Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Envelope-based Partial Partial Least Squares with Application to Cytokine-based Biomarker Analysis for COVID-19

Yeonhee Park*[1] | Zhihua Su[2] | Dongjun Chung[3]

[1]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Wisconsin, USA
[2]Department of Statistics, University of Florida, Florida, USA
[3]Department of Biomedical Informatics, The Ohio State University, Ohio, USA

Correspondence
*Yeonhee Park, Department of Biostatistics and Medical Informatics. Email: ypark56@wisc.edu

Present Address
*Yeonhee Park, 610 Walnut street, 207B WARF Building, Madison, WI 53726

Abstract

Partial least squares (PLS) regression is a popular alternative to ordinary least squares regression because of its superior prediction performance demonstrated in many cases. In various contemporary applications, the predictors include both continuous and categorical variables. A common practice in PLS regression is to treat the categorical variable as continuous. However, studies find that this practice may lead to biased estimates and invalid inferences[1]. Based on a connection between the envelope model and PLS, we develop an envelope-based partial PLS estimator that considers the PLS regression on the conditional distributions of the response(s) and continuous predictors on the categorical predictors. Root-n consistency and asymptotic normality are established for this estimator. Numerical study shows that this approach can achieve more efficiency gains in estimation and produce better predictions. The method is applied for the identification of cytokine-based biomarkers for COVID-19 patients, which reveals the association between the cytokine-based biomarkers and patients' clinical information including disease status at admission and demographical characteristics. The efficient estimation leads to a clear scientific interpretation of the results.

KEYWORDS:
Dimension reduction; envelope model; Grassmann manifold; multivariate regression; partial least squares

1 | INTRODUCTION

COVID-19 is a worldwide pandemic. As of April 2021, it has infected more than 147 million people and caused more than 3.1 million deaths worldwide[2]. Despite tremendous efforts to improve the diagnosis and treatment of COVID-19, we still have a limited understanding of the associations between the key immunologic factors and the clinical information of the COVID-19 patients. These associations can aid in the treatment and management of the disease. Many studies on COVID-19 patients have collected data on various biomarkers, such as the COVID-IP project[3]. It would be of great scientific and medical interest to develop a new statistical tool that facilitates the identification of such associations from the COVID-19 datasets.

The multivariate linear regression model is a common tool for the investigation of the association between key immunologic factors (such as cytokines) and COVID-19 patients' clinical information[4]. Compared to the traditional ordinary least squares (OLS) fitting, partial least squares (PLS) is a popular alternative known for its superior prediction performance[5,6]. It is originated in econometrics[7] and is now widely used in many applied disciplines including chemometrics, social science, food science, and genetics. Among the various applications, it is common to have both categorical and continuous variables in the predictors. For

example, in the COVID-19 dataset, to study the impact of clinical variables on cytokines levels, categorical predictors include the patient's sex, ethnicity, and indicators for underlying diseases such as asthma and diabetes, and continuous predictors include patient's clinical status such as temperature, respiratory rate, and oxygen saturation. When the data have both continuous and categorical predictors, a common practice is to treat the categorical predictors as continuous. However, practitioners discovered that this can "lead to biased estimates and therefore to invalid inferences and erroneous conclusions"[1], see also Lohmoller (2013)[8] and Hair et al. (2012)[9].

In this paper, we resolve the issue via the link between PLS and the envelope model. The envelope model was first proposed in Cook et al. (2010)[10] which achieves estimation efficiency in multivariate linear regression using dimension reduction techniques. Cook et al. (2013)[11] discovers a link between PLS and the envelope model that in a population they are estimating the same parameter but use different sample estimation methods. Since its first introduction, PLS stands as an iterative moment-based algorithm instead of a model-based method. It is easy to use and fast to compute, but it is difficult to obtain a complete understanding of its properties and make improvements to overcome its disadvantages. On the other hand, the estimation of the envelope model uses a model-based objective function, which facilitates the theoretical investigation of its estimator. The link between PLS and envelope model enables us to study PLS via the envelope model and design new variants to make it more adaptive to different data structures.

The article aims to develop an envelope-based partial PLS (EPPLS) estimator. 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. This provides us with a $\sqrt{n}$-consistent estimator for the regression coefficients, and this estimator achieves more efficient gains and better prediction performance than OLS, PLS, and principal component regression (PCR) in our numerical study and the COVID-19 dataset. We also establish consistency and the asymptotic distribution of this estimator. In addition, using the link of PLS and the envelope model, we derive a partial PLS (PPLS) algorithm, which is analogous to the PLS algorithm.

The rest of the article is organized as follows. A review of the envelope methodology, as well as the link between PLS and the envelope model, is provided in Section 2. We propose the EPPLS, and discuss the estimation, theoretical properties, and order determination in Section 3. Based on the link between PLS and the envelope model, Section 4 derives a moment-based iterative algorithm that yields a PPLS estimator. The numerical performance of the proposed estimators is investigated in Section 5 via simulations. The analysis of a COVID-19 dataset is elaborated in Section 6. We conclude the paper with a discussion in Section 7.

2 | REVIEW OF THE ENVELOPE MODEL AND ITS CONNECTION TO PLS

The envelope model is first introduced by Cook et al. (2010)[10] as an efficient method to estimate the regression coefficients under the context of multivariate linear regression. It uses sufficient dimension reduction techniques to identify the part of the data that is immaterial to the estimation goal. The subsequent estimation is only based on the material part and is thus more efficient. The envelope model has since been adapted to many areas including PLS[11,12,13], generalized linear models[14], spatial regression model[15], variable selection[16], Bayesian analysis[17,18] and tensor regression[19,20]. Codes for fitting the envelope models are included in `R` package `Renvlp` available in CRAN. A complete review of the envelope model is in Cook (2018)[21].

Among all the envelope models, the predictor envelope model[11] is most related to the background of our discussion. Thus we review the envelope model under the context of the predictor envelope model. Consider a linear regression model

$$\mathbf{Y} = \boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\beta}^T(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}) + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{Y} \in \mathbb{R}^r$ is the univariate response ($r = 1$) or multivariate response vector ($r > 1$) with mean $\boldsymbol{\mu}_{\mathbf{Y}}$, $\mathbf{X}$ is a $p\times 1$ predictor with mean $\boldsymbol{\mu}_{\mathbf{X}}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{X}}$, $\boldsymbol{\epsilon} \in \mathbb{R}^r$ denotes the error vector with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$, and $\boldsymbol{\beta} \in \mathbb{R}^{p\times r}$ denotes unknown regression coefficients. The predictor envelope model assumes that part of $\mathbf{X}$ is immaterial to the regression and does not affect the distribution of $\mathbf{Y}$ directly or indirectly. Specifically the predictor envelope model assumes that there is a subspace $\mathcal{S} \subseteq \mathbb{R}^p$ such that

$$\text{(a) } \operatorname{cov}(\mathbf{Y}, \mathbf{Q}_{\mathcal{S}}\mathbf{X} \mid \mathbf{P}_{\mathcal{S}}\mathbf{X}) = 0 \quad \text{and} \quad \text{(b) } \operatorname{cov}(\mathbf{P}_{\mathcal{S}}\mathbf{X}, \mathbf{Q}_{\mathcal{S}}\mathbf{X}) = 0, \tag{2}$$

where $\mathbf{P}_{\cdot}$ denotes the projection matrix, $\mathbf{I}$ denotes the identity matrix, and $\mathbf{Q}_{\cdot} = \mathbf{I} - \mathbf{P}_{\cdot}$. Assumption (2a) states that $\mathbf{Q}_{\mathcal{S}}\mathbf{X}$ provides no information about $\mathbf{Y}$ given $\mathbf{P}_{\mathcal{S}}\mathbf{X}$, and (2b) implies that $\mathbf{Q}_{\mathcal{S}}\mathbf{X}$ is uncorrelated with $\mathbf{P}_{\mathcal{S}}\mathbf{X}$. Cook et al. (2013)[11] proved that assumptions in (2) are equivalent to imposing the following structure to the model parameters

$$\text{(c) } \operatorname{span}(\boldsymbol{\beta}) \subseteq \mathcal{S} \quad \text{and} \quad \text{(d) } \boldsymbol{\Sigma}_{\mathbf{X}} = \mathbf{P}_{\mathcal{S}}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{P}_{\mathcal{S}} + \mathbf{Q}_{\mathcal{S}}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{Q}_{\mathcal{S}}. \tag{3}$$

The structure of $\boldsymbol{\beta}$ in (3c) asserts that the span of $\boldsymbol{\beta}$ is contained in $\mathcal{S}$. When $\boldsymbol{\Sigma}_{\mathbf{X}}$ can be decomposed as in (3d), then $\mathcal{S}$ is called a reducing subspace of $\boldsymbol{\Sigma}_{\mathbf{X}}$ [22]. The $\boldsymbol{\Sigma}_{\mathbf{X}}$-envelope of $\boldsymbol{\beta}$, denoted by $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta})$, is the smallest reducing subspace that contains span($\boldsymbol{\beta}$). In other words, $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta})$ is the smallest subspace that satisfies (3), or equivalently the assumptions in (2). If appears in subscripts, $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta})$ is abbreviated to $\mathcal{E}$. We call $\mathbf{P}_{\mathcal{E}}\mathbf{X}$ the material part of $\mathbf{X}$ and $\mathbf{Q}_{\mathcal{E}}\mathbf{X}$ the immaterial part. Let u ($0 \leq u \leq p$) denote the dimension of $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta})$, $\mathbf{G} \in \mathbb{R}^{p\times u}$ an orthonormal basis of $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta})$, and $\mathbf{G}_0 \in \mathbb{R}^{p\times(p-u)}$ an orthonormal basis of $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta})^{\perp}$, i.e. the orthogonal complement of $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta})$. Then the coordinate form of the predictor envelope model is

$$\mathbf{Y} = \boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\xi}^T\mathbf{G}^T(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}) + \boldsymbol{\epsilon}, \qquad \boldsymbol{\Sigma}_{\mathbf{X}} = \mathbf{G}\boldsymbol{\Delta}\mathbf{G}^T + \mathbf{G}_0\boldsymbol{\Delta}_0\mathbf{G}_0^T, \tag{4}$$

where $\boldsymbol{\beta} = \mathbf{G}\boldsymbol{\xi}$, $\boldsymbol{\xi} \in \mathbb{R}^{u\times r}$ carries the coordinates of $\boldsymbol{\beta}$ with respect to $\mathbf{G}$, $\boldsymbol{\Delta} \in \mathbb{R}^{u\times u}$ and $\boldsymbol{\Delta}_0 \in \mathbb{R}^{(p-u)\times(p-u)}$ carries the coordinates of $\boldsymbol{\Sigma}_{\mathbf{X}}$ with respect to $\mathbf{G}$ and $\mathbf{G}_0$. From (4), $\boldsymbol{\Sigma}_{\mathbf{X}}$ is partitioned into the variation of the material part $\mathbf{P}_{\mathcal{E}}\mathbf{X}$ and the variation of the immaterial part $\mathbf{Q}_{\mathcal{E}}\mathbf{X}$.

