Simultaneous envelopes for multivariate linear regression

R. Dennis Cook* and Xin Zhang†

December 2, 2013

Abstract

We introduce envelopes for simultaneously reducing the predictors and the responses in multivariate linear regression, so the regression then depends only on estimated linear combinations of $\mathbf{X}$ and $\mathbf{Y}$. We use a likelihood-based objective function for estimating envelopes and then propose algorithms for estimation of a simultaneous envelope as well as for basic Grassmann manifold optimization. The asymptotic properties of the resulting estimator are studied under normality and extended to general distributions. We also investigate likelihood ratio tests and information criteria for determining the simultaneous envelope dimensions. Simulation studies and real data examples show substantial gain over the classical methods like, partial least squares, canonical correlation analysis and reduced-rank regression. This article has supplementary material available online.

Key Words: Canonical correlations; envelope model; Grassmann manifold; partial least squares; principal component analysis; reduced-rank regression; sufficient dimension reduction.

1 Introduction

Multivariate linear regression with predictors $\mathbf{X} \in \mathbb{R}^p$ and responses $\mathbf{Y} \in \mathbb{R}^r$ is a cornerstone of multivariate statistics. When p and r are not small, it is widely recognized that reducing the dimensionalities of $\mathbf{X}$ and $\mathbf{Y}$ may often result in improved performance. Perhaps the most popular methods for reducing the number of predictors and responses are principal component analysis (PCA), partial least squares (PLS) regression, canonical correlation analysis (CCA) and reduced-rank regression (RRR).

*R. Dennis Cook is Professor, School of Statistics, University of Minnesota, Minneapolis, MN 55455 (E-mail: dennis@stat.umn.edu).

†Xin Zhang is Ph.D student, School of Statistics, University of Minnesota, Minneapolis, MN, 55455 (Email: zhan0648@umn.edu).

Principal component analysis (Jolliffe, 1986, 2005) is an un-supervised dimension reduction method designed to select orthogonal linear combinations of $\mathbf{X}$ and $\mathbf{Y}$ with maximal variation. However, it does not use any information about the relationship between $\mathbf{X}$ and $\mathbf{Y}$, and thus separate PCA reductions of $\mathbf{X}$ and $\mathbf{Y}$ could be ineffective for regression. Partial least squares was proposed as iterative algorithms, NIPALS (Wold, 1966) and SIMPLS (de Jong, 1993), for predictor reduction. These methods, which are used extensively in chemometrics, reduce the predictors by iteratively estimating linear combinations of them that have maximal covariance with the response vector. Canonical correlation analysis (Hotelling, 1936; Anderson, 1984) is used to investigate the overall correlations between the two sets of variables $\mathbf{X}$ and $\mathbf{Y}$. It can simultaneously reduce $\mathbf{X}$ and $\mathbf{Y}$ by finding pairs of linear combinations such that the correlations of these pairs are in descending order and components are uncorrelated across different pairs. A probabilistic interpretation of CCA was given by Bach and Jordan (2005) as a latent variable model for two normal random vectors. Reduced-rank regression (Izenman 1975; Reinsel and Velu 1998) restricts on the rank of the regression coefficient matrix and therefore improves prediction by reducing the number of parameters in the model.

Sufficient dimension reduction methods for multivariate responses are also available for reducing dimensionality. For example, sliced inverse regression (Li, 1991) was extended to multivariate response data by Li, et al. (2003). Li, Wen and Zhu (2008) proposed "projective resampling" to deal with a multivariate regression by using univariate reduction methods. However, such methods are beyond the scope of this work since they are designed to estimate only a subspace as a preliminary step in an analysis. In contrast, we work in the context of the multivariate linear model with a view toward prediction and coefficient estimation.

Informally, multivariate linear regression can involve both material and immaterial variation in the responses and in the predictors. Material variation provides information that is directly relevant to the regression, while the immaterial variation is essentially irrelevant to the regression and serves to increase estimative variation. Envelopes, which were introduced by Cook, Li and Chiaromonte (2010) for response reduction, use a subspace to envelop the material information and thereby exclude the immaterial variation.

Essentially a form of targeted dimension reduction, this process can lead to substantial efficiency gains when the immaterial variation is large relative to the material variation. Cook, Helland and Su (2013) adapted envelopes to the predictors, and showed that the SIMPLS algorithm for partial least squares regression converges to an envelope in the predictor space. Following Cook, et al. (2010), they demonstrated that using a likelihood-based objective function to separate the material and immaterial variation and to provide an estimator of the coefficient matrix produces clear and often substantial estimative and predictive advantages over SIMPLS.

However, little is known about using envelopes for joint reduction of the responses and predictors. The previous developments kindle a hope that we can combine their advantages to produce efficiency gains than are greater than those possible by reducing either the responses or the predictors alone. In this article, we develop likelihood-based envelope methods for simultaneously separating the material and immaterial variation in the responses and in the predictors. We show a potential for synergy in a synchronized reduction, producing an overall reduction in estimative variation surpassing that indicated by the marginal reductions. Finding a likelihood-based envelope can be computationally challenging, and so we propose a novel and fast optimization algorithm.

The rest of this paper is organized as follows. In Section 2, we briefly review the algebraic definition of an envelope as a prelude to our development of simultaneous reduction that starts in Section 3. We depart a bit from convention and integrate our review of individual response and predictor envelopes into our development of simultaneous envelopes. We also link the simultaneous envelope method with partial least squares and canonical correlation analysis. In Section 4 we introduce a likelihood-based objective function that includes the objective functions used by Cook, et al. (2010) and Cook, et al. (2013) as special cases. Novel algorithms for estimating a simultaneous envelope are also given in this section. In Section 5, asymptotic properties of the simultaneous envelope estimators are studied under normality and under general distributional assumptions. Encouraging simulation results on prediction and on determine the dimensions of envelopes are given in Section 7. In Section 8, we demonstrate the superiority of the simultaneous envelope estimator compared to classical methods by simulations and by predicting the contents of

biscuit dough samples. Proofs and other technical details are included in the Supplement to this article.

The following notations and definitions will be used in our exposition. Let $\mathbf{A} \sim \mathbf{B}$ denote that $\mathbf{A}$ has the same distribution as $\mathbf{B}$, let $\mathbf{A} \perp\!\!\!\perp \mathbf{B}$ denote that $\mathbf{A}$ is independent of $\mathbf{B}$ and let $\mathbf{A} \perp\!\!\!\perp \mathbf{B}|\mathbf{C}$ indicate that $\mathbf{A}$ is conditionally independent of $\mathbf{B}$ given $\mathbf{C}$. Let $\mathbb{R}^{m\times n}$ be the set of all real $m \times n$ matrices and let $\mathbb{S}^{k\times k}$ be the set of all real symmetric $k \times k$ matrices. If $\mathbf{M} \in \mathbb{R}^{m\times n}$ then $\mathrm{span}(\mathbf{M}) \subseteq \mathbb{R}^m$ is the subspace spanned by columns of $\mathbf{M}$. We use $\widehat{\boldsymbol{\theta}}_u$ to denote sample estimator for $\boldsymbol{\theta}$ with known parameters u. We use $\mathbf{P}_{\mathbf{A}(\mathbf{V})} = \mathbf{A}(\mathbf{A}^T\mathbf{V}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{V}$ to denote the projection onto $\mathrm{span}(\mathbf{A})$ with the $\mathbf{V}$ inner product and use $\mathbf{P}_{\mathbf{A}}$ to denote projection onto $\mathrm{span}(\mathbf{A})$ with the identity inner product. Let $\mathbf{Q}_{\mathbf{A}(\mathbf{V})} = \mathbf{I} - \mathbf{P}_{\mathbf{A}(\mathbf{V})}$. For an $m \times n$ matrix $\mathbf{A}$ and a $p \times q$ matrix $\mathbf{B}$, their direct sum is defined as the $(m+p)\times(n+q)$ block diagonal matrix $\mathbf{A}\oplus\mathbf{B} = \mathrm{diag}(\mathbf{A}, \mathbf{B})$. We will also use the $\oplus$ operator for two subspaces. If $\mathcal{S} \subseteq \mathbb{R}^p$ and $\mathcal{R} \subseteq \mathbb{R}^q$ then $\mathcal{S}\oplus\mathcal{R} = \mathrm{span}(\mathbf{S}\oplus\mathbf{R})$ where $\mathbf{S}$ and $\mathbf{R}$ are basis matrices for $\mathcal{S}$ and $\mathcal{R}$. The sum of two subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$ of $\mathbb{R}^m$ is defined as $\mathcal{S}_1 + \mathcal{S}_2 = \{\mathbf{v}_1 + \mathbf{v}_2|\mathbf{v}_1 \in \mathcal{S}_1, \mathbf{v}_2 \in \mathcal{S}_2\}$. We will use operators $\mathrm{vec} : \mathbb{R}^{a\times b} \to \mathbb{R}^{ab}$, which vectorizes an arbitrary matrix by stacking its columns, and $\mathrm{vech} : \mathbb{R}^{a\times a} \to \mathbb{R}^{a(a+1)/2}$, which vectorizes a symmetric matrix by extracting its columns of elements below or on the diagonal.

The estimation of envelopes will eventually be done by optimizing over Grassmann manifolds. A Grassmann manifold is the collection of all linear subspaces of a vector space of a given dimension. We use $\mathcal{G}_{(k,n)}$ to denote all the k-dimensional subspaces in $\mathbb{R}^n$, which is a Grassmann manifold with dimension $k(n-k)$. For more background in Grassmann manifolds and Grassmann optimizations, see Edelman, Tomas and Smith (1998). There are two currently available packages for Grassmann manifold optimization: R package *GrassmannOptim* by Adragni, Cook and Wu (2012) and the MATLAB package *sg_min* by Ross A. Lippert (http://web.mit.edu/~ripper/www/software/).

2 Definition of an envelope

In this section we briefly review definitions and a property of reducing subspaces and envelopes.

Definition 1. *A subspace $\mathcal{R} \subseteq \mathbb{R}^p$ is said to be a reducing subspace of $\mathbf{M} \in \mathbb{R}^{p\times p}$ if and only if $\mathcal{R}$ decomposes $\mathbf{M}$ as $\mathbf{M} = \mathbf{P}_\mathcal{R}\mathbf{M}\mathbf{P}_\mathcal{R} + \mathbf{Q}_\mathcal{R}\mathbf{M}\mathbf{Q}_\mathcal{R}$. If $\mathcal{R}$ is a reducing subspace of $\mathbf{M}$, we say that $\mathcal{R}$ reduces $\mathbf{M}$.*

This definition of a reducing subspace is equivalent to the usual definition found in functional analysis (Conway, 1990) and in the literature on invariant subspaces, but the underlying notion of reduction is incompatible with how it is usually understood in statistics. Nevertheless, it is common terminology in those areas and is the basis for the definition of an envelope (Cook, et al., 2010) which is central to our developments.

Definition 2. *Let $\mathbf{M} \in \mathbb{S}^{p\times p}$ and let $\mathcal{B} \subseteq \operatorname{span}(\mathbf{M})$. Then the $\mathbf{M}$-envelope of $\mathcal{B}$, denoted by $\mathcal{E}_\mathbf{M}(\mathcal{B})$, is the intersection of all reducing subspaces of $\mathbf{M}$ that contain $\mathcal{B}$.*

The intersection of two reducing subspaces of $\mathbf{M}$ is still a reducing subspace of $\mathbf{M}$. This means that $\mathcal{E}_\mathbf{M}(\mathcal{B})$, which is unique by its definition, is the smallest reducing subspace containing $\mathcal{B}$. Also, the $\mathbf{M}$-envelope of $\mathcal{B}$ always exist because of the requirement $\mathcal{B} \subseteq \operatorname{span}(\mathbf{M})$.

The following proposition from Cook, et al. (2010) gives a characterization of envelopes.

Proposition 1. *If $\mathbf{M} \in \mathbb{S}^{p\times p}$ has $q \leq p$ eigenspaces, then the $\mathbf{M}$-envelope of $\mathcal{B} \subseteq \operatorname{span}(\mathbf{M})$ can be constructed as $\mathcal{E}_\mathbf{M}(\mathcal{B}) = \sum_{i=1}^{q} \mathbf{P}_i\mathcal{B}$, where $\mathbf{P}_i$ is the projection onto the i-th eigenspace of $\mathbf{M}$.*

From this proposition, we see that the $\mathbf{M}$-envelope of $\mathcal{B}$ is the sum of the eigenspaces of $\mathbf{M}$ that are not orthogonal to $\mathcal{B}$; that is, the eigenspaces of $\mathbf{M}$ onto which $\mathcal{B}$ projects non-trivially. This implies that the envelope is the span of some subset of the eigensvectors of $\mathbf{M}$. In the regression context, $\mathcal{B}$ is typically the span of a regression coefficient matrix or a matrix of covariances, and $\mathbf{M}$ is chosen as a covariance matrix which is usually positive definite.

3 Simultaneous envelopes

3.1 Definition and structure

The standard multivariate linear model can be written as

$$\mathbf{Y} = \boldsymbol{\mu}_Y + \boldsymbol{\beta}^T(\mathbf{X} - \boldsymbol{\mu}_X) + \boldsymbol{\epsilon}, \tag{3.1}$$

where $\boldsymbol{\mu}_Y$ is the mean for $\mathbf{Y}$, $\boldsymbol{\mu}_X$ is the mean for $\mathbf{X}$, $\boldsymbol{\epsilon}$ is the error vector that has mean 0, variance $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}} > 0$ and is independent of $\mathbf{X}$, and $\boldsymbol{\beta} \in \mathbb{R}^{p\times r}$ is the regression coefficient matrix in which we are primarily interested. We use $\boldsymbol{\Sigma}_{\mathbf{X}} > 0$ and $\boldsymbol{\Sigma}_{\mathbf{Y}} > 0$ to denote the population covariance matrices of $\mathbf{X}$ and $\mathbf{Y}$, and use $\boldsymbol{\Sigma}_{\mathbf{XY}}$ to denote their cross-covariance matrix. The covariance matrix of the population residual vector $\boldsymbol{\epsilon}$ is then $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}} = \boldsymbol{\Sigma}_{\mathbf{Y}} - \boldsymbol{\Sigma}_{\mathbf{XY}}^T\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\boldsymbol{\Sigma}_{\mathbf{XY}}$. Similarly, let $\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}} = \boldsymbol{\Sigma}_{\mathbf{X}} - \boldsymbol{\Sigma}_{\mathbf{XY}}\boldsymbol{\Sigma}_{\mathbf{Y}}^{-1}\boldsymbol{\Sigma}_{\mathbf{XY}}^T$. The sample counterparts of these population covariance matrices are denoted by $\mathbf{S}_{\mathbf{X}}$, $\mathbf{S}_{\mathbf{Y}}$, $\mathbf{S}_{\mathbf{XY}}$, $\mathbf{S}_{\mathbf{Y}|\mathbf{X}}$ and $\mathbf{S}_{\mathbf{X}|\mathbf{Y}}$. We also define $\boldsymbol{\Sigma}_{\mathbf{A}|\mathbf{B}}$ and $\mathbf{S}_{\mathbf{A}|\mathbf{B}}$ in the same way for two arbitrary random vectors $\mathbf{A}$ and $\mathbf{B}$.

Envelope methods have the potential to increase efficiency in estimation of $\boldsymbol{\beta}$ after reducing $\mathbf{Y}$ (Cook et al., 2010) and to improve prediction of $\mathbf{Y}$ after reducing $\mathbf{X}$ (Cook et al., 2013). Our goal is to combine their advantages by simultaneously reducing $\mathbf{X}$ and $\mathbf{Y}$ to decrease both predictive and estimative variation. We next give a coordinate representation of simultaneous envelopes.

Let $d \leq \min(r, p)$ denote the rank of $\boldsymbol{\beta}$ and consider the singular value decomposition of $\boldsymbol{\beta} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, where $\mathbf{U} \in \mathbb{R}^{p\times d}$ and $\mathbf{V} \in \mathbb{R}^{r\times d}$ are orthogonal matrices, $\mathbf{U}^T\mathbf{U} = \mathbf{I}_d = \mathbf{V}^T\mathbf{V}$, and $\mathbf{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$ is a diagonal matrix with elements $\lambda_1 \geq \cdots \geq \lambda_d > 0$ being the d singular values of $\boldsymbol{\beta}$. We define the column space (the left eigenspace) and the row space (the right eigenspace) of $\boldsymbol{\beta}$ as $\operatorname{span}(\boldsymbol{\beta}) = \operatorname{span}(\mathbf{U}) \equiv \mathcal{L}$ and $\operatorname{span}(\boldsymbol{\beta}^T) = \operatorname{span}(\mathbf{V}) \equiv \mathcal{R}$. Then two envelopes can be constructed for simultaneously reducing the predictor space and the response space:

1. $\mathbf{X}$-envelope (left envelope): $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{L})$ with $\dim(\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{L})) = d_X$, $d \leq d_X \leq p$.

2. $\mathbf{Y}$-envelope (right envelope): $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\mathcal{R})$ with $\dim(\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\mathcal{R})) = d_Y$, $d \leq d_Y \leq r$.

We know from Cook, et al. (2010; Proposition 3.1) that $\mathcal{E}_{\mathbf{\Sigma_Y}}(\mathcal{R}) = \mathcal{E}_{\mathbf{\Sigma_{Y|X}}}(\mathcal{R})$. Consequently, an alternative definition of the $\mathbf{Y}$-envelope is $\mathcal{E}_{\mathbf{\Sigma_Y}}(\mathcal{R})$, which then has the same form as $\mathbf{X}$-envelope, by replacing $\mathbf{X}$ with $\mathbf{Y}$ and $\mathcal{L}$ with $\mathcal{R}$. We use $\mathcal{E}_{\mathbf{\Sigma_{Y|X}}}(\mathcal{R})$ as the definition of the $\mathbf{Y}$-envelope to facilitate later parameterizations. If we imaging that the elements of $\boldsymbol{\beta}$ are generated with respect to Lebesque measure then it follows from Propostion 1 that no reduction is possible: $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{L}) = \mathbb{R}^p$ and $\mathcal{E}_{\mathbf{\Sigma_{Y|X}}}(\mathcal{R}) = \mathbb{R}^r$ with probability one. Later in this section we show that proper envelopes imply certain relations between $\mathbf{X}$ and $\mathbf{Y}$ that may reasonably hold in practice.

From the definitions of the $\mathbf{X}$- and $\mathbf{Y}$-envelopes, if $d = r < p$ then we can reduce $\mathbf{X}$ only and $\mathcal{E}_{\mathbf{\Sigma_{Y|X}}}(\mathcal{R}) = \mathbb{R}^r$. Similarly, if $d = p < r$ then $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{L}) = \mathbb{R}^p$ and reduction is possible only in the response space. Hence, we will assume $d < \min(r,\ p)$ from now on and discuss the general situation where simultaneous reduction is possible. Let $\mathbf{L} \in \mathbb{R}^{p\times d_X}$ be an orthogonal basis for $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{L})$ and let $\mathbf{R} \in \mathbb{R}^{r\times d_Y}$ be an orthogonal basis for $\mathcal{E}_{\mathbf{\Sigma_{Y|X}}}(\mathcal{R})$. Also, $(\mathbf{L}, \mathbf{L}_0)$ is an orthogonal basis for $\mathbb{R}^p$ and $(\mathbf{R}, \mathbf{R}_0)$ is an orthogonal basis for $\mathbb{R}^r$. Then from Definition 1 we can write the covariance matrices as

$$\begin{aligned} \mathbf{\Sigma_X} &= \mathbf{L\Omega L}^T + \mathbf{L}_0\mathbf{\Omega}_0\mathbf{L}_0^T, \qquad (3.2) \\ \mathbf{\Sigma_{Y|X}} &= \mathbf{R\Phi R}^T + \mathbf{R}_0\mathbf{\Phi}_0\mathbf{R}_0^T. \qquad (3.3) \end{aligned}$$

The covariance matrix decomposition in (3.2) indicates that the eigenvectors of $\mathbf{\Sigma_X}$ fall in either $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{L})$ or $\mathcal{E}^{\perp}_{\mathbf{\Sigma_X}}(\mathcal{L})$ with corresponding eigenvalues being the eigenvalues of $\mathbf{\Omega}$ and $\mathbf{\Omega}_0$. No relationship is assumed between the eigenvalues of $\mathbf{\Omega}$ and $\mathbf{\Omega}_0$: The eigenvalues of $\mathbf{\Omega}$ could be any subset of the eigenvalues of $\mathbf{\Sigma_X}$. Similar results hold for the eigenvalues and eigenvectors of $\mathbf{\Sigma_{Y|X}}$, as seen in (3.3).

All of the information that is available about $\boldsymbol{\beta}$ is carried by the reduced variables $\mathbf{R}^T\mathbf{Y}$ and $\mathbf{L}^T\mathbf{X}$, which can be seen as follows. Recall the singular value decomposition of $\boldsymbol{\beta}$: $\boldsymbol{\beta} = \mathbf{UDV}^T$, where $\mathrm{span}(\mathbf{U}) = \mathcal{L} \subseteq \mathrm{span}(\mathbf{L})$ and $\mathrm{span}(\mathbf{V}) = \mathcal{R} \subseteq \mathrm{span}(\mathbf{R})$ by the definition of the $\mathbf{X}$- and $\mathbf{Y}$-envelopes. Hence, $\boldsymbol{\beta} = \mathbf{UDV}^T = (\mathbf{LA})\mathbf{D}(\mathbf{B}^T\mathbf{R}^T)$ for some semi-orthogonal matrices $\mathbf{A} \in \mathbb{R}^{d_X\times d}$ and $\mathbf{B} \in \mathbb{R}^{d_Y\times d}$. Then $\mathbf{R}_0^T\boldsymbol{\beta}^T = 0$, $\boldsymbol{\beta}^T\mathbf{L}_0 = 0$ and model (3.1) can be reduced to

$$\begin{cases} \mathbf{R}^T\mathbf{Y} = & \mathbf{R}^T\boldsymbol{\mu}_\mathbf{Y} + \boldsymbol{\eta}^T\{\mathbf{L}^T(\mathbf{X} - \boldsymbol{\mu}_\mathbf{X})\} + \mathbf{R}^T\boldsymbol{\epsilon}, \\ \mathbf{R}_0^T\mathbf{Y} = & \mathbf{R}_0^T\boldsymbol{\mu}_\mathbf{Y} + \mathbf{R}_0^T\boldsymbol{\epsilon}, \end{cases} \qquad (3.4)$$

where $\boldsymbol{\eta} = \mathbf{B}\mathbf{D}\mathbf{A}^T \in \mathbb{R}^{d_X \times d_Y}$ has rank d. The *simultaneous envelope model* then becomes

$$\begin{aligned} \mathbf{Y} &= \mathbf{R}\mathbf{R}^T\mathbf{Y} + \mathbf{R}_0\mathbf{R}_0^T\mathbf{Y} \\ &= \boldsymbol{\mu}_{\mathbf{Y}} + \mathbf{R}\boldsymbol{\eta}^T\mathbf{L}^T(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}) + \boldsymbol{\epsilon}. \end{aligned} \tag{3.5}$$

with $\boldsymbol{\Sigma}_{\mathbf{X}}$ and $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$ given by (3.2) and (3.3).

Comparing to (3.1), we see that the regression coefficient matrix is now $\boldsymbol{\beta} = \mathbf{L}\boldsymbol{\eta}\mathbf{R}^T$, where $\boldsymbol{\eta}$ contains the coordinates of $\boldsymbol{\beta}$ relative to $\mathbf{L}$ and $\mathbf{R}$. This implies that the columns and rows of $\boldsymbol{\beta}$ vary only within the $\mathbf{Y}$-envelope and the $\mathbf{X}$-envelope. By letting $\mathbf{L} = \mathbf{I}_p$ or $\mathbf{R} = \mathbf{I}_r$, there will be reductions only in the column space or the row space of $\boldsymbol{\beta}$. These are the two special situations studied by Cook et al. (2010) and Cook et al. (2013).

If $\mathbf{L} = \mathbf{I}_p$, it follows from (3.3) and (3.4) that $\mathrm{cov}(\mathbf{R}^T\mathbf{Y}, \mathbf{R}_0^T\mathbf{Y}|\mathbf{X}) = 0$ and $\mathbf{R}_0^T\mathbf{Y}|\mathbf{X} \sim \mathbf{R}_0^T\mathbf{Y}$, which motivated the construction of response envelopes (Cook, et al., 2010). If $\boldsymbol{\epsilon}$ is normally distributed, then this pair of conditions is equivalent to $\mathbf{R}_0^T\mathbf{Y} \perp\!\!\!\perp \mathbf{R}^T\mathbf{Y}|\mathbf{X}$ and $\mathbf{R}_0^T\mathbf{Y}|\mathbf{X} \sim \mathbf{R}_0^T\mathbf{Y}$. If $(\mathbf{X}, \mathbf{Y})$ is jointly normal then this pair of conditions is equivalent to $\mathbf{R}_0^T\mathbf{Y} \perp\!\!\!\perp (\mathbf{R}^T\mathbf{Y}, \mathbf{X})$. We refer to $\mathbf{R}^T\mathbf{Y}$ and $\mathbf{R}_0^T\mathbf{X}$ as the material and immaterial parts of the responses. This is because $\mathbf{R}_0^T\mathbf{Y}$ is neither affected by the predictors nor correlated with the complementary part of the responses $\mathbf{R}^T\mathbf{Y}$ and in this sense has no contribution to linear regression.

