# Partial Predictor Envelope Model with application to Cytokine-based Biomarker Analysis for COVID-19

Yeonhee Park

Department of Biostatistics and Medical Informatics
University of Wisconsin-Madison

*ypark56@wisc.edu*

KSS Spring Conference 2021

May 28, 2021

Joint work with Zhihua Su and Dongjun Chung

# Outline

1. Introduction
2. Methods
   - Review of Envelope Models
   - Predictor Envelope Models
   - Partial Predictor Envelope Models
3. Data Application
4. Discussion

# COVID-19

It is a global pandemic that
has affected 223 countries, areas or territories.

As of April 2021,
it has infected more than 147 million people and
caused more than 3.1 million deaths worldwide.

- Many studies on COVID-19 patients have collected data on various biomarkers (COVID-IP project `www.immunophenotype.org`).
- Cytokines are associated with COVID-19 severity and survival (Lu et al. 2020; Laing et al. 2020; Takahashi et al. 2020).
- The identification of the association between the cytokine-based biomarkers and COVID-19 severity and demographics leads to a better understanding and management of the disease.

# MULTIVARIATE LINEAR REGRESSION MODEL

The multivariate linear regression model is a common tool for the investigation of the association between key immunologic factors and COVID-19 patients' clinical information (Genser et al. 2007).

- Traditional ordinal least squares fitting
- Partial least squares

It is common to have both continuous and categorical variables in the predictors, e.g., in the COVID-19 dataset,

- Continuous predictors: Patients' clinical status such as temperature, respiratory rate and oxygen saturation
- Categorical predictors: Patients' sex, ethnicity and indicators for underlying disease such as asthma and diabetes.

A common practice: to treat the categorical predictors as continuous.

Problem: This can “lead to biased estimates and therefore to invalid inferences and erroneous conclusion” (Schuberth et al. 2018; Lohmoller. 2013; Hair et al. 2012).

It is common to have both continuous and categorical variables in the predictors, e.g., in the COVID-19 dataset,

- Continuous predictors: Patients' clinical status such as temperature, respiratory rate and oxygen saturation
- Categorical predictors: Patients' sex, ethnicity and indicators for underlying disease such as asthma and diabetes.

A common practice: to treat the categorical predictors as continuous.

Approach: instead of treating the categorical predictors as continuous, we condition both the response(s) and the continuous predictors on the categorical predictors, and then perform the envelope estimation based on the conditional distributions.

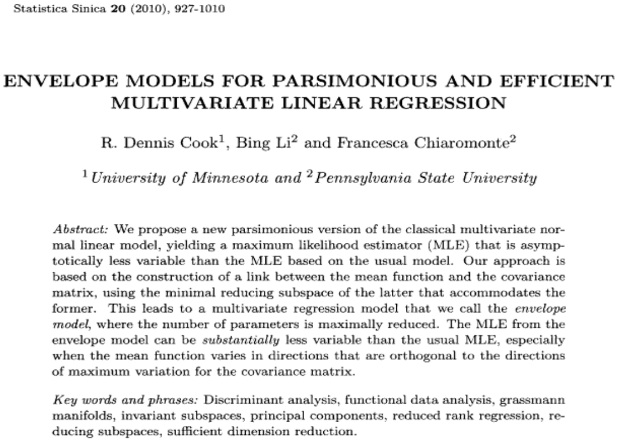
# Review of Envelope Models

Consider multivariate linear regression:

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\epsilon}, \tag{1}$$

- $\mathbf{Y} \in \mathbb{R}^{r}$ is multivariate response vector
- $\mathbf{X} \in \mathbb{R}^{p}$ is a non-stochastic predictor vector
- $\boldsymbol{\epsilon} \in \mathbb{R}^{r}$ is error with mean **0** and positive definite covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{r \times r}$
- $\boldsymbol{\mu} \in \mathbb{R}^{r}$ is an unknown intercept
- $\boldsymbol{\beta} \in \mathbb{R}^{r \times p}$ is an unknown regression coefficients.

Cook et al. (2010) introduced the envelope model to achieve efficient estimation in multivariate linear regression.



Statistica Sinica **20** (2010), 927-1010

## ENVELOPE MODELS FOR PARSIMONIOUS AND EFFICIENT MULTIVARIATE LINEAR REGRESSION

R. Dennis Cook[1], Bing Li[2] and Francesca Chiaromonte[2]

[1] *University of Minnesota and* [2] *Pennsylvania State University*

*Abstract:* We propose a new parsimonious version of the classical multivariate normal linear model, yielding a maximum likelihood estimator (MLE) that is asymptotically less variable than the MLE based on the usual model. Our approach is based on the construction of a link between the mean function and the covariance matrix, using the minimal reducing subspace of the latter that accommodates the former. This leads to a multivariate regression model that we call the *envelope model*, where the number of parameters is maximally reduced. The MLE from the envelope model can be *substantially* less variable than the usual MLE, especially when the mean function varies in directions that are orthogonal to the directions of maximum variation for the covariance matrix.