Estimation of the predictor envelope model uses normal likelihood as an objective function, and performs a manifold optimization to obtain the estimator of $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta})$. Let $\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}$ be the sample variance matrix of $\mathbf{X}$, and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}|\mathbf{Y}}$ the sample conditional covariance matrix of $\mathbf{X}$ given $\mathbf{Y}$, then

$$\widehat{\mathcal{E}}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}) = \underset{\text{span}(\mathbf{G})\in\mathcal{L}_{p\times u}}{\text{argmin}} \{\log|\mathbf{G}^T\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}^{-1}\mathbf{G}| + \log|\mathbf{G}^T\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}|\mathbf{Y}}\mathbf{G}|\}, \tag{5}$$

where $\mathcal{L}_{p\times u}$ denotes $p \times u$ Grassmann manifold, which is the set of all u dimensional subspace of a p dimensional space. Once we have $\widehat{\mathcal{E}}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta})$, $\widehat{\mathbf{G}}$, the estimator of $\mathbf{G}$, can be taken as any orthonormal basis of $\widehat{\mathcal{E}}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta})$. Then the predictor envelope estimator of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}} = \widehat{\mathbf{G}}(\widehat{\mathbf{G}}^T\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}\widehat{\mathbf{G}})^{-1}\widehat{\mathbf{G}}^T\widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}}$, where $\widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}}$ is the sample covariance matrix of $\mathbf{X}$ and $\mathbf{Y}$. Cook et al. (2013) [11] shows that the predictor envelope estimator is asymptotically more efficient or as efficient as the OLS estimator.

The predictor envelope model has a close connection with PLS. PLS aims to find a reduction of $\mathbf{X}$, i.e. $\mathbf{W}^T\mathbf{X}$, where $\mathbf{W} \in \mathbb{R}^{p\times u}$, and then estimates $\boldsymbol{\beta}$ based on the regression of $\mathbf{Y}$ on $\mathbf{W}^T\mathbf{X}$. PLS uses sequential moment-based method to estimate $\mathbf{W}$ columnwise, and different variants of PLS use slightly different algorithms. We take SIMPLS [23] as an example, which is a popular variant implemented in various software packages. Suppose that $\mathbf{w}_i \in \mathbb{R}^p$ is the vector obtained in the ith step ($i \leq u$). Let $\boldsymbol{\Sigma}_{\mathbf{XY}}$ denote the covariance matrix between $\mathbf{X}$ and $\mathbf{Y}$. Set $\mathbf{W}_0 = 0$. At the $k+1$th step, let $\mathbf{W}_k = (\mathbf{w}_1, \ldots, \mathbf{w}_k) \in \mathbb{R}^{p\times k}$, then $\mathbf{w}_{k+1}$ is obtained by

$$\mathbf{w}_{k+1} = \arg\max_{\mathbf{w}} \mathbf{w}^T\boldsymbol{\Sigma}_{\mathbf{XY}}\boldsymbol{\Sigma}_{\mathbf{XY}}^T\mathbf{w} \quad \text{subject to} \quad \mathbf{w}^T\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{W}_k = 0 \text{ and } \mathbf{w}^T\mathbf{w} = 1. \tag{6}$$

Let $\mathcal{W}_k = \text{span}(\mathbf{W}_k)$. If we change the length constrains in (6) to $\mathbf{w}^T\mathbf{Q}_{\mathcal{W}_k}\mathbf{w} = 1$, then we obtain another popular PLS variant NIPALS [24]. Once an estimator of $\mathbf{W}$ is obtained, denoted by $\widehat{\mathbf{W}}$, the PLS estimator of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}} = \widehat{\mathbf{W}}(\widehat{\mathbf{W}}^T\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}\widehat{\mathbf{W}})^{-1}\widehat{\mathbf{W}}^T\widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}}$.

Cook et al. (2013) [11] shows a close connection between SIMPLS and the predictor envelope model: $\mathcal{W}_1 \subset \mathcal{W}_2 \subset \cdots \subset \mathcal{W}_u = \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta})$. This indicates that at the population level SIMPLS seeks the same reduction as the predictor envelope model. At the sample level, SIMPLS and the predictor envelope model use different algorithms to estimate $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta})$. SIMPLS uses the moment-based algorithm (6) while the predictor envelope model uses a likelihood-based method (5). Based on this connection, we are able to study the properties of the SIMPLS estimator or develop its extensions through the predictor envelope model, and we call the predictor envelope model (4) envelope-based partial least squares (EPLS) hereafter.

3 | ENVELOPE-BASED PARTIAL PARTIAL LEAST SQUARES

3.1 | Formulation

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 the linear regression model in (1) can be written as

$$\mathbf{Y} = \boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\beta}_1^T(\mathbf{X}_1 - \boldsymbol{\mu}_1) + \boldsymbol{\beta}_2^T(\mathbf{X}_2 - \boldsymbol{\mu}_2) + \boldsymbol{\epsilon}, \tag{7}$$

where $\boldsymbol{\beta}_1 \in \mathbb{R}^{p_1\times r}$ denotes the regression coefficients for the continuous predictors and $\boldsymbol{\beta}_2 \in \mathbb{R}^{p_2\times r}$ denotes the regression coefficients for the categorical predictors. We further assume a working model between $\mathbf{X}_1$ and $\mathbf{X}_2$, $\mathbf{X}_1 = \boldsymbol{\mu}_1 + \boldsymbol{\gamma}^T(\mathbf{X}_2 - \boldsymbol{\mu}_2) + \mathbf{e}$, where $\boldsymbol{\gamma}$ is a $p_2\times p_1$ matrix, and $\mathbf{e} \in \mathbb{R}^{p_1}$ has mean $\mathbf{0}$ and is independent of $\boldsymbol{\epsilon}$ and $\mathbf{X}_2$. Let $\boldsymbol{\mu}_{1|2} = \text{E}(\mathbf{X}_1|\mathbf{X}_2)$ and $\boldsymbol{\Sigma}_{1|2} = \text{cov}(\mathbf{X}_1|\mathbf{X}_2)$. Then we have $\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \boldsymbol{\gamma}^T(\mathbf{X}_2 - \boldsymbol{\mu}_2)$ and $\text{cov}(\mathbf{e}) = \boldsymbol{\Sigma}_{1|2}$.

We impose similar assumptions as EPLS, but on the conditional distribution given $\mathbf{X}_2$. More specifically, it assumes that there is a subspace $\mathcal{S}$ of $\mathbb{R}^{p_1}$ such that

$$\text{(i) } \operatorname{cov}(\mathbf{Y}, \mathbf{Q}_{\mathcal{S}}\mathbf{X}_1 \mid \mathbf{P}_{\mathcal{S}}\mathbf{X}_1, \mathbf{X}_2) = 0 \quad \text{and} \quad \text{(ii) } \operatorname{cov}(\mathbf{Q}_{\mathcal{S}}\mathbf{X}_1, \mathbf{P}_{\mathcal{S}}\mathbf{X}_1 \mid \mathbf{X}_2) = 0. \tag{8}$$

Condition (i) indicates that given $\mathbf{P}_{\mathcal{S}}\mathbf{X}_1$ and $\mathbf{X}_2$, $\mathbf{Q}_{\mathcal{S}}\mathbf{X}_1$ provides no information about $\mathbf{Y}$, and condition (ii) implies that after removing the effects of $\mathbf{X}_2$, $\mathbf{Q}_{\mathcal{S}}\mathbf{X}_1$ is uncorrelated with $\mathbf{P}_{\mathcal{S}}\mathbf{X}_1$. Condition (i) implies that $\operatorname{span}(\boldsymbol{\beta}_1) \subseteq \mathcal{S}$ and condition (ii) implies $\mathcal{S}$ is a reducing subspace of $\boldsymbol{\Sigma}_{1|2}$. The smallest subspace $\mathcal{S}$ that satisfies both (i) and (ii) in (8) 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. Thus we have (iii) $\operatorname{span}(\boldsymbol{\beta}_1) \subseteq \mathcal{E}_{1|2}(\boldsymbol{\beta}_1)$ and (iv) $\boldsymbol{\Sigma}_{1|2} = \mathbf{P}_{\mathcal{E}_{1|2}}\boldsymbol{\Sigma}_{1|2}\mathbf{P}_{\mathcal{E}_{1|2}} + \mathbf{Q}_{\mathcal{E}_{1|2}}\boldsymbol{\Sigma}_{1|2}\mathbf{Q}_{\mathcal{E}_{1|2}}$. Note that just as the EPLS model, conditions (iii) and (iv) are equivalent to conditions (i) and (ii). Let d denote the dimension of $\mathcal{E}_{1|2}$ with $0 \le d \le p_1$, $\boldsymbol{\Gamma} \in \mathbb{R}^{p_1 \times d}$ an orthonormal basis of $\mathcal{E}_{\boldsymbol{\Sigma}_{1|2}}(\boldsymbol{\beta}_1)$ and $\boldsymbol{\Gamma}_0 \in \mathbb{R}^{p_1 \times (p_1 - d)}$ an orthonormal basis of $\mathcal{E}_{\boldsymbol{\Sigma}_{1|2}}(\boldsymbol{\beta}_1)^{\perp}$. When (iii) and (iv) are satisfied, the coordinate form of the linear regression model (7) is

$$\begin{aligned} \mathbf{Y} &= \boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\eta}^T\boldsymbol{\Gamma}^T(\mathbf{X}_1 - \boldsymbol{\mu}_1) + \boldsymbol{\beta}_2^T(\mathbf{X}_2 - \boldsymbol{\mu}_2) + \boldsymbol{\epsilon}, \\ \mathbf{X}_1 &= \boldsymbol{\mu}_1 + \boldsymbol{\gamma}^T(\mathbf{X}_2 - \boldsymbol{\mu}_2) + \mathbf{e} \quad \text{and} \quad \boldsymbol{\Sigma}_{1|2} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T, \end{aligned} \tag{9}$$

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 $\boldsymbol{\Gamma}_0$. Since the envelope structure is imposed on part of the predictors, we call (9) the envelope-based partial PLS (EPPLS) model. When $d = p_1$, the EPPLS model reduces to the standard linear regression model (7).