If $\mathbf{R} = \mathbf{I}_r$ then, from (3.2) and (3.5) , $\mathrm{cov}(\mathbf{Y}, \mathbf{L}_0^T\mathbf{X}|\mathbf{L}^T\mathbf{X}) = 0$ and $\mathrm{cov}(\mathbf{L}^T\mathbf{X}, \mathbf{L}_0^T\mathbf{X}) = 0$, which are the two conditions used by Cook, et al. (2013, Proposition 2.1) for predictor reduction. If $(\mathbf{X}, \mathbf{Y})$ has a joint normal distribution, then this pair of conditions is equivalent to $\mathbf{L}_0^T\mathbf{X} \perp\!\!\!\perp (\mathbf{L}^T\mathbf{X}, \mathbf{Y})$. We refer to $\mathbf{L}^T\mathbf{Y}$ and $\mathbf{L}_0^T\mathbf{X}$ as the material and immaterial parts of the predictors. Similar to the response, this is because $\mathbf{L}_0^T\mathbf{Y}$ is neither affected by the response nor correlated with the rest of the predictors.

The above conditions for $\mathbf{L}$ and $\mathbf{R}$ are not stated symmetrically because of the natural of regression. Cook et al. (2010) treated $\mathbf{X}$ as fixed since $\mathbf{S}_{\mathbf{X}}$ is an ancillary statistic for the $\mathbf{Y}$-envelope, while Cook et al. (2013) treated $\mathbf{X}$ as random because $\mathbf{S}_{\mathbf{X}}$ is not ancillary for $\mathbf{X}$ reduction. For simultaneous reductions, we assume that $\mathbf{X}$ and $\mathbf{Y}$ have a joint distribution throughout this article. The covariance decompositions (3.2) and (3.3)

play a critical role in obtaining the above relationships, and they distinguish the envelope reductions from other methods for reducing the column and row dimensions of $\boldsymbol{\beta}$.

The previous relationships follow from the marginal response and predictor envelopes. The following lemma describes additional relationships between the material part in $\mathbf{X}$ (or $\mathbf{Y}$) and the immaterial part in $\mathbf{Y}$ (or $\mathbf{X}$) that come with the simultaneous envelopes.

Lemma 1. *Assume the simultaneous envelope model (3.5). Then* $\mathrm{cov}(\mathbf{R}^T\mathbf{Y}, \mathbf{L}_0^T\mathbf{X}) = 0$ *and* $\mathrm{cov}(\mathbf{L}^T\mathbf{X}, \mathbf{R}_0^T\mathbf{Y}) = 0$.

This lemma, which does not require normality of $\mathbf{X}$ or $\mathbf{Y}$, is implied by the previous discussion if $\mathbf{L} = \mathbf{I}_p$ or $\mathbf{R} = \mathbf{I}_r$. It shows a similarity between simultaneous envelope reduction and canonical correlation analysis: the selected components are uncorrelated with the rest of the components. Additional discussion of the connection between envelopes and canonical correlation is given in Section 3.3.

3.2 A visualized example of simultaneous envelope

Figure 3.1 (a) and (b) illustrate the working mechanism of the simultaneous envelope reduction for a multivariate regression with two response $\mathbf{Y} = (Y_1, Y_2)^T$ and two predictors $\mathbf{X} = (X_1, X_2)^T$. For ease of illustration, we assume that $\mathbf{Y} = \boldsymbol{\beta}^T\mathbf{X} + \boldsymbol{\epsilon}$ with rank one regression coefficient matrix $\boldsymbol{\beta} = \mathbf{L}\mathbf{R}^T$ for some 2×1 matrices $\mathbf{L}$ and $\mathbf{R}$ such that $(\mathbf{L}, \mathbf{L}_0) \in \mathbb{R}^{2\times 2}$ and $(\mathbf{R}, \mathbf{R}_0) \in \mathbb{R}^{2\times 2}$ are orthogonal matrices. Then the plots demonstrate the set-up where $\mathbf{L}$ and $\mathbf{R}$ span the $\mathbf{X}$- and $\mathbf{Y}$-envelope.

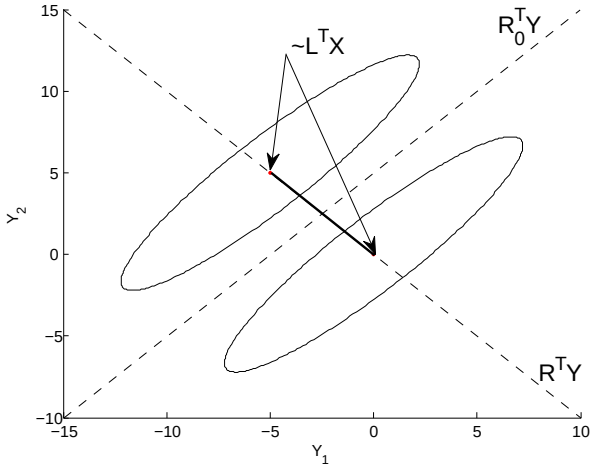
In the first plot, the conditional distribution of $\mathbf{Y}|\mathbf{X}$ is represented by the ellipses, whose axes are the directions of the eigenvectors of $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$. The shift from one contour to another is captured by $\boldsymbol{\beta}^T\mathbf{X} = \mathbf{R}\mathbf{L}^T\mathbf{X}$, which is in the direction of $\mathbf{R}$ and has magnitude proportional to $\mathbf{L}^T\mathbf{X}$. From Proposition 1, the $\mathbf{Y}$-envelope is the sum of eigenspaces of $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$ that are not orthogonal to $\mathbf{R}$. In this plot, the eigenvector corresponds to the larger eigenvalue of $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$ is orthogonal to $\mathbf{R}$ and hence represents the immaterial information of the regression. By projecting the data onto $\mathbf{R}^T\mathbf{Y}$, we will eliminate the immaterial variation in the response. The response envelope reduction in this case is very efficient because $\mathbf{R}$ lies in the eigenspace corresponds to the small eigenvalue of $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$.

The second plot represents the marginal distribution of $\mathbf{X}$. From Proposition 1, the reduction by $\mathbf{X}$-envelope is available when $\mathbf{L}$ is contained in a subset of the eigenspaces of $\mathbf{\Sigma_X}$. In this plot, $\mathbf{L}$ happens to be the first eigenvector of $\mathbf{\Sigma_X}$, which means it spans the $\mathbf{X}$-envelope and $\mathbf{L}_0^T\mathbf{X}$ will be immaterial information of the regression. It should be pointed out that if we assign the Lebesgue measure to this 2-dimensional space, the probability of $\mathbf{L}$ being one of the two eigenvector of $\mathbf{\Sigma_X}$ will be zero. But statistically, this event could happen because it is equivalent to requiring (a) $\mathrm{cov}(\mathbf{Y}, \mathbf{L}_0^T\mathbf{X}|\mathbf{L}^T\mathbf{X}) = 0$ and (b) $\mathrm{cov}(\mathbf{L}^T\mathbf{X}, \mathbf{L}_0^T\mathbf{X}) = 0$, see discussion in Section 3.1. As represented by this plot, the predictor envelope reduction has great advantage over OLS when $\mathbf{L}$ corresponds to the larger eigenvalue of $\mathbf{\Sigma_X}$. This can be seen from the first plot, where the magnitude of $\mathbf{L}^T\mathbf{X}$ is proportional to the strength of the linear relationship.

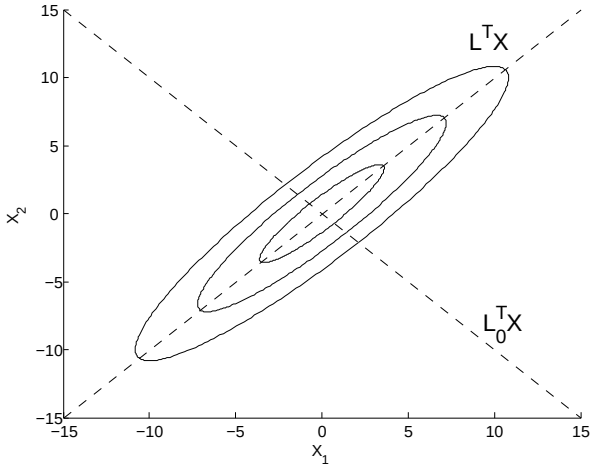
For a toy data example, we use the meat data analyzed by Cook et al. (2013) for envelope predictor reduction in multivariate linear regression. This dataset consists of spectral measurements from infrared transmittance for 103 meat samples. There are three response variables: percentages of protein, fat and water. The sum of the three percentages is not one because of the other chemical content in the sample. We take Y_1 to be protein and Y_2 to be the sum of water and fat. We take two spectral measurements at 910nm and 960nm for illustration. The estimated $\mathbf{X}$-envelope and the $\mathbf{Y}$-envelope are both one-dimensional. The left plot was constructed by conditioning on the high, median and low values of $\widehat{\mathbf{L}}^T\mathbf{X}$, we can clearly see that the estimated envelope direction $\widehat{\mathbf{R}}$ matches the major axes of the contour of each sub-sample. So in this example the $\mathbf{Y}$-envelope reduces the dimension but not very much immaterial information. On the other hand, the right plot closely resembles the schematic representation shown in top-right. Consequently, the simultaneous envelope method offers a more precise estimate of $\boldsymbol{\beta}$ than the standard method. The standard errors of the OLS estimated coefficients in $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ are 1.2, 12.7, 12.8 and 49.5 times of that of the simultaneous envelope estimator.

3.3 Links to PCA, PLS, CCA and RRR

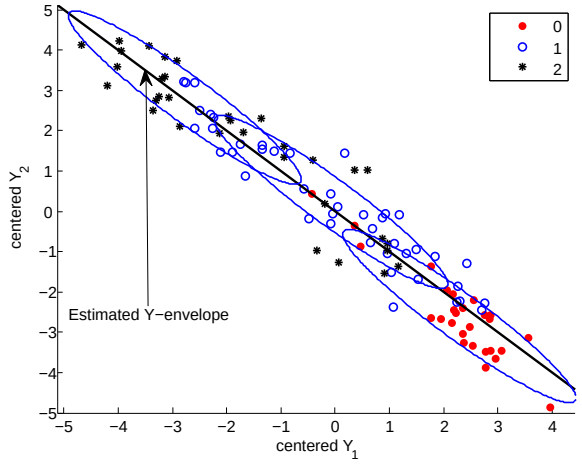
As stated previously, the eigenvalues of $\mathbf{\Omega}_0$ could be any subset of the eigenvalues of $\mathbf{\Sigma_X}$. When some or all of the largest few eigenvalues of $\mathbf{\Sigma_X}$ come from $\mathbf{\Omega}_0$, the first few



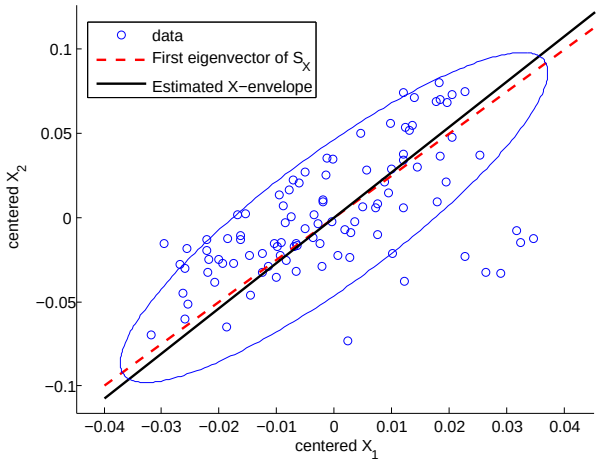
(a) $\mathbf{Y}$-envelope



(b) $\mathbf{X}$-envelope



(c) Estimated $\mathbf{Y}$-envelope



(d) Estimated $\mathbf{X}$-envelope

Figure 3.1: Working mechanism of simultaneous envelope reduction. (a) Schematic representation of response envelope reduction; (b) schematic representation of envelope in the predictor space; (c) the meat data with estimated $\mathbf{Y}$-envelope, where the data points were marked differently according to their values in the predictor envelope $\widehat{\mathbf{L}}^T\mathbf{X}_i$.
; (d) the meat data with estimated $\mathbf{X}$-envelope and the first eigenvector of $\mathbf{S_X}$. To help visualization, elliptical contours that cover 90% of the data points were added in (c) and (d).

principal components of $\mathbf{X}$ will be from the immaterial part of $\mathbf{X}$, which is ineffective for the regression as we can see from the above relationships. Principal components may be effective only if the larger eigenvalues of $\mathbf{\Sigma_X}$ all come from $\mathbf{\Omega}$; that is, from the material variation. Similar problems of principal component analysis arise for reducing $\mathbf{Y}$.

As mentioned in the Introduction, the partial least squares method reduces only the predictors and so it is comparable to simultaneous envelopes method when setting $\mathbf{R} = \mathbf{I}_r$. Cook, et al. (2013) showed that in this case the SIMPLS algorithm (de Jong, 1993) produces a $\sqrt{n}$-consistent estimator of $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{L})$, and that a likelihood-based estimator can do much better for prediction than the SIMPLS estimator.

Canonical correlation analysis is widely used for the purpose of simultaneously reducing the predictors and the responses. In the population, it finds canonical pairs of directions $\{\mathbf{a}_i, \mathbf{b}_i\}$, $i = 1, \ldots, d$, such that the correlations between $\mathbf{a}_i^T\mathbf{X}$ and $\mathbf{b}_i^T\mathbf{Y}$ are maximized. The maximization is over the constraints $\mathbf{a}_j^T\mathbf{\Sigma_X}\mathbf{a}_k = 0$, $\mathbf{a}_j^T\mathbf{\Sigma_{XY}}\mathbf{b}_k = 0$ and $\mathbf{b}_j^T\mathbf{\Sigma_Y}\mathbf{b}_k = 0$ for all $j \neq k$, and $\mathbf{a}_j^T\mathbf{\Sigma_X}\mathbf{a}_j = 1$ and $\mathbf{b}_j^T\mathbf{\Sigma_Y}\mathbf{b}_j = 1$ for all j. The solution is then $\{\mathbf{a}_i, \mathbf{b}_i\} = \{\mathbf{\Sigma_X}^{-1/2}\mathbf{e}_i, \mathbf{\Sigma_Y}^{-1/2}\mathbf{f}_i\}$, where $\{\mathbf{e}_i, \mathbf{f}_i\}$ is the i-th left-right eigenvector pair of the correlation matrix $\boldsymbol{\rho} = \mathbf{\Sigma_X}^{-1/2}\mathbf{\Sigma_{XY}}\mathbf{\Sigma_Y}^{-1/2}$.

Lemma 2. *Under the simultaneous envelope model (3.5), canonical correlation analysis can find at most d directions in the population, where $d = \mathrm{rank}(\boldsymbol{\beta}) \leq \min(d_X, d_Y)$. Moreover, the directions are contained in the simultaneous envelope as*

$$\mathrm{span}(\mathbf{a}_1, \ldots, \mathbf{a}_d) \subseteq \mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{L}),\ \mathrm{span}(\mathbf{b}_1, \ldots, \mathbf{b}_d) \subseteq \mathcal{E}_{\mathbf{\Sigma_{Y|X}}}(\mathcal{R}). \tag{3.6}$$

Hence, canonical correlation analysis may miss some information about the regression by ignoring some material parts of $\mathbf{X}$ and/or $\mathbf{Y}$. For example, when r is small, it can find at most r linear combinations of $\mathbf{X}$, which can be insufficient for regression. Moreover, the most correlated pairs are not, in general, the most predictable pairs for regression. Canonical correlations are often used for data visualization instead of regression, so it may be expected that they can fail in prediction. In the simulation studies of Section 7.1 and Section J in the Supplement, we found that the prediction performances based on canonical correlations varied widely for different covariance structures.

The maximum likelihood estimator of RRR is obtained as

$$\widehat{\boldsymbol{\beta}}_{\mathrm{RRR}} = \arg \min_{\mathrm{rank}(\boldsymbol{\beta})=d} \{\sum_{i=1}^{n} (\mathbf{Y}_i - \boldsymbol{\beta}^T \mathbf{X}_i)^T \mathbf{S}_{\mathbf{Y}|\mathbf{X}}^{-1} (\mathbf{Y}_i - \boldsymbol{\beta}^T \mathbf{X}_i)\}, \tag{3.7}$$

which is the same as the regression coefficient matrix obtained by using the canonical variables (see Reinsel and Velu 1998; Section 2.4.2). Therefore, in the purpose of estimation and prediction, the maximum likelihood RRR estimator is equivalent to the estimator based on CCA estimator. RRR can also be applied with identity inner product in (3.7) instead of $\mathbf{S}_{\mathbf{Y}|\mathbf{X}}$. And for any minimization criteria of RRR, the estimators always have the form $\widehat{\boldsymbol{\beta}} = \widehat{\mathbf{A}}\widehat{\mathbf{B}}$ where $\widehat{\mathbf{A}} \in \mathbb{R}^{p\times d}$ and $\widehat{\mathbf{B}} \in \mathbb{R}^{d\times r}$ are both $\sqrt{n}$-consistent estimators for their population counterparts $\mathbf{A}$ and $\mathbf{B}$. Similar to Lemma 2, we have the following relations by definition,

$$\mathrm{span}(\mathbf{A}) = \mathrm{span}(\boldsymbol{\beta}) \subseteq \mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\mathcal{L}),\ \mathrm{span}(\mathbf{B}^T) = \mathrm{span}(\boldsymbol{\beta}^T) \subseteq \mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{Y}}(\mathcal{R}). \tag{3.8}$$

Therefore, RRR has the same drawback as CCA that it may loss some material information in regression. Simulation studies in Section 7 includes RRR estimator based on identity inner product.

3.4 Potential gain

To gain intuition about the potential advantages of simultaneous envelopes, we next consider the case where $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\mathcal{L})$ and $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{Y}}(\mathcal{R})$ are known. Estimation of these envelopes in practice will mitigate the findings in this section, but we have nevertheless found them to be useful qualitative indicators of the benefits of simultaneous reduction.

The envelope estimator for $\boldsymbol{\beta}$ with known semi-orthogonal basis matrices $\mathbf{L}$ and $\mathbf{R}$, denoted by $\widehat{\boldsymbol{\beta}}_{\mathbf{L},\mathbf{R}}$, can be written as

$$\widehat{\boldsymbol{\beta}}_{\mathbf{L},\mathbf{R}} = \mathbf{L}\widehat{\boldsymbol{\eta}}_{\mathbf{L},\mathbf{R}}\mathbf{R}^T = \mathbf{L}(\mathbf{L}^T\mathbf{S}_\mathbf{X}\mathbf{L})^{-1}\mathbf{L}^T\mathbf{S}_{\mathbf{XY}}\mathbf{R}\mathbf{R}^T = \mathbf{P}_{\mathbf{L}(\mathbf{S}_\mathbf{X})}\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}\mathbf{P}_\mathbf{R}, \tag{3.9}$$

where $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} = \mathbf{S}_\mathbf{X}^{-1}\mathbf{S}_{\mathbf{XY}}$ is the ordinary least squares estimator. Clearly, the estimator $\widehat{\boldsymbol{\beta}}_{\mathbf{L},\mathbf{R}}$ uses only the material variation in $\mathbf{Y}$ and $\mathbf{X}$. It can be obtained by projecting $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ to the reduced predictor space and the reduced response space, and so does not depend on the particular bases $\mathbf{L}$ and $\mathbf{R}$ selected. The estimator $\widehat{\boldsymbol{\eta}}_{\mathbf{L},\mathbf{R}}$ is the ordinary least squares estimator of the coefficient matrix for the regression of $\mathbf{R}^T\mathbf{Y}$ on $\mathbf{L}^T\mathbf{X}$.

The next proposition shows how the variance of the ordinary least squares estimator with normal predictors can be potentially reduced by using simultaneous envelopes. Let $f_p = n - p - 2$ and $f_x = n - d_X - 2$.

Proposition 2. *Assume that* $\mathbf{X} \sim N_p(\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Sigma}_{\mathbf{X}})$, $n > p + 2$ *and that semi-orthogonal basis matrices* $\mathbf{L}$, $\mathbf{R}$ *for the left and right envelopes are known. Then* $\mathrm{var}(\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})) = f_p^{-1}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}} \otimes \boldsymbol{\Sigma}_{\mathbf{X}}^{-1}$ *and*

$$\begin{aligned}\mathrm{var}(\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathbf{L},\mathbf{R}})) &= f_x^{-1}\left(\mathbf{R}\boldsymbol{\Phi}\mathbf{R}^T\right) \otimes \left(\mathbf{L}\boldsymbol{\Omega}^{-1}\mathbf{L}^T\right) \\ &= f_p f_x^{-1}\mathrm{var}(\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})) - f_x^{-1}\mathbf{R}_0\boldsymbol{\Phi}_0\mathbf{R}_0^T \otimes \mathbf{L}\boldsymbol{\Omega}^{-1}\mathbf{L}^T \\ &\quad - f_x^{-1}\mathbf{R}\boldsymbol{\Phi}\mathbf{R}^T \otimes \mathbf{L}_0\boldsymbol{\Omega}_0^{-1}\mathbf{L}_0^T - f_x^{-1}\mathbf{R}_0\boldsymbol{\Phi}_0\mathbf{R}_0^T \otimes \mathbf{L}_0\boldsymbol{\Omega}_0^{-1}\mathbf{L}_0^T,\end{aligned}$$

where $\boldsymbol{\Omega} = \mathbf{L}^T\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{L}$, $\boldsymbol{\Omega}_0 = \mathbf{L}_0^T\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{L}_0$, $\boldsymbol{\Phi}_0 = \mathbf{R}_0^T\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}\mathbf{R}_0$ *and* $\boldsymbol{\Phi} = \mathbf{R}^T\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}\mathbf{R}$.

This proposition shows that the variation in $\widehat{\boldsymbol{\beta}}_{\mathbf{L},\mathbf{R}}$ can be seen in two parts: the first part is the variation in $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ times a constant $f_p f_x^{-1} \leq 1$, and the second consists of terms that reduce this value depending on the variances associated with the immaterial information $\mathbf{L}_0^T\mathbf{X}$ and $\mathbf{R}_0^T\mathbf{Y}$.

If $\mathbf{R} = \mathbf{I}_r$ then $\boldsymbol{\Phi} = \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$ and we get the multivariate version of Proposition 2.3 by Cook, et al. (2013) for univariate response regression:

$$\mathrm{var}(\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathbf{L}})) = f_p f_x^{-1}\mathrm{var}(\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})) - f_x^{-1}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}} \otimes \mathbf{L}_0\boldsymbol{\Omega}_0^{-1}\mathbf{L}_0^T. \tag{3.10}$$

When p is close to n and the $\mathbf{X}$-envelope dimension d_X is small, the constant $f_p f_x^{-1}$ could be small and the gain by $\widehat{\boldsymbol{\beta}}_{\mathbf{L},\mathbf{R}}$ over $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ could be substantial. If there is substantial co-linearity in the predictors, so $\boldsymbol{\Sigma}_{\mathbf{X}}$ has some small eigenvalues, and if the corresponding eigenvectors of $\boldsymbol{\Sigma}_{\mathbf{X}}$ fall in $\mathcal{E}^{\perp}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{L})$, then the variance of $\widehat{\boldsymbol{\beta}}_{\mathbf{L}}$ could be reduced considerably since $\boldsymbol{\Omega}_0^{-1}$ will be large. It is widely known that colinearity in $\mathbf{X}$ can increase the variance of $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$. However, when the eigenvectors of $\boldsymbol{\Sigma}_{\mathbf{X}}$ corresponding to these small eigenvalues lie in $\mathcal{E}^{\perp}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{L})$, the variance of $\widehat{\boldsymbol{\beta}}_{\mathbf{L},\mathbf{R}}$ is not affected by colinearity.

Similarly, if $\mathbf{L} = \mathbf{I}_p$ then $\boldsymbol{\Omega} = \boldsymbol{\Sigma}_{\mathbf{X}}$ and we get the following new expression for $\mathbf{Y}$ reduction:

$$\mathrm{var}(\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathbf{R}})) = f_p f_x^{-1}\mathrm{var}(\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})) - f_x^{-1}\mathbf{R}_0\boldsymbol{\Phi}_0\mathbf{R}_0^T \otimes \boldsymbol{\Sigma}_{\mathbf{X}}. \tag{3.11}$$

If the eigenvectors with larger eigenvalues of $\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}}$ lie in $\mathcal{E}^{\perp}_{\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\mathcal{R})$, then the variance of $\widehat{\boldsymbol{\beta}}_{\mathbf{R}}$ may be reduced considerably since then $\mathbf{\Phi}_0$ will be large.