*Key words and phrases:* Discriminant analysis, functional data analysis, grassmann manifolds, invariant subspaces, principal components, reduced rank regression, reducing subspaces, sufficient dimension reduction.

- We partition the response vector **Y** into a material part and an immaterial part.
- The distribution of the material part changes with the predictor **X** and the distribution of the immaterial part does not.

Specifically,

- $\mathcal{S}$ = a subspace of $\mathbb{R}^r$.
- $\boldsymbol{\Gamma}$ = an orthogonal basis of $\mathcal{S}$ and $\boldsymbol{\Gamma}_0$ = an orthogonal basis of $\mathcal{S}^\perp$, where $\mathcal{S}^\perp$ denotes the orthogonal complement of the subspace $\mathcal{S}$.
- $\boldsymbol{\Gamma}^T\mathbf{Y}$ and $\boldsymbol{\Gamma}_0^T\mathbf{Y}$ are called the material part and the immaterial part if
  1. $\boldsymbol{\Gamma}_0^T\mathbf{Y}|\mathbf{X} \sim \boldsymbol{\Gamma}_0^T\mathbf{Y}$
  2. $\text{cov}(\boldsymbol{\Gamma}^T\mathbf{Y}, \boldsymbol{\Gamma}_0^T\mathbf{Y}|\mathbf{X}) = \mathbf{0}$

  are satisfied.

Let $\mathcal{B} = \text{span}(\beta)$. Such two conditions are equivalent to

$$\mathcal{B} \subseteq \mathcal{S} \qquad \text{and} \qquad \mathbf{\Sigma} = \mathbf{P}_{\mathcal{S}}\mathbf{\Sigma}\mathbf{P}_{\mathcal{S}} + \mathbf{Q}_{\mathcal{S}}\mathbf{\Sigma}\mathbf{Q}_{\mathcal{S}}. \tag{2}$$

- A subspace $\mathcal{S}$ is said to be a reducing subspace of $\mathbf{\Sigma}$ if $\mathbf{\Sigma}\mathcal{S} \subseteq \mathcal{S}$ and $\mathbf{\Sigma}\mathcal{S}^{\perp} \subseteq \mathcal{S}^{\perp}$.
- $\mathcal{S}$ is a reducing subspace of $\mathbf{\Sigma}$ (Conway, 1990).
- Let $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ denote the $\mathbf{\Sigma}$-envelope of $\mathcal{B}$, the smallest reducing subspace of $\mathbf{\Sigma}$ containing $\mathcal{B}$.
- Model (1) with (2) is called the envelope model.

Let $u$ denote the dimension of $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ and $\boldsymbol{\Gamma} \in \mathbb{R}^{r \times u}$ be an orthogonal basis of $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$. The coordinate form of the envelope model is

$$\begin{aligned} \mathbf{Y} &= \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\eta}\mathbf{X} + \boldsymbol{\epsilon} \\ \boldsymbol{\Sigma} &= \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^\top + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^\top, \end{aligned}$$

- $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$, $\boldsymbol{\eta} \in \mathbb{R}^{u \times p}$
- $\boldsymbol{\Omega} = \boldsymbol{\Gamma}^\top\boldsymbol{\Sigma}\boldsymbol{\Gamma}$ and $\boldsymbol{\Omega}_0 = \boldsymbol{\Gamma}_0^\top\boldsymbol{\Sigma}\boldsymbol{\Gamma}_0$

Note when $u = r$, $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}) = \mathbb{R}^r$.

- Cook et al. (2010) shows that the envelope estimator of $\boldsymbol{\beta}$ is more efficient than or at least as efficient as the standard estimator.
- The efficiency gains can be substantial when $\|\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^\top\| \leq \|\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^\top\|$, where $\|\cdot\|$ denotes the spectral norm of a matrix or vector.