The EPPLS model (9) has a close connection with the EPLS model (4). Let $\mathbf{r}_{1|2}$ denote the population residuals from the linear regression of $\mathbf{X}_1$ on $\mathbf{X}_2$, i.e. $\mathbf{r}_{1|2} = \mathbf{X}_1 - \boldsymbol{\mu}_1 - \boldsymbol{\gamma}^T(\mathbf{X}_2 - \boldsymbol{\mu}_2)$. Then the linear model (7) can be reparameterized as $\mathbf{Y} = \boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\beta}_1^T\mathbf{r}_{1|2} + \boldsymbol{\beta}_2^{*T}(\mathbf{X}_2 - \boldsymbol{\mu}_2) + \boldsymbol{\epsilon}$, where $\boldsymbol{\beta}_2^* = \boldsymbol{\gamma}\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2$ is a linear combination of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. Let $\mathbf{r}_{\mathbf{Y}|2}$ denote the population residuals from the regression of $\mathbf{Y}$ on $\mathbf{X}_2$, then $\mathbf{r}_{\mathbf{Y}|2} = \mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}} - \boldsymbol{\beta}_2^{*T}(\mathbf{X}_2 - \boldsymbol{\mu}_2)$. Based on the reparametrization, we have

$$\mathbf{r}_{\mathbf{Y}|2} = \boldsymbol{\beta}_1^T\mathbf{r}_{1|2} + \boldsymbol{\epsilon}, \tag{10}$$

which presents a multivariate linear regression model of $\mathbf{r}_{\mathbf{Y}|2}$ on $\mathbf{r}_{1|2}$. Now we impose the EPLS structure (4) on (10). Let $\boldsymbol{\Sigma}_{\mathbf{r}}$ be the covariance matrix of $\mathbf{r}_{1|2}$. Then the $\boldsymbol{\Sigma}_{\mathbf{r}}$-envelope of $\boldsymbol{\beta}_1$, denoted by $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{r}}}(\boldsymbol{\beta}_1)$, is the smallest reducing subspace of $\boldsymbol{\Sigma}_{\mathbf{r}}$ that contains $\operatorname{span}(\boldsymbol{\beta}_1)$. Since $\boldsymbol{\Sigma}_{\mathbf{r}} = \boldsymbol{\Sigma}_{1|2}$, $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{r}}}(\boldsymbol{\beta}_1)$ is the same as $\mathcal{E}_{\boldsymbol{\Sigma}_{1|2}}(\boldsymbol{\beta}_1)$ in the EPPLS model (9). This relationship is analogous to the connection between the partial envelope model and the response envelope model for the residuals described in Su and Cook (2011)[25].

3.2 | Estimation

We use the normal likelihood as an objective function for 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 EPPLS model. Let $\mathbb{X}_1^T = (\mathbf{X}_{11}^T, \mathbf{X}_{12}^T, \ldots, \mathbf{X}_{1n}^T) \in \mathbb{R}^{p_1 \times n}$, $\mathbb{X}_2^T = (\mathbf{X}_{21}^T, \mathbf{X}_{22}^T, \ldots, \mathbf{X}_{2n}^T) \in \mathbb{R}^{p_2 \times n}$ and $\mathbb{Y}^T = (\mathbf{Y}_1^T, \mathbf{Y}_2^T, \ldots, \mathbf{Y}_n^T) \in \mathbb{R}^{r \times n}$ be the data matrices, and $\bar{\mathbf{X}}_1$, $\bar{\mathbf{X}}_2$ and $\bar{\mathbf{Y}}$ the sample means of $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{Y}$. Then $\mathbb{X}_{1c} = \mathbb{X}_1 - \mathbf{1}_n\bar{\mathbf{X}}_1^T$, $\mathbb{X}_{2c} = \mathbb{X}_2 - \mathbf{1}_n\bar{\mathbf{X}}_2^T$, and $\mathbb{Y}_c = \mathbb{Y} - \mathbf{1}_n\bar{\mathbf{Y}}^T$ are the centered data matrices for $\mathbb{X}_1$, $\mathbb{X}_2$, and $\mathbb{Y}$, respectively. The parameters under the EPPLS model are $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_{\mathbf{Y}}$, $\boldsymbol{\eta}$, $\boldsymbol{\beta}_2$, $\boldsymbol{\Omega}$, $\boldsymbol{\Omega}_0$, $\operatorname{span}(\boldsymbol{\Gamma})$, $\boldsymbol{\gamma}$ and $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$. Note that $\boldsymbol{\Gamma}$ is not identifiable, only $\operatorname{span}(\boldsymbol{\Gamma})$ is identifiable. We first fix an orthonormal basis $\boldsymbol{\Gamma}$ and estimate other parameters by maximizing the objective function. Based on the derivations in Supplemental Material, the estimators of these parameters can be written as explicit functions of $\boldsymbol{\Gamma}$. We substitute them back to the objective function, which now only has one parameter $\operatorname{span}(\boldsymbol{\Gamma})$, i.e. $\mathcal{E}_{\boldsymbol{\Sigma}_{1|2}}(\boldsymbol{\beta}_1)$. 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 $\mathbf{S}_{(\mathbf{Y},1)|2} = (1/n)\mathbb{Y}_c^T\mathbf{Q}_{\mathbb{X}_{2c}}\mathbb{X}_{1c}$ denote the sample conditional variance of $\mathbf{X}_1$ given $\mathbf{X}_2$, the sample conditional variance of $\mathbf{Y}$ given $\mathbf{X}_2$ and the sample conditional covariance between $\mathbf{Y}$ and $\mathbf{X}_1$ given $\mathbf{X}_2$, respectively. The estimator of the EPPLS can be obtained by solving the following optimization problem

$$\widehat{\mathcal{E}}_{\boldsymbol{\Sigma}_{1|2}}(\boldsymbol{\beta}_1) = \underset{\operatorname{span}(\boldsymbol{\Gamma}) \in \mathcal{L}_{p_1 \times d}}{\operatorname{argmin}} \{\log|\mathbf{S}_{\mathbf{Y}|2}| + \log|\boldsymbol{\Gamma}^T\mathbf{S}_{1|2}^{-1}\boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}^T\mathbf{S}_{1|(\mathbf{Y},2)}\boldsymbol{\Gamma}|\}, \tag{11}$$

where $\mathcal{L}_{p_1 \times d}$ denotes a $p_1 \times d$ Grassmann manifold. Details are provided in Section A of the Supplemental materials. Note that the objective function in (11) has the same form as the objective function for the EPLS model in (5) with $\mathbf{r}_{\mathbf{Y}|2}$ being the response and $\mathbf{r}_{1|2}$ being the predictor, which echoes the relationship between EPPLS model and the EPLS model discussed at the end of Section 3.1.

The optimization in (11) can be solved using the computing algorithm in Cook et al. (2016)[26], or applying existing softwares such as the `R` package `Renvlp`. Once we have $\widehat{\mathcal{E}}_{\boldsymbol{\Sigma}_{1|2}}(\boldsymbol{\beta}_1)$, $\widehat{\boldsymbol{\Gamma}}$ can be taken to be any orthonormal basis of $\widehat{\mathcal{E}}_{\boldsymbol{\Sigma}_{1|2}}(\boldsymbol{\beta}_1)$, and $\widehat{\boldsymbol{\Gamma}}_0$ can be

taken to be any orthonormal basis of $\widehat{\mathcal{E}}_{\mathbf{\Sigma}_{1|2}}(\boldsymbol{\beta}_1)^{\perp}$. 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 $\mathbf{S}_{1|2} = \mathbf{R}_{1|2}^T\mathbf{R}_{1|2}/n$ and $\mathbf{S}_{(\mathbf{Y},1)|2} = \mathbf{R}_{\mathbf{Y}|2}^T\mathbf{R}_{1|2}/n$. The estimators of the EPPLS parameters are

$$\begin{aligned}
&\hat{\boldsymbol{\mu}}_{\mathbf{Y}} = \bar{\mathbf{Y}}, && \hat{\boldsymbol{\mu}}_1 = \bar{\mathbf{X}}_1 && \hat{\boldsymbol{\mu}}_2 = \bar{\mathbf{X}}_2\\
&\hat{\boldsymbol{\gamma}} = (\mathbb{X}_{2c}^T\mathbb{X}_{2c})^{-1}\mathbb{X}_{2c}^T\mathbb{X}_{1c}, && \hat{\boldsymbol{\eta}} = (\widehat{\boldsymbol{\Gamma}}^T\mathbf{R}_{1|2}^T\mathbf{R}_{1|2}\widehat{\boldsymbol{\Gamma}})^{-1}\widehat{\boldsymbol{\Gamma}}^T\mathbf{R}_{1|2}^T\mathbf{R}_{\mathbf{Y}|2},\\
&\widehat{\boldsymbol{\Omega}} = (1/n)\widehat{\boldsymbol{\Gamma}}^T\mathbf{R}_{1|2}^T\mathbf{R}_{1|2}\widehat{\boldsymbol{\Gamma}} = \widehat{\boldsymbol{\Gamma}}^T\mathbf{S}_{1|2}\widehat{\boldsymbol{\Gamma}}, && \widehat{\boldsymbol{\Omega}}_0 = (1/n)\widehat{\boldsymbol{\Gamma}}_0^T\mathbf{R}_{1|2}^T\mathbf{R}_{1|2}\widehat{\boldsymbol{\Gamma}}_0 = \widehat{\boldsymbol{\Gamma}}_0^T\mathbf{S}_{1|2}\widehat{\boldsymbol{\Gamma}}_0,\\
&\widehat{\boldsymbol{\beta}}_2 = (\mathbb{X}_{2c}^T\mathbb{X}_{2c})^{-1}\mathbb{X}_{2c}^T(\mathbb{Y}_c - \mathbb{X}_{1c}\widehat{\boldsymbol{\beta}}_1), && \widehat{\mathbf{\Sigma}}_{\mathbf{Y}|\mathbf{X}} = (1/n)\mathbf{R}_{\mathbf{Y}|2}^T\mathbf{Q}_{\mathbf{R}_{1|2}\boldsymbol{\Gamma}}\mathbf{R}_{\mathbf{Y}|2}.
\end{aligned} \tag{12}$$

Then

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_1 &= \widehat{\boldsymbol{\Gamma}}\hat{\boldsymbol{\eta}} = \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}(\mathbf{S}_{1|2})}\mathbf{S}_{1|2}^{-1}\mathbf{S}_{(1,\mathbf{Y})|2} = \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}(\mathbf{S}_{1|2})}\widehat{\boldsymbol{\beta}}_{1,\text{ols}},\\
\widehat{\mathbf{\Sigma}}_{1|2} &= \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Omega}}\widehat{\boldsymbol{\Gamma}}^T + \widehat{\boldsymbol{\Gamma}}_0\widehat{\boldsymbol{\Omega}}_0\widehat{\boldsymbol{\Gamma}}_0^T = \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}\mathbf{S}_{1|2}\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}} + \mathbf{Q}_{\widehat{\boldsymbol{\Gamma}}}\mathbf{S}_{1|2}\mathbf{Q}_{\widehat{\boldsymbol{\Gamma}}},
\end{aligned} \tag{13}$$

where $\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}(\mathbf{S}_{1|2})}$ denotes the projection matrix onto span($\widehat{\boldsymbol{\Gamma}}$) with $\mathbf{S}_{1|2}$ inner product, and $\widehat{\boldsymbol{\beta}}_{1,\text{ols}} = \mathbf{S}_{1|2}^{-1}\mathbf{S}_{(1,\mathbf{Y})|2}$ is the OLS estimator of $\boldsymbol{\beta}_1$. Thus the EPPLS estimator $\widehat{\boldsymbol{\beta}}_1$ is obtained by projecting the OLS estimator onto $\widehat{\mathcal{E}}_{\mathbf{\Sigma}_{1|2}}(\boldsymbol{\beta}_1)$ with the $\mathbf{S}_{1|2}$ inner product. The estimator $\widehat{\boldsymbol{\beta}}_2$ has the same expression as its OLS estimator, except that $\widehat{\boldsymbol{\beta}}_{1,\text{ols}}$ is replaced by the EPPLS estimator $\widehat{\boldsymbol{\beta}}_1$.