More importantly, the last term of the expansion in Proposition 2 represents a synergy between the $\mathbf{X}$ and $\mathbf{Y}$ reductions that is not present in the marginal reductions. If the eigenvectors of $\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}}$ with large eigenvalues lie in $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\mathcal{R})$ or if the eigenvectors of $\mathbf{\Sigma}_{\mathbf{X}}$ with small eigenvalues lie in $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{X}}}(\mathcal{L})$, then the variance reductions in either (3.11) or (3.10) could be insignificant. However, the synergy in simultaneous $\mathbf{X}$ and $\mathbf{Y}$ reductions may still reduce the variance substantially because one of factors in the Kronecker product $f_x^{-1}\mathbf{R}_0\mathbf{\Phi}_0\mathbf{R}_0^T \otimes \mathbf{L}_0\mathbf{\Omega}_0^{-1}\mathbf{L}_0^T$ could be still be large.

Let $\mathbf{x}_N$ denote a new observation from $N(\boldsymbol{\mu}_{\mathbf{X}}, \mathbf{\Sigma}_{\mathbf{X}})$, let $\mathbf{z}_N = \mathbf{x}_N - \boldsymbol{\mu}_{\mathbf{X}}$ be held fixed and consider $\mathrm{var}(\widehat{\boldsymbol{\beta}}^T\mathbf{z}_N) = \mathrm{var}\left((\mathbf{z}_N^T \otimes \mathbf{I})\mathrm{vec}(\widehat{\boldsymbol{\beta}}^T)\right)$, which is the variance of a fitted vector for $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ and $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{\mathbf{L},\mathbf{R}}$. It is straightforward from Proposition 2 that,

$$\begin{aligned} f_p\mathrm{var}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}^T\mathbf{z}_N) &= f_x\mathrm{var}(\widehat{\boldsymbol{\beta}}_{\mathbf{L},\mathbf{R}}^T\mathbf{z}_N) + \mathbf{R}_0\mathbf{\Phi}_0\mathbf{R}_0^T\left(\mathbf{z}_N^T\mathbf{L}\mathbf{\Omega}^{-1}\mathbf{L}^T\mathbf{z}_N\right) \\ &\quad + \mathbf{R}\mathbf{\Phi}\mathbf{R}^T(\mathbf{z}_N^T\mathbf{L}_0\mathbf{\Omega}_0^{-1}\mathbf{L}_0^T\mathbf{z}_N) + \mathbf{R}_0\mathbf{\Phi}_0\mathbf{R}_0^T\left(\mathbf{z}_N^T\mathbf{L}_0\mathbf{\Omega}_0^{-1}\mathbf{L}_0^T\mathbf{z}_N\right). \end{aligned}$$

We see from the above equality that only the part of $\mathbf{z}_N$ that lies in $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{X}}}(\mathcal{L})$ will contribute to the variance in prediction from $\widehat{\boldsymbol{\beta}}_{\mathbf{L},\mathbf{R}}$, while the prediction variance from $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ depends on the whole of $\mathbf{z}_N$. If, in an extreme case, $\mathbf{z}_N \in \mathcal{E}^{\perp}_{\mathbf{\Sigma}_{\mathbf{X}}}(\mathcal{L})$, then $\mathrm{var}(\widehat{\boldsymbol{\beta}}_{\mathbf{L},\mathbf{R}}^T\mathbf{z}_N) = 0$.

4 Estimating envelopes

Let $\mathbf{C} = (\mathbf{X}^T, \mathbf{Y}^T)^T$ denote the concatenated random vectors, which has mean $\boldsymbol{\mu}_{\mathbf{C}}$ and covariance $\mathbf{\Sigma}_{\mathbf{C}}$, and let $\mathbf{S}_{\mathbf{C}}$ be the sample covariance matrix of $\mathbf{C}$. In order to estimate the parameters of the simultaneous envelope model (3.5), we introduce and probe a likelihood-based objective function that includes the objective functions in Cook, et al. (2010) and Cook, et al. (2013) as special cases. Variations on the objective function and their corresponding algorithms are also studied. We first give coordinate dependent and coordinate independent representations of $\mathbf{\Sigma}_{\mathbf{C}}$ in Section 4.1 to facilitate estimation.

4.1 Structure of $\boldsymbol{\Sigma}_{\mathbf{C}}$

Since we already have the structure of $\boldsymbol{\Sigma}_{\mathbf{X}}$ and $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$ in (3.2) and (3.3), and $\boldsymbol{\Sigma}_{\mathbf{XY}} = \boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\beta} = \mathbf{L}\boldsymbol{\Omega}\boldsymbol{\eta}\mathbf{R}^T$, we need only the following expression for $\boldsymbol{\Sigma}_{\mathbf{Y}}$ to complete the necessary ingredients of $\boldsymbol{\Sigma}_{\mathbf{C}}$,

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathbf{Y}} &= \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}} + \boldsymbol{\Sigma}_{\mathbf{XY}}^T\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\boldsymbol{\Sigma}_{\mathbf{XY}} \\ &= \mathbf{R}\boldsymbol{\Phi}\mathbf{R}^T + \mathbf{R}_0\boldsymbol{\Phi}_0\mathbf{R}_0^T + \mathbf{R}\boldsymbol{\eta}^T\boldsymbol{\Omega}\mathbf{L}^T\left(\mathbf{L}\boldsymbol{\Omega}\mathbf{L}^T + \mathbf{L}_0\boldsymbol{\Omega}_0\mathbf{L}_0^T\right)^{-1}\mathbf{L}\boldsymbol{\Omega}\boldsymbol{\eta}\mathbf{R}^T \\ &= \mathbf{R}(\boldsymbol{\Phi} + \boldsymbol{\eta}^T\boldsymbol{\Omega}\boldsymbol{\eta})\mathbf{R}^T + \mathbf{R}_0\boldsymbol{\Phi}_0\mathbf{R}_0^T.\end{aligned}$$

Then we get the coordinate representation of the covariance $\boldsymbol{\Sigma}_{\mathbf{C}}$ as

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathbf{C}} &= \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{X}} & \boldsymbol{\Sigma}_{\mathbf{XY}} \\ \boldsymbol{\Sigma}_{\mathbf{XY}}^T & \boldsymbol{\Sigma}_{\mathbf{Y}} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{L}\boldsymbol{\Omega}\mathbf{L}^T + \mathbf{L}_0\boldsymbol{\Omega}_0\mathbf{L}_0^T & \mathbf{L}\boldsymbol{\Omega}\boldsymbol{\eta}\mathbf{R}^T \\ \mathbf{R}\boldsymbol{\eta}^T\boldsymbol{\Omega}\mathbf{L}^T & \mathbf{R}(\boldsymbol{\Phi} + \boldsymbol{\eta}^T\boldsymbol{\Omega}\boldsymbol{\eta})\mathbf{R}^T + \mathbf{R}_0\boldsymbol{\Phi}_0\mathbf{R}_0^T \end{pmatrix}. \end{aligned} \tag{4.1}$$

By noticing that $\boldsymbol{\Sigma}_{\mathbf{X}} = \mathbf{P}_{\mathbf{L}}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{P}_{\mathbf{L}} + \mathbf{Q}_{\mathbf{L}}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{Q}_{\mathbf{L}}$, $\boldsymbol{\Sigma}_{\mathbf{XY}} = \mathbf{P}_{\mathbf{L}}\boldsymbol{\Sigma}_{\mathbf{XY}}\mathbf{P}_{\mathbf{R}}$ and $\boldsymbol{\Sigma}_{\mathbf{Y}} = \mathbf{P}_{\mathbf{R}}\boldsymbol{\Sigma}_{\mathbf{Y}}\mathbf{P}_{\mathbf{R}} + \mathbf{Q}_{\mathbf{R}}\boldsymbol{\Sigma}_{\mathbf{Y}}\mathbf{Q}_{\mathbf{R}}$ from the above expression, we can further obtain the coordinate independent representation

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathbf{C}} &= \mathbf{P}_{\mathbf{L}\oplus\mathbf{R}}\boldsymbol{\Sigma}_{\mathbf{C}}\mathbf{P}_{\mathbf{L}\oplus\mathbf{R}} + \mathbf{Q}_{\mathbf{L}\oplus\mathbf{R}}\boldsymbol{\Sigma}_{\mathbf{C}}\mathbf{Q}_{\mathbf{L}\oplus\mathbf{R}}, \\ &= \mathbf{P}_{\mathbf{L}\oplus\mathbf{R}}\boldsymbol{\Sigma}_{\mathbf{C}}\mathbf{P}_{\mathbf{L}\oplus\mathbf{R}} + \mathbf{Q}_{\mathbf{L}\oplus\mathbf{R}}\boldsymbol{\Sigma}_{\mathbf{D}}\mathbf{Q}_{\mathbf{L}\oplus\mathbf{R}}, \end{aligned} \tag{4.2}$$

where $\boldsymbol{\Sigma}_{\mathbf{D}} \equiv \boldsymbol{\Sigma}_{\mathbf{X}} \oplus \boldsymbol{\Sigma}_{\mathbf{Y}}$ and $\mathbf{P}_{\mathbf{L}\oplus\mathbf{R}} = \mathbf{P}_{\mathbf{L}} \oplus \mathbf{P}_{\mathbf{R}}$ is the projection onto the direct sum of two envelopes, $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{L}) \oplus \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\mathcal{R})$.

So far we have considered $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{L})$ and $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\mathcal{R})$ as separate subspaces. Motivated by (4.2), the next lemma states that the direct sum of two arbitrary envelopes is itself an envelope. Let $\mathbf{M}_1 \in \mathbb{S}^{p_1\times p_1}$, $\mathbf{M}_2 \in \mathbb{S}^{p_2\times p_2}$, and let $\mathcal{S}_1$ and $\mathcal{S}_2$ be subspaces of $\mathrm{span}(\mathbf{M}_1)$ and $\mathrm{span}(\mathbf{M}_2)$, which is required by Definition 2. Then

Lemma 3. $\mathcal{E}_{\mathbf{M}_1}(\mathcal{S}_1) \oplus \mathcal{E}_{\mathbf{M}_2}(\mathcal{S}_2) = \mathcal{E}_{\mathbf{M}_1\oplus\mathbf{M}_2}(\mathcal{S}_1 \oplus \mathcal{S}_2)$.

From this lemma, we have $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{L})\oplus\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\mathcal{R}) = \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{L})\oplus\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}}}(\mathcal{R}) = \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}\oplus\boldsymbol{\Sigma}_{\mathbf{Y}}}(\mathcal{L}\oplus\mathcal{R}) = \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{D}}}(\mathcal{L} \oplus \mathcal{R})$. We call this the simultaneous envelope for $\boldsymbol{\beta}$.

4.2 The estimation criterion and resulting estimators

Assuming multivariate normality for $\mathbf{C}$, the negative log-likelihood minimized over $\boldsymbol{\mu}_{\mathbf{C}}$ leads to the following objective function for estimation of $\boldsymbol{\Sigma}_{\mathbf{C}}$,

$$\mathrm{F}(\boldsymbol{\Sigma}_{\mathbf{C}}) = \log|\boldsymbol{\Sigma}_{\mathbf{C}}| + \mathrm{trace}(\mathbf{S}_{\mathbf{C}}\boldsymbol{\Sigma}_{\mathbf{C}}^{-1}). \tag{4.3}$$

We use this as a multi-purpose objective function, which gives the maximum likelihood estimator of $\boldsymbol{\beta}$ under normality of $\mathbf{C}$ and a $\sqrt{n}$-consistent simultaneous envelope estimator of $\boldsymbol{\beta}$ when $\mathbf{C}$ has finite fourth moments. We also use $\mathrm{F}(\cdot)$ as a generic objective function whose definition changes and is implied by its own argument. This should cause no confusion since $\mathrm{F}(\cdot)$ will always be written with its arguments.

Substituting the coordinate form for $\boldsymbol{\Sigma}_{\mathbf{C}}$ from (4.1) into $\mathrm{F}(\boldsymbol{\Sigma}_{\mathbf{C}})$ leads to an objective function that can be minimized explicitly over $\boldsymbol{\Omega}$, $\boldsymbol{\Omega}_0$, $\boldsymbol{\Phi}$, $\boldsymbol{\Phi}_0$ and $\boldsymbol{\eta}$ with $\mathbf{R}$ and $\mathbf{L}$ held fixed. The resulting partially minimized objective function for simultaneous envelopes can be expressed as follows.

$$\mathrm{F}(\mathbf{L}\oplus\mathbf{R}) = \log|(\mathbf{L}^T\oplus\mathbf{R}^T)\mathbf{S}_{\mathbf{C}}(\mathbf{L}\oplus\mathbf{R})| + \log|(\mathbf{L}^T\oplus\mathbf{R}^T)\mathbf{S}_{\mathbf{D}}^{-1}(\mathbf{L}\oplus\mathbf{R})|, \tag{4.4}$$

where $\mathbf{S}_{\mathbf{D}} = \mathbf{S}_{\mathbf{X}}\oplus\mathbf{S}_{\mathbf{Y}}$ is the sample version of $\boldsymbol{\Sigma}_{\mathbf{D}}$. Let $|\mathbf{A}|_0$ denote the product of all nonzero eigenvalues of a matrix $\mathbf{A}$. Then we have the following coordinate independent representation of (4.4),

$$\mathrm{F}(\mathbf{P}_{\mathbf{L}\oplus\mathbf{R}}) = \log|\mathbf{P}_{\mathbf{L}\oplus\mathbf{R}}\mathbf{S}_{\mathbf{C}}\mathbf{P}_{\mathbf{L}\oplus\mathbf{R}}|_0 + \log|\mathbf{P}_{\mathbf{L}\oplus\mathbf{R}}\mathbf{S}_{\mathbf{D}}^{-1}\mathbf{P}_{\mathbf{L}\oplus\mathbf{R}}|_0. \tag{4.5}$$

Moreover, if we substitute the coordinate free representation of $\boldsymbol{\Sigma}_{\mathbf{C}}$ from (4.2) into $\mathrm{F}(\boldsymbol{\Sigma}_{\mathbf{C}})$, we immediately get (4.5).

Objective function (4.4) is invariant under orthogonal transformations: for any orthogonal $(d_X + d_Y)\times(d_X + d_Y)$ matrix $\mathbf{O}$, $\mathrm{F}(\mathbf{L}\oplus\mathbf{R}) = \mathrm{F}((\mathbf{L}\oplus\mathbf{R})\mathbf{O})$. The same result holds if we replace $\mathbf{S}_{\mathbf{C}}$ and $\mathbf{S}_{\mathbf{D}}$ with their population counterparts. Minimization of $\mathrm{F}(\mathbf{L}\oplus\mathbf{R})$ is thus a Grassmann optimization problem with special direct sum structure. Consequently, neither $\mathbf{L}$ nor $\mathbf{R}$ is identified. However, $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{L})$ and $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\mathcal{R})$ are identified, and these are all that is needed to estimate $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}_{\mathbf{X}}$ and $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$. While the estimators of $\boldsymbol{\eta}$, $\boldsymbol{\Omega}$, $\boldsymbol{\Omega}_0$, $\boldsymbol{\Phi}$ and $\boldsymbol{\Phi}_0$ depend on the particular bases chosen to minimize $\mathrm{F}(\mathbf{L}\oplus\mathbf{R})$,

the estimators of $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}_{\mathbf{X}}$ and $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$ are independent of the choice since the estimated projection matrices $\mathbf{P}_{\widehat{\mathbf{L}}}$ and $\mathbf{P}_{\widehat{\mathbf{R}}}$ do not depend on the basis. In short, any values $\widehat{\mathbf{L}}$ of $\mathbf{L}$ and $\widehat{\mathbf{R}}$ of $\mathbf{R}$ that minimize $\mathrm{F}(\mathbf{L} \oplus \mathbf{R})$ are allowed. Estimators of $\mathbf{L}_0$ and $\mathbf{R}_0$ are then any semi-orthogonal matrices $\widehat{\mathbf{L}}_0$ and $\widehat{\mathbf{R}}_0$ so that $(\widehat{\mathbf{L}}, \widehat{\mathbf{L}}_0)$ and $(\widehat{\mathbf{R}}, \widehat{\mathbf{R}}_0)$ are orthogonal matrices.

The next lemma summarizes the estimators that result from this process.

Lemma 4. *Let* $\widehat{\mathbf{L}} \oplus \widehat{\mathbf{R}} = \arg\min \mathrm{F}(\mathbf{L} \oplus \mathbf{R})$*, where* $\mathbf{L} \in \mathbb{R}^{p \times d_X}$ *and* $\mathbf{R} \in \mathbb{R}^{r \times d_Y}$ *are semi-orthogonal basis matrices. Then the estimators of the remaining parameters are* $\widehat{\boldsymbol{\Omega}} = \widehat{\mathbf{L}}^T \mathbf{S}_{\mathbf{X}} \widehat{\mathbf{L}}$, $\widehat{\boldsymbol{\Omega}}_0 = \widehat{\mathbf{L}}_0^T \mathbf{S}_{\mathbf{X}} \widehat{\mathbf{L}}_0$, $\widehat{\boldsymbol{\eta}} = (\widehat{\mathbf{L}}^T \mathbf{S}_{\mathbf{X}} \widehat{\mathbf{L}})^{-1} (\widehat{\mathbf{L}}^T \mathbf{S}_{\mathbf{XY}} \widehat{\mathbf{R}})$, $\widehat{\boldsymbol{\Phi}}_0 = \widehat{\mathbf{R}}_0^T \mathbf{S}_{\mathbf{Y}} \widehat{\mathbf{R}}_0$ *and*

$$\widehat{\boldsymbol{\Phi}} = \widehat{\mathbf{R}}^T \left(\mathbf{S}_{\mathbf{Y}} - \mathbf{S}_{\mathbf{XY}}^T \widehat{\mathbf{L}} \left(\widehat{\mathbf{L}}^T \mathbf{S}_{\mathbf{X}} \widehat{\mathbf{L}} \right)^{-1} \widehat{\mathbf{L}}^T \mathbf{S}_{\mathbf{XY}} \right) \widehat{\mathbf{R}}.$$

The simultaneous envelope estimator for $\boldsymbol{\beta}$ *is then*

$$\widehat{\boldsymbol{\beta}} = \widehat{\mathbf{L}} \widehat{\boldsymbol{\eta}} \widehat{\mathbf{R}}^T = \mathbf{P}_{\widehat{\mathbf{L}}(\mathbf{S}_{\mathbf{X}})} \widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} \mathbf{P}_{\widehat{\mathbf{R}}}. \tag{4.6}$$

The simultaneous envelope estimator for $\boldsymbol{\beta}$ in (4.6) coincides with the plug-in envelope estimator $\widehat{\boldsymbol{\beta}}_{\widehat{\mathbf{L}},\widehat{\mathbf{R}}}$ obtained by regarding estimated $\widehat{\mathbf{L}}$ and $\widehat{\mathbf{R}}$ as the known values of $\mathbf{L}$ and $\mathbf{R}$. We next turn to methods for minimizing (4.4).

4.3 Alternating algorithm

If we fix $\mathbf{L}$ as an arbitrary orthogonal basis, then the objective function $\mathrm{F}(\mathbf{L} \oplus \mathbf{R})$ in (4.4) can be re-expressed as an objective function in $\mathbf{R}$:

$$\mathrm{F}(\mathbf{R}|\mathbf{L}) = \log |\mathbf{R}^T \mathbf{S}_{\mathbf{Y}|\mathbf{L}^T\mathbf{X}} \mathbf{R}| + \log |\mathbf{R}^T \mathbf{S}_{\mathbf{Y}}^{-1} \mathbf{R}|. \tag{4.7}$$

Similarly, if we fix $\mathbf{R}$, the objective function $\mathrm{F}(\mathbf{L} \oplus \mathbf{R})$ reduces to the conditional objective function

$$\mathrm{F}(\mathbf{L}|\mathbf{R}) = \log |\mathbf{L}^T \mathbf{S}_{\mathbf{X}|\mathbf{R}^T\mathbf{Y}} \mathbf{L}| + \log |\mathbf{L}^T \mathbf{S}_{\mathbf{X}}^{-1} \mathbf{L}|. \tag{4.8}$$

We use the following alternating algorithm based on (4.7) and (4.8) to obtain a minimizer of the objective function $\mathrm{F}(\mathbf{L} \oplus \mathbf{R})$ in (4.4).

1. Initialization. Set the starting value $\mathbf{L}^{(0)}$ and get $\mathbf{R}^{(0)} = \arg\min_{\mathbf{R}} \mathrm{F}(\mathbf{R}|\mathbf{L}^{(0)})$.

2. Alternating. For the k-th stage, obtain $\mathbf{L}^{(k)} = \arg\min_{\mathbf{L}} \mathrm{F}(\mathbf{L}|\mathbf{R} = \mathbf{R}^{(k-1)})$ and $\mathbf{R}^{(k)} = \arg\min_{\mathbf{R}} \mathrm{F}(\mathbf{R}|\mathbf{L} = \mathbf{L}^{(k)})$.

3. Convergence criterion. Evaluate $\left\{\mathrm{F}(\mathbf{L}^{(k-1)} \oplus \mathbf{R}^{(k-1)}) - \mathrm{F}(\mathbf{L}^{(k)} \oplus \mathbf{R}^{(k)})\right\}$ and return to the alternating step if it is bigger than a tolerance; otherwise, stop the iteration and use $\mathbf{L}^{(k)} \oplus \mathbf{R}^{(k)}$ as the final estimator.

In the initialization step of the algorithm, we could also set $\mathbf{R}^{(0)}$ to be some initial value and get $\mathbf{L}^{(0)} = \arg\min_{\mathbf{L}} \mathrm{F}(\mathbf{L}|\mathbf{R}^{(0)})$. Then interchanging the roles of $\mathbf{L}$ and $\mathbf{R}$ in the alternating step will give us another algorithm, which has the same performance as the alternating algorithm outlined above. Comparing to the objective function used by Cook et al. (2010), we see that $\mathrm{F}(\mathbf{R}|\mathbf{L})$ is the objective function for estimating the $\mathbf{Y}$-envelope in the regression of $\mathbf{Y}$ on the reduced predictors $\mathbf{L}^T\mathbf{X}$ for fixed $\mathbf{L}$. Similarly from the objective function used by Cook et al. (2013), we notice that $\mathrm{F}(\mathbf{L}|\mathbf{R})$ is the objective function for estimating the $\mathbf{X}$-envelope in the regression of the reduced responses $\mathbf{R}^T\mathbf{Y}$ on the predictors $\mathbf{X}$ for fixed $\mathbf{R}$.

In our experience, as long as we use good initial values, the alternating algorithm, which monotonically decreases $\mathrm{F}(\mathbf{L}\oplus\mathbf{R})$, will converge after only a few cycles, typically less than four. Root-n consistent starting values are particularly important to mitigate potential problems caused by multiple local minima and to ensure efficient estimation. For instance, under joint normality of $\mathbf{X}$ and $\mathbf{Y}$, one Newton-Raphson iteration from any $\sqrt{n}$-consistent estimator will be asymptotically as efficient as the maximum likelihood estimator, even if local minima are present (Small et al., 2000; Lehmann and Casella, 1998, Theorem 4.3).

4.4 One-dimensional optimization

In this section we propose a fast 1D algorithm for $\sqrt{n}$-consistent envelope estimation. The algorithm, which does not itself require starting values, can be used for standalone estimation or to obtain starting values for the alternating algorithm. We state our algorithm in terms of estimating $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{L})$ from $f(\mathbf{L}) \equiv \mathrm{F}(\mathbf{L}|\mathbf{R} = \mathbf{I}_r)$. It works similarly for estimating $\mathcal{E}_{\mathbf{\Sigma_{Y|X}}}(\mathcal{R})$ and for estimating a general envelope $\mathcal{E}_{\mathbf{M}}(\mathcal{S})$.

Optimization over $\mathcal{G}_{(p,d_X)}$, as required in each cycle of the alternating algorithm, can be computationally expensive. The next lemma offers a way to break down the d_X-dimensional manifold into fast 1D components.

Lemma 5. *Let $\mathbf{M} \in \mathbb{S}^{p\times p}$ be a positive definite matrix and let $\mathcal{S} \subset \mathbb{R}^p$. Suppose $(\mathbf{B}, \mathbf{B}_0)$ is an orthogonal basis of $\mathbb{R}^p$ where $\mathbf{B} \in \mathbb{R}^{p\times q}$, $\mathbf{B}_0 \in \mathbb{R}^{p\times(p-q)}$ and $\operatorname{span}(\mathbf{B}) \subseteq \mathcal{E}_{\mathbf{M}}(\mathcal{S})$. Then $\mathbf{v} \in \mathcal{E}_{\mathbf{B}_0^T\mathbf{M}\mathbf{B}_0}(\mathbf{B}_0^T\mathcal{S})$ implies $\mathbf{B}_0\mathbf{v} \in \mathcal{E}_{\mathbf{M}}(\mathcal{S})$.*

Suppose we have an estimate $\mathbf{B}$ of q directions in the envelope $\mathcal{E}_{\mathbf{M}}(\mathcal{S})$. Then we can estimate the rest of $\mathcal{E}_{\mathbf{M}}(\mathcal{S})$ by focusing our attention on $\mathcal{E}_{\mathbf{B}_0^T\mathbf{M}\mathbf{B}_0}(\mathbf{B}_0^T\mathcal{S})$, which is a lower dimensional envelope. This leads to the following 1D algorithm.