# Graphical Illustration

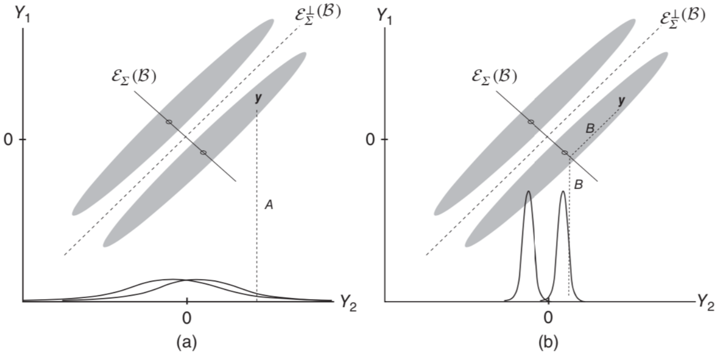

$Y_1$
$\mathcal{E}_\Sigma^\perp(\mathcal{B})$
$\mathcal{E}_\Sigma(\mathcal{B})$
y
0
A
$Y_2$
0
(a)
$Y_1$
$\mathcal{E}_\Sigma^\perp(\mathcal{B})$
$\mathcal{E}_\Sigma(\mathcal{B})$
y
B
0
B
$Y_2$
0
(b)

Figure: Graphical illustration envelope estimation: (a) standard analysis (b) envelope analysis (Cook, 2018).

# Predictor Envelope Models

Consider multivariate linear regression:

$$\mathbf{Y} = \boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\beta}^{\top}(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}) + \boldsymbol{\epsilon},$$

- $\mathbf{Y} \in \mathbb{R}^r$ is univariate response or multivariate response vector with mean $\boldsymbol{\mu}_{\mathbf{Y}}$
- $\mathbf{X} \in \mathbb{R}^p$ is a predictor vector with mean $\boldsymbol{\mu}_{\mathbf{X}}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{X}}$
- $\boldsymbol{\epsilon} \in \mathbb{R}^r$ is error with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$
- $\boldsymbol{\beta} \in \mathbb{R}^{p \times r}$ is an unknown regression coefficients.

The predictor envelope model assumes that there is a subspace $\mathcal{S} \subseteq \mathbb{R}^p$ such that

1. $\text{cov}(\mathbf{Y}, \mathbf{Q}_{\mathcal{S}}\mathbf{X} | \mathbf{P}_{\mathcal{S}}\mathbf{X}) = 0$
2. $\text{cov}(\mathbf{P}_{\mathcal{S}}\mathbf{X}, \mathbf{Q}_{\mathcal{S}}\mathbf{X}) = 0$,

which are equivalent to imposing the following structure to the model parameters

1. $\text{span}(\beta) \subseteq \mathcal{S}$
2. $\Sigma_{\mathbf{X}} = \mathbf{P}_{\mathcal{S}}\Sigma_{\mathbf{X}}\mathbf{P}_{\mathcal{S}} + \mathbf{Q}_{\mathcal{S}}\Sigma_{\mathbf{X}}\mathbf{Q}_{\mathcal{S}}$

# Partial Predictor Envelope Models

Suppose that $\mathbf{X}$ is partitioned into $\mathbf{X}_1$ and $\mathbf{X}_2$, where $\mathbf{X}_1 \in \mathbb{R}^{p_1}$ denotes a vector of continuous predictors and $\mathbf{X}_2 \in \mathbb{R}^{p_2}$ denotes a vector of categorical predictors ($p_1 + p_2 = p$).

Let $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ be the mean of $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively. Then,

$$\mathbf{Y} = \boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\beta}_1^\top(\mathbf{X}_1 - \boldsymbol{\mu}_1) + \boldsymbol{\beta}_2^\top(\mathbf{X}_2 - \boldsymbol{\mu}_2) + \boldsymbol{\epsilon},$$

We further assume a linear relationship between $\mathbf{X}_1$ and $\mathbf{X}_2$:

$$\mathbf{X}_1 = \boldsymbol{\mu}_1 + \gamma^\top(\mathbf{X}_2 - \boldsymbol{\mu}_2) + \boldsymbol{e},$$

where $\gamma$ is a $p_2 \times p_1$ matrix and $\boldsymbol{e} \in \mathbb{R}^{p_1}$ has mean $\mathbf{0}$ and is independent of $\boldsymbol{\epsilon}$.
Let $\boldsymbol{\mu}_{1|2} = E(\mathbf{X}_1|\mathbf{X}_2)$ and $\Sigma_{1|2} = \text{cov}(\mathbf{X}_1|\mathbf{X}_2)$. Then, we have $\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \gamma^\top(\mathbf{X}_2 - \boldsymbol{\mu}_2)$ and $\text{cov}(\boldsymbol{e}) = \Sigma_{1|2}$.

Impose similar assumptions as the predictor envelope models: it assumes that there is a subspace $\mathcal{S} \subseteq \mathbb{R}^{p_1}$ such that

1. $\text{cov}(\mathbf{Y}, \mathbf{Q}_{\mathcal{S}}\mathbf{X}_1 | \mathbf{P}_{\mathcal{S}}\mathbf{X}_1, \mathbf{X}_2) = 0$
2. $\text{cov}(\mathbf{P}_{\mathcal{S}}\mathbf{X}_1, \mathbf{Q}_{\mathcal{S}}\mathbf{X}_1 | \mathbf{X}_2) = 0$.