When $p_1 > n$, the matrices $\mathbf{S}_{1|2}$ and $\mathbf{S}_{(1,\mathbf{Y})|2}$ in (11) are singular. Since the objective function in (11) depends on the inverse of $\mathbf{S}_{1|2}$, and the inverse of $\mathbf{S}_{(1,\mathbf{Y})|2}$ is required in the algorithm to solve (11), we use a high-dimensional precision matrix estimator to replace $\mathbf{S}_{1|2}^{-1}$ and $\mathbf{S}_{(1,\mathbf{Y})|2}^{-1}$. While many precision matrix estimators are applicable, e.g., Sun and Zhang (2013)[27], Zhang and Zou (2014)[28], Khare et al. (2015)[29], we adopt the sparse permutation invariant covariance estimator[30] SPICE since it guarantees to have a positive definite matrix and its consistency does not rely on any sparsity assumption. We use the `R` package `PDSCE` to compute the SPICE estimators of $\mathbf{S}_{1|2}^{-1}$ and $\mathbf{S}_{(1,\mathbf{Y})|2}^{-1}$, and denote the resulting estimators as $\mathbf{S}_{1|2,\text{sp}}^{-1}$ and $\mathbf{S}_{(1,\mathbf{Y})|2,\text{sp}}^{-1}$. Then $\mathbf{S}_{1|2,\text{sp}}$ and $\mathbf{S}_{(1,\mathbf{Y})|2,\text{sp}}$ replace $\mathbf{S}_{1|2}$ and $\mathbf{S}_{(1,\mathbf{Y})|2}$ in (11), as well as in the estimators in (12) and (13).

3.3 | Theoretical Properties

In this section, we establish consistency and asymptotic distribution of the EPPLS estimator. Let vec denote the vector operator that stacks the columns of a matrix to a vector, and let vech denote the vector half operator that stacks the lower triangle of a symmetric matrix to a vector. We use $\otimes$ to denote the Kronecker product, $\dagger$ to denote the Moore-Penrose generalized inverse, and $\xrightarrow{d}$ to denote convergence in distribution. The parameters in (9) include $\mathbf{h} = \{\boldsymbol{\mu}_{\mathbf{Y}}^T, \boldsymbol{\mu}_1^T, \text{vec}^T(\boldsymbol{\beta}_1), \text{vec}^T(\boldsymbol{\beta}_2), \text{vec}^T(\boldsymbol{\gamma}), \text{vech}^T(\mathbf{\Sigma}_{1|2}), \text{vech}^T(\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}})\}^T$, and the constituent parameters of the EPPLS model are $\boldsymbol{\phi} = \{\boldsymbol{\mu}_{\mathbf{Y}}^T, \boldsymbol{\mu}_1^T, \text{vec}^T(\boldsymbol{\eta}), \text{vec}^T(\boldsymbol{\Gamma}), \text{vec}^T(\boldsymbol{\beta}_2), \text{vec}^T(\boldsymbol{\gamma}), \text{vech}^T(\boldsymbol{\Omega}), \text{vech}^T(\boldsymbol{\Omega}_0), \text{vech}^T(\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}})\}^T$. Under the EPPLS model, $\mathbf{h}$ is a function of $\boldsymbol{\phi}$. Proposition 1 indicates that the EPPLS estimator is $\sqrt{n}$ consistent and asymptotically normal even the errors are not normally distributed.

Proposition 1. Suppose that the EPPLS model (9) holds, $(\boldsymbol{\epsilon}^T, \mathbf{e}^T)^T$ has finite fourth moments and is independently and identically distributed in the sample. Let $\widehat{\mathbf{h}}$ denote the EPPLS estimator of $\mathbf{h}$, then we have

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

where $\boldsymbol{\Delta} = \partial\mathbf{h}/\partial\boldsymbol{\phi}^T$ 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 EPPLS estimator is more efficient than or as efficient as the standard estimator asymptotically.

The finite fourth moment condition is required for the $\sqrt{n}$ consistency of the estimators of $\mathbf{\Sigma}_{1|2}$ and $\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}}$. If we further assume normality, then we can obtain the explicit expression of the asymptotic covariance matrix for the EPPLS estimators vec($\widehat{\boldsymbol{\beta}}_1$) and vec($\widehat{\boldsymbol{\beta}}_2$), as shown in Proposition 2.

Proposition 2. Assume that the conditions in Proposition 1 hold, and we further assume that $(\boldsymbol{\epsilon}^T, \mathbf{e}^T)^T$ is normally distributed. Then,

$$\sqrt{n}\{\text{vec}(\widehat{\boldsymbol{\beta}}_1) - \text{vec}(\boldsymbol{\beta}_1)\} \xrightarrow{d} N(0, \mathbf{V}_1), \qquad \sqrt{n}\{\text{vec}(\widehat{\boldsymbol{\beta}}_2) - \text{vec}(\boldsymbol{\beta}_2)\} \xrightarrow{d} N(0, \mathbf{V}_2),$$

where

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

Note that the expression of $\mathbf{V}_1$ is the same as the asymptotic covariance matrix of $\boldsymbol{\beta}$ under the EPLS model (10) with $\mathbf{r}_{\mathbf{Y}|2}$ being the response and $\mathbf{r}_{1|2}$ being the predictor (see Proposition 9 in Cook et al. (2013)[11]), except that according to Proposition 9, we should have $\text{cov}(\mathbf{r}_{\mathbf{Y}|2}|\mathbf{r}_{1|2})$ instead of $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$ in $\mathbf{V}_1$. However, Lemma 1 below asserts that they are actually equal. The asymptotic variance $\mathbf{V}_1$ again echoes the connection between the EPPLS model (9) and the EPLS model (10).

Lemma 1. Under the EPPLS model (9), $\text{cov}(\mathbf{Y}|\mathbf{X}) = \text{cov}(\mathbf{r}_{\mathbf{Y}|2}|\mathbf{r}_{1|2})$.

3.4 | Order determination

To implement the EPPLS model, we first need to select d, the dimension of $\mathcal{E}_{\boldsymbol{\Sigma}_{1|2}}(\boldsymbol{\beta}_1)$. While many methods such as cross validation, likelihood ratio testing can be used for the selection of d, we find that BIC has the best performance especially when the sample size is moderate to large. The BIC is constructed as $\text{BIC}(d) = -2l^*(d) + \log(n)N(d)$, where $l^*(d)$ is the maximized log likelihood 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 of the EPPLS model with the dimension of $\mathcal{E}_{\boldsymbol{\Sigma}_{1|2}}(\boldsymbol{\beta}_1)$ being d. We compute BIC for all possible d and choose the one that minimizes BIC. The consistency of BIC is given in the following proposition.

Proposition 3. Assume that the EPPLS model (9) holds and that $(\boldsymbol{\epsilon}^T, \mathbf{e}^T)^T$ is normally distributed. Let $\hat{d}$ be the dimension selected by BIC. Then $P(\hat{d} = d) \to 1$ as n tends to infinity.

Proposition 3 indicates that when the sample size increases, BIC chooses the correct model with probability tending to 1. Normality is assumed here since BIC is a likelihood-based method, thus l^* is inaccurate when the data distribution widely differs from normal. However, numerical analysis (not shown here) indicates that BIC still performs well under a moderate departure from normality.

4 | PARTIAL SIMPLS ALGORITHM

Based on the connection between SIMPLS and the EPLS model, we develop a moment-based iterative algorithm, called the partial SIMPLS (PPLS) algorithm, for estimating the EPPLS subspace $\mathcal{E}_{\boldsymbol{\Sigma}_{1|2}}(\boldsymbol{\beta}_1)$. Its result can be a standalone estimator or a starting value for the optimization in (11).

Since the envelope $\mathcal{E}_{\boldsymbol{\Sigma}_{1|2}}(\boldsymbol{\beta}_1)$ in EPPLS (9) is the same as the predictor envelope $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{r}}(\boldsymbol{\beta}_1)$ in (10), PPLS estimates a basis of $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{r}}(\boldsymbol{\beta}_1)$ using the same algorithm (6) except by replacing $\mathbf{X}$ by $\mathbf{r}_{1|2}$ and $\mathbf{Y}$ by $\mathbf{r}_{\mathbf{Y}|2}$. Note that $\boldsymbol{\Sigma}_\mathbf{X}$ and $\boldsymbol{\Sigma}_{\mathbf{XY}}$ in (6) are $\boldsymbol{\Sigma}_{1|2}$ and $\boldsymbol{\Sigma}_{(1,\mathbf{Y})|2}$ in the context of (10), where $\boldsymbol{\Sigma}_{(1,\mathbf{Y})|2}$ is the covariance matrix between $\mathbf{r}_{1|2}$ and $\mathbf{r}_{\mathbf{Y}|2}$. The sample estimator of $\boldsymbol{\Sigma}_{1|2}$ and $\boldsymbol{\Sigma}_{(1,\mathbf{Y})|2}$ are $\mathbf{S}_{1|2}$ and $\mathbf{S}_{(1,\mathbf{Y})|2}$. Given the sample, PPLS estimates each column of the basis of $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{r}}(\boldsymbol{\beta}_1)$ sequentially. Set $\mathbf{W}_0 = 0$. At the $k+1$th step, let $\mathbf{W}_k = (\mathbf{w}_1, \ldots, \mathbf{w}_k) \in \mathbb{R}^{p_1 \times k}$, then $\mathbf{w}_{k+1}$ is obtained by

$$\mathbf{w}_{k+1} = \arg\max_{\mathbf{w}} \mathbf{w}^T\mathbf{S}_{(1,\mathbf{Y})|2}\mathbf{S}_{(1,\mathbf{Y})|2}^T\mathbf{w} \quad \text{subject to} \quad \mathbf{w}^T\mathbf{S}_{1|2}\mathbf{W}_k = 0 \text{ and } \mathbf{w}^T\mathbf{w} = 1. \tag{14}$$

The algorithm (14) is terminated when $k = d$. Based on Cook et al. (2013)[11], $\text{span}(\mathbf{W}_d)$ estimates $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{r}}(\boldsymbol{\beta}_1)$, and thus is an estimator for $\mathcal{E}_{\boldsymbol{\Sigma}_{1|2}}(\boldsymbol{\beta}_1)$. Once we obtain $\mathbf{W}_d$, the PPLS estimator of $\boldsymbol{\beta}_1$ is obtained by $\widehat{\boldsymbol{\beta}}_{1,\text{PPLS}} = \mathbf{P}_{\mathbf{W}_d(\mathbf{S}_{1|2})}\widehat{\boldsymbol{\beta}}_{1,\text{OLS}}$, which has the same form as the EPPLS estimator of $\boldsymbol{\beta}_1$ in (13). The estimators for other parameters including $\boldsymbol{\beta}_2$, $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$, $\boldsymbol{\Sigma}_{1|2}$ have the same form as the EPPLS estimators in (12) and (13) except that the estimators of the bases $\widehat{\boldsymbol{\Gamma}}$ and $\widehat{\boldsymbol{\Gamma}}_0$ are replaced by the PPLS estimator $\mathbf{W}_d$ and $\mathbf{W}_{d,0}$, where $\mathbf{W}_{d,0}$ is any orthonormal basis of $\text{span}(\mathbf{W}_d)^\perp$. The dimension d can be chosen by cross validation, which is a common practice for SIMPLS.