Let $\ell_k \in \mathbb{R}^p$, $k = 1, \ldots, d_X$, be the stepwise directions obtained. Let $\mathbf{L}_k = (\ell_1, \ldots, \ell_k)$, let $(\mathbf{L}_k, \mathbf{L}_{0k})$ be an orthogonal basis for $\mathbb{R}^p$ and set initial value $\ell_0 = \mathbf{L}_0 = 0$, then for $k = 0, \ldots, d_X - 1$, get

$$\begin{aligned} \mathbf{v}_{k+1} &= \arg\min_{\mathbf{v}\in\mathbb{R}^{p-k}} f_k(\mathbf{v}), \text{ subject to } \mathbf{v}^T\mathbf{v} = 1, \\ \ell_{k+1} &= \mathbf{L}_{0k}\mathbf{v}_{k+1}, \end{aligned} \tag{4.9}$$

where $f_k(\mathbf{v}) = \log|\mathbf{v}^T\mathbf{L}_{0k}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\mathbf{L}_{0k}\mathbf{v}| + \log|\mathbf{v}^T(\mathbf{L}_{0k}^T\mathbf{S}_{\mathbf{X}}\mathbf{L}_{0k})^{-1}\mathbf{v}|$ for $\mathbf{v} \in \mathbb{R}^{p-k}$. This algorithm does not require an initial value search. It starts simply from $\ell_0 = \mathbf{L}_0 = 0$ and builds the estimator sequentially.

Suppose we know $\operatorname{span}(\mathbf{L}_k) \subset \mathcal{E}_{\mathbf{\Sigma}_\mathbf{X}}(\mathcal{L})$, then we can estimate the remaining part of $\mathcal{E}_{\mathbf{\Sigma}_\mathbf{X}}(\mathcal{L})$ by substituting $\mathbf{L} = (\mathbf{L}_k, \mathbf{L}_{0k}\mathbf{V})$, where $\mathbf{V} \in \mathbb{R}^{(p-k)\times(d_X-k)}$, into $f(\mathbf{L})$, leading to the objective function $\log|\mathbf{V}^T\mathbf{L}_{0k}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\mathbf{L}_{0k}\mathbf{V}| + \log|\mathbf{V}^T(\mathbf{L}_{0k}^T\mathbf{S}_{\mathbf{X}}\mathbf{L}_{0k})^{-1}\mathbf{V}|$, whose one dimensional version is exactly $f_k(\mathbf{v})$. This describes how the 1D algorithm is minimizing the full objective function $f(\mathbf{L})$ sequentially. Let $\widehat{\mathbf{L}}_{1\mathrm{D}}$ denote the estimator from this algorithm.

The following proposition formally states the $\sqrt{n}$-consistency of the 1D algorithm for estimating $\mathcal{E}_{\mathbf{\Sigma}_\mathbf{X}}(\mathcal{L})$. Similar results hold for estimating $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\mathcal{R})$ and for estimating an arbitrary envelope using the 1D algorithm.

Proposition 3. *Assume that $\mathbf{S}_\mathbf{X}$ and $\mathbf{S}_{\mathbf{X}|\mathbf{Y}}$ are $\sqrt{n}$-consistent estimators of $\mathbf{\Sigma}_\mathbf{X}$ and $\mathbf{\Sigma}_{\mathbf{X}|\mathbf{Y}}$. Let $\widehat{\mathbf{L}}_{1\mathrm{D}}$ denote the estimator obtained from the 1D algorithm. Then $\mathbf{P}_{\widehat{\mathbf{L}}_{1\mathrm{D}}}$ is $\sqrt{n}$-consistent for the projection onto $\mathcal{E}_{\mathbf{\Sigma}_\mathbf{X}}(\mathcal{L})$.*

We have also considered sequentially minimizing $f(\ell) = \log|\ell^T \mathbf{S}_{\mathbf{X}|\mathbf{Y}}\ell| + \log|\ell^T \mathbf{S}_{\mathbf{X}}^{-1}\ell|$, so that $\ell_i^T \ell_j = 0$ if $i \neq j$ and equals one if $i = j$. Although this seems to be a reasonable way of sequentially constructing the envelope, we do not know if this procedure results in a consistent estimator. Our numerical experience is that the 1D algorithm (4.9) performs much better than such sequential optimization.

The 1D algorithm can be used to get fast $\sqrt{n}$-consistent starting values for the alternating algorithm in Section 4.3 by separately estimating the $\mathbf{X}$-envelope and the $\mathbf{Y}$-envelope bases. Since the 1D algorithm turns the optimizations over d_X and d_Y-dimensional manifolds to d_X and d_Y sequential optimizations over 1D manifolds, the computation complexity is reduced drastically. For the simulation examples we studied, the computation costs of the 1D algorithm are from tens to hundreds times less than the computation costs of using d_X and d_Y-dimensional Grassmann manifold optimizations.

Perhaps more importantly, we have found $\widehat{\mathbf{L}}_{1\mathrm{D}}$ and $\widehat{\mathbf{R}}_{1\mathrm{D}}$ to be practically as efficient as the final estimators obtained by alternating because the alternating algorithm nearly always converges after only a few iterations and produces only small changes. To emphasize the utility of the 1D estimators, we use them in the simulations and real data examples that follow in later sections. In Section 7.3, we use a simulation example to demonstrate that estimators from the 1D manifold algorithm may have the same behavior as maximum likelihood estimators.

5 Asymptotic properties

The parameters involved in the coordinate representation of the simultaneous envelope model are $\boldsymbol{\eta}$, $\boldsymbol{\Omega}$, $\boldsymbol{\Omega}_0$, $\boldsymbol{\Phi}$, $\boldsymbol{\Phi}_0$, $\mathbf{L}$ and $\mathbf{R}$. In Sections 5.1 and 5.2, we focus on the asymptotic properties of the estimable functions $\boldsymbol{\beta} = \mathbf{L}\boldsymbol{\eta}\mathbf{R}^T$, $\boldsymbol{\Sigma}_{\mathbf{X}} = \mathbf{L}\boldsymbol{\Omega}\mathbf{L}^T + \mathbf{L}_0\boldsymbol{\Omega}_0\mathbf{L}_0^T$, and $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}} = \mathbf{R}\boldsymbol{\Phi}\mathbf{R}^T + \mathbf{R}_0\boldsymbol{\Phi}_0\mathbf{R}_0^T$. Specifically, we study the asymptotic covariances of the following

parameters $\boldsymbol{\phi}$ and estimable functions $\boldsymbol{h}$.

$$\boldsymbol{\phi} = \begin{pmatrix} \boldsymbol{\phi}_1 \\ \boldsymbol{\phi}_2 \\ \boldsymbol{\phi}_3 \\ \boldsymbol{\phi}_4 \\ \boldsymbol{\phi}_5 \\ \boldsymbol{\phi}_6 \\ \boldsymbol{\phi}_7 \end{pmatrix} \equiv \begin{pmatrix} \mathrm{vec}(\boldsymbol{\eta}) \\ \mathrm{vec}(\mathbf{L}) \\ \mathrm{vec}(\mathbf{R}) \\ \mathrm{vech}(\boldsymbol{\Omega}) \\ \mathrm{vech}(\boldsymbol{\Omega}_0) \\ \mathrm{vech}(\boldsymbol{\Phi}) \\ \mathrm{vech}(\boldsymbol{\Phi}_0) \end{pmatrix}, \ \boldsymbol{h} = \begin{pmatrix} \boldsymbol{h}_1(\boldsymbol{\phi}) \\ \boldsymbol{h}_2(\boldsymbol{\phi}) \\ \boldsymbol{h}_3(\boldsymbol{\phi}) \end{pmatrix} \equiv \begin{pmatrix} \mathrm{vec}(\boldsymbol{\beta}) \\ \mathrm{vech}(\boldsymbol{\Sigma}_{\mathbf{X}}) \\ \mathrm{vech}(\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}) \end{pmatrix}.$$

Since $\boldsymbol{h} \in \mathbb{R}^{\frac{1}{2}(p+r)(p+r+1)}$ and $\boldsymbol{\phi} \in \mathbb{R}^{\frac{1}{2}p(p+1)+\frac{1}{2}r(r+1)+d_X d_Y}$, the simultaneous envelope model reduces the total number of parameters by $\frac{1}{2}(p+r)(p+r+1) - \frac{1}{2}p(p+1) - \frac{1}{2}r(r+1) - d_X d_Y = pr - d_X d_Y$.

5.1 Without the normality assumption

Let $\widehat{\boldsymbol{h}}_{\text{full}} = \left(\mathrm{vec}^T(\widehat{\boldsymbol{\beta}}_{\text{OLS}}), \mathrm{vech}^T(\mathbf{S}_{\mathbf{X}}), \mathrm{vech}^T(\mathbf{S}_{\mathbf{Y}|\mathbf{X}})\right)^T$ denote the OLS estimator of $\boldsymbol{h}$ under the standard model (3.1), and let $\widehat{\boldsymbol{h}}$ denote the estimator from Lemma 4 under the simultaneous envelope model (3.5). The true values of $\boldsymbol{h}$ and $\boldsymbol{\phi}$ are denoted as $\boldsymbol{h}_0$ and $\boldsymbol{\phi}_0$. Define $\boldsymbol{\Delta} \equiv \partial \boldsymbol{h}(\boldsymbol{\phi})/\partial \boldsymbol{\phi}$ to be the gradient matrix, whose explicit form is given in the Supplement, Section I.1. We use "avar" to denote an asymptotic covariance matrix: $\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{h}}) = \mathbf{A}$ is equivalent to $\sqrt{n}(\widehat{\boldsymbol{h}} - \boldsymbol{h}_0) \to N(0, \mathbf{A})$ in distribution. We use the expansion matrix (Henderson and Searle, 1979) $\mathbf{E}_p \in \mathbb{R}^{p^2 \times p(p+1)/2}$ to relate the vec operation and vech operation: for any symmetric matrix $\mathbf{M} \in \mathbb{S}^{p\times p}$, $\mathrm{vec}(\mathbf{M}) = \mathbf{E}_p \mathrm{vech}(\mathbf{M})$. Then $\sqrt{n}(\widehat{\boldsymbol{h}}_{\text{full}} - \boldsymbol{h}_0) \to N(0, \boldsymbol{\Gamma})$, for some positive definite covariance matrix $\boldsymbol{\Gamma}$. Since there is a one-to-one relationship between h and $\boldsymbol{\Sigma}_{\mathbf{C}}$, we treat the objective function $\mathrm{F}(\boldsymbol{\Sigma}_{\mathbf{C}})$ in (4.3) as a function of $\boldsymbol{h}$ and $\widehat{\boldsymbol{h}}_{\text{full}}$ and write it as $\mathrm{F}(\boldsymbol{h}, \widehat{\boldsymbol{h}}_{\text{full}})$. Let $\mathbf{J}_{\boldsymbol{h}} = 1/2 \times \partial^2 \mathrm{F}(\boldsymbol{h}, \widehat{\boldsymbol{h}}_{\text{full}})/\partial \boldsymbol{h} \partial \boldsymbol{h}^T$ evaluated at $\widehat{\boldsymbol{h}}_{\text{full}} = \boldsymbol{h} = \boldsymbol{h}_0$, which is the Fisher information matrix for $\boldsymbol{h}$ when $\mathbf{C}$ is normal,

$$\mathbf{J}_{\boldsymbol{h}} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{\mathbf{X}} & 0 & 0 \\ 0 & \frac{1}{2}\mathbf{E}_p^T(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{\mathbf{X}}^{-1})\mathbf{E}_p & 0 \\ 0 & 0 & \frac{1}{2}\mathbf{E}_r^T(\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1})\mathbf{E}_r \end{pmatrix}. \quad (5.1)$$

The following proposition formally states the asymptotic distribution of $\widehat{\boldsymbol{h}}$.

Proposition 4. *Assume that the data $(\mathbf{x}_i, \mathbf{y}_i)$ are i.i.d. from a joint distribution with finite fourth moments. Then $\sqrt{n}(\mathrm{vec}(\widehat{\boldsymbol{h}}) - \mathrm{vec}(\boldsymbol{h}_0))$ converges in distribution to a normal*

random variable with mean $\mathbf{0}$ *and covariance matrix*

$$\mathbf{W} = \boldsymbol{\Delta}\left(\boldsymbol{\Delta}^T\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\Delta}\right)^{\dagger}\boldsymbol{\Delta}^T\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\Gamma}\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\Delta}\left(\boldsymbol{\Delta}^T\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\Delta}\right)^{\dagger}\boldsymbol{\Delta}^T,$$

where $\dagger$ *indicates the Moore-Penrose inverse. In particular,* $\sqrt{n}(\mathrm{vec}(\widehat{\boldsymbol{\beta}}) - \mathrm{vec}(\boldsymbol{\beta}))$ *converges in distribution to a normal random variable with mean* $\mathbf{0}$ *and covariance* $\mathbf{W}_{11}$*, the upper-left* $pr \times pr$ *block of* $\mathbf{W}$*. Moreover,* $\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{h}}_{\mathrm{full}}) \geq \mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{h}})$ *if* $\mathrm{span}(\mathbf{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{\Delta})$ *is a reducing subspace of* $\mathbf{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{\Gamma}\mathbf{J}_{\boldsymbol{h}}^{1/2}$.

The $\sqrt{n}$-consistency of the estimator $\widehat{\boldsymbol{\beta}}$ is essentially because $\mathbf{S}_{\mathbf{X}}$, $\mathbf{S}_{\mathbf{Y}}$ and $\mathbf{S}_{\mathbf{XY}}$ are $\sqrt{n}$-consistent and because of the properties of $\mathrm{F}(\boldsymbol{h}, \widehat{\boldsymbol{h}}_{\mathrm{full}})$. The asymptotic covariance matrix $\mathbf{W}_{11}$ can be computed straightforwardly, but its accuracy for any fixed sample size may depend on the distribution of $\mathbf{C}$. Fortunately, bootstrap methods can provide a good approximation of $\mathbf{W}_{11}$, as discussed in Section 5.3.

5.2 Under the normality assumption

The asymptotic covariance $\mathbf{W}$ simplifies when $\mathbf{C} \sim N(\boldsymbol{\mu}_{\mathbf{C}}, \boldsymbol{\Sigma}_{\mathbf{C}})$ because then objective function $\mathrm{F}(\boldsymbol{h}, \widehat{\boldsymbol{h}}_{\mathrm{full}})$ agrees with the negative log-likelihood function and $\boldsymbol{\Gamma} = \mathbf{J}_{\boldsymbol{h}}^{-1} = \mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{h}}_{\mathrm{full}})$. And $\{\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{h}}_{\mathrm{full}}) - \mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{h}})\}$ will be positive semi-definite because $\mathrm{span}(\mathbf{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{\Delta})$ is always a reducing subspace of $\mathbf{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{\Gamma}\mathbf{J}_{\boldsymbol{h}}^{1/2} = \mathbf{I}$.

Proposition 5. *Assume that* $\mathbf{C} \sim N(\boldsymbol{\mu}_{\mathbf{C}}, \boldsymbol{\Sigma}_{\mathbf{C}})$*. Then* $\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{h}}) = \boldsymbol{\Delta}(\boldsymbol{\Delta}^T\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\Delta})^{\dagger}\boldsymbol{\Delta}^T$*. Moreover,* $\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{h}}) \leq \mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{h}}_{\mathrm{full}})$,

$$\mathbf{J}_{\boldsymbol{h}}^{-1} - \boldsymbol{\Delta}(\boldsymbol{\Delta}^T\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\Delta})^{\dagger}\boldsymbol{\Delta}^T = \mathbf{J}_{\boldsymbol{h}}^{-\frac{1}{2}}\mathbf{Q}_{\mathbf{J}_{\boldsymbol{h}}^{\frac{1}{2}}\boldsymbol{\Delta}}\mathbf{J}_{\boldsymbol{h}}^{-\frac{1}{2}} \geq 0.$$

In particular, $\mathrm{avar}(\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}})) \leq \mathrm{avar}(\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}))$.

This proposition is obtained by direct computation, and is consistent with Cook et al. (2010) and Cook et al. (2013). Also, because the estimator $\widehat{\boldsymbol{\beta}}$ is the MLE under normality, its asymptotic variance is no larger than that of $\mathbf{X}$-envelope estimator or $\mathbf{Y}$-envelope estimator. Explicit expressions can be found in the Supplement, Section I. To further interpret this result, we next consider the asymptotic variance of $\mathrm{vec}(\widehat{\boldsymbol{\beta}})$, which can lead to the asymptotic variances for predictions and fitted values.

Proposition 6. *Assume that* $\mathbf{C} \sim N(\boldsymbol{\mu}_{\mathbf{C}}, \boldsymbol{\Sigma}_{\mathbf{C}})$. *Then*

$$\begin{aligned} \operatorname{avar}(\sqrt{n}\operatorname{vec}(\widehat{\boldsymbol{\beta}})) &= \operatorname{avar}(\sqrt{n}\operatorname{vec}(\widehat{\boldsymbol{\beta}}_{\mathbf{L},\mathbf{R}})) + \operatorname{avar}\left(\sqrt{n}\operatorname{vec}(\mathbf{Q}_{\mathbf{L}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\eta},\mathbf{R}})\right) \\ &+ \operatorname{avar}\left(\sqrt{n}\operatorname{vec}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\eta},\mathbf{L}}\mathbf{Q}_{\mathbf{R}})\right), \end{aligned}$$

where we use $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\eta},\mathbf{R}}$, $\widehat{\boldsymbol{\beta}}_{\mathbf{L},\mathbf{R}}$ *and* $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\eta},\mathbf{L}}$ *to denote the maximum likelihood estimators when the parameters in their subscripts are known. Explicit expressions for the asymptotic variances of* $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\eta},\mathbf{R}}$, $\widehat{\boldsymbol{\beta}}_{\mathbf{L},\mathbf{R}}$ *and* $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\eta},\mathbf{L}}$ *are given in the Supplement, Section I.*

The first term in the above decomposition, $\operatorname{avar}(\sqrt{n}\operatorname{vec}(\widehat{\boldsymbol{\beta}}_{\mathbf{L},\mathbf{R}}))$, is the same as that in Proposition 2, which gave the asymptotic variance if we knew bases for the true envelope. If we set $\mathbf{L} = \mathbf{I}_p$ in the above decomposition, so we are pursuing $\mathbf{Y}$ reduction only, then the decomposition reduces to that given by Cook et al. (2010; Theorem 6.1). Setting $\mathbf{R} = \mathbf{I}_r$, which indicates $\mathbf{X}$ reduction only gives the corresponding result of Cook et al. (2013; Proposition 4.4). The projections $\mathbf{Q}_{\mathbf{L}}$ and $\mathbf{Q}_{\mathbf{R}}$ serve to orthogonalize the random vectors so that the asymptotic variances are additive.

5.3 Residual bootstrap

To illustrate the application of the above bootstrap method, we consider a simple model with $p = r = 3$ and $d_X = d_Y = 1$. We generated $\mathbf{L}$, $\mathbf{R}$ and $\boldsymbol{\eta}$ by filling in random numbers from uniform$(0, 1)$ distribution. Then we orthonormalized $\mathbf{L}$, $\mathbf{R}$ and obtained corresponding $\mathbf{L}_0$, $\mathbf{R}_0$. The covariance matrices were $\boldsymbol{\Omega} = 5$, $\boldsymbol{\Omega}_0 = \mathbf{I}_2$, $\boldsymbol{\Phi} = 1$ and $\boldsymbol{\Phi}_0 = 10\mathbf{I}_2$. The data vectors $\mathbf{C}_i = (\mathbf{X}_i^T, \mathbf{Y}_i^T)^T$, $i = 1, \ldots, n$, were simulated from $\boldsymbol{\Sigma}_{\mathbf{C}}^{1/2}\mathbf{U}_i$ where $\mathbf{U}_i$ is a vector of i.i.d uniform random variables with mean 0 standard deviation 1. Therefore, $\mathbf{C}_i$ would follow a distribution with mean 0, covariance $\boldsymbol{\Sigma}_{\mathbf{C}}$ and finite fourth moments. A dataset with $n = 100$ observations was generated and $B = 100$ bootstrap datasets were used throughout. We used the bootstrap method to estimate the variances of two estimators: the OLS estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ and the simultaneous envelope estimator $\widehat{\boldsymbol{\beta}}_{1\mathrm{D}}$ which was obtained by using the 1D algorithm without alternating.

Table 1 summarizes all the $p \times r = 9$ elements of $\operatorname{vec}(\boldsymbol{\beta})$, $\operatorname{vec}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})$ and $\operatorname{vec}(\widehat{\boldsymbol{\beta}}_{1\mathrm{D}})$ and their asymptotic, bootstrap and actual standard errors. We included the asymptotic standard errors of the elements in $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$, which were the square roots of diagonals in

True parameter									
$\text{vec}(\boldsymbol{\beta}) \times 100$	30	30	32	36	36	38	14	14	15
OLS estimator									
$\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{OLS}}) \times 100$	37	37	14	29	42	53	39	2	15
$\text{E}_N \left\{ \text{vec}(\widehat{\boldsymbol{\beta}}_{\text{OLS}}) \right\} \times 100$	29	29	32	37	36	38	13	17	13
Asymptotic se$\times 100$	22	22	22	20	20	19	26	26	26
Bootstrap se$\times 100$	21	20	17	17	17	16	25	28	25
Actual se$\times 100$	22	23	22	21	22	18	30	24	23
Simultaneous envelope estimator									
$\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{1D}}) \times 100$	31	34	30	39	43	38	16	18	16
$\text{E}_N \left\{ \text{vec}(\widehat{\boldsymbol{\beta}}_{\text{1D}}) \right\} \times 100$	30	30	32	36	35	38	14	14	15
Asymptotic se$\times 100$	3	3	3	3	3	3	2	2	2
Bootstrap se$\times 100$	3	2	3	3	3	3	2	2	2
Actual se$\times 100$	3	2	3	3	3	3	2	2	2

Table 1: Bootstrap and estimated asymptotic standard errors of the 9 elements in $\widehat{\boldsymbol{\beta}}$ under the OLS estimator and the simultaneous envelope estimator with fixed $d_X = d_Y = 1$.

$\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}} \otimes \boldsymbol{\Sigma}_{\mathbf{X}}^{-1}/n$ and the asymptotic standard errors of the MLE under normality for the simultaneous envelope model, which is given in the Supplement, in order to compare with the estimator from the 1D manifold algorithm. We also repeatedly simulated $N = 100$ datasets under the same setting, and estimated the averaged coefficient estimates $\text{E}_N(\widehat{\boldsymbol{\beta}})$ and the actual standard errors of $\text{vec}(\widehat{\boldsymbol{\beta}})$ from the $N = 100$ replicates.

The asymptotic standard errors and the actual standard errors were well estimated by the bootstrap method. Moreover, the estimated asymptotic standard errors for the 1D algorithm are really close to the asymptotic standard errors of the MLE although the normality assumption was violated. As expected, the envelope estimator had much smaller standard errors than that of the OLS estimator.

6 Selecting the dimensions

6.1 Rank of β

Since the simultaneous envelope contains the row and the column space of $\boldsymbol{\beta}$, the dimensions d_X and d_Y are bounded below by $d = \text{rank}(\boldsymbol{\beta})$. Thus when determine the dimensions d_X and d_Y, it is helpful to first have some guidance on d. Bura and Cook

(2003) developed a chi-squared test for the rank d that requires only that the response variables have finite second moments. The test statistic is $\Lambda_d = n\sum_{j=d+1}^{\min(p,r)}\varphi_j^2$, where $\varphi_1 \geq \cdots \geq \varphi_{\min(p,r)}$ are eigenvalues of the $p \times r$ matrix

$$\widehat{\boldsymbol{\beta}}_{\text{std}} = \{(n-p-1)\}/n\}^{1/2}\mathbf{S}_{\mathbf{X}}^{1/2}\widehat{\boldsymbol{\beta}}_{\text{OLS}}\mathbf{S}_{\mathbf{Y}|\mathbf{X}}^{-1/2}. \tag{6.1}$$

Then they derived that $\boldsymbol{\Lambda}_d$ is asymptotically distributed as a $\chi^2_{(p-d)(r-d)}$ random variable under the null hypothesis that $\text{rank}(\boldsymbol{\beta}) = d$. The rank is then determined by comparing a sequence of test statistics Λ_k, $k = 0, \ldots, \min(p,r)-1$, to the percentiles of their null distribution $\chi^2_{(p-k)(r-k)}$. The first non-significant value of k will serve as our estimate of the rank of $\boldsymbol{\beta}$.