The smallest subspace $\mathcal{S}$ satisfying the two conditions above is the $\boldsymbol{\Sigma}_{1|2}$-envelope of $\boldsymbol{\beta}_1$, denoted by $\mathcal{E}_{\boldsymbol{\Sigma}_{1|2}}(\boldsymbol{\beta}_1)$, or $\mathcal{E}_{1|2}$ for short.

- Let $d$ denote the dimension of $\mathcal{E}_{1|2}$ with $0 \leq d \leq p_1$
- Let $\boldsymbol{\Gamma} \in \mathbb{R}^{p_1 \times d}$ be an orthogonal basis of $\mathcal{E}_{1|2}$
- Let $\boldsymbol{\Gamma}_0 \in \mathbb{R}^{p_1 \times (p_1 - d)}$ be an orthogonal basis $\mathcal{E}_{1|2}^{\perp}$

Under the conditions, the coordinate form of the linear regression model is

$$\mathbf{Y} = \boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\eta}^\top \boldsymbol{\Gamma}^\top (\mathbf{X}_1 - \boldsymbol{\mu}_1) + \boldsymbol{\beta}_2^\top (\mathbf{X}_2 - \boldsymbol{\mu}_2) + \boldsymbol{\epsilon},$$

$$\mathbf{X}_1 = \boldsymbol{\mu}_1 + \gamma^\top (\mathbf{X}_2 - \boldsymbol{\mu}_2) + \boldsymbol{e} \quad \text{and} \quad \boldsymbol{\Sigma}_{1|2} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^\top + \boldsymbol{\Gamma}_0 \boldsymbol{\Omega}_0 \boldsymbol{\Gamma}_0^\top,$$

where $\boldsymbol{\beta}_1 = \boldsymbol{\Gamma}\boldsymbol{\eta}$ and $\boldsymbol{\eta} \in \mathbb{R}^{d \times r}$ carries the coordinates of $\boldsymbol{\beta}_1$ with respect to $\boldsymbol{\Gamma}$. The matrices $\boldsymbol{\Omega} \in \mathbb{R}^{d \times d}$ and $\boldsymbol{\Omega}_0 \in \mathbb{R}^{(p_1 - d) \times (p_1 - d)}$ are positive definite and contain the coordinates of $\boldsymbol{\Sigma}_{1|2}$ with respect to $\boldsymbol{\Gamma}$ and $\boldsymbol{\Gamma}_0$.

# CLOSE CONNECTION WITH PREDICTOR ENVELOPE MODELS

Let $\mathbf{r}_{1|2}$ denote the population residuals from the linear regression of $\mathbf{X}_1$ on $\mathbf{X}_2$:

$$\mathbf{r}_{1|2} = \mathbf{X}_1 - \boldsymbol{\mu}_1 - \gamma^\top(\mathbf{X}_2 - \boldsymbol{\mu}_2)$$

Let $\mathbf{r}_{\mathbf{Y}|2}$ denote the population residuals from the regression of $\mathbf{Y}$ on $\mathbf{X}_2$. Then,

$$\mathbf{r}_{\mathbf{Y}|2} = \mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}} - \boldsymbol{\beta}_2^{*\top}(\mathbf{X}_2 - \boldsymbol{\mu}_2),$$

where $\boldsymbol{\beta}_2^{*\top} = \gamma\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2$. We have $\mathbf{r}_{\mathbf{Y}|2} = \boldsymbol{\beta}_1^\top \mathbf{r}_{1|2} + \boldsymbol{\epsilon}$.

The $\boldsymbol{\Sigma}_{\mathbf{r}}$-envelope of $\boldsymbol{\beta}_1$ is the smallest reducing subspace of $\boldsymbol{\Sigma}_{\mathbf{r}}$ that contains span$(\boldsymbol{\beta}_1)$.
Note: $\boldsymbol{\Sigma}_{\mathbf{r}} = \boldsymbol{\Sigma}_{1|2}$ and $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{r}}}(\boldsymbol{\beta}_1) = \mathcal{E}_{\boldsymbol{\Sigma}_{1|2}}(\boldsymbol{\beta}_1)$.