Remark 1. It is feasible to derive the asymptotic distribution for PPLS estimator as in Propositions 1 and 2, but the form of the asymptotic variance is too complicated to be useful in practice. Hence we suggest using the bootstrap approach to estimate the

TABLE 1 Computing time for methods used in simulation studies.

(n, r)	EPPLS	EPLS	PPLS	PLS	PCR	PRINCALS	CA	OLS
(100, 1)	6.58 secs	2.18 secs	1.24 mins	2.26 mins	0.07 secs	1.21 secs	2.70 secs	0.05 secs
(100, 10)	8.69 secs	1.98 secs	2.23 mins	4.17 mins	0.06 secs	0.97 secs	2.31 secs	0.18 secs
(100, 30)	8.14 secs	3.11 secs	3.05 mins	5.78 mins	0.06 secs	0.97 secs	1.86 secs	0.07 secs
(300, 1)	4.48 secs	1.16 secs	1.94 mins	3.83 mins	0.05 secs	2.04 secs	3.60 secs	0.06 secs
(300, 10)	6.37 secs	1.55 secs	4.22 mins	7.34 mins	0.06 secs	2.06 secs	3.76 secs	0.07 secs
(300, 30)	7.84 secs	4.08 secs	6.90 mins	12.61 mins	0.07 secs	2.07 secs	3.62 secs	0.10 secs
(1000, 1)	5.79 secs	0.77 secs	4.94 mins	9.76 mins	0.06 secs	4.50 secs	7.95 secs	0.16 secs
(1000, 10)	6.12 secs	1.31 secs	10.78 mins	19.95 mins	0.08 secs	4.55 secs	8.29 secs	0.24 secs
(1000, 30)	9.47 secs	2.46 secs	23.59 mins	41.55 mins	0.09 secs	4.63 secs	8.20 secs	0.56 secs

variability of the PPLS estimator. Note that for the envelop-based method EPPLS, we have the explicit form of the asymptotic variance, which is an advantage of EPPLS from an inferential statistical perspective.

5 | SIMULATION STUDY

In this section, we compared EPPLS and PPLS with existing methods including OLS, PCR, categorical principal component analysis (PRINCALS[31]), correspondence analysis (CA[32]), PLS, and EPLS. PCR regards categorical variables as continuous variables while PRINCALS and CA use the mixture of continuous and categorical variables to fit the multivariate linear regression model. Specifically, PRINCALS considers continuous transformation of categorical variables through monotone spline function with degree 2 and CA uses multiple correspondence analysis for categorical variables. The envelope dimension d was chosen by BIC for the envelope methods such as EPPLS and EPLS, and by cross-validation for PPLS and PLS. For PCR, PRINCALS, and CA, the number of principal components (PC) is chosen such that the selected PCs explain at least 90% of the total variation of all predictors.

We first investigated a low-dimensional case where OLS is used as a benchmark. The data were generated from model (9), with $p_1 = 6$, $p_2 = 3$, $d = 1$, $\boldsymbol{\mu}_{\mathbf{Y}} = 0$ and $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}} = 10\mathbf{I}_r$. The dimension of $\mathbf{Y}$ was varied from $r = 1$, 10 and 30. The matrix $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0)$ was obtained by normalizing a $p_1 \times p_1$ matrix of independent normal $(0, 2^2)$ variates, $\boldsymbol{\eta}$ was a $d \times r$ matrix with each element being independent normal $(1, 10^2)$ variates, and $\boldsymbol{\beta}_2 = (1.5\mathbf{1}_r, 1.2\mathbf{1}_r, 2\mathbf{1}_r)^T$, where $\mathbf{1}_r \in \mathbb{R}^r$ denotes a vector of 1. Let $\mathbf{A} \in \mathbb{R}^{d\times d}$ be an independent normal $(8, 2^2)$ variates and $\mathbf{B} \in \mathbb{R}^{(p_1-d)\times(p_1-d)}$ be a matrix of independent normal $(0.6, 1^2)$ variates, $\boldsymbol{\Omega} = \mathbf{A}\mathbf{A}^T$ and $\boldsymbol{\Omega}_0 = \mathbf{B}\mathbf{B}^T$. We have $\|\boldsymbol{\Omega}\| = 31.20$ and $\|\boldsymbol{\Omega}_0\| = 15.15$, where $\|\cdot\|$ denotes the spectral norm. To generate the continuous predictors $\mathbf{X}_1$, we let $\boldsymbol{\mu}_1 = \mathbf{1}_{p_1}$ and $\boldsymbol{\gamma} = (-1.4\mathbf{1}_{p_1}, 0.8\mathbf{1}_{p_1}, 2.5\mathbf{1}_{p_1})^T$. We let $\mathbf{e}$ follow a multivariate normal distribution with a zero mean vector and variance matrix $\boldsymbol{\Sigma}_{1|2} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$. The errors $\boldsymbol{\epsilon}$ was generated from a multivariate normal distribution with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$. The categorical predictors were $\mathbf{X}_2 = 10(W_{21}, W_{22}, W_{23})$, where W_{21}, W_{22} and W_{23} were independent Bernoulli variates that take value 1 with probability 0.4, 0.5 and 0.8, respectively.

We considered the sample size from 50 to 1000. For each sample size, 100 replications were simulated. First we investigated the computing time of each method. The computing time was calculated by the average of 10 replications, and it included the selection of the number of components. The results were displayed in Table 1. PCR and OLS are the fastest methods to compute, followed by EPLS, PRINCALS, CA and EPPLS. PLS and PPLS are methods that take the longest to compute. The computing time was measured with 2.3GHx Quad-core intel core i7 processor and 32GB memory.

For each replication, we estimated $\boldsymbol{\beta}_1$ using methods EPPLS, EPLS, PPLS, PLS, PCR, PRINCALS, CA, and OLS, and calculated $\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\|_F$, where $\|\cdot\|_F$ denotes the Frobenius norm. For EPLS, PLS, PCR, PRINCALS, CA, and OLS, we fitted the response $\mathbf{Y}$ on all predictors $\mathbf{X} = (\mathbf{X}_1^T, \mathbf{X}_2^T)^T$, and obtained $\widehat{\boldsymbol{\beta}}_1$ by extracting the submatrix of $\widehat{\boldsymbol{\beta}}$ that corresponds to $\mathbf{X}_1$. The average and standard deviation of $\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\|_F$ based on the 100 replications are summarized in Table 2. Note that $\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\|_F$ is the square root of the mean square error (MSE) of $\widehat{\boldsymbol{\beta}}_1$. Among all the methods, EPPLS has the smallest MSE, followed by PPLS. Note that PPLS and EPPLS estimate the same parameter in population, but they use different sampling algorithms. EPPLS is likelihood-based and is usually more efficient than PPLS. EPLS performs much better than OLS. However, it loses substantial efficiency compared to EPPLS. For each (n, r) pair, the $\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\|_F$ from EPLS is at least three times as large as EPPLS. This is because EPPLS treats categorical and continuous predictors differently in estimation, and is, therefore, more

TABLE 2 Results of average (standard deviation / (number of replications)$^{1/2}$) of $\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\|_F$ based on 100 replications.

r	Methods	$n = 50$	$n = 100$	$n = 300$	$n = 1000$
1	EPPLS	0.49(0.029)	0.58(0.054)	0.33(0.012)	0.07(0.005)
	EPLS	3.87(0.844)	3.27(0.539)	2.06(0.275)	0.23(0.005)
	PPLS	0.50(0.023)	1.13(0.065)	1.02(0.052)	0.07(0.003)
	PLS	2.64(0.001)	10.60(0.003)	15.64(0.002)	1.07(0.0001)
	PCR	2.60(0.015)	$10.61(3.3 * 10^{-4})$	$15.65(3.2 * 10^{-4})$	$1.07(1.7 * 10^{-4})$
	PRINCALS	1.37(0.088)	$10.62(8.8 * 10^{-4})$	$15.65(2.6 * 10^{-4})$	$1.08(2.2 * 10^{-4})$
	CA	2.67(0.008)	$10.61(3.3 * 10^{-4})$	$15.65(2.9 * 10^{-4})$	$1.08(1.9 * 10^{-4})$
	OLS	12.91(0.812)	8.06(0.500)	4.37(0.284)	2.29(0.140)
10	EPPLS	1.27(0.031)	1.06(0.029)	0.63(0.018)	0.26(0.006)
	EPLS	10.20(1.692)	8.37(1.259)	7.98(0.767)	1.23(0.121)
	PPLS	3.43(0.200)	4.31(0.241)	3.36(0.175)	0.73(0.036)
	PLS	21.36(0.011)	40.62(0.002)	51.76(0.008)	20.93(0.002)
	PCR	20.91(0.060)	40.66(0.002)	51.81(0.001)	20.96(0.0002)
	PRINCALS	10.57(0.310)	40.68(0.002)	51.82(0.001)	$20.96(2.2 * 10^{-4})$
	CA	21.39(0.012)	40.66(0.001)	$51.81(8.1 * 10^{-4})$	$20.96(2.1 * 10^{-4})$
	OLS	44.40(1.013)	29.19(0.692)	16.27(0.348)	8.58(0.206)
30	EPPLS	2.58(0.072)	1.49(0.034)	0.89(0.021)	0.45(0.011)
	EPLS	40.88(3.521)	25.95(2.567)	6.16(0.999)	3.88(0.517)
	PPLS	8.62(0.956)	6.36(0.366)	5.07(0.316)	1.76(0.108)
	PLS	47.49(0.024)	60.59(0.017)	74.61(0.011)	54.68(0.005)
	PCR	46.47(0.136)	60.66(0.002)	74.69(0.001)	54.74(0.001)
	PRINCALS	23.06(0.697)	60.69(0.004)	74.70(0.002)	$54.74(6.5 * 10^{-4})$
	CA	47.57(0.026)	60.66(0.002)	74.69(0.001)	$54.74(5.8 * 10^{-4})$
	OLS	77.08(1.150)	50.84(0.744)	28.85(0.360)	15.95(0.168)

efficient. Most of the time, PCR, PRINCALS, CA, and PLS perform worse than OLS. This is because these methods seek for the linear combinations of $\mathbf{X}$ that provide either the largest variance or the largest covariance with $\mathbf{Y}$. These directions are not necessarily the ones that provide information to the estimation $\boldsymbol{\beta}$. So the estimators from these methods may have large bias and underperform OLS (see Figure 1). Note that although EPLS and PLS are estimating the same parameter, EPLS is more stable than PLS, since it is a model-based method and is proved to be $\sqrt{n}$ consistent[11].