6.2 Selecting d_X and d_Y

One way to determine d_X and d_Y is by using a sequence of likelihood ratio tests based on the statistic $\Lambda_{d_X,d_Y} = 2(\widehat{L}_{\text{full}} - \widehat{L}_{d_X,d_Y})$, where $\widehat{L}_{\text{full}} = -(n/2)\log|\mathbf{S}_{\mathbf{C}}| - n(p+r)/2$ is the log likelihood for the standard model and $\widehat{L}_{d_X,d_Y} = -(n/2)\log|\widehat{\boldsymbol{\Sigma}}_{\mathbf{C}}| - (n/2)\text{trace}(\mathbf{S}_{\mathbf{C}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{C}}^{-1})$ is the maximum log likelihood for simultaneous envelope model, where $\widehat{\boldsymbol{\Sigma}}_{\mathbf{C}} = \mathbf{P}_{\widehat{\mathbf{L}}\oplus\widehat{\mathbf{R}}}\mathbf{S}_{\mathbf{C}}\mathbf{P}_{\widehat{\mathbf{L}}\oplus\widehat{\mathbf{R}}} + \mathbf{Q}_{\widehat{\mathbf{L}}\oplus\widehat{\mathbf{R}}}\mathbf{S}_{\mathbf{D}}\mathbf{Q}_{\widehat{\mathbf{L}}\oplus\widehat{\mathbf{R}}}$ is the MLE of the simultaneous envelope model. Under the null hypothesis, the test statistic Λ_{d_X,d_Y} is asymptotically distributed as a $\chi^2_{(p-d_X)(r-d_Y)}$ random variable. We start testing with $d_X = d_Y = d_0$ where the choice of d_0 can be guided by the Bura-Cook estimator. Then sequentially test $d_X = d_0,\ d_0+1, \ldots,$ until first non-significant value, and sequentially test $d_Y = d_0,\ d_0+1, \ldots,$ until first non-significant value.

Information criteria such as Akaike's information criterion (AIC) and Bayesian information criterion (BIC) could also be used to select the dimension d_X and d_Y, similar to Cook et al. (2010). For dimensions d_X and d_Y, AIC is $\text{A}(d_X,d_Y) = 2k_{d_X,d_Y} - 2\widehat{L}_{d_X,d_Y}$, where $k_{d_X,d_Y} = p(p+3)/2 + r(r+3)/2 + d_Xd_Y$ is the number of parameters in the model; and BIC is $\text{B}(d_X,d_Y) = k_{d_X,d_Y}\log(n) - 2\widehat{L}_{d_X,d_Y}$. We search (d_X,d_Y) from (d_0,d_0) to (p,r) and choose the pair that has the smallest AIC or BIC. See Section 7.3 for a simulation to illustrate determining the dimensions by likelihood ratio testing, AIC and BIC.

Our experience suggests that BIC is the most favorable criterion for determining dimensions when sample size is not too small. Since the true dimension always exist, BIC is consistent in the sense that the probability of selecting the correct dimension approaches 1. According to Shao (1997) and Yang (2005), BIC would perform better than AIC when the true model has a simple finite dimensional structure. However, if one is concerned more about the correctness of the envelope model itself over its simplicity, AIC is more favorable since it is more conservative on choosing the dimensions.

In practice, we use cross-validation to determine the optimal dimensions for prediction, although the best predictive dimensions can be different from the true envelope dimensions d_X and d_Y.

7 Simulations

In practice, we found the estimator from the 1D algorithm essentially as efficient as the final estimator by alternating because the alternating procedure usually converged after only a few iterations and produced only small changes. To emphasize the 1D algorithm, we use it for all the simulations and examples in this section. In Section 7.2, we use a simulation example to demonstrate that estimators from the 1D manifold algorithm may have the same behavior as maximum likelihood estimators.

In Section 7.1, we compared the simultaneous envelope estimator to many other methods including OLS, PLS, CCA, RRR, $\mathbf{X}$-envelope and $\mathbf{Y}$-envelope, where the dimensions for each method were chosen purely based on the simulated data with no prior knowledge. In the Supplement Section J, the true dimensions were used for each method and we also used two different criterion for prediction and estimation. Given the true dimension, the simultaneous envelope estimators always had the best performance in both prediction and estimation among all estimators and always performed substantially better than OLS.

7.1 Prediction with cross-validation

Prediction error of a multivariate linear model is a composite of the variability in $\widehat{\boldsymbol{\beta}}$ and the intrinsic variance in $\boldsymbol{\epsilon}$. As illustrated by Table 1, the simultaneous envelope method can substantially increase estimation accuracy in the regression coefficient matrix and

thus it may also increase prediction accuracy. We simulated such situations in the following examples where we generated data from joint normal distribution of $\mathbf{X}$ and $\mathbf{Y}$.

We simulated data from two envelope models (E1 and E2) and two reduced-rank regression models (R1 and R2) as follows. The dimensions were $p = 10$, $r = 15$ and varied d, d_X and d_Y. For the envelope models, we generated $\mathbf{L}$, $\mathbf{R}$ and $\boldsymbol{\eta}$ by filling in random numbers from the uniform(0,1) distribution. Then we orthogonalized $\mathbf{L}$ and $\mathbf{R}$ and obtained corresponding $\mathbf{L}_0$ and $\mathbf{R}_0$. For the RRR models, we simulated $\boldsymbol{\beta} = c\mathbf{A}\mathbf{B}^T$ where $c \in \mathbb{R}^1$, $\mathbf{A} \in \mathbb{R}^{p\times d}$, $\mathbf{B} \in \mathbb{R}^{r\times d}$ were filled with uniform$(0, 1)$ random numbers, then $\mathbf{A}$ and $\mathbf{B}$ were standardized to semi-orthogonal matrices. The covariance matrices $\boldsymbol{\Omega}$, $\boldsymbol{\Omega}_0$, $\boldsymbol{\Phi}$, $\boldsymbol{\Phi}_0$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_{\mathbf{X}}$ were generated as follows.

(E1) Isotropic envelope covariances. $\{\boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\Phi}, \boldsymbol{\Phi}_0\} = \{5\mathbf{I}_{d_X}, \mathbf{I}_{p-d_X}, 5\mathbf{I}_{d_Y}, \mathbf{I}_{r-d_Y}\}$.

(E2) Randomized envelope covariances. All $\boldsymbol{\Omega}$, $\boldsymbol{\Omega}_0$, $\boldsymbol{\Phi}$ and $\boldsymbol{\Phi}_0$ were randomly generated as $\mathbf{M}\mathbf{M}^T$ where $\mathbf{M}$ had the same size as the corresponding covariances and were filled with uniform$(0, 1)$ numbers.

(R1) Identity covariances. $\boldsymbol{\Sigma}_{\mathbf{X}} = 5\mathbf{I}_{10}$ and $\boldsymbol{\Sigma} = 5\mathbf{I}_{15}$.

(R2) Randomized covariances. Both $\boldsymbol{\Sigma}_{\mathbf{X}}$ and $\boldsymbol{\Sigma}$ were randomly generated as $\mathbf{M}\mathbf{M}^T$ in the similar way as in model (E2).

We simulated 200 data sets for each setting and used 100 for training and the others for testing. For each method, after obtaining the estimator $\widehat{\boldsymbol{\beta}}$ from one training data set, we used the testing data set to compute the following quantity, which is an estimate of the variance of $\widehat{\mathbf{Y}} - \mathbf{Y}_{\text{fit}} = \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^T \mathbf{X}_{\text{new}}$ for a new observation $\mathbf{X}_{\text{new}}$.

$$\boldsymbol{\alpha} = \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^T \widehat{\boldsymbol{\Sigma}}_{\mathbf{X}} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \in \mathbb{R}^{r\times r}, \tag{7.1}$$

where $\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}$ is the sample covariance matrix of $\mathbf{X}$ obtained from the testing set. Notice that $\text{cov}(\widehat{\mathbf{Y}} - \mathbf{Y}) = \text{cov}(\widehat{\mathbf{Y}} - \mathbf{Y}_{\text{fit}}) + \text{cov}(\boldsymbol{\epsilon})$ for all unbiased estimators, so we could make our comparison in the magnitudes of $\text{cov}(\widehat{\mathbf{Y}} - \mathbf{Y}_{\text{fit}})$. Therefore the Frobenius norm $||\boldsymbol{\alpha}||_F$ was used as a measure of overall prediction error. We also considered the Frobenius norm

$||\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}||_F$ as a measure of overall parameter estimation accuracy in the simulation studies in Supplement Section J.

We used five-fold cross-validation to find the best predictive dimensions for each of the methods. The results were summarized in Table 2. Our simulation results suggested that simultaneous envelopes with dimensions determined by cross-validation could be used in practice as a powerful prediction method. The simultaneous envelope methods had significantly better performance than all the other methods in most cases. For each of the other methods, it could perform as well as the simultaneous envelope in some cases but could have much worse performances in other cases.

In model (R2), where the true envelope dimensions were $(d_X, d_Y) = (p, r)$, the envelope estimators with its true dimensions will be equivalent to OLS. Nevertheless, the $\mathbf{X}$-envelope and the simultaneous envelope showed significant improvement over OLS in prediction by using the best predictive dimensions. This was probably due to the variance-bias tradeoff in finite samples. Model (R1) is both an envelope model and a RRR model with $d = d_X = d_Y$. It has the simplest structure, but the simultaneous envelope estimators still produced significant improvement over OLS. RRR outperformed OLS but was still not as efficient as the simultaneous envelope estimators. Model (E2) produced immaterial covariances $\boldsymbol{\Phi}_0$ and $\boldsymbol{\Omega}_0$ that had much larger eigenvalues than the material parts $\boldsymbol{\Phi}$ and $\boldsymbol{\Omega}$. The simultaneous envelope method efficiently discarded the immaterial information and performed drastically better than other methods in models (E2). Moreover, CCA aimed for the most correlated pairs of $\widehat{\mathbf{L}}^T\mathbf{X}$ and $\widehat{\mathbf{R}}^T\mathbf{Y}$ and mistakenly found the immaterial pairs in model (E2).

To further study how the relative magnitude of immaterial variation over material variation can affect the performance of envelope methods, we repeated the above simulations in model (E2) for varying $d = d_X = d_Y$ from 1 to 8 with the same sample size $N = 200$. The true envelope dimensions were used instead of using cross-validation. We then plotted the averaged prediction error $||\boldsymbol{\alpha}||_F$ over that of OLS in Figure 7.1. When the rank of $\boldsymbol{\beta}$ was smaller than 5, which is the half of its maximum possible rank, the immaterial variance $\boldsymbol{\Phi}_0$ contained the larger eigenvalues of $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$ hence the $\mathbf{Y}$-envelope and the simultaneous envelope estimators had drastic gains in efficiency over OLS estimators.

(d_X, d_Y, N) or (d, N)		OLS	PLS	CCA	RRR	$\mathbf{X}$-env.	$\mathbf{Y}$-env.	S. env.	S. E.$\leq$
E1	$(3, 3, 100)$	3.15	1.55	**1.14**	2.46	1.50	2.53	**0.63**	0.06
	$(3, 3, 400)$	0.70	0.46	**0.22**	2.23	0.46	0.58	**0.12**	0.01
	$(3, 3, 900)$	0.30	0.14	**0.08**	2.28	0.14	0.25	**0.05**	0.006
E2	$(2, 5, 400)$	0.82	0.40	2.17	0.28	**0.22**	0.47	**0.17**	0.04
	$(5, 3, 400)$	1.00	0.51	1.76	**0.36**	0.55	0.65	**0.27**	0.07
	$(7, 4, 400)$	1.01	0.82	1.75	0.68	0.75	**0.59**	**0.32**	0.08
R1	$(1, 200)$	1.68	**0.85**	1.01	1.05	0.97	1.53	**0.73**	0.03
	$(2, 200)$	1.68	0.98	1.49	1.89	**0.97**	1.52	**0.97**	0.04
	$(3, 200)$	1.68	1.35	2.00	2.38	**1.19**	1.52	**1.18**	0.05
R2	$(2, 200)$	2.63	1.23	8.94	**0.90**	1.08	2.63	**1.08**	0.11*
	$(3, 200)$	2.86	1.54	26.6	1.52	**1.08**	2.79	**1.08**	0.15*
	$(4, 200)$	2.96	2.67	16.4	1.97	**1.45**	2.96	**1.45**	0.13*

Table 2: Prediction performances measured by $||\boldsymbol{\alpha}||_F$, where * in the last column means the corresponding S.E. upper bound were computed without the S.E. of CCA. The best two methods for each setting were emphasized with boldface.

With increased rank of $\boldsymbol{\beta}$, the immaterial variance parts were reduced but the simultaneous envelope reduction can still gain significantly over the OLS estimator. Even for $d = 8$, the prediction error measurement $||\boldsymbol{\alpha}||_F$ of the simultaneous envelope estimator was only 70% of that of the OLS estimator.

7.2 Optimization with 1D manifold algorithm

The simulation results in Section 7.1 were all based on the 1D algorithm. The full Grassmann manifold optimization, which gives the MLE solution, would give indistinguishable performances for small dimensions but could have trouble converging in larger dimensions. To gain some insight on the computation cost of 1D algorithm versus the full Grassmann manifold optimization, we simulated 100 data sets from model (E2) with $N = 400$ and computed the averaged CPU time for estimating an envelope using the two methods. When $d_X = d_Y$ increased from 2 to 3, the ratio of the averaged time costs on full Grassmannian optimization over that of the 1D algorithm were increasing from 1.1 to 2.8 for estimating $\mathbf{X}$-envelopes and from 1.4 to 3.8 for estimating $\mathbf{Y}$-envelopes.

To examine how well the 1D manifold algorithm estimators can approximate the MLE, we studied probability plots of the asymptotic χ^2 likelihood ratio test statistics. Using model (E2), we simulated 100 dataset with sample size $N = 200$ with $p = r = 7$

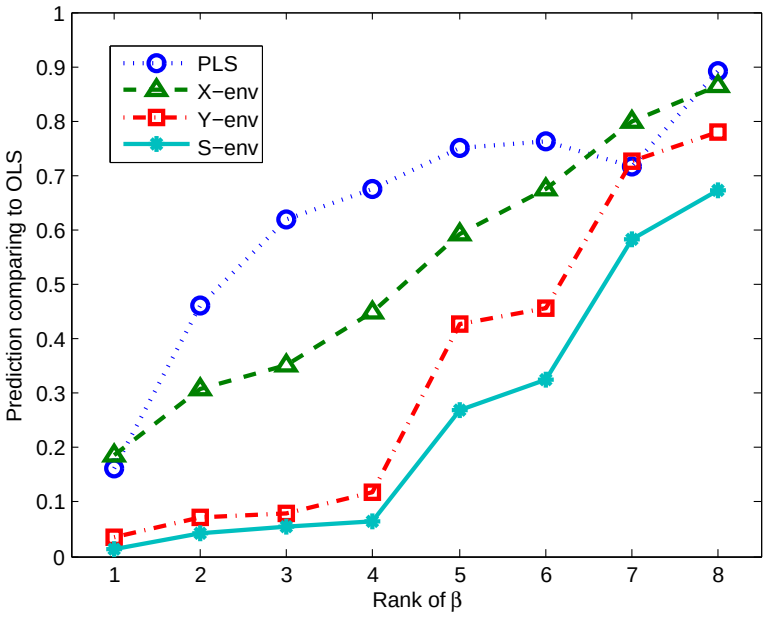


Figure 7.1: Prediction versus OLS for varying $d = d_X = d_Y$. Each point is the ratio of $||\boldsymbol{\alpha}||_F$ for a specified estimator and for OLS estimator. The lines for CCA and RRR (not shown in the plot) were always larger than or close to 1.

and $d_X = d_Y = 3$. Consider the following hypothesis test given d_X and d_Y,

$$H_0 : \mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\mathcal{L}) \oplus \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\mathcal{R}) = \text{span}(\mathbf{L} \oplus \mathbf{R});$$

$$H_a : \mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\mathcal{L}) \oplus \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\mathcal{R}) \neq \text{span}(\mathbf{L} \oplus \mathbf{R}).$$

The likelihood ratio test statistic for this hypothesis is $\boldsymbol{\Lambda} = -2\log(L(\mathbf{L}, \mathbf{R})) + 2\log(L(\widehat{\mathbf{L}}, \widehat{\mathbf{R}}))$, where $\widehat{\mathbf{L}}$ and $\widehat{\mathbf{R}}$ are MLEs, is asymptotically χ^2 distributed. The degrees of freedom is the sum of the two Grassmann manifolds dimensions: $(p - d_X)d_X + (r - d_Y)d_Y = 24$. We computed the likelihood ratio test statistics using the 1D algorithm (4.9) to estimate $\widehat{\mathbf{L}}$ and $\widehat{\mathbf{R}}$. Figure 7.2 summarized the calculated sample quantiles of the one hundred statistics and the expected quantiles of the χ^2_{24} distribution. All the points fall near a straight line, indicating that the estimators from the 1D algorithm behave as MLEs.

7.3 Determining the dimensions (d_X, d_Y)

We simulated 100 datasets for each of the following settings, and performed the chi-square test for the rank of $\boldsymbol{\beta}$ with level 0.05, the likelihood ratio test with level 0.01, AIC model selection and BIC model selection. We used model (E2) in Section 7.1 with $p = 6$

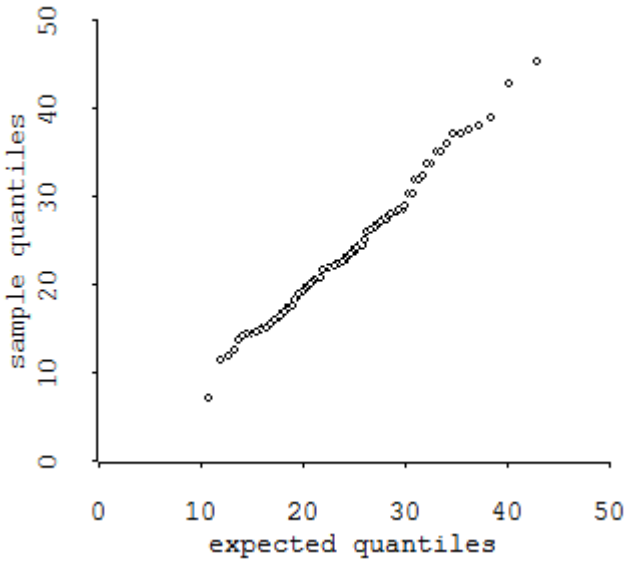


Figure 7.2: Probability plot based on the 1D algorithm

predictors and $r = 8$ responses in two settings: (I) $N = 100$ observations for each dataset with $\mathbf{X}$- and $\mathbf{Y}$-envelope dimensions $d_X = d_Y = 2$. (II) $N = 300$ observations for each dataset with $\mathbf{X}$- and $\mathbf{Y}$-envelope dimensions $d_X = 2$, $d_Y = 3$. Table 3 summarizes the simulation performances of various tests. Clearly, the χ^2 test quite often reveals the correct dimension of $\boldsymbol{\beta}$ and BIC has the best performance on determining the envelope dimensions.

We further studied the likelihood ratio test statistics obtained by the 1D algorithm and SIMPLS. We used $p = r = 7$, $d_X = d_Y = 3$ and simulated 100 dataset with $N = 300$ observations from model (E2) in Section 7.1. The likelihood ratio test was computed for the hypothesis $d_X = d_Y = 3$. The test statistic follows a χ^2-distribution with $pr - d_X d_Y = 40$ degrees of freedom as shown in Section 6.2. The probability plot of expected quantiles versus sample quantiles from the 100 datasets are given in Figure 7.3. Since the test statistic depends on our estimators of $\mathbf{L}$ and $\mathbf{R}$, the probability plot should follow a straight line if the estimators behave as MLEs. From this plot, we see again that the behavior of the 1D algorithm is as expected for MLEs. However, the SIMPLS algorithm produced noticeable curvature in its probability plot.

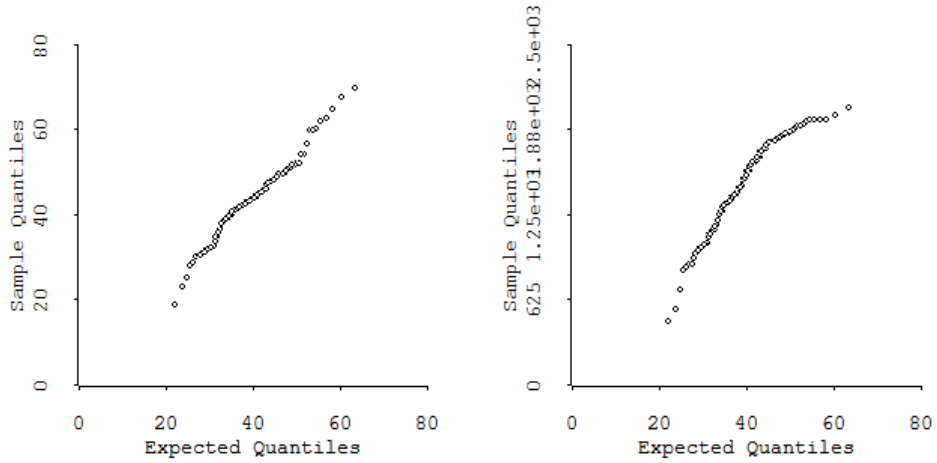


Figure 7.3: Chi-squared distribution probability plots. The 1D algorithm results the left plot, and the SIMPLS algorithm results the right plot. The reference distribution is the χ^2_{40}-distribution.

	correctness on	setting (I)	setting (II)
χ^2	d	89%	97%
LRT	d_X	98%	97%
	d_Y	100%	90%
	d_X and d_Y	98%	89%
AIC	d_X	75%	92%
	d_Y	67%	63%
	d_X and d_Y	56%	61%
BIC	d_X	98%	99%
	d_Y	100%	100%
	d_X and d_Y	98%	99%

Table 3: The numbers indicate how many times each true/estimated dimensions occurs among the 100 datasets.

8 Biscuit NIR spectroscopy data

This experiment involved varying the composition of biscuit dough. Two sets of dough pieces were measured, a calibration set and a prediction set. They were created and measured as two distinct sets, on separate occasions, and do not result from a random (or any other) split of a larger set. We have $r = 4$ response variables: fat, sucrose, flour and water percentages of biscuit dough. To keep the number of predictors less than the sample size of the training set, we used the spectral range 1200-1280nm with 4nm steps, which gave $p = 20$ predictors with $n_1 = 39$ training samples and $n_2 = 31$ testing samples. The prediction accuracy was based on the sum of squared error for predicting the given testing set.

The rank test of $\boldsymbol{\beta}$ indicated that $d = \text{rank}(\boldsymbol{\beta}) = 2$ at the 0.05 significant level. Simultaneous envelope dimension $d_Y = 2$ was selected by AIC, BIC and LRT; $d_X = 10, 3, 8$ was selected by AIC, BIC and LRT respectively. Based on the prediction error, AIC chose the best dimension, while BIC made a bad choice, which was probably due to the small sample size of the training set.

The sum of squared prediction errors (prediction SSE) on the testing sets are summarized in the Figure 8.1. When the number of components was less than 8, the envelope and SIMPLS estimators had similar performances and were both inferior to OLS because the dimension was too small to cover all the material information. Starting at 8 components, the envelope estimator performed much better than the SIMPLS and OLS estimators. It converged to the OLS estimator at 13 components and stayed the same thereafter. However, SIMPLS performed much worse than OLS unless the number of component was larger than 16. To aid visualization, prediction errors from CCA and RRR were not plotted because they were more than twice of that of OLS. We also considered predictions based on $\mathbf{X}$- or $\mathbf{Y}$-envelope reduction alone. The prediction errors for the $\mathbf{X}$-envelope estimator traced the errors from the simultaneous envelope estimator but were uniformly larger. The prediction errors for the $\mathbf{Y}$-envelope estimators were indistinguishable to that of the OLS estimator for $d_Y > 1$ and were larger than that for $d_Y = 1$.

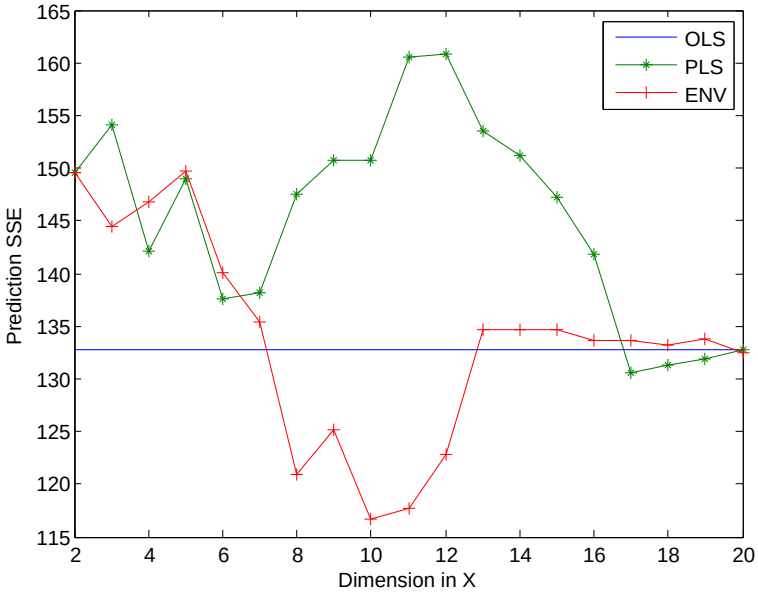


Figure 8.1: Prediction sum of square errors on the testing set. X-axis denote the numbers of components for PLS and simultaneous envelope $\mathbf{X}$-dimension, d_X, where the $\mathbf{Y}$-dimension of simultaneous envelope is fixed at $d_Y = 2$. To help visualization, the CCA performance is not included and the SSE for CCA are all greater than OLS no matter how many components to use.