# Estimation

Let $(\mathbf{X}_{11}, \mathbf{X}_{21}, \mathbf{Y}_1), \ldots, (\mathbf{X}_{1n}, \mathbf{X}_{2n}, \mathbf{Y}_n)$ be $n$ independent observations from the model.
Let $\mathbb{X}_1^\top = (\mathbf{X}_{11}^\top, \mathbf{X}_{12}^\top, \ldots, \mathbf{X}_{1n}^\top) \in \mathbb{R}^{n \times p_1}$, $\mathbb{X}_2^\top = (\mathbf{X}_{21}^\top, \mathbf{X}_{22}^\top, \ldots, \mathbf{X}_{2n}^\top) \in \mathbb{R}^{n \times p_2}$ and $\mathbb{Y}^\top = (\mathbf{Y}_1^\top, \mathbf{Y}_2^\top, \ldots, \mathbf{Y}_n^\top) \in \mathbb{R}^{n \times r}$ be the data matrices.
For a fixed dimension $d$, the normal log likelihood of the partial predictor envelope model is given by

$$
\begin{aligned}
l = {} & \frac{n(r+p_1)}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Omega}| - \frac{n}{2}\log|\boldsymbol{\Omega}_0| - \frac{1}{2}\operatorname{tr}[\{\mathbb{X}_1 - \mathbf{1}_n\boldsymbol{\mu}_1^\top - (\mathbb{X}_2 - \mathbf{1}_n\bar{\mathbf{X}}_2^\top)\boldsymbol{\gamma}\} \\
& (\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^\top + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^\top)^{-1}\{\mathbb{X}_1 - \mathbf{1}_n\boldsymbol{\mu}_1^\top - (\mathbb{X}_2 - \mathbf{1}_n\bar{\mathbf{X}}_2^\top)\boldsymbol{\gamma}\}^\top] - \frac{n}{2}\log|\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}| \\
& - \frac{1}{2}\operatorname{tr}[\{\mathbb{Y} - \mathbf{1}_n\boldsymbol{\mu}_{\mathbf{Y}}^\top - (\mathbb{X}_1 - \mathbf{1}_n\boldsymbol{\mu}_1^\top)\boldsymbol{\Gamma}\boldsymbol{\eta} - (\mathbb{X}_2 - \mathbf{1}_n\bar{\mathbf{X}}_2^\top)\boldsymbol{\beta}_2\}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\{\mathbb{Y} - \mathbf{1}_n\boldsymbol{\mu}_{\mathbf{Y}}^\top \\
& - (\mathbb{X}_1 - \mathbf{1}_n\boldsymbol{\mu}_1^\top)\boldsymbol{\Gamma}\boldsymbol{\eta} - (\mathbb{X}_2 - \mathbf{1}_n\bar{\mathbf{X}}_2^\top)\boldsymbol{\beta}_2\}^\top],
\end{aligned}
$$

where $\bar{\mathbf{X}}_2 = (1/n)\sum_{i=1}^n \mathbf{X}_{2i}$ denotes the sample mean of $\mathbf{X}_2$, and $\mathbf{1}_n \in \mathbb{R}^n$ denotes an $n$-dimensional vector of 1's.

Let $\mathbf{S}_{1|2} = (1/n)\mathbb{X}_{1c}^T\mathbf{Q}_{\mathbb{X}_{2c}}\mathbb{X}_{1c}$, $\mathbf{S}_{\mathbf{Y}|2} = (1/n)\mathbb{Y}_c^T\mathbf{Q}_{\mathbb{X}_{2c}}\mathbb{Y}_c$, and $\mathbf{S}_{(\mathbf{Y},1)|2} = (1/n)\mathbb{Y}_c^T\mathbf{Q}_{\mathbb{X}_{2c}}\mathbb{X}_{1c}$, where $\mathbb{X}_{1c}$, $\mathbb{X}_{2c}$, and $\mathbb{Y}_c$ are the centered data matrices for $\mathbb{X}_1$, $\mathbb{X}_2$, and $\mathbb{Y}$, respectively.

- $\hat{\Gamma}$ is obtained from

$$\underset{\mathrm{span}(\boldsymbol{\Gamma})\in\mathcal{L}_{p_1\times d}}{\operatorname{argmin}} \{\log|\boldsymbol{\Gamma}^\top\mathbf{S}_{1|2}\boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}^\top\mathbf{S}_{1|2}^{-1}\boldsymbol{\Gamma}| + \log|\mathbf{S}_{\mathbf{Y}|2} - \mathbf{S}_{(\mathbf{Y},1)|2}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^\top\mathbf{S}_{1|2}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^\top\mathbf{S}_{(\mathbf{Y},1)|2}^\top|\},$$

  where $\mathcal{L}_{p_1\times d}$ denotes a $p_1 \times d$ Grassmann manifold (i.e., the collection of all $d$-dimensional subspaces of $\mathbb{R}^{p_1}$.