Figure 1 takes on a close look at the bias and variance of a randomly chosen element of $\boldsymbol{\beta}_1$. From the left panel, we noticed that the PLS estimator indeed carries a large bias, as indicated in Schuberth et al. (2018)[1], when it treats the discrete predictors as continuous. The estimator of PCR, PRINCALS, and CA also bear a large bias since it does not take the information of $\mathbf{Y}$ into account in the construction of the principal components. OLS and EPLS are consistent methods and do not have a large bias. But their estimators are more variant than the EPPLS and PPLS estimators as shown in the right panel of Figure 1. PPLS has about the same bias as EPPLS and a slightly larger variance compared to EPPLS, but the difference is dwarfed by the magnitude of the variance of EPLS or OLS.

Moreover, we investigated the performance of hypothesis testing for the coefficients $\boldsymbol{\beta}_1$ based on the asymptotic distribution established in Proposition 2. To perform the hypothesis testing, we followed the simulation setting that generated Table 2 except that we set the first three rows of $\boldsymbol{\Gamma}$ to zero. It implies that the first three rows of $\boldsymbol{\beta}_1$ are zero vectors and the remaining elements of $\boldsymbol{\beta}_1$ are nonzero. Then we test the hypothesis if each element in $\boldsymbol{\beta}_1$ is zero. Specifically, let $\beta_{1,ij}$ denote the (i, j)th element in $\boldsymbol{\beta}_1$, and we test the hypothesis $H_0 : \beta_{1,ij} = 0$. The standard error of estimators of $\hat{\beta}_{1,ij}$ and the p-value of each test were calculated using the asymptotic distribution in Proposition 2. The simulation was replicated 100 times. We reported the 5th, 50th, and 95th percentiles of the average p-values in Table 3. The p-values for the zero elements in $\boldsymbol{\beta}_1$ and non-zero elements in $\boldsymbol{\beta}_1$ were reported separately. The results show that with the asymptotic distribution in Proposition 2, the hypothesis testing procedure is able to detect the nonzero elements in $\boldsymbol{\beta}_1$ with high power. For the zero elements in $\boldsymbol{\beta}_1$, the testing procedure can also control the Type I error under the desired level.

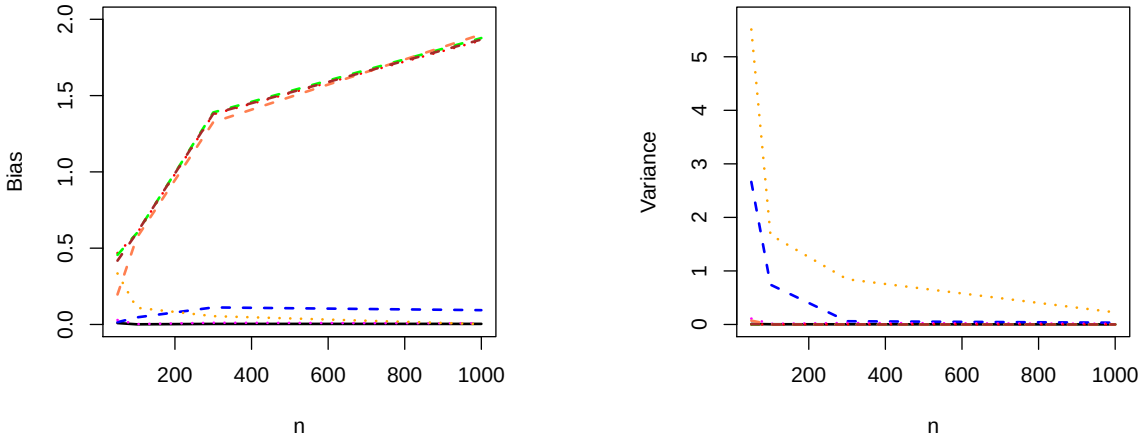

FIGURE 1 Bias (left panel) and variance (right panel) for a random picked element of $\boldsymbol{\beta}_1$ when $r = 30$. The black solid line marks for EPPLS, the blue dashed line marks for EPLS, the magenta dotted line marks for PPLS, the red dotted line marks for PLS, the green dashed line marks for PCR, the coral dashed line marks for PRINCALS, the brown dashed line marks for CA, and the orange dotted line marks for OLS.

TABLE 3 Results of the 5th, 50th, and 95th percentiles of the average p-values based on 100 replications.

		The 5th, 50th, and 95th percentiles of the p-values	
r	n	zero element of $\boldsymbol{\beta}_1$	nonzero element of $\boldsymbol{\beta}_1$
1	50	0.336, 0.907, 0.992	$4.5 * 10^{-8}$, $1.4 * 10^{-3}$, 0.469
	100	0.394, 0.929, 0.994	0, 0, 0.002
	300	0.366, 0.923, 0.993	0, 0, $1.3 * 10^{-14}$
	1000	0.731, 0.930, 0.993	0, 0, 0.003
10	50	0.191, 0.889, 0.990	0, $1.2 * 10^{-9}$, 0.098
	100	0.275, 0.889, 0.989	0, 0, $6.3 * 10^{-8}$
	300	0.279, 0.880, 0.993	0, 0, 0
	1000	0.344, 0.916, 0.990	0, 0, $6.5 * 10^{-12}$
30	50	0.065, 0.840, 0.988	0, 0, $2.1 * 10^{-4}$
	100	0.085, 0.854, 0.989	0, 0, $2.1 * 10^{-13}$
	300	0.183, 0.861, 0.995	0, 0, $4.7 * 10^{-8}$
	1000	0.305, 0.877, 0.993	0, 0, 0

For sensitivity analysis, we considered a situation where the immaterial part of $\mathbf{X}_1$ has a larger variation than the material part ($\|\boldsymbol{\Omega}\| < \|\boldsymbol{\Omega}_0\|$). The results of both scenarios (i.e., $\|\boldsymbol{\Omega}\| > \|\boldsymbol{\Omega}_0\|$ and $\|\boldsymbol{\Omega}\| < \|\boldsymbol{\Omega}_0\|$) presented in Table 2 and Web Table 1 show that EPPLS yields the most efficiency gains, and its performance is quite stable. In addition, we considered the case where $\mathbf{X}_1$ and $\mathbf{X}_2$ do not have a linear relationship. The results are in Web Table 2. The performance of EPPLS is very stable and is still the best among all models under comparison. However, the performance of PPLS deteriorates a lot, that PPLS even underperforms OLS most of the time. The details of the sensitivity analyses are provided in Section C of Supplemental Materials.

We also investigated a high-dimensional setting where $p_1 > n$. The data were generated in the same way as that produced Table 2, with n fixed at 100, p_2 fixed at 10 and $p_1 = 150, 300$ and 600. The ten binary predictors were drawn from Bernoulli distributions that take value 1 with probabilities $0.2, 0.25, 0.3, 0.35, 0.4, 0.5, 0.5, 0.55, 0.75$, and 0.8. The coefficient matrix $\boldsymbol{\beta}_2$ had structure $\boldsymbol{\beta}_2 = \mathbf{b}\mathbf{1}_r^T$, where each element in $\mathbf{b} \in \mathbb{R}^{p_2}$ was independent normal $(0, 0.5^2)$ random variates. The coefficient matrix $\boldsymbol{\gamma}$ was $\boldsymbol{\gamma} = (1.2\mathbf{1}_{p_1}, 0.8\mathbf{1}_{p_1}, 0.5\mathbf{1}_{p_1}, 2\mathbf{1}_{p_1}, -0.5\mathbf{1}_{p_1}, -1.2\mathbf{1}_{p_1}, -0.8\mathbf{1}_{p_1}, 1.8\mathbf{1}_{p_1}, 2.5\mathbf{1}_{p_1}, 1.5\mathbf{1}_{p_1})^T$. In high-dimensional settings,

prediction is a more common criterion than MSE for comparison of methods, we then computed the prediction errors, which is the square root of mean squared residuals, for methods EPPLS, PPLS, EPLS, PLS, PCR, PRINCALS, and CA using the five-fold cross validation, with 100 replications for each sample size. Note that the OLS is not applicable when $n < p$.

The results are provided in Table 4. PLS is known for its stable performance in high dimensional settings, and it performs better than PCR, PRINCALS, and CA as shown in Table 4. By conditioning on the categorical variables, PPLS further reduces the prediction errors compared to PLS. The mechanism of the envelope methods EPLS and EPPLS is to remove the variation from the immaterial part, and they have the lowest prediction errors. Between the two envelope methods, EPPLS treats $\mathbf{X}_1$ and $\mathbf{X}_2$ separately by conditioning $\mathbf{Y}$ and $\mathbf{X}_1$ on $\mathbf{X}_2$ and has the best performance in all cases in Table 4.

TABLE 4 Results of average (standard deviation / (number of replications)$^{1/2}$) of the prediction errors based on 100 replications for high dimensional setting.

r	Methods	$p_1 = 150$		$p_1 = 300$		$p_1 = 600$	
1	EPPLS	5.93	(0.097)	21.89	(0.300)	28.68	(0.606)
	EPLS	9.05	(0.127)	33.61	(0.432)	48.51	(1.073)
	PPLS	9.09	(0.104)	41.09	(0.442)	60.98	(1.278)
	PLS	13.17	(0.107)	60.43	(0.409)	96.24	(1.383)
	PCR	15.28	(0.154)	68.23	(0.640)	104.76	(1.758)
	PRINCALS	13.83	(0.116)	63.92	(0.476)	102.48	(0.819)
	CA	13.90	(0.123)	64.17	(0.484)	102.49	(0.810)
10	EPPLS	67.49	(1.060)	76.41	(1.814)	79.24	(1.694)
	EPLS	69.40	(0.601)	96.78	(0.697)	125.36	(1.947)
	PPLS	99.20	(1.070)	143.23	(1.462)	175.75	(3.214)
	PLS	146.04	(1.186)	211.45	(1.436)	279.92	(4.209)
	PCR	166.11	(1.617)	239.42	(2.239)	302.39	(5.555)
	PRINCALS	152.96	(1.358)	223.65	(1.680)	294.26	(2.350)
	CA	153.83	(1.252)	224.54	(1.668)	294.30	(2.317)
30	EPPLS	93.04	(1.022)	106.61	(1.064)	112.93	(2.394)
	EPLS	129.00	(1.032)	157.74	(1.047)	184.22	(2.509)
	PPLS	195.76	(2.099)	238.98	(2.400)	256.36	(5.496)
	PLS	288.34	(2.373)	353.39	(2.423)	414.83	(5.923)
	PCR	327.77	(3.207)	399.96	(3.730)	444.60	(6.878)
	PRINCALS	301.71	(2.694)	373.43	(2.809)	436.59	(3.480)
	CA	303.35	(2.495)	375.06	(2.799)	436.66	(3.420)

6 | DATA APPLICATION

COVID-19 is a global pandemic that has affected 223 countries, areas, or territories. Study shows that cytokines are associated with COVID-19 severity and survival[33,3,34], and 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. For this purpose, we analyzed the data from a study investigated in Laing et al. (2020)[3], which included 63 COVID-19 patients. In addition, the data also contained 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 the 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. Data and detailed protocols for this study are publicly available on the COVID-IP project website (www.immunophenotype.org).