9 Discussion

The partial envelope model (Su and Cook, 2011) is an extension of the envelope model in which some predictors are of special interest. Suppose $\mathbf{X}$ can be partitioned into $\mathbf{X}_1 \in \mathbb{R}^{p_1}$ and $\mathbf{X}_2 \in \mathbb{R}^{p_2}$, $p_1 + p_2 = p$, and the regression coefficient becomes $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T\right)^T$ correspondingly. The partial envelope model focus on estimating $\boldsymbol{\beta}_1$ instead of $\boldsymbol{\beta}$, and can be written as,

$$\mathbf{Y} = \boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\beta}_1^T(\mathbf{X}_1 - \boldsymbol{\mu}_{\mathbf{X}1}) + \boldsymbol{\beta}_2^T(\mathbf{X}_2 - \boldsymbol{\mu}_{\mathbf{X}2}) + \boldsymbol{\epsilon}, \tag{9.1}$$

where $\boldsymbol{\mu}_{\mathbf{X}1} = \mathrm{E}(\mathbf{X}_1)$ and $\boldsymbol{\mu}_{\mathbf{X}2} = \mathrm{E}(\mathbf{X}_2)$. The partial envelope in this model is $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\mathrm{span}(\boldsymbol{\beta}_1^T))$, which is contained in the envelope $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\mathcal{R}) = \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\mathrm{span}(\boldsymbol{\beta}^T))$. Since the partial envelope model focuses on a smaller subspace, there is potentially more efficiency gain than with the envelope model.

It is straightforward to extent partial envelope to simultaneous envelopes. Let $\boldsymbol{\Sigma}_{\mathbf{X}1} \in \mathbb{R}^{p_1 \times p_1}$ be the covariance matrix of $\mathbf{X}_1$, and let $\boldsymbol{\Sigma}_{\mathbf{X}2} \in \mathbb{R}^{p_2 \times p_2}$ be the covariance matrix of $\mathbf{X}_2$. We can then study the simultaneous partial envelope $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\mathrm{span}(\boldsymbol{\beta}_1^T)) \oplus \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}_1}}(\mathrm{span}(\boldsymbol{\beta}_1))$ to reduce $\mathbf{Y}$ and to reduce $\mathbf{X}_1$.

The simultaneous envelope model contains the envelope models in Cook et al. (2010) and Cook et al. (2013) as special cases. Extensions to the simultaneous envelope methods provide unified treatments for these methods. For instance, variable selection in multivariate linear regression has started to become an interesting and important research area. Multivariate linear methods that have well performances in classical problems were extended to high-dimensional situations. To deal with high-dimensional problems where $p, r > n$, Chun and Keles (2010) introduced sparse PLS and Chen and Huang (2012) proposed sparse RRR. We conjecture that an variable selection extension of the simultaneous envelope methodology will have nice applications in high-dimensional settings. Alternatively, we can first apply the SIMPLS algorithm, which is actually an envelope algorithm shown by Cook et al. (2013), to reduce the dimensionality and then apply the simultaneous envelope methods to gain efficiency in estimation.

Acknowledgment

We thank the Editor, the Associate Editor and three referees for their insightful comments and constructive suggestions that have greatly improved the presentation of the paper. Research for this article was supported in part by grant DMS-1007547 from the National Science Foundation.

Supplementary Materials

Proofs and Technical Details: Detailed proofs for all Lemmas and Propositions are provided in the online supplement to this article. (PDF file)

References

[1] ADRAGNI, K., COOK, R.D. AND WU, S. (2012), GrassmannOptim: An R Package for Grassmann Manifold Optimization, *Journal of Statistical Software*, **50**, 1–18.

[2] BACH, F.R. AND JORDAN, M.I. (2005), A probabilistic interpretation of canonical correlation analysis, *Technical Report*, **688**, 1–11.

[3] BURA, E. AND COOK, R.D. (2003), Rank estimation in reduced-rank regression, *Journal of Multivariate Analysis*, **87**, 159–176.

[4] CHUN, H. AND KELES, S. (2010), Sparse partial least squares regression for simultaneous dimension reduction and variable selection, *Journal of the Royal Statistical Society: Series B*, **72**, 3–25.

[5] CHEN, L. AND HUANG, J.Z. (2012), Sparse reduced-rank regression for simultaneous dimension reduction and variable selection, *Journal of the American Statistical Association*, **107**, 1533–1545.

[6] CONWAY, J. (1990). A course in functional analysis. Second edition. *Springer, New York*.

[7] COOK, R.D., HELLAND, I.S. AND SU, Z. (2013), Envelopes and partial least squares regression. *To appear in JRSS-B*.

[8] COOK, R.D., LI, B. AND CHIAROMONTE, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression (with discussion). *Statistica Sinica*, **20**,927–1010.

[9] EDELMAN, A., TOMAS, A.A. AND SMITH, S.T. (1998), The geometry of algorithms with orthogonality constraints, *SIAM Journal of Matrix Analysis and Applications*, **20**, 303–353.

[10] HENDERSON, H.V. AND SEARLE, S.R. (1979), Vec and Vech operators for matrices, with some uses in Jacobians and multivariate statistics. *Canadian Journal of Statistics*, **7**, 65–81.

[11] HOTELLING, H. (1936), Relations between two sets of variates. *Biometrika*. **28**, 321–377.

[12] IZENMAN, A. J. (1975), Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*. **5**, 248–264.

[13] JOLLIFFE, I. (2005), Principal Component Analysis. *Encyclopedia of Statistics in Behavioral Science.*

[14] JOLLIFFE, I. (1986), Principal component analysis, *Springer Verlag, New York.*

[15] DE JONG, S. (1993), SIMPLS: an alternative approach to partial least squares regression. *Chemometr. Intell. Lab. Syst.*,**18**, 251–26.

[16] LEHMANN, E.L. AND CASELLA G. (1998) Theory of Point Estimation, Second Edition *New York: Springer*.

[17] LI, B., WEN, S. AND ZHU, L.X. (2008) On a Projective Resampling Method for Dimension Reduction With Multivariate Responses, *Journal of the American Statistical Association*, **103**, 1177–1186.

[18] LI, K.C., ARAGON, Y., SHEDDEN, K. AND AGAN, C.T. (2003), Dimension reduction for multivariate response data, *Journal of the American Statistical Association*, **98**, 99–109.

[19] REINSEL, G. C. AND VELU, R. P. (1998), Multivariate reduced-rank regression: theory and applications. *Springer, New York*.

[20] SHAO, J. (1997), An asymptotic theory for linear model selection (with discussion). *Statistica sinica*, **7**, 221–264.

[21] SHAPIRO, A. (1986), Asymptotic theory of overparameterized structural models, *Journal of the American Statistical Association*, **81**, 142–149.

[22] SMALL, C.G., WANG, J. AND YANG, Z. (2000) Eliminating multiple root problems in estimation. *Statistical Science*, **15**, 313–341.

[23] SU, Z. AND COOK, R.D. (2011). Partial envelopes for efficient estimation in multivariate linear regression. *Biometrika*, **98**, 133–146.

[24] TYLER, D.E. (1981). Asymptotic inference for eigenvectors. *Annals of Statistics*, **9**, 725–736.

[25] WOLD, H. (1966), Estimation of Principal Components and Related Models by Iterative Least Squares. *New York: Academic Press.*

[26] YANG, Y. (2005), Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika*. **92**, 934–950.

Supplement: Proofs and Technical Details for "Simultaneous envelopes for multivariate linear regression"

R. Dennis Cook and Xin Zhang

The referred equations in the Supplement are labeled as (A1), (A2) and so on, whereas labels such as (1), (2), etc. refer to equations the main text. We also include here additional lemmas and corollaries (and when necessary, their proofs), sometimes within the proof of an assertion in the main text. In the following proofs, we use f and F as generic functions whose definitions change.

A Proof of Lemma 1 and Lemma 2

Lemma 1 and Lemma 2 follow directly from the discussions in Section 3 and the definition of envelopes.

B Proof of Proposition 2

Proof. Let $\mathbb{X} \in \mathbb{R}^{n\times p}$ and $\mathbb{Y} \in \mathbb{R}^{n\times r}$ be the centered data matrices from i.i.d. random vectors $\mathbf{X}$ and $\mathbf{Y}$. Then $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$ and $\widehat{\boldsymbol{\beta}}_{\mathbf{L},\mathbf{R}} = \mathbf{L}(\mathbf{L}^T\mathbb{X}^T\mathbb{X}\mathbf{L})^{-1}\mathbf{L}^T\mathbb{X}^T\mathbb{Y}\mathbf{R}\mathbf{R}^T$.

For the variance of $\mathrm{vec}(\widehat{\boldsymbol{\beta}}^T_{\mathrm{OLS}})$, we decompose it into two terms, conditioning on $\mathbb{X}$.

$$\begin{aligned}
\mathrm{var}(\mathrm{vec}(\widehat{\boldsymbol{\beta}}^T_{\mathrm{OLS}})) &= \mathrm{var}\left\{\mathrm{vec}(\mathbb{Y}^T\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1})\right\} \\
&= \mathrm{var}\left\{\left((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T \otimes \mathbf{I}_r\right)\mathrm{vec}(\mathbb{Y}^T)\right\} \\
&= \mathrm{var}\left\{\mathrm{E}\{\left((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T \otimes \mathbf{I}_r\right)\mathrm{vec}(\mathbb{Y}^T)|\mathbb{X}\}\right\} \\
&\quad +\mathrm{E}\left\{\mathrm{var}(\left((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T \otimes \mathbf{I}_r\right)\mathrm{vec}(\mathbb{Y}^T)|\mathbb{X})\right\} \\
&\equiv VE + EV.
\end{aligned}$$

The first term is

$$\begin{aligned} VE &= \operatorname{var}\left\{\mathrm{E}\{\left((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T \otimes \mathbf{I}_r\right)\operatorname{vec}(\mathbb{Y}^T)|\mathbb{X}\}\right\} \\ &= \operatorname{var}\left\{\left((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T \otimes \mathbf{I}_r\right)\mathrm{E}\{\operatorname{vec}(\mathbb{Y}^T)|\mathbb{X}\}\right\} \\ &= \operatorname{var}\left\{\left((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T \otimes \mathbf{I}_r\right)\operatorname{vec}(\boldsymbol{\beta}^T\mathbb{X}^T)\right\} \\ &= \operatorname{var}\left\{\operatorname{vec}(\boldsymbol{\beta}^T\mathbb{X}^T\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1})\right\} = 0. \end{aligned}$$

Then the variance is just the second term which is

$$\begin{aligned} \operatorname{var}(\operatorname{vec}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}^T)) &= EV = \mathrm{E}\left\{\operatorname{var}(\left((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T \otimes \mathbf{I}_r\right)\operatorname{vec}(\mathbb{Y}^T)|\mathbb{X})\right\} \\ &= \mathrm{E}\{\left((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T \otimes \mathbf{I}_r\right)\operatorname{var}(\operatorname{vec}(\mathbb{Y}^T)|\mathbb{X})\left((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T \otimes \mathbf{I}_r\right)^T\} \\ &= \mathrm{E}\{\left((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T \otimes \mathbf{I}_r\right)(\mathbf{I}_n \otimes \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}})\left((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T \otimes \mathbf{I}_r\right)^T\} \\ &= \mathrm{E}\{(\mathbb{X}^T\mathbb{X})^{-1} \otimes \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}\} \\ &= f_p^{-1}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}, \end{aligned}$$

where the last equality come from the fact that $\mathbb{X}^T\mathbb{X} \sim W_p(\boldsymbol{\Sigma}_{\mathbf{X}}, n-1)$. The degrees of freedom $n-1$ are greater than $p-1$. Then $(\mathbb{X}^T\mathbb{X})^{-1}$ follows an inverse Wishart distribution $W_p^{-1}(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}, n-1)$, and its mean equals $\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}/(n-p-2) = f_p^{-1}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}$.

Similarly for $\widehat{\boldsymbol{\beta}}_{\mathbf{L},\mathbf{R}}$, we have the following expressions by noticing that $\operatorname{vec}(\widehat{\boldsymbol{\beta}}_{\mathbf{L},\mathbf{R}}^T) = (\mathbf{L} \oplus \mathbf{R})\operatorname{vec}(\widehat{\boldsymbol{\eta}}^T)$ and $\widehat{\boldsymbol{\eta}}$ is the OLS estimator of $\mathbf{R}^T\mathbf{Y}$ on $\mathbf{L}^T\mathbf{X}$.

$$\begin{aligned} \operatorname{var}(\operatorname{vec}(\widehat{\boldsymbol{\beta}}_{\mathbf{L},\mathbf{R}}^T)) &= f_x^{-1}(\mathbf{L} \otimes \mathbf{R})(\boldsymbol{\Sigma}_{\mathbf{L}^T\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{\mathbf{R}^T\mathbf{Y}|\mathbf{L}^T\mathbf{X}})(\mathbf{L}^T \otimes \mathbf{R}^T) \\ &= f_x^{-1}(\mathbf{L} \otimes \mathbf{R})((\mathbf{L}^T\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{L})^{-1} \otimes (\mathbf{R}^T\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}\mathbf{R}))(\mathbf{L}^T \otimes \mathbf{R}^T) \\ &= f_x^{-1}\left(\mathbf{L}(\mathbf{L}^T\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{L})^{-1}\mathbf{L}^T\right) \otimes \left(\mathbf{R}\mathbf{R}^T\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}\mathbf{R}\mathbf{R}^T\right) \\ &= f_x^{-1}\left(\mathbf{L}\boldsymbol{\Omega}^{-1}\mathbf{L}^T\right) \otimes \left(\mathbf{P}_{\mathbf{R}}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}\mathbf{P}_{\mathbf{R}}\right). \end{aligned}$$

Next, by noticing that

$$\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} = \mathbf{L}(\mathbf{L}^T\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{L})^{-1}\mathbf{L}^T + \mathbf{L}_0(\mathbf{L}_0^T\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{L}_0)^{-1}\mathbf{L}_0^T, \tag{B.1}$$

$$\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}} = \mathbf{P}_{\mathbf{R}}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}\mathbf{P}_{\mathbf{R}} + \mathbf{P}_{\mathbf{R}_0}\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}\mathbf{P}_{\mathbf{R}_0}, \tag{B.2}$$

$$\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}} = \mathbf{R}\boldsymbol{\Phi}\mathbf{R}^T + \mathbf{R}_0\boldsymbol{\Phi}_0\mathbf{R}_0^T, \tag{B.3}$$

we have

$$\begin{aligned}\mathrm{var}(\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}^T)) &= f_p^{-1}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\otimes\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}\\ &= f_x f_p^{-1}\mathrm{var}(\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathbf{L},\mathbf{R}}^T)) + f_p^{-1}\mathbf{L}\boldsymbol{\Omega}^{-1}\mathbf{L}^T\otimes\mathbf{R}_0\boldsymbol{\Phi}_0\mathbf{R}_0^T\\ &\quad + f_p^{-1}\mathbf{L}_0\boldsymbol{\Omega}_0^{-1}\mathbf{L}_0^T\otimes\mathbf{R}\boldsymbol{\Phi}\mathbf{R}^T + f_p^{-1}\mathbf{L}_0\boldsymbol{\Omega}_0^{-1}\mathbf{L}_0^T\otimes\mathbf{R}_0\boldsymbol{\Phi}_0\mathbf{R}_0^T,\end{aligned}$$

where $\boldsymbol{\Omega}=\mathbf{L}^T\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{L}$, $\boldsymbol{\Omega}_0=\mathbf{L}_0^T\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{L}_0$, $\boldsymbol{\Phi}_0=\mathbf{R}_0^T\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}\mathbf{R}_0$ and $\boldsymbol{\Phi}=\mathbf{R}^T\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}\mathbf{R}$. Then the result follows directly. □

C Proof for Lemma 3

From Proposition 2.2 in Cook et al. (2010), we have $\mathcal{E}_{\mathbf{M}_i}(\mathcal{S}_i)=\sum_{j=1}^{q_i}\mathbf{P}_{ij}\mathcal{S}_i$, $i=1,2$, where $\mathbf{P}_{ij}$ is the projection onto the j-th eigenspace of $\mathbf{M}_i$. Since $\mathbf{M}=\mathbf{M}_1\oplus\mathbf{M}_2$, then the eigen-projections of $\mathbf{M}$ will be $\mathbf{P}_{1j}\oplus\mathbf{0}_{p_2}$, $j=1,\ldots,q_1$, and $\mathbf{0}_{p_1}\oplus\mathbf{P}_{2k}$, $k=1,\ldots,q_2$. Therefore, applying Proposition 2.2 of Cook et al. (2010) again, we have

$$\begin{aligned}\mathcal{E}_{\mathbf{M}}(\mathcal{S}) &= \left\{\sum_{j=1}^{q_1}[(\mathbf{P}_{1j}\oplus\mathbf{0})(\mathcal{S}_1\oplus\mathcal{S}_2)]\right\}+\left\{\sum_{k=1}^{q_2}[(\mathbf{0}\oplus\mathbf{P}_{2k})(\mathcal{S}_1\oplus\mathcal{S}_2)]\right\}\\ &= \{\mathcal{E}_{\mathbf{M}_1}(\mathcal{S}_1)\oplus\mathbf{0}\}+\{\mathbf{0}\oplus\mathcal{E}_{\mathbf{M}_2}(\mathcal{S}_2)\}\\ &= \mathcal{E}_{\mathbf{M}_1}(\mathcal{S}_1)\oplus\mathcal{E}_{\mathbf{M}_2}(\mathcal{S}_2).\end{aligned}$$

D Proof for Lemma 4

D.1 The log-det term: $\log|\boldsymbol{\Sigma}_{\mathbf{C}}|$

Apply the following orthogonal transformation to $\boldsymbol{\Sigma}_{\mathbf{C}}$,

$$\begin{aligned}\mathbf{O}^T\boldsymbol{\Sigma}_{\mathbf{C}}\mathbf{O} &= \begin{pmatrix}\begin{pmatrix}\mathbf{L} & \mathbf{L}_0\end{pmatrix}^T & \mathbf{0}\\ \mathbf{0} & \begin{pmatrix}\mathbf{R} & \mathbf{R}_0\end{pmatrix}^T\end{pmatrix}\begin{pmatrix}\boldsymbol{\Sigma}_{\mathbf{X}} & \boldsymbol{\Sigma}_{\mathbf{XY}}\\ \boldsymbol{\Sigma}_{\mathbf{XY}}^T & \boldsymbol{\Sigma}_{\mathbf{Y}}\end{pmatrix}\begin{pmatrix}\begin{pmatrix}\mathbf{L} & \mathbf{L}_0\end{pmatrix} & \mathbf{0}\\ \mathbf{0} & \begin{pmatrix}\mathbf{R} & \mathbf{R}_0\end{pmatrix}\end{pmatrix}\\ &= \begin{pmatrix}\begin{pmatrix}\boldsymbol{\Omega} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Omega}_0\end{pmatrix} & \begin{pmatrix}\boldsymbol{\Omega}\boldsymbol{\eta} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\\ \begin{pmatrix}\boldsymbol{\eta}^T\boldsymbol{\Omega} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix} & \begin{pmatrix}\boldsymbol{\Phi}+\boldsymbol{\eta}^T\boldsymbol{\Omega}\boldsymbol{\eta} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Phi}_0\end{pmatrix}\end{pmatrix}.\end{aligned}$$

We can further transform it into a block diagonal matrix as in the following general case:

$$\begin{pmatrix}\mathbf{I} & \mathbf{0}\\ -\mathbf{B}^T\mathbf{A}^{-1} & \mathbf{I}\end{pmatrix}\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{B}^T & \mathbf{C}\end{pmatrix}\begin{pmatrix}\mathbf{I} & -\mathbf{A}^{-1}\mathbf{B}\\ \mathbf{0} & \mathbf{I}\end{pmatrix}=\begin{pmatrix}\mathbf{A} & \mathbf{0}\\ \mathbf{0} & \mathbf{C}-\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}\end{pmatrix}. \quad \text{(D.1)}$$

By taking the transformation matrix $\mathbf{T} = \begin{pmatrix} \mathbf{I} & \mathbf{T}_{12} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}$ where $\mathbf{T}_{12} = \begin{pmatrix} -\boldsymbol{\eta} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$, we have the block diagonal matrix

$$\mathbf{T}^T\mathbf{O}^T\boldsymbol{\Sigma}_{\mathbf{C}}\mathbf{O}\mathbf{T} = \mathrm{diag}\{\boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\Phi}, \boldsymbol{\Phi}_0\} \equiv \boldsymbol{\Sigma}_{\mathbf{diag}}. \tag{D.2}$$

Also, because of the determinants of $\mathbf{T}$ and $\mathbf{O}$ are both 1, $|\boldsymbol{\Sigma}_{\mathbf{C}}| = |\boldsymbol{\Sigma}_{\mathbf{diag}}|$, and

$$\log|\boldsymbol{\Sigma}_{\mathbf{C}}| = \log|\boldsymbol{\Omega}| + \log|\boldsymbol{\Omega}_0| + \log|\boldsymbol{\Phi}| + \log|\boldsymbol{\Phi}_0|. \tag{D.3}$$

D.2 The trace term: $\mathrm{trace}(\mathbf{S}_{\mathbf{C}}\boldsymbol{\Sigma}_{\mathbf{C}}^{-1})$

$$\mathrm{trace}(\mathbf{S}_{\mathbf{C}}\boldsymbol{\Sigma}_{\mathbf{C}}^{-1}) = \mathrm{trace}(\mathbf{S}_{\mathbf{C}}\mathbf{O}\mathbf{T}\boldsymbol{\Sigma}_{\mathbf{diag}}^{-1}\mathbf{T}^T\mathbf{O}^T) = \mathrm{trace}(\mathbf{T}^T\mathbf{O}^T\mathbf{S}_{\mathbf{C}}\mathbf{O}\mathbf{T}\boldsymbol{\Sigma}_{\mathbf{diag}}^{-1}).$$

Write $\mathbf{O} \equiv \begin{pmatrix} \mathbf{O}_g & 0 \\ 0 & \mathbf{O}_h \end{pmatrix}$, where $\mathbf{O}_g = (\mathbf{L}, \mathbf{L}_0)$ and $\mathbf{O}_h = (\mathbf{R}, \mathbf{R}_0)$. Next, we compute the matrix

$$\mathbf{T}^T\mathbf{O}^T\mathbf{S}_{\mathbf{C}}\mathbf{O}\mathbf{T} = \mathbf{T}^T \begin{pmatrix} \mathbf{O}_g^T\mathbf{S}_{\mathbf{X}}\mathbf{O}_g & \mathbf{O}_g^T\mathbf{S}_{\mathbf{XY}}\mathbf{O}_h \\ \mathbf{O}_h^T\mathbf{S}_{\mathbf{XY}}^T\mathbf{O}_g & \mathbf{O}_h^T\mathbf{S}_{\mathbf{Y}}\mathbf{O}_h \end{pmatrix} \mathbf{T} = \begin{pmatrix} \mathbf{O}_g^T\mathbf{S}_{\mathbf{X}}\mathbf{O}_g & * \\ * & \mathbf{M}_1 \end{pmatrix}.$$

Since we only need the trace of $\mathbf{T}^T\mathbf{O}^T\mathbf{S}_{\mathbf{C}}\mathbf{O}\mathbf{T}\boldsymbol{\Sigma}_{\mathbf{diag}}^{-1}$, and $\boldsymbol{\Sigma}_{\mathbf{diag}}$ is block diagonal, so the off-diagonal blocks of $\mathbf{T}^T\mathbf{O}^T\mathbf{S}_{\mathbf{C}}\mathbf{O}\mathbf{T}$ are not needed. And the matrix $\mathbf{M}_1$ can be computed to be

$$\mathbf{M}_1 = \mathbf{O}_h^T\mathbf{S}_{\mathbf{Y}}\mathbf{O}_h + (\mathbf{O}_g\mathbf{T}_{12})^T\mathbf{S}_{\mathbf{X}}(\mathbf{O}_g\mathbf{T}_{12}) + (\mathbf{O}_g\mathbf{T}_{12})^T\mathbf{S}_{\mathbf{XY}}\mathbf{O}_h + \left[(\mathbf{O}_g\mathbf{T}_{12})^T\mathbf{S}_{\mathbf{XY}}\mathbf{O}_h\right]^T.$$

Therefore,

$$\begin{aligned} \mathrm{trace}(\mathbf{S}_{\mathbf{C}}\boldsymbol{\Sigma}_{\mathbf{C}}^{-1}) &= \mathrm{trace}(\mathbf{T}^T\mathbf{O}^T\mathbf{S}_{\mathbf{C}}\mathbf{O}\mathbf{T}\boldsymbol{\Sigma}_{\mathbf{diag}}^{-1}) \\ &= \mathrm{trace}(\mathbf{O}_g^T\mathbf{S}_{\mathbf{X}}\mathbf{O}_g\mathrm{diag}(\boldsymbol{\Omega}^{-1}, \boldsymbol{\Omega}_0^{-1})) + \mathrm{trace}(\mathbf{M}_1\mathrm{diag}(\boldsymbol{\Phi}^{-1}, \boldsymbol{\Phi}_0^{-1})) \\ &\equiv \mathrm{trace}_1 + \mathrm{trace}_2. \end{aligned}$$