Let $\mathbf{R}_{1|2} = \mathbf{Q}_{\mathbb{X}_{2c}}\mathbb{X}_{1c}$ denote the sample residuals from the regression of $\mathbf{X}_1$ on $\mathbf{X}_2$, and let $\mathbf{R}_{\mathbf{Y}|2} = \mathbf{Q}_{\mathbb{X}_{2c}}\mathbb{Y}_c$ denote the sample residuals from the regression of $\mathbf{Y}$ on $\mathbf{X}_2$. Thus we have $\mathbf{S}_{1|2} = \mathbf{R}_{1|2}^\top\mathbf{R}_{1|2}/n$ and $\mathbf{S}_{(\mathbf{Y},1)|2} = \mathbf{R}_{\mathbf{Y}|2}^\top\mathbf{R}_{1|2}/n$. The estimators of the PPEM parameters are given.

- $\hat{\mu}_{\mathbf{Y}} = \bar{\mathbf{Y}}$, $\hat{\mu}_1 = \bar{\mathbf{X}}_1$
- $\hat{\gamma} = (\mathbb{X}_{2c}^\top \mathbb{X}_{2c})^{-1} \mathbb{X}_{2c}^\top \mathbb{X}_{1c}$
- $\hat{\eta} = (\hat{\boldsymbol{\Gamma}}^\top \mathbf{R}_{1|2}^\top \mathbf{R}_{1|2} \hat{\boldsymbol{\Gamma}})^{-1} \hat{\boldsymbol{\Gamma}}^\top \mathbf{R}_{1|2}^\top \mathbf{R}_{\mathbf{Y}|2}$
- $\hat{\boldsymbol{\Omega}} = (1/n) \hat{\boldsymbol{\Gamma}}^\top \mathbf{R}_{1|2}^\top \mathbf{R}_{1|2} \hat{\boldsymbol{\Gamma}} = \hat{\boldsymbol{\Gamma}}^\top \mathbf{S}_{1|2} \hat{\boldsymbol{\Gamma}}$
- $\hat{\boldsymbol{\Omega}}_0 = (1/n) \hat{\boldsymbol{\Gamma}}_0^\top \mathbf{R}_{1|2}^\top \mathbf{R}_{1|2} \hat{\boldsymbol{\Gamma}}_0 = \hat{\boldsymbol{\Gamma}}_0^\top \mathbf{S}_{1|2} \hat{\boldsymbol{\Gamma}}_0$
- $\widehat{\beta}_2 = (\mathbb{X}_{2c}^\top \mathbb{X}_{2c})^{-1} \mathbb{X}_{2c}^\top (\mathbb{Y}_c - \mathbb{X}_{1c} \widehat{\beta}_1)$
- $\hat{\boldsymbol{\Sigma}}_{\mathbf{Y}|\mathbf{X}} = (1/n) \mathbf{R}_{\mathbf{Y}|2}^\top \mathbf{Q}_{\mathbf{R}_{1|2}\boldsymbol{\Gamma}} \mathbf{R}_{\mathbf{Y}|2}$
- $\widehat{\beta}_1 = \hat{\boldsymbol{\Gamma}} \hat{\eta} = \mathbf{P}_{\hat{\boldsymbol{\Gamma}}(\mathbf{S}_{1|2})} \mathbf{S}_{1|2}^{-1} \mathbf{S}_{(1,\mathbf{Y})|2} = \mathbf{P}_{\hat{\boldsymbol{\Gamma}}(\mathbf{S}_{1|2})} \widehat{\beta}_{1,\mathrm{ols}}$
- $\hat{\boldsymbol{\Sigma}}_{1|2} = \hat{\boldsymbol{\Gamma}} \hat{\boldsymbol{\Omega}} \hat{\boldsymbol{\Gamma}}^\top + \hat{\boldsymbol{\Gamma}}_0 \hat{\boldsymbol{\Omega}}_0 \hat{\boldsymbol{\Gamma}}_0^\top = \mathbf{P}_{\hat{\boldsymbol{\Gamma}}} \mathbf{S}_{1|2} \mathbf{P}_{\hat{\boldsymbol{\Gamma}}} + \mathbf{Q}_{\hat{\boldsymbol{\Gamma}}} \mathbf{S}_{1|2} \mathbf{Q}_{\hat{\boldsymbol{\Gamma}}}$

where $\mathbf{P}_{\hat{\boldsymbol{\Gamma}}(\mathbf{S}_{1|2})}$ denotes the projection matrix onto $\mathrm{span}(\hat{\boldsymbol{\Gamma}})$ with $\mathbf{S}_{1|2}$ inner product, and $\widehat{\beta}_{1,\mathrm{ols}} = \mathbf{S}_{1|2}^{-1} \mathbf{S}_{(1,\mathbf{Y})|2}$ is the OLS estimator of $\beta_1$.