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 including temperature, blood glucose, National Early Warning Score 2 (NEWS2) score, serum lactate, the 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 demographic information and indicators for underlying disease status. Demographics information contained age, BMI, ethnicity, and sex. Age was a binary variable taking value 1 for patients 45 years and older, and 0 otherwise. BMI was measured in ordinal scale based on categories of below 20, 20–24, 25–29, 30–34, and 35 and above. The ethnicity variable included three categories asian, black, and caucasian. We created two binary indicators, one for asian and one for black. The sex indicator took value 1 for males and 2 for females. For underlying diseases, hypertension, ischaemic heart disease, non-asthma chronic lung disease, asthma, diabetes, and active malignancy were considered. This gave a total of 11 categorical variables. All variables were standardized.

We fitted the data with EPPLS, PPLS, EPLS, PLS, PCR and OLS, and computed the prediction errors as the root mean square error. The prediction error was obtained by five-fold cross-validations with 50 random splits of the data. OLS had the largest prediction error of 38.74, followed by PCR, which had a prediction error of 6.041. PLS and EPLS had similar prediction errors: 5.120 for PLS and 5.247 for EPLS. PPLS and EPPLS had the lowest prediction errors: 2.194 for PPLS and 2.192 for EPPLS. The efficiency gains obtained from EPPLS and PPLS also led to better prediction performance.

The estimation efficiency also led to a clear scientific interpretation of the results. Based on the regression coefficient estimators, we investigated the associations between cytokines and covariates. Figure 2 shows the heatmaps of $\widehat{\boldsymbol{\beta}}_1$ from all six methods, and Web Figure 1 shows the clustering structure of the responses ($\mathbf{Y}$) and continuous variables ($\mathbf{X}_1$). Recall that $\widehat{\boldsymbol{\beta}}_1$ presents the associations of the cytokines with patient status at admission. It was noteworthy to observe that under EPPLS, interleukin 10 (IL10) stands out to be the most important cytokine, highlighted by a clear strong association across multiple patient statuses at admission, including severity (admission_WHO_ordinal_scale), blood pressure (admission_BP_diastolic, admission_BP_systolic), serum lactate (admission_lactate_venous), and oxygen saturation (admission_os_sats). Under the normality assumption, Proposition 2 was applied to perform the hypothesis test of $\boldsymbol{\beta}_1 = \mathbf{0}$. The regression coefficients for the association of IL10 (IL10_Th_cyto_cyto) across admission_WHO_ordnial_scale, admission_BP_diastolic, admission_BP_systolic, admission_lactate_venous, and admission_os_sats are statistically significant with p-value 2.02×10^{-8}, 2.07×10^{-8}, 2.54×10^{-8}, 3.25×10^{-8}, and 7.82×10^{-8}, respectively. This is consistent with the report that IL10 is associated with COVID-19 severity and mortality, cytokine storm, and intensive care unit (ICU) stay in COVID-19 patients[33]. The importance of IL10 was not as evident in competing approaches based on the absolute values of $\widehat{\boldsymbol{\beta}}_1$. The OLS and PCR estimators were very variable, and hard to extract much information from the coefficients. EPLS and PLS both showed a few influential cytokines including IL10, but it was not obvious that IL10 was the most important one as in EPPLS. Although PPLS and EPPLS have the same estimation goal in population, their sample performance can vary. In this example, PPLS also noticed the strong association between IL10 and patient admission status, but the leading role of IL10 was not as obvious as in EPPLS.

In addition to IL10, interleukin 6 (IL6) and CXCL10 (IP10) are determined to be co-leading cytokines. Interestingly, Laing et al. (2020)[3] reported that the status of COVID-19 patients is characterized by a severity-related triad of IL10, IL6, and IP10. The triad/block of IL10, IL6, IP10 was most obvious under PLS but also shown in EPPLS from the clustering structure of $\mathbf{Y}$ in Web Figure 1. However, the triad/block was missed by EPLS, PPLS, PCR, and OLS. We also noted that under EPPLS, interferon-γ or type II interferon (IFNg) had coefficient estimates similar to IP10, which is consistent with the observation that IFNg levels are correlated with IP10[3]. This similarity was not present in EPLS, PPLS, PLS, PCR, or OLS.

Figure 3 shows the heatmaps for the estimators of $\boldsymbol{\beta}_2$, which present the associations of cytokines with demographics and underlying diseases. Firstly, we noticed the strong association of IP10 with sex. It has been reported that men have a higher risk of infection, mortality, and comorbidities from COVID-19 compared to women[35]. Thus it is important to investigate the sex difference in COVID-19. Recently, Takahashi et al. (2020)[34] reported the association of IP10 with the sex difference in immune responses that underlie COVID-19 disease outcomes. This association was also captured by all models, although it appeared weaker under OLS. Secondly, we observed a clear association of interferon-γ (INFg) with both the asian and black populations and a strong association of type III interferon (IFNl2.3) with the black population under EPPLS. This association was not observed under EPLS and PLS, and the association between IFNl2.3 and the black population was weak under OLS. Significant racial/ethnic disparities have been reported for COVID-19, with the disproportionate burden on African and Latino population[36]. Hence, cytokine markers IP10, INFg, and IFNl2.3 can potentially be important for understanding the biological mechanisms associated with sex bias and racial/ethnic disparities in COVID-19. Thirdly, we observed the association of interleukin 2 (IL2) with multiple pre-existing disease statuses, including ischemic heart disease (IHD), asthma, and hypertension (HTN), which was not clear under EPLS and PLS. Association of IL2 with asthma was previously reported[37]. Therefore IL2 can potentially

be considered as a marker for pre-existing disease status. Finally, we observed the strong association of interferon-λ1 (IFNl1) with age under EPPLS, EPLS, PPLS, and OLS, but not under PLS or PCR. Recently, Dinnon et al. (2020)[38] developed a mouse model for COVID-19, which can be used to study age-related disease pathogenesis of COVID-19. IFNl1 is a potential clinical target for the treatment of human COVID-19 using this mouse model[38]. Heatmaps with uniform color scale are in Web Figures 3 and 4 in the Supplementary Materials.

7 | DISCUSSION

We have proposed an EPPLS model which achieves estimation efficiency when both continuous and categorical predictors are present. EPPLS is proposed when the categorical predictors $\mathbf{X}_2$ are assumed to be fixed in the model formulation, but it is also applicable to cases when $\mathbf{X}_2$ is random and follows a certain distribution. If all predictors are continuous, the idea of EPPLS can be applied when part of the predictors is of main interest. The proposed model can potentially be applied to generalized linear regression where the response variable is categorical. EPPLS can also incorporate heteroscedastic structure[39], spatial correlation[15] or time dependence[40]. A Bayesian approach can also be derived for these models which allow users to incorporate prior information for estimation.

Theoretical properties in Section 3.3, such as consistency and asymptotic normality, have been established when the number of predictors is smaller than the sample size. In high-dimensional settings where $p > n$, theoretical properties are not valid without further assumptions such as sparsity, low-rank structure, or other parametric structures. Numerically EPPLS shows a better prediction performance compared to other methods in our simulation settings. Development of variants of EPPLS that better adapts to the high-dimensional settings is an interesting topic for future study.

ACKNOWLEDGEMENTS

The authors thank the associate editor and three referees for their valuable comments. Dr. Park is partially supported by University of Wisconsin-Madison Office of the Vice Chancellor for Research and Graduate Education. Dr. Su is partially supported by a grant from Simons Foundation. Dr. Chung is partially supported by NIH/NIGMS grant R01-GM122078, NIH/NCI grant R21-CA209848, and NIH/NIDA grant U01-DA045300.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on COVID-IP project website at https://www.immunophenotype.org/index.php/data/bulk-data-downloads/. These data were derived from the following resources available in the public domain: www.immunophenotype.org.

SUPPORTING INFORMATION

All proofs, technical details, and sensitivity analysis are available with this paper. The programming code is available on the first author's personal webpage. The supplemental material with an R markdown provides a tutorial on the programming code used in the simulation study.

References

1. Schuberth F, Henseler J, Dijkstra TK. Partial least squares path modeling using ordinal categorical indicators. *Quality & Quantity* 2018; 52(1): 9–35.

2. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases* 2020; 20(5): 533–534.

3. Laing AG, Lorenc A, Del Barrio IDM, et al. A dynamic COVID-19 immune signature includes associations with poor prognosis. *Nature Medicine* 2020; 26(10): 1623–1635.

4. Genser B, Cooper PJ, Yazdanbakhsh M, Barreto ML, Rodrigues LC. A guide to modern statistical analysis of immunological data. *BMC immunology* 2007; 8(1): 1–15.

5. Martens H, Nï T. *Multivariate calibration*. John Wiley & Sons . 1992.

6. Frank LE, Friedman JH. A statistical view of some chemometrics regression tools. *Technometrics* 1993; 35(2): 109–135.

7. Wold H. Estimation of principal components and related models by iterative least squares.. *In Multivariate Analysis* 1966; 59: 391-420.

8. Lohmöller JB. *Latent variable path modeling with partial least squares*. Springer Science & Business Media . 2013.

9. Hair JF, Sarstedt M, Ringle CM, Mena JA. An assessment of the use of partial least squares structural equation modeling in marketing research. *Journal of the academy of marketing science* 2012; 40(3): 414–433.

10. Cook RD, Li B, Chiaromonte F. Envelope models for parsimonious and efficient multivariate linear regression (with discussion). *Statistica Sinica* 2010; 20: 927–1010.

11. Cook RD, Helland IS, Su Z. Envelopes and partial least squares regression. *Journal of the Royal Statistical Society: Series B* 2013; 75(5): 851–877.

12. Cook RD, Su Z. Scaled predictor envelopes and partial least-squares regression. *Technometrics* 2016; 58(2): 155–165.

13. Zhu G, Su Z. Envelope-based sparse partial least squares. *The Annals of Statistics* 2020; 48(1): 161–182.

14. Cook RD, Zhang X. Foundations for envelope models and methods. *Journal of the American Statistical Association* 2015; 110(510): 599–611.

15. Rekabdarkolaee HM, Wang Q, Naji Z, Fuentes M. New parsimonious multivariate spatial model: Spatial envelope. *Statistica Sinica* 2021: To appear.

16. Su Z, Zhu G, Chen X, Yang Y. Sparse envelope model: efficient estimation and response variable selection in multivariate linear regression. *Biometrika* 2016; 103(3): 579–593.

17. Khare K, Pal S, Su Z. A Bayesian Approach for Envelope Models. *Annals of Statistics* 2017; 45(1): 196–222.

18. Lee M, Chakraborty S, Su Z. A Bayesian approach to envelope quantile regression. *Statistica Sinica* 2021: To appear.

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

20. Ding S, Cook RD. Matrix variate regressions and envelope models. *Journal of the Royal Statistical Society: Series B* 2018; 80(2): 387-408.

21. Cook RD. *An Introduction to Envelopes: Dimension Reduction for Efficient Estimation in Multivariate Statistics*. Hoboken, NJ: John Wiley & Sons . 2018.