The first trace term,

$$\begin{aligned} \mathrm{trace}_1 &= \mathrm{trace}(\mathbf{O}_g^T\mathbf{S}_{\mathbf{X}}\mathbf{O}_g\mathrm{diag}(\boldsymbol{\Omega}^{-1}, \boldsymbol{\Omega}_0^{-1})) \\ &= \mathrm{trace}\left\{ \begin{pmatrix} \mathbf{L}^T \\ \mathbf{L}_0^T \end{pmatrix} \mathbf{S}_{\mathbf{X}} \begin{pmatrix} \mathbf{L} & \mathbf{L}_0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\Omega}^{-1} & 0 \\ 0 & \boldsymbol{\Omega}_0^{-1} \end{pmatrix} \right\} \\ &= \mathrm{trace}\{\mathbf{L}^T\mathbf{S}_{\mathbf{X}}\mathbf{L}\boldsymbol{\Omega}^{-1}\} + \mathrm{trace}\{\mathbf{L}_0^T\mathbf{S}_{\mathbf{X}}\mathbf{L}_0\boldsymbol{\Omega}_0^{-1}\} = \mathrm{trace}\{\mathbf{S}_{\mathbf{X}}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\}. \end{aligned}$$

Similarly for the second trace term,

$$\begin{aligned}\text{trace}_2 &= \text{trace}(\mathbf{M}_1\text{diag}(\mathbf{\Phi}^{-1}, \mathbf{\Phi}_0^{-1})) \\ &= \text{trace}\{\mathbf{R}^T\mathbf{S_Y}\mathbf{R}\mathbf{\Phi}^{-1}\} + \text{trace}\{\mathbf{R}_0^T\mathbf{S_Y}\mathbf{R}_0\mathbf{\Phi}_0^{-1}\} \\ &\quad +\text{trace}\{(\mathbf{L}\boldsymbol{\eta})^T\mathbf{S_X}(\mathbf{L}\boldsymbol{\eta})\mathbf{\Phi}^{-1}\} - 2 \times \text{trace}\{(\mathbf{L}\boldsymbol{\eta})^T\mathbf{S_{XY}}\mathbf{R}\mathbf{\Phi}^{-1}\}.\end{aligned}$$

D.3 Partial derivatives of the objective function

The objective function $\mathrm{F}(\mathbf{\Sigma_C})$ now becomes an objective function of all the parameters

$$\begin{aligned}\mathrm{F}(\mathbf{L}, \mathbf{R}, \boldsymbol{\eta}, \mathbf{\Phi}, \mathbf{\Phi}_0, \mathbf{\Omega}, \mathbf{\Omega}_0) &= \log|\mathbf{\Omega}| + \log|\mathbf{\Omega}_0| + \log|\mathbf{\Phi}| + \log|\mathbf{\Phi}_0| \\ &\quad +\text{trace}\{\mathbf{L}^T\mathbf{S_X}\mathbf{L}\mathbf{\Omega}^{-1}\} + \text{trace}\{\mathbf{L}_0^T\mathbf{S_X}\mathbf{L}_0\mathbf{\Omega}_0^{-1}\} \\ &\quad +\text{trace}\{\mathbf{R}^T\mathbf{S_Y}\mathbf{R}\mathbf{\Phi}^{-1}\} + \text{trace}\{\mathbf{R}_0^T\mathbf{S_Y}\mathbf{R}_0\mathbf{\Phi}_0^{-1}\} \\ &\quad +\text{trace}\{(\mathbf{L}\boldsymbol{\eta})^T\mathbf{S_X}(\mathbf{L}\boldsymbol{\eta})\mathbf{\Phi}^{-1}\} - 2\text{trace}\{(\mathbf{L}\boldsymbol{\eta})^T\mathbf{S_{XY}}\mathbf{R}\mathbf{\Phi}^{-1}\}.\end{aligned}$$

We will apply the following result repeatedly in this section.

$$\arg\min_{\mathbf{A}}\{\log|\mathbf{A}| + \text{trace}(\mathbf{A}^{-1}\mathbf{B})\} = \mathbf{B}, \tag{D.4}$$

where $\mathbf{B}$ is positive definite symmetric matrix and the minimization is over all positive definite symmetric matrices.

D.3.1 Minimization over $\mathbf{\Omega}$, $\mathbf{\Omega}_0$, $\mathbf{\Phi}$ and $\mathbf{\Phi}_0$

Applying (D.4), we have

$$\widehat{\mathbf{\Omega}} = \widehat{\mathbf{L}}^T\mathbf{S_X}\widehat{\mathbf{L}}; \tag{D.5}$$

$$\widehat{\mathbf{\Omega}}_0 = \widehat{\mathbf{L}}_0^T\mathbf{S_X}\widehat{\mathbf{L}}_0; \tag{D.6}$$

$$\widehat{\mathbf{\Phi}}_0 = \widehat{\mathbf{R}}_0^T\mathbf{S_Y}\widehat{\mathbf{R}}_0. \tag{D.7}$$

After we obtain the minimizer of $\widehat{\boldsymbol{\eta}}$ and substitute it into the objective function, we can also use (D.4) to get the following.

$$\widehat{\mathbf{\Phi}} = \widehat{\mathbf{R}}^T\left(\mathbf{S_Y} - \mathbf{S}_{\mathbf{XY}}^T\widehat{\mathbf{L}}\left(\widehat{\mathbf{L}}^T\mathbf{S_X}\widehat{\mathbf{L}}\right)^{-1}\widehat{\mathbf{L}}^T\mathbf{S_{XY}}\right)\widehat{\mathbf{R}}. \tag{D.8}$$

D.3.2 Partial derivative w.r.t. $\boldsymbol{\eta}$

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\eta}}\mathrm{F}(\mathbf{L},\mathbf{R},\boldsymbol{\eta},\boldsymbol{\Phi},\boldsymbol{\Phi}_0,\boldsymbol{\Omega},\boldsymbol{\Omega}_0) &= \frac{\partial}{\partial \boldsymbol{\eta}}\left(\mathrm{trace}\{(\mathbf{L}\boldsymbol{\eta})^T\mathbf{S}_{\mathbf{X}}(\mathbf{L}\boldsymbol{\eta})\boldsymbol{\Phi}^{-1}\}\right)\\ &\quad -2\frac{\partial}{\partial \boldsymbol{\eta}}\left(\mathrm{trace}\{(\mathbf{L}\boldsymbol{\eta})^T\mathbf{S}_{\mathbf{XY}}\mathbf{R}\boldsymbol{\Phi}^{-1}\}\right)\\ &= 2\mathbf{L}^T\mathbf{S}_{\mathbf{X}}\mathbf{L}\boldsymbol{\eta}\boldsymbol{\Phi}^{-1}-2\mathbf{L}^T\mathbf{S}_{\mathbf{XY}}\mathbf{R}\boldsymbol{\Phi}^{-1}.\end{aligned}$$

Set it to 0, we have

$$\widehat{\boldsymbol{\eta}} = (\widehat{\mathbf{L}}^T\mathbf{S}_{\mathbf{X}}\widehat{\mathbf{L}})^{-1}(\widehat{\mathbf{L}}^T\mathbf{S}_{\mathbf{XY}}\widehat{\mathbf{R}}). \tag{D.9}$$

Substituting (D.5), (D.6), (D.7) and (D.9) into the full objective function, we have the following partially optimized objective function

$$\begin{aligned}\mathrm{F}(\mathbf{L},\mathbf{R},\boldsymbol{\Phi}) &= \log|\mathbf{L}^T\mathbf{S}_{\mathbf{X}}\mathbf{L}|+\log|\mathbf{L}_0^T\mathbf{S}_{\mathbf{X}}\mathbf{L}_0|+\log|\boldsymbol{\Phi}|+\log|\mathbf{R}_0^T\mathbf{S}_{\mathbf{Y}}\mathbf{R}_0|\\ &\quad +\mathrm{trace}\{\mathbf{I}_l\}+\mathrm{trace}\{\mathbf{I}_{p-l}\}+\mathrm{trace}\{\mathbf{R}^T\mathbf{S}_{\mathbf{Y}}\mathbf{R}\boldsymbol{\Phi}^{-1}\}+\mathrm{trace}\{\mathbf{I}_{r-m}\}\\ &\quad +\mathrm{trace}\{(\mathbf{L}^T\mathbf{S}_{\mathbf{XY}}\mathbf{R})^T(\mathbf{L}^T\mathbf{S}_{\mathbf{X}}\mathbf{L})^{-1}(\mathbf{L}^T\mathbf{S}_{\mathbf{XY}}\mathbf{R})\boldsymbol{\Phi}^{-1}\}\\ &\quad -2\times\mathrm{trace}\{(\mathbf{L}^T\mathbf{S}_{\mathbf{XY}}\mathbf{R})^T(\mathbf{L}^T\mathbf{S}_{\mathbf{X}}\mathbf{L})^{-1}(\mathbf{L}^T\mathbf{S}_{\mathbf{XY}}\mathbf{R})\boldsymbol{\Phi}^{-1}\}.\end{aligned}$$

If we ignore the terms that do not change with $\boldsymbol{\Phi}$, the rest part of the objective function becomes

$$\mathrm{F}(\boldsymbol{\Phi}) = \log|\boldsymbol{\Phi}|+\mathrm{trace}\{\mathbf{R}^T\mathbf{S}_{\mathbf{Y}}\mathbf{R}\boldsymbol{\Phi}^{-1}\}-\mathrm{trace}\{(\mathbf{L}^T\mathbf{S}_{\mathbf{XY}}\mathbf{R})^T(\mathbf{L}^T\mathbf{S}_{\mathbf{X}}\mathbf{L})^{-1}(\mathbf{L}^T\mathbf{S}_{\mathbf{XY}}\mathbf{R})\boldsymbol{\Phi}^{-1}\},$$

which leads us to (D.8)

D.4 MLE

The MLE for $\widehat{\mathbf{L}}$ and $\widehat{\mathbf{R}}$ are obtained by minimizing over the Grassmann manifolds $\mathcal{G}_{(d_X,p)}$ and $\mathcal{G}_{(d_Y,r)}$, $(\widehat{\mathbf{L}},\widehat{\mathbf{R}}) = \arg\min_{(\mathbf{L},\mathbf{R})}\{\mathrm{F}(\mathbf{L},\mathbf{R})\}$,where the objective function $\mathrm{F}(\mathbf{L},\mathbf{R})$ is

$$\begin{aligned}\mathrm{F}(\mathbf{L},\mathbf{R}) &= \log|\mathbf{L}^T\mathbf{S}_{\mathbf{X}}\mathbf{L}|+\log|\mathbf{L}_0^T\mathbf{S}_{\mathbf{X}}\mathbf{L}_0|+\log|\mathbf{R}_0^T\mathbf{S}_{\mathbf{Y}}\mathbf{R}_0|\\ &\quad +\log|\mathbf{R}^T\left(\mathbf{S}_{\mathbf{Y}}-\mathbf{S}_{\mathbf{XY}}^T\mathbf{L}(\mathbf{L}^T\mathbf{S}_{\mathbf{X}}\mathbf{L})^{-1}\mathbf{L}^T\mathbf{S}_{\mathbf{XY}}\right)\mathbf{R}|\\ &= \log\left|\begin{array}{cc}\mathbf{L}^T\mathbf{S}_{\mathbf{X}}\mathbf{L} & \mathbf{L}^T\mathbf{S}_{\mathbf{XY}}\mathbf{R}\\ \mathbf{R}^T\mathbf{S}_{\mathbf{XY}}^T\mathbf{L} & \mathbf{R}^T\mathbf{S}_{\mathbf{Y}}\mathbf{R}\end{array}\right|+\log|\mathbf{L}_0^T\mathbf{S}_{\mathbf{X}}\mathbf{L}_0|+\log|\mathbf{R}_0^T\mathbf{S}_{\mathbf{Y}}\mathbf{R}_0|\\ &= \log\left|\left(\mathbf{L}^T\oplus\mathbf{R}^T\right)\begin{pmatrix}\mathbf{S}_{\mathbf{X}} & \mathbf{S}_{\mathbf{XY}}\\ \mathbf{S}_{\mathbf{XY}}^T & \mathbf{S}_{\mathbf{Y}}\end{pmatrix}(\mathbf{L}\oplus\mathbf{R})\right|\\ &\quad +\log\left|\left(\mathbf{L}^T\oplus\mathbf{R}^T\right)\left(\mathbf{S}_{\mathbf{X}}^{-1}\oplus\mathbf{S}_{\mathbf{Y}}^{-1}\right)(\mathbf{L}\oplus\mathbf{R})\right| \equiv \mathrm{F}(\mathbf{L}\oplus\mathbf{R}).\end{aligned}$$

Then the MLEs for $\boldsymbol{\eta},\boldsymbol{\Omega},\boldsymbol{\Omega}_0,\boldsymbol{\Phi},\boldsymbol{\Phi}_0$ are as given in the lemma.

E Proof for Lemma 5

From our set-up, we know that $\mathbf{B}_0^T\mathbf{M}\mathbf{B}_0 > 0$ thus $\mathcal{E}_{\mathbf{B}_0^T\mathbf{M}\mathbf{B}_0}(\mathbf{B}_0^T\mathcal{S})$ exists. Let $\mathbf{\Gamma}$ be a basis of $\mathcal{E}_{\mathbf{M}}(\mathcal{S})$, and $(\mathbf{\Gamma}, \mathbf{\Gamma}_0)$ be a orthogonal basis of $\mathbb{R}^p$, then $\mathbf{M} = \mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T$ and $\mathcal{S} \subseteq \text{span}(\mathbf{\Gamma})$ for some $\mathbf{\Omega} > 0$ and $\mathbf{\Omega}_0 > 0$. Therefore,

$$\mathbf{B}_0^T\mathbf{M}\mathbf{B}_0 = (\mathbf{B}_0^T\mathbf{\Gamma})\mathbf{\Omega}(\mathbf{B}_0^T\mathbf{\Gamma})^T + (\mathbf{B}_0^T\mathbf{\Gamma}_0)\mathbf{\Omega}_0(\mathbf{B}_0^T\mathbf{\Gamma}_0)^T \tag{E.1}$$

$$\mathbf{B}_0^T\mathcal{S} \subseteq \text{span}(\mathbf{B}_0^T\mathbf{\Gamma}), \tag{E.2}$$

where $\text{span}(\mathbf{B}_0^T\mathbf{\Gamma})$ is the orthogonal compliment of $\text{span}(\mathbf{B}_0^T\mathbf{\Gamma}_0)$ in $\mathbb{R}^{p-q}$ since $\text{span}(\mathbf{B}) \subseteq \text{span}(\mathbf{\Gamma})$. Then we see

$$\mathbf{B}_0^T\mathbf{M}\mathbf{B}_0 = \mathbf{P}_{\mathbf{B}_0^T\mathbf{\Gamma}}\mathbf{B}_0^T\mathbf{M}\mathbf{B}_0\mathbf{P}_{\mathbf{B}_0^T\mathbf{\Gamma}} + \mathbf{Q}_{\mathbf{B}_0^T\mathbf{\Gamma}}\mathbf{B}_0^T\mathbf{M}\mathbf{B}_0\mathbf{Q}_{\mathbf{B}_0^T\mathbf{\Gamma}},$$

which means $\text{span}(\mathbf{B}_0^T\mathbf{\Gamma})$ is a reducing subspace of $\mathbf{B}_0^T\mathbf{M}\mathbf{B}_0$ which also contains $\mathbf{B}_0^T\mathcal{S}$ by (E.2). By definition, we know $\mathcal{E}_{\mathbf{B}_0^T\mathbf{M}\mathbf{B}_0}(\mathbf{B}_0^T\mathcal{S})$ is the smallest reducing subspace of $\mathbf{B}_0^T\mathbf{M}\mathbf{B}_0$ that contains $\mathbf{B}_0^T\mathcal{S}$. Hence, $\mathcal{E}_{\mathbf{B}_0^T\mathbf{M}\mathbf{B}_0}(\mathbf{B}_0^T\mathcal{S}) \subseteq \text{span}(\mathbf{B}_0^T\mathbf{\Gamma})$. Therefore, $\mathbf{v} \in \mathcal{E}_{\mathbf{B}_0^T\mathbf{M}\mathbf{B}_0}(\mathbf{B}_0^T\mathcal{S})$ implies $\mathbf{B}_0\mathbf{v} \in \mathbf{B}_0\text{span}(\mathbf{B}_0^T\mathbf{\Gamma}) \subseteq \text{span}(\mathbf{\Gamma}) = \mathcal{E}_{\mathbf{M}}(\mathcal{S})$.

F Proof for Proposition 3

Let $\mathbf{L} \in \mathbb{R}^{p\times d_X}$ be a semi-orthogonal basis of $\mathcal{E}_{\mathbf{\Sigma}_\mathbf{X}}(\mathcal{L})$, then we have $\mathbf{\Sigma}_\mathbf{X} = \mathbf{L}\mathbf{\Omega}\mathbf{L}^T + \mathbf{L}_0\mathbf{\Omega}_0\mathbf{L}_0$, $\mathbf{\Sigma}_{\mathbf{XY}}\mathbf{\Sigma}_\mathbf{Y}^{-1/2} = \mathbf{L}\boldsymbol{\xi}$ and $\mathbf{\Sigma}_{\mathbf{X}|\mathbf{Y}} = \mathbf{L}(\mathbf{\Omega} - \boldsymbol{\xi}\boldsymbol{\xi}^T)\mathbf{L}^T + \mathbf{L}_0\mathbf{\Omega}_0\mathbf{L}_0$, where $\boldsymbol{\xi}$ is $d_X \times r$ matrix with rank d. We also assume that $\mathbf{\Omega} > 0$, $\mathbf{\Omega}_0 > 0$ and $\mathbf{\Omega} - \boldsymbol{\xi}\boldsymbol{\xi}^T > 0$. In this section, we first prove that the population 1D algorithm is estimating the envelope, then we show the $\sqrt{n}$-consistency of the sample algorithm as stated in Proposition 3.

F.1 The first direction ℓ_1 in the population

We first solve the optimization for the first direction $\ell_1 = \arg\min_{\ell\in\mathbb{R}^p} f_0(\ell)$ subject to $\ell^T\ell = 1$, where $f_0(\ell) = \log|\ell^T\mathbf{\Sigma}_{\mathbf{X}|\mathbf{Y}}\ell| + \log|\ell^T\mathbf{\Sigma}_\mathbf{X}^{-1}\ell|$. We want to show that $\ell_1 \in \text{span}(\mathbf{L})$.

Let $\ell = \mathbf{L}\mathbf{h} + \mathbf{L}_0\mathbf{h}_0$ for some $\mathbf{h} \in \mathbb{R}^{d_X}$ and $\mathbf{h}_0 \in \mathbb{R}^{(p-d_X)}$. Consider the optimization problem as unconstrained problem, $\ell_1 = \arg\min_{\ell\in\mathbb{R}^p}\left\{f_0(\ell) - 2\log|\ell^T\ell|\right\}$. Then we

will have the same solution as the original problem up to an arbitrary scaling constant. Next, we plug-in these expressions for ℓ, $\mathbf{\Sigma_X}$ and $\mathbf{\Sigma_{X|Y}}$,

$$\begin{aligned} f_0(\ell) - 2\log|\ell^T\ell| &= \log\{\mathbf{h}^T(\mathbf{\Omega} - \boldsymbol{\xi}\boldsymbol{\xi}^T)\mathbf{h} + \mathbf{h}_0^T\mathbf{\Omega}_0\mathbf{h}_0\} + \log\{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h} + \mathbf{h}_0^T\mathbf{\Omega}_0^{-1}\mathbf{h}_0\} \\ &\quad -2\log\{\mathbf{h}^T\mathbf{h} + \mathbf{h}_0^T\mathbf{h}_0\} \\ &\equiv f(\mathbf{h}, \mathbf{h}_0) \end{aligned}$$

Taking partial derivative with respect to $\mathbf{h}_0$ and $\mathbf{h}$, we have

$$\begin{aligned} \frac{\partial}{\partial \mathbf{h}_0} f(\mathbf{h}, \mathbf{h}_0) &= \frac{2\mathbf{\Omega}_0\mathbf{h}_0}{\mathbf{h}^T(\mathbf{\Omega} - \boldsymbol{\xi}\boldsymbol{\xi}^T)\mathbf{h} + \mathbf{h}_0^T\mathbf{\Omega}_0\mathbf{h}_0} + \frac{2\mathbf{\Omega}_0^{-1}\mathbf{h}_0}{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h} + \mathbf{h}_0^T\mathbf{\Omega}_0^{-1}\mathbf{h}_0} - \frac{4\mathbf{h}_0}{\mathbf{h}^T\mathbf{h} + \mathbf{h}_0^T\mathbf{h}_0} \\ \frac{\partial}{\partial \mathbf{h}} f(\mathbf{h}, \mathbf{h}_0) &= \frac{2(\mathbf{\Omega} - \boldsymbol{\xi}\boldsymbol{\xi}^T)\mathbf{h}}{\mathbf{h}^T(\mathbf{\Omega} - \boldsymbol{\xi}\boldsymbol{\xi}^T)\mathbf{h} + \mathbf{h}_0^T\mathbf{\Omega}_0\mathbf{h}_0} + \frac{2\mathbf{\Omega}^{-1}\mathbf{h}}{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h} + \mathbf{h}_0^T\mathbf{\Omega}_0^{-1}\mathbf{h}_0} - \frac{4\mathbf{h}}{\mathbf{h}^T\mathbf{h} + \mathbf{h}_0^T\mathbf{h}_0} \end{aligned}$$

To get local minimums we need to set $\frac{\partial}{\partial \mathbf{h}_0} f(\mathbf{h}, \mathbf{h}_0) = 0$ and $\frac{\partial}{\partial \mathbf{h}} f(\mathbf{h}, \mathbf{h}_0) = 0$ which become the following.

$$\begin{aligned} \left\{\frac{2\mathbf{\Omega}_0}{\mathbf{h}^T(\mathbf{\Omega} - \boldsymbol{\xi}\boldsymbol{\xi}^T)\mathbf{h} + \mathbf{h}_0^T\mathbf{\Omega}_0\mathbf{h}_0} + \frac{2\mathbf{\Omega}_0^{-1}}{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h} + \mathbf{h}_0^T\mathbf{\Omega}_0^{-1}\mathbf{h}_0}\right\}\mathbf{h}_0 &= \left\{\frac{4}{\mathbf{h}^T\mathbf{h} + \mathbf{h}_0^T\mathbf{h}_0}\right\}\mathbf{h}_0 \\ \left\{\frac{2(\mathbf{\Omega} - \boldsymbol{\xi}\boldsymbol{\xi}^T)}{\mathbf{h}^T(\mathbf{\Omega} - \boldsymbol{\xi}\boldsymbol{\xi}^T)\mathbf{h} + \mathbf{h}_0^T\mathbf{\Omega}_0\mathbf{h}_0} + \frac{2\mathbf{\Omega}^{-1}}{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h} + \mathbf{h}_0^T\mathbf{\Omega}_0^{-1}\mathbf{h}_0}\right\}\mathbf{h} &= \left\{\frac{4}{\mathbf{h}^T\mathbf{h} + \mathbf{h}_0^T\mathbf{h}_0}\right\}\mathbf{h} \end{aligned}$$

Since the matrices involved in the above equations are all positive definite: $\mathbf{\Omega}^{-1} > 0$, $\mathbf{\Omega} - \boldsymbol{\xi}\boldsymbol{\xi}^T > 0$, $\mathbf{\Omega}_0 > 0$ and $\mathbf{\Omega}_0^{-1} > 0$. The solutions to the above two equations can only be eigenvectors of the matrices

$$\mathbf{A}_0 = \frac{2\mathbf{\Omega}_0}{\mathbf{h}^T(\mathbf{\Omega} - \boldsymbol{\xi}\boldsymbol{\xi}^T)\mathbf{h} + \mathbf{h}_0^T\mathbf{\Omega}_0\mathbf{h}_0} + \frac{2\mathbf{\Omega}_0^{-1}}{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h} + \mathbf{h}_0^T\mathbf{\Omega}_0^{-1}\mathbf{h}_0} \tag{F.1}$$

$$\mathbf{A} = \frac{2(\mathbf{\Omega} - \boldsymbol{\xi}\boldsymbol{\xi}^T)}{\mathbf{h}^T(\mathbf{\Omega} - \boldsymbol{\xi}\boldsymbol{\xi}^T)\mathbf{h} + \mathbf{h}_0^T\mathbf{\Omega}_0\mathbf{h}_0} + \frac{2\mathbf{\Omega}^{-1}}{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h} + \mathbf{h}_0^T\mathbf{\Omega}_0^{-1}\mathbf{h}_0}. \tag{F.2}$$

The eigenvectors of $\mathbf{A}_0$ are the same as those of $\mathbf{\Omega}_0$. Hence, we have $\mathbf{h}_0$ equals 0 or any eigenvector of $\mathbf{\Omega}_0$(and of course $\mathbf{\Omega}_0^{-1}$). Therefore, the minimum value of $f(\mathbf{h}, \mathbf{h}_0)$ has to be obtained by 0 or some eigenvector of $\mathbf{\Omega}_0$ (since $\mathbf{h}_0 = \infty$ can be easily eliminated). If $\mathbf{h}_0 = 0$ then our conclusion follows.