# Dimension selection

To estimate the dimension of $\mathcal{E}_{1|2}$, we apply the Bayesian information criterion (BIC).

We choose a value $d_{\text{opt}}$ by minimizing

$$\text{BIC}(d) = -2l^*(d) + \log(n)N(d),$$

where $l^*(d)$ is the maximized $l$ for a fixed $d$, and

$$N(d) = r + p_1 + p_2 + p_1p_2 + p_1(p_1+1)/2 + dr + p_2r + r(r+1)/2$$

is the number of parameters.

# THEORETICAL PROPERTIES

## Proposition 1

*Under the PPEM model, suppose that* $(\boldsymbol{\epsilon}^\top, \mathbf{e}^\top)^\top$ *has finite fourth moments and is independently and identically distributed in the sample. Let* $\widehat{\mathbf{h}}$ *denote the PPEM estimator of* $\mathbf{h} = \{\boldsymbol{\mu}_{\mathbf{Y}}^\top, \boldsymbol{\mu}_1^\top, vec^\top(\boldsymbol{\beta}_1), vec^\top(\boldsymbol{\beta}_2), vec^\top(\gamma), vech^\top(\boldsymbol{\Sigma}_{1|2}), vech^\top(\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}})\}^\top$*, then we have*

$$\sqrt{n}(\widehat{\mathbf{h}} - \mathbf{h}) \xrightarrow{d} N(0, \mathbf{U}), \qquad \mathbf{U} = \Delta(\Delta^\top \mathbf{V} \Delta)^\dagger \Delta,$$

*where* $\Delta = \partial \mathbf{h} / \partial \boldsymbol{\phi}^\top$, $\boldsymbol{\phi} = \{\boldsymbol{\mu}_{\mathbf{Y}}^\top, \boldsymbol{\mu}_1^\top, vec^\top(\boldsymbol{\eta}), vec^\top(\boldsymbol{\Gamma}), vec^\top(\boldsymbol{\beta}_2), vec^\top(\gamma), vech^\top(\boldsymbol{\Omega}), vech^\top(\boldsymbol{\Omega}_0), vech^\top(\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}})\}^\top$*, is the gradient matrix, and* $\mathbf{V}$ *is the Fisher information matrix from the standard estimation (performed by OLS). In other words,* $\mathbf{V}^{-1}$ *is the asymptotic covariance matrix of the OLS estimator of* $\mathbf{h}$*. Furthermore, since* $\mathbf{V}^{-1} - \mathbf{U}$ *is a positive semi-definite matrix, the PPEM estimator is more efficient than or as efficient as the standard estimator asymptotically.*

## Proposition 2

*Assume that the conditions in Proposition 1 hold, and we further assume that* $(\boldsymbol{\epsilon}^\top, \mathbf{e}^\top)^\top$ *is normally distributed. Then,*

$$\sqrt{n}\{vec(\widehat{\boldsymbol{\beta}}_1) - vec(\boldsymbol{\beta}_1)\} \overset{d}{\rightarrow} N(0, \mathbf{V}_1),$$

$$\sqrt{n}\{vec(\widehat{\boldsymbol{\beta}}_2) - vec(\boldsymbol{\beta}_2)\} \overset{d}{\rightarrow} N(0, \mathbf{V}_2),$$

*where*

$$\begin{aligned}\mathbf{V}_1 &= \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}} \otimes \boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^\top + (\boldsymbol{\eta}^\top \otimes \boldsymbol{\Gamma}_0)\{\boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\boldsymbol{\eta}^\top \otimes \boldsymbol{\Omega}_0 + \boldsymbol{\Omega} \otimes \boldsymbol{\Omega}_0^{-1} \\ &\quad + \boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Omega}_0 - 2\mathbf{I}_d \otimes \mathbf{I}_{p_1-d}\}^{-1}(\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_0^\top), \\ \mathbf{V}_2 &= \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}} \otimes \boldsymbol{\Sigma}_2^{-1} + \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}} \otimes \boldsymbol{\gamma}^T\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^\top\boldsymbol{\gamma} + (\boldsymbol{\eta}^\top \otimes \boldsymbol{\gamma}^\top\boldsymbol{\Gamma}_0)\{\boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\boldsymbol{\eta}^\top \otimes \boldsymbol{\Omega}_0 \\ &\quad + \boldsymbol{\Omega} \otimes \boldsymbol{\Omega}_0^{-1} + \boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Omega}_0 - 2\mathbf{I}_d \otimes \mathbf{I}_{p_1-d}\}^{-1}(\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_0^\top\boldsymbol{\gamma}).\end{aligned}$$