22. Conway J. *A course in functional analysis*. New York: Springer . 1990.

23. De Jong S. SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and intelligent laboratory systems* 1993; 18(3): 251–263.

24. Wold H. Path models with latent variables: The NIPALS approach. In: New York: Academic Press. 1975 (pp. 307–357).

25. Su Z, Cook RD. Partial envelopes for efficient estimation in multivariate linear regression. *Biometrika* 2011; 98: 133–146.

26. Cook RD, Forzani L, Su Z. A note on fast envelope estimation. *Journal of Multivariate Analysis* 2016; 150: 42–54.

27. Sun T, Zhang CH. Sparse matrix inversion with scaled lasso. *The Journal of Machine Learning Research* 2013; 14(1): 3385–3418.

28. Zhang T, Zou H. Sparse precision matrix estimation via lasso penalized D-trace loss. *Biometrika* 2014; 101(1): 103–120.

29. Khare K, Oh SY, Rajaratnam B. A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2015; 77(4): 803–825.

30. Rothman AJ. Positive definite estimators of large covariance matrices. *Biometrika* 2012; 99(3): 733–740.

31. De Leeuw J. Multivariate analysis with optimal scaling. 2011.

32. Chavent M, Kuentz-Simonet V, Labenne A, Saracco J. Multivariate analysis of mixed data: The R Package PCAmixdata. *arXiv preprint arXiv:1411.4911* 2014.

33. Lu L, Zhang H, Dauphars DJ, He YW. A Potential Role of Interleukin-10 in COVID-19 Pathogenesis. *Trends in Immunology* 2020.

34. Takahashi T, Ellingson MK, Wong P, et al. Sex differences in immune responses that underlie COVID-19 disease outcomes. *Nature* 2020; 588(7837): 315–320.

35. Chakravarty D, Nair SS, Hammouda N, et al. Sex differences in SARS-CoV-2 infection rates and the potential link to prostate cancer. *Communications Biology* 2020; 3(1): 1–12.

36. Hooper MW, Nápoles AM, Pérez-Stable EJ. COVID-19 and racial/ethnic disparities. *JAMA* 2020; 323(24): 2466–2467.

37. Boonpiyathad S, Pornsuriyasak P, Buranapraditkun S, Klaewsongkram J. Interleukin-2 levels in exhaled breath condensates, asthma severity, and asthma control in nonallergic asthma.. In: . 34. ; 2013.

38. Dinnon KH, Leist SR, Schäfer A, et al. A mouse-adapted model of SARS-CoV-2 to test COVID-19 countermeasures. *Nature* 2020; 586(7830): 560–566.

39. Park Y, Su Z, Zhu H. Groupwise envelope models for imaging genetic analysis. *Biometrics* 2017; 73(4): 1243-1253.

40. Wang L, Ding S. Vector autoregression and envelope model. *Stat* 2018; 7(1): e203.

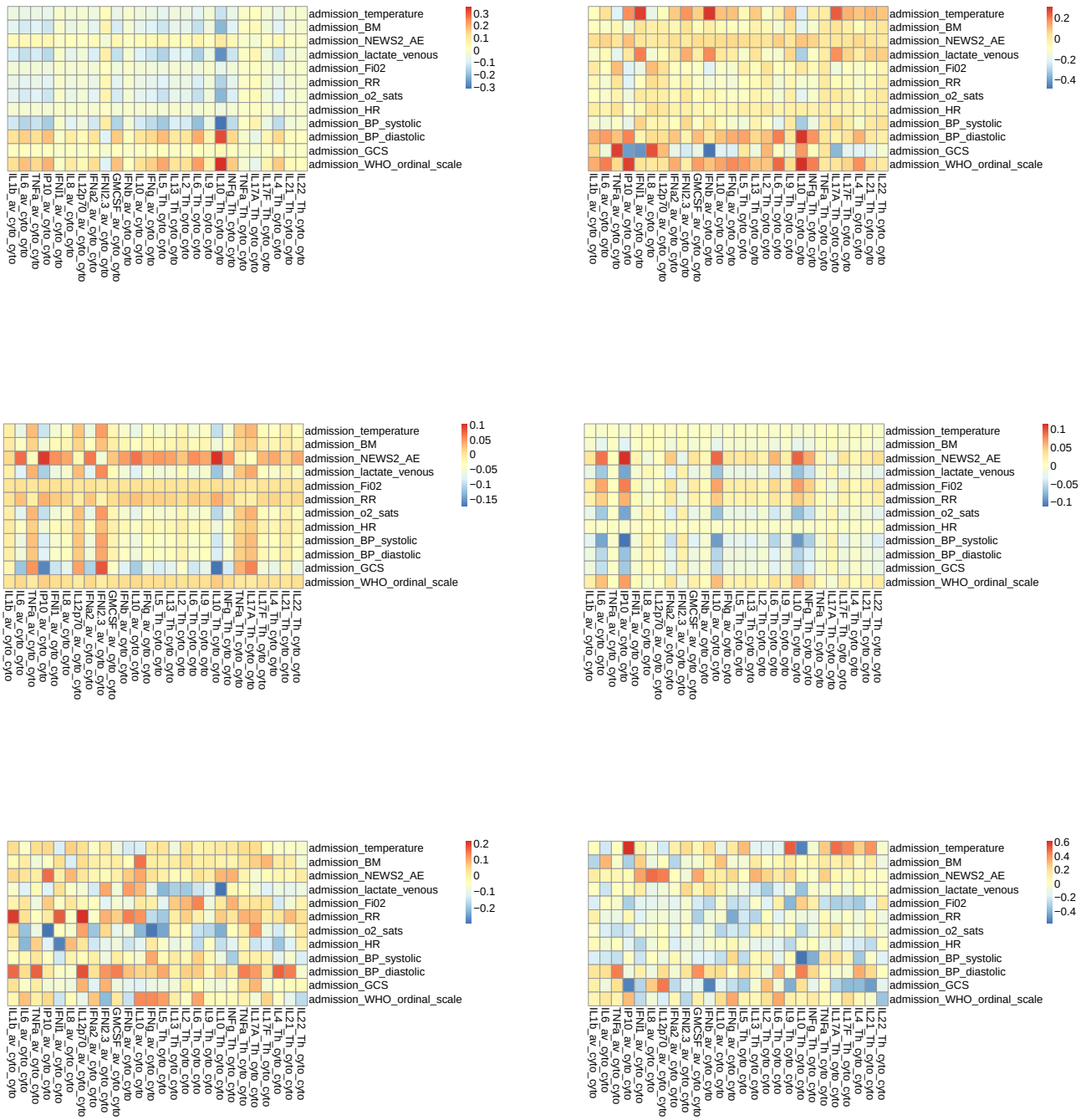

FIGURE 2 Heatmaps of the regression coefficients of $\widehat{\boldsymbol{\beta}}_1$ under EPPLS (left of 1st row), EPLS (right of 1st row), PPLS (left of 2nd row), PLS (right of 2nd row), PCR (left of 3rd row), and OLS (right of 3rd row).

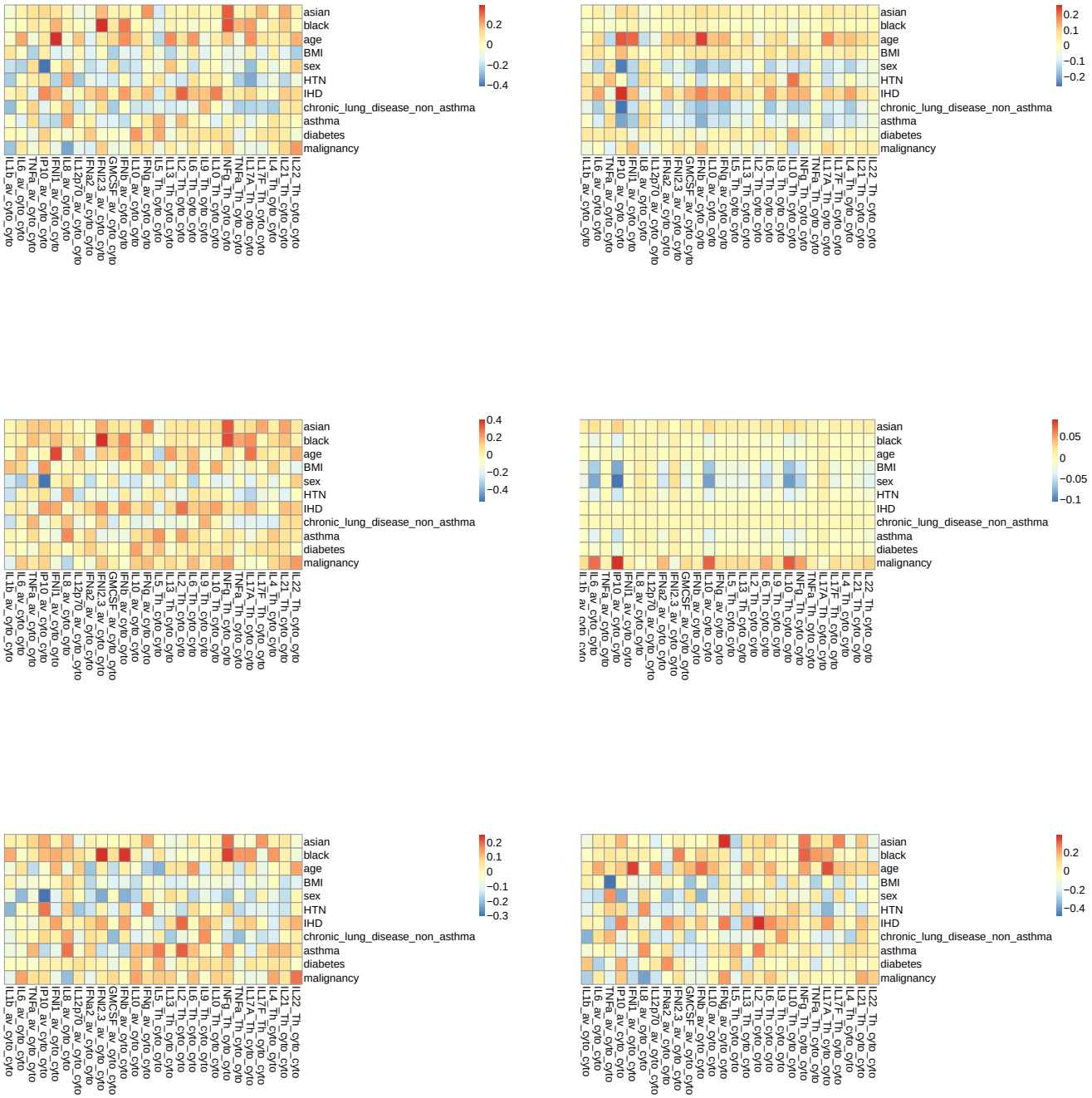

FIGURE 3 Heatmaps of the regression coefficients of $\widehat{\boldsymbol{\beta}}_2$ under EPPLS (left of 1st row), EPLS (right of 1st row), PPLS (left of 2nd row), PLS (right of 2nd row), PCR (left of 3rd row), and OLS (right of 3rd row).