf{S}_{\mathbf{E}}^{-1})$, and
- Wilks's lambda (likelihood ratio criterion): $|\mathbf{S}_{\mathbf{E}}|/|\mathbf{S}_h + \mathbf{S}_{\mathbf{E}}|$.

Under appropriate distribution assumptions, we reject H_0 if Hotelling-Lawley's trace statistic and Roy's largest root are small, or if Wilk's lambda is large.

References

Izenman, Alan J (Mar. 2009). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. en. Springer Science & Business Media.