# DATA APPLICATION

- We analyzed the data from a study investigated in Laing et al. (2020) (COVID-IP project website `www.immunophenotype.org`).
- It included 63 COVID-19 patients and 10 non-COVID-19 patients who were hospitalized for lower respiratory tract infections as controls.
- For each patient, measurements were obtained for 26 cytokines, as well as a set of clinical information including demographics, patient status at admission, and underlying disease status.
- Among the 73 patients, 9 had missing data on BMI, ethnicity, or cytokines, and were excluded from analysis.
- Thus our analysis was based on a dataset containing 64 patients, including 26 severe cases, 22 moderate cases, 6 low cases, and 10 non-COVID patients.

- We took the logarithm of the cytokine measurements as a multivariate response vector.
- The continuous variables were 12 measurements of the patient status at admission: temperature, blood glucose, National Early Warning Score 2 (NEWS2) score, serum lactate, fraction of inspired oxygen, respiratory rate, oxygen saturation, heart rate, systolic blood pressure, diastolic blood pressure, coma score, WHO score for severity of illness.
- The categorical variables were demographics information (age, BMI, ethnicity and sex) and indicators for underlying disease status. This gave a total of 11 categorical variables.
- All variables were standardized.

We fitted the data with PPEM, PEM, PLS, PCR and OLS, and computed the prediction errors. The prediction error was obtained by five-fold cross validations with 50 random splits of the data.

- PPEM: 2.192
- PEM: 5.247
- PLS: 5.120
- PCR: 6.041
- OLS: 38.74

The estimation efficiency also led to a clear scientific interpretation of the results.

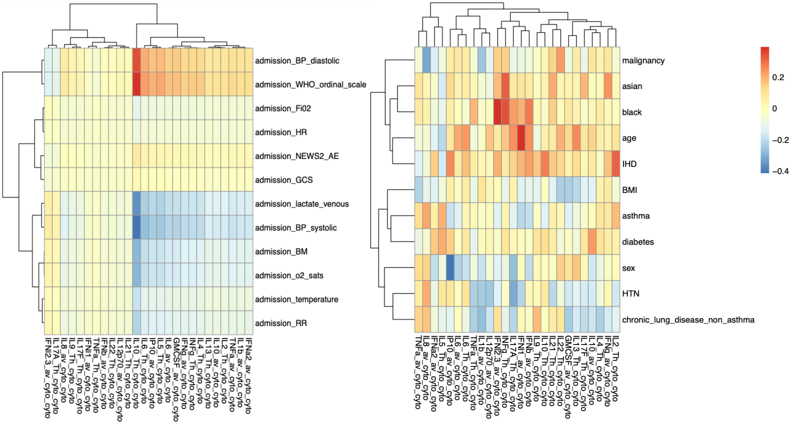


Figure: Heatmaps of the regression coefficients under PPEM.

# Take Home Messages

- We proposed a PPEM model to achieve estimation efficiency when both continuous and categorical predictors are present.
- We established root-n consistency and asymptotic normality for the PPEM estimator.
- Numerical study (not present in this presentation) and real data analysis show that PPEM can achieve more efficiency gains in estimation and produce better predictions than OLS, PLS, and PCR.
- This method is helpful to reveal the association between the cytokine-based biomarkers for COVID-19 patients and patients' clinical information including disease status at admission and demographical characteristics.

# References

- Lu et al. (2020). A potential role of interleukin-10 in COVID-19 pathogenesis. Trends in Immunology.
- Laing et al. (2020). A dynamic covid-19 immune signature includes associations with poor prognosis. Nature Medicine 26, 1623–1635.
- Takahashi et al. (2020). Sex differences in immune responses that underlie COVID-19 disease outcomes. Nature 588, 315–320.
- Genser et al. (2007). A guide to modern statistical analysis of immunological data. BMC immunology 8, 1–15.
- Schuberth et al. (2018). Partial least squares path modeling using ordinal categorical indicators. Quality & Quantity 52, 9–35.
- Lohmoller. (2013). Latent variable path modeling with partial least squares. Springer Science & Business Media.

- Hair et al. (2012). An assessment of the use of partial least squares structural equation modeling in marketing research. Journal of the academy of marketing science 40, 414–433.
- Cook et al. (2010). Envelope models for parsimonious and efficient multivariate linear regression (with discussion). Statistica Sinica 20, 927–1010.
- Conway. (1990). A course in functional analysis. New York: Springer.
- Cook. (2018). An Introduction to Envelopes: Dimension Reduction for Efficient Estimation in Multivariate Statistics. Hoboken, NJ: John Wiley & Sons.

# Thank you!