Let's assume $\mathbf{h}_0 \neq 0$ and $\mathbf{\Omega}_0\mathbf{h}_0 = \lambda_k\mathbf{h}_0$.

Then,

$$\begin{aligned} f(\mathbf{h}, \mathbf{h}_0) &= \log\{\frac{\mathbf{h}^T(\mathbf{\Omega} - \boldsymbol{\xi}\boldsymbol{\xi}^T)\mathbf{h} + \lambda_k\mathbf{h}_0^T\mathbf{h}_0}{\mathbf{h}^T\mathbf{h} + \mathbf{h}_0^T\mathbf{h}_0}\} + \log\{\frac{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h} + \frac{1}{\lambda_k}\mathbf{h}_0^T\mathbf{h}_0}{\mathbf{h}^T\mathbf{h} + \mathbf{h}_0^T\mathbf{h}_0}\} \\ &= \log\{\frac{\mathbf{h}^T(\mathbf{\Omega} - \boldsymbol{\xi}\boldsymbol{\xi}^T)\mathbf{h}}{\mathbf{h}^T\mathbf{h}}W_h + \lambda_k(1 - W_h)\} + \log\{\frac{\mathbf{h}^T\mathbf{\Omega}^{-1}\mathbf{h}}{\mathbf{h}^T\mathbf{h}}W_h + \frac{1}{\lambda_k}(1 - W_h)\} \end{aligned}$$

where $W_h = \frac{\mathbf{h}^T\mathbf{h}}{\mathbf{h}^T\mathbf{h}+\mathbf{h}_0^T\mathbf{h}_0}$ is the weight between 0 and 1. Because $\log()$ is concave, we have $\log(aW_h + b(1-W_h)) \geq W_h \log(a) + (1-W_h)\log(b)$, hence,

$$\begin{aligned} f(\mathbf{h}, \mathbf{h}_0) &\geq W_h \left\{ \log \frac{\mathbf{h}^T(\boldsymbol{\Omega} - \boldsymbol{\xi}\boldsymbol{\xi}^T)\mathbf{h}}{\mathbf{h}^T\mathbf{h}} + \log \frac{\mathbf{h}^T\boldsymbol{\Omega}^{-1}\mathbf{h}}{\mathbf{h}^T\mathbf{h}} \right\} + (1-W_h) \left\{ \log(\lambda_k) + \log(\frac{1}{\lambda_k}) \right\} \\ &= W_h \left\{ \log \frac{\mathbf{h}^T(\boldsymbol{\Omega} - \boldsymbol{\xi}\boldsymbol{\xi}^T)\mathbf{h}}{\mathbf{h}^T\mathbf{h}} + \log \frac{\mathbf{h}^T\boldsymbol{\Omega}^{-1}\mathbf{h}}{\mathbf{h}^T\mathbf{h}} \right\} \\ &\geq W_h \cdot \min_{\mathbf{h}\in\mathbb{R}^{d_X}} \left\{ \log \frac{\mathbf{h}^T(\boldsymbol{\Omega} - \boldsymbol{\xi}\boldsymbol{\xi}^T)\mathbf{h}}{\mathbf{h}^T\mathbf{h}} + \log \frac{\mathbf{h}^T\boldsymbol{\Omega}^{-1}\mathbf{h}}{\mathbf{h}^T\mathbf{h}} \right\} \\ &\geq \min_{\mathbf{h}\in\mathbb{R}^{d_X}} \left\{ \log \frac{\mathbf{h}^T(\boldsymbol{\Omega} - \boldsymbol{\xi}\boldsymbol{\xi}^T)\mathbf{h}}{\mathbf{h}^T\mathbf{h}} + \log \frac{\mathbf{h}^T\boldsymbol{\Omega}^{-1}\mathbf{h}}{\mathbf{h}^T\mathbf{h}} \right\} \end{aligned}$$

The last inequality can be seen by noticing $\min_{\mathbf{h}\in\mathbb{R}^{d_X}} \left\{ \log \frac{\mathbf{h}^T(\boldsymbol{\Omega} - \boldsymbol{\xi}\boldsymbol{\xi}^T)\mathbf{h}}{\mathbf{h}^T\mathbf{h}} + \log \frac{\mathbf{h}^T\boldsymbol{\Omega}^{-1}\mathbf{h}}{\mathbf{h}^T\mathbf{h}} \right\} < 0$. This is because that

$$\log \frac{\mathbf{h}^T(\boldsymbol{\Omega} - \boldsymbol{\xi}\boldsymbol{\xi}^T)\mathbf{h}}{\mathbf{h}^T\mathbf{h}} + \log \frac{\mathbf{h}^T\boldsymbol{\Omega}^{-1}\mathbf{h}}{\mathbf{h}^T\mathbf{h}} < \log \frac{\mathbf{h}^T\boldsymbol{\Omega}\mathbf{h}}{\mathbf{h}^T\mathbf{h}} + \log \frac{\mathbf{h}^T\boldsymbol{\Omega}^{-1}\mathbf{h}}{\mathbf{h}^T\mathbf{h}},$$

where the minimum of the right hand side is zero by taking $\mathbf{h}$ equals to any eigenvector of $\boldsymbol{\Omega}$. The inequality holds because the strict concaveness of logarithm function and because we have $\mathbf{h}^T\boldsymbol{\xi}\boldsymbol{\xi}^T\mathbf{h} > 0$ for all $\mathbf{h} \neq 0$. (If $\mathbf{h}^T\boldsymbol{\xi}\boldsymbol{\xi}^T\mathbf{h} = 0$ for some $\mathbf{h}$ then it will contradict with the dimension of the envelope.)

Moreover, the lower bound of $f(\mathbf{h}, \mathbf{h}_0)$, which is negative, will be attained if we let $W_h = 1$ and let $\mathbf{h} = \arg\min_{\mathbf{h}\in\mathbb{R}^{d_X}} \left\{ \log \frac{\mathbf{h}^T(\boldsymbol{\Omega} - \boldsymbol{\xi}\boldsymbol{\xi}^T)\mathbf{h}}{\mathbf{h}^T\mathbf{h}} + \log \frac{\mathbf{h}^T\boldsymbol{\Omega}^{-1}\mathbf{h}}{\mathbf{h}^T\mathbf{h}} \right\}$. So we have the minimum found at $W_h = \frac{\mathbf{h}^T\mathbf{h}}{\mathbf{h}^T\mathbf{h}+\mathbf{h}_0^T\mathbf{h}_0} = 1$, or equivalently, $\ell = \mathbf{L}\mathbf{h} \in \text{span}(\mathbf{L})$.

F.2 The $(k+1)$-th direction ℓ_{k+1} in the population

Suppose $\mathbf{L}_k = (\ell_1, \ldots, \ell_k) \in \text{span}(\mathbf{L})$, $k < d_X$, then we want to show that $\ell_{k+1} \in \mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\mathcal{L})$. Since the minimization is over

$$f_k(\mathbf{v}) = \log(\mathbf{v}^T\mathbf{L}_{0k}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\mathbf{L}_{0k}\mathbf{v}) + \log(\mathbf{v}^T(\mathbf{L}_{0k}^T\boldsymbol{\Sigma}_\mathbf{X}\mathbf{L}_{0k})^{-1}\mathbf{v}), \tag{F.3}$$

subject to $\mathbf{v}^T\mathbf{v} = 1$. We know that this gives $\mathbf{v}_{k+1} \in \mathcal{E}_{\mathbf{L}_{0k}^T\boldsymbol{\Sigma}_\mathbf{X}\mathbf{L}_{0k}}(\mathbf{L}_{0k}^T\mathcal{L})$ which will lead to $\ell_{k+1} = \mathbf{L}_{0k}\mathbf{v}_{k+1} \in \mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\mathcal{L})$ by Lemma 5.

F.3 The $\sqrt{n}$-consistency for the sample algorithm

The proof hinges on Amemiya's (1985) results on the asymptotic properties of extremum estimators. We first state these results and then sketch how they can be used to prove the $\sqrt{n}$-consistency for our algorithm.

Let $Q_n(\mathbf{y}, \boldsymbol{\theta})$ be a real-valued function of the random variables $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^T$ and the parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)^T$. We shall sometimes write $Q_n(\mathbf{y}, \boldsymbol{\theta})$ more compactly as $Q_n(\boldsymbol{\theta})$. Let the parameter space be $\boldsymbol{\Theta}$ and let the true value of $\boldsymbol{\theta}$ be $\boldsymbol{\theta}_0$ which is in $\boldsymbol{\Theta}$. Then Theorem 4.1.1 in Amemiya (1985) gave the consistency of the extremum estimator, $\widehat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} Q_n(\mathbf{y}, \boldsymbol{\theta})$, as follows.

Theorem 1. *Under the following assumptions, $\widehat{\boldsymbol{\theta}}_n$ converges to $\boldsymbol{\theta}_0$ in probability.*
(A) The parameter space $\boldsymbol{\Theta}$ is a compact subset of $\mathbb{R}^K$.
(B) $Q_n(\mathbf{y}, \boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ for all $\mathbf{y}$ and is a measurable function of $\mathbf{y}$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.
(C) $n^{-1}Q_n(\boldsymbol{\theta})$ converges to a nonstochastic function $Q(\boldsymbol{\theta})$ in probability uniformly in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ as n goes to infinity, and $Q(\boldsymbol{\theta})$ attains a unique global maximum at $\boldsymbol{\theta}_0$.

The continuity of $Q(\boldsymbol{\theta})$ then follows from assumptions $(A) - (C)$. For a consistent $\widehat{\boldsymbol{\theta}}_n$, Theorem 4.1.3 in Amemiya (1985) established the asymptotic normality of $\widehat{\boldsymbol{\theta}}_n$ as follows.

Theorem 2. *In addition to the assumptions in Theorem 1, we assume the following conditions.*
(D) $\partial^2 Q_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T$ exists and is continuous in an open, convex neighborhood of $\boldsymbol{\theta}_0$.
(E) $n^{-1}\left\{\partial^2 Q_n(\mathbf{y}, \boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T\right\}_{\boldsymbol{\theta}=\boldsymbol{\theta}_n^*}$ converges to a finite nonsingular matrix

$$\mathbf{A}(\boldsymbol{\theta}_0) = \lim_{n\to\infty} \mathrm{E}\left\{n^{-1}\left\{\partial^2 Q_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T\right\}_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right\},$$

for any random sequences $\boldsymbol{\theta}_n^*$ such that $\operatorname{plim}(\boldsymbol{\theta}_n^*) = \boldsymbol{\theta}_0$.
(F) $n^{-1/2}\left\{\partial Q_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\right\}_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \to N(0, \mathbf{B}(\boldsymbol{\theta}_0))$, where

$$\mathbf{B}(\boldsymbol{\theta}_0) = \lim_{n\to\infty} \mathrm{E}\left\{n^{-1}\left\{\partial Q_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\right\}_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\left\{\partial Q_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^T\right\}_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right\}.$$

Then $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \to N(0, \mathbf{A}(\boldsymbol{\theta}_0)^{-1}\mathbf{B}(\boldsymbol{\theta}_0)\mathbf{A}(\boldsymbol{\theta}_0)^{-1})$.

In our adaptation of Theorem 1 and Theorem 2, we let $\boldsymbol{\theta} \equiv \ell$ whose true value is denoted by ℓ_0 and let the random variables $\mathbf{y} = \{(\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_n, \mathbf{Y}_n)\}$ where $(\mathbf{X}_i, \mathbf{Y}_i)$'s, $i = 1, \ldots, n$, are i.i.d. random variables. The parameter space is the 1D manifold $\boldsymbol{\Theta} = \mathcal{G}_{(p,1)}$ which is a compact set of $\mathbb{R}^{p\times 1}$, so condition (A) in Theorem 1 is satisfied. The function to be maximized is defined as follows.

$$Q_n(\ell) = -n/2 \log(\ell^T \mathbf{S}_{\mathbf{X}}^{-1} \ell) - n/2 \log(\ell^T \mathbf{S}_{\mathbf{X}|\mathbf{Y}} \ell) + n \log(\ell^T \ell). \tag{F.4}$$

Condition (B) holds because that $\mathbf{S}_{\mathbf{X}|\mathbf{Y}}$ and $\mathbf{S}_{\mathbf{X}}^{-1}$ are measurable functions of the i.i.d. variables $\{(\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_n, \mathbf{Y}_n)\}$, and because Q_n is continuous in $\boldsymbol{\theta}$, $\mathbf{S}_{\mathbf{X}}^{-1}$ and $\mathbf{S}_{\mathbf{X}|\mathbf{Y}}$.

We next verify condition (C) that $n^{-1}Q_n(\ell)$ converges uniformly to

$$Q(\ell) = -1/2 \log(\ell^T \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \ell) - 1/2 \log(\ell^T \boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}} \ell) + \log(\ell^T \ell). \tag{F.5}$$

It follows from Section F.1 that $Q(\ell)$ attains the unique global maximum at ℓ_0. For simplicity, we assume $\boldsymbol{\Sigma}_{\mathbf{X}}$ and $\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}$ both have distinct eigenvalues so that ℓ_0 is the unique maximum of $Q_n(\ell)$ in the 1D manifold $\boldsymbol{\Theta}$. For the case where there are multiple local maxima of $Q(\ell)$, we can obtain similar results by applying Theorem 4.1.2 in Amemiya (1985) as an alternative of Theorem 1. When p and d_X are fixed and n goes to infinity, the eigenvectors and eigenvalues of $\mathbf{S}_{\mathbf{X}}$, $\mathbf{S}_{\mathbf{X}}^{-1}$ and $\mathbf{S}_{\mathbf{X}|\mathbf{Y}}$ are $\sqrt{n}$-consistent for the eigenvectors and eigenvalues of $\boldsymbol{\Sigma}_{\mathbf{X}}$, $\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}$ and $\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}$. Then $n^{-1}Q_n(\ell)$ converge in probability to $Q(\ell)$ uniformly in ℓ, as can be seen from the following argument.

$$\begin{aligned} n^{-1}Q_n(\ell) - Q(\ell) &= -1/2 \left(\log(\ell^T \mathbf{S}_{\mathbf{X}}^{-1} \ell) - \log(\ell^T \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \ell)\right) \\ &\quad -1/2 \left(\log(\ell^T \mathbf{S}_{\mathbf{X}|\mathbf{Y}} \ell) - \log(\ell^T \boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}} \ell)\right) \\ &= -1/2 \log(\ell^T \mathbf{S}_{\mathbf{X}}^{-1} \ell / \ell^T \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \ell) - 1/2 \log(\ell^T \mathbf{S}_{\mathbf{X}|\mathbf{Y}} \ell / \ell^T \boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}} \ell). \end{aligned}$$

Hence, $\sup_{\ell\in\boldsymbol{\Theta}} \log(\ell^T \mathbf{S}_{\mathbf{X}}^{-1} \ell / \ell^T \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \ell) = \sup_{\ell\in\boldsymbol{\Theta}} \log(\ell^T \boldsymbol{\Sigma}_{\mathbf{X}}^{1/2} \mathbf{S}_{\mathbf{X}}^{-1} \boldsymbol{\Sigma}_{\mathbf{X}}^{1/2} \ell / \ell^T \ell)$. Notice that $\sup_{\ell\in\boldsymbol{\Theta}} \log(\ell^T \boldsymbol{\Sigma}_{\mathbf{X}}^{1/2} \mathbf{S}_{\mathbf{X}}^{-1} \boldsymbol{\Sigma}_{\mathbf{X}}^{1/2} \ell / \ell^T \ell)$ equals to the logarithm of the largest eigenvalue of $\boldsymbol{\Sigma}_{\mathbf{X}}^{1/2} \mathbf{S}_{\mathbf{X}}^{-1} \boldsymbol{\Sigma}_{\mathbf{X}}^{1/2}$ which converges to 0 in probability. Similarly, $\sup_{\ell\in\boldsymbol{\Theta}} \log(\ell^T \mathbf{S}_{\mathbf{X}|\mathbf{Y}} \ell / \ell^T \boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}} \ell)$ converges to zero in probability. Therefore, $n^{-1}Q_n(\ell)$ converges to $Q(\ell)$ in probability uniformly in $\ell \in \boldsymbol{\Theta}$. Note that we have assumed $\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}} > 0$ and $\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} > 0$, so their eigenvalues will be bounded away from zero.

We next verify conditions $(D)-(F)$. By straightforward calculation, condition (D) follows from the second derivative matrix

$$\begin{aligned} n^{-1}\frac{\partial^2 Q_n(\ell)}{\partial\ell\partial\ell^T} &= 2(\ell^T\mathbf{S}_{\mathbf{X}}^{-1}\ell)^{-2}(\mathbf{S}_{\mathbf{X}}^{-1}\ell\ell^T\mathbf{S}_{\mathbf{X}}^{-1}) - (\ell^T\mathbf{S}_{\mathbf{X}}^{-1}\ell)^{-1}\mathbf{S}_{\mathbf{X}}^{-1} \\ &\quad +2(\ell^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\ell)^{-2}(\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\ell\ell^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}) - (\ell^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\ell)^{-1}\mathbf{S}_{\mathbf{X}|\mathbf{Y}} \\ &\quad -2(\ell^T\ell)^{-2}\mathbf{P}_\ell + (\ell^T\ell)^{-1}\mathbf{I}_p. \end{aligned} \tag{F.6}$$

Condition (E) holds because the above quantity is a smooth function of ℓ, $\mathbf{S}_{\mathbf{X}}^{-1}$ and $\mathbf{S}_{\mathbf{X}|\mathbf{Y}}$.

Last, we need to verify condition (F). From the proof of Theorem 2, we only need to show that $n^{-1}\left\{\partial Q_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\right\}_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = O_p(1/\sqrt{n})$ for $\sqrt{n}$-consistency of the estimator $\widehat{\boldsymbol{\theta}}_n$.

$$n^{-1}\left\{\partial Q_n(\ell)/\partial\ell\right\}_{\ell=\ell_0} = -(\ell_0^T\mathbf{S}_{\mathbf{X}}^{-1}\ell_0)^{-1}\mathbf{S}_{\mathbf{X}}^{-1}\ell_0 - (\ell_0^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\ell_0)^{-1}\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\ell_0 + 2\ell_0. \tag{F.7}$$

Following the derivation in Section F.1, we know that $\left\{\partial Q(\ell)/\partial\ell\right\}_{\ell=\ell_0} = 0$. Then the result follows from the fact that $n^{-1}\partial Q_n(\ell)/\partial\ell$ is a smooth function of $\mathbf{S}_{\mathbf{X}}^{-1}$ and $\mathbf{S}_{\mathbf{X}|\mathbf{Y}}$ which are $\sqrt{n}$-consistent estimators.

So far, we have verified the conditions $(A)-(F)$ so that the sample estimator $\widehat{\ell}$ will be $\sqrt{n}$-consistent for the population estimator. For the $(k+1)$-th direction, $k < d_X$, let $\widehat{\mathbf{L}}_k$ denote an $\sqrt{n}$-consistent estimator of the first k directions and let $(\widehat{\mathbf{L}}_k, \widehat{\mathbf{L}}_{0k})$ be an orthogonal matrix. The $(k+1)$-th direction is defined by $\ell_{k+1} = \widehat{\mathbf{L}}_{0k}\mathbf{v}$ where the parameters are $\mathbf{v} \in \boldsymbol{\Theta} \subset \mathbb{R}^{p-k}$ and the parameter space is $\boldsymbol{\Theta} = \mathcal{G}_{p-k,1}$. We show that we can obtain a $\sqrt{n}$-consistent estimator $\widehat{\mathbf{v}}_n$, so the $\sqrt{n}$-consistency of $\widehat{\ell}_{k+1} = \widehat{\mathbf{L}}_{0k}\widehat{\mathbf{v}}_n$ then follows. From Section F.2, we define our function $Q_n(\mathbf{v})$ and $Q(\mathbf{v})$ as

$$\begin{aligned} Q_n(\mathbf{v}) &= -n/2\log(\mathbf{v}^T\widehat{\mathbf{L}}_{0k}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\widehat{\mathbf{L}}_{0k}\mathbf{v}) - n/2\log(\mathbf{v}^T\widehat{\mathbf{L}}_{0k}^T\mathbf{S}_{\mathbf{X}}^{-1}\widehat{\mathbf{L}}_{0k}\mathbf{v}) + n\log(\mathbf{v}^T\mathbf{v}) \\ Q(\mathbf{v}) &= -1/2\log(\mathbf{v}^T\mathbf{L}_{0k}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\mathbf{L}_{0k}\mathbf{v}) - 1/2\log(\mathbf{v}^T\mathbf{L}_{0k}^T\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\mathbf{L}_{0k}\mathbf{v}) + \log(\mathbf{v}^T\mathbf{v}) \end{aligned}$$

Following the same logic as verifying the conditions for the first direction, we can see that $\widehat{\mathbf{v}}_n = \arg\max Q_n(\mathbf{v})$ will be $\sqrt{n}$-consistent for $\mathbf{v}_0 = \arg\max Q(\mathbf{v})$ by noticing that $\widehat{\mathbf{L}}_{0k}^T\mathbf{S}_{\mathbf{X}|\mathbf{Y}}\widehat{\mathbf{L}}_{0k}$ and $\widehat{\mathbf{L}}_{0k}^T\mathbf{S}_{\mathbf{X}}^{-1}\widehat{\mathbf{L}}_{0k}$ are $\sqrt{n}$-consistent estimators for $\mathbf{L}_{0k}^T\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}}\mathbf{L}_{0k}$ and $\mathbf{L}_{0k}^T\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\mathbf{L}_{0k}$. Since all the d_X directions will be $\sqrt{n}$-consistent, the projection onto $\widehat{\mathbf{L}}_{1\mathrm{D}} = (\widehat{\ell}_1, \ldots, \widehat{\ell}_{d_X})$ will be a $\sqrt{n}$-consistent estimator for the projection onto the envelope $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{L})$.

G Proof for Proposition 4

The Jacobian matrix $\mathbf{J}_{\boldsymbol{h}}$ can be obtained by following Cook et al. (2010), and we make use of the following lemma from Cook et al. (2013).

Lemma 6. *Suppose that* $\mathbf{A} \in \mathbb{S}^{t\times t}$ *is non-singular and that the column-partition* $(\mathbf{B}, \mathbf{B}_0) \in \mathbb{R}^{t\times t}$ *is orthogonal. Then (1)* $|\mathbf{B}^T\mathbf{A}\mathbf{B}| = |\mathbf{A}| \cdot |\mathbf{B}_0^T\mathbf{A}^{-1}\mathbf{B}_0|$*, and (2)* $|\mathbf{A}| \leq |\mathbf{B}^T\mathbf{A}\mathbf{B}| \cdot |\mathbf{B}_0^T\mathbf{A}\mathbf{B}_0|$ *with equality if and only if* $\mathrm{span}(\mathbf{B})$ *reduces* $\mathbf{A}$.

In the population, the objective function $\mathrm{F}(\mathbf{L}\oplus\mathbf{R})$ from (4.4) can be written as,

$$\begin{aligned}
\mathrm{F}(\mathbf{L}\oplus\mathbf{R}) &= \log|\mathbf{L}^T\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{L}| + \log|\mathbf{L}_0^T\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{L}_0| + \log|\mathbf{R}_0^T\boldsymbol{\Sigma}_{\mathbf{Y}}\mathbf{R}_0| \\
&\quad + \log|\mathbf{R}^T\left(\boldsymbol{\Sigma}_{\mathbf{Y}} - \boldsymbol{\Sigma}_{\mathbf{XY}}^T\mathbf{L}(\mathbf{L}^T\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{L})^{-1}\mathbf{L}^T\boldsymbol{\Sigma}_{\mathbf{XY}}\right)\mathbf{R}|. \\
&\geq \log|\boldsymbol{\Sigma}_{\mathbf{X}}| + \log|\mathbf{R}_0^T\boldsymbol{\Sigma}_{\mathbf{Y}}\mathbf{R}_0| + \log|\mathbf{R}^T\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{L}^T\mathbf{X}}\mathbf{R}| \\
&\geq \log|\boldsymbol{\Sigma}_{\mathbf{X}}| + \log|\mathbf{R}_0^T\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{L}^T\mathbf{X}}\mathbf{R}_0| + \log|\mathbf{R}^T\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{L}^T\mathbf{X}}\mathbf{R}| \\
&\geq \log|\boldsymbol{\Sigma}_{\mathbf{X}}| + \log|\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{L}^T\mathbf{X}}| \\
&\geq \log|\boldsymbol{\Sigma}_{\mathbf{X}}| + \log|\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}|,
\end{aligned}$$

where $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{L}^T\mathbf{X}} = \boldsymbol{\Sigma}_{\mathbf{Y}} - \boldsymbol{\Sigma}_{\mathbf{XY}}^T\mathbf{L}(\mathbf{L}^T\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{L})^{-1}\mathbf{L}^T\boldsymbol{\Sigma}_{\mathbf{XY}}$. The first and third inequalities came directly from Lemma 6, which also says that the first inequality turns to equality if and only if (C1) $\mathrm{span}(\mathbf{L})$ reduces $\boldsymbol{\Sigma}_{\mathbf{X}}$ and that the third inequality becomes equality if and only if (C3) $\mathrm{span}(\mathbf{R})$ reduces $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{L}^T\mathbf{X}}$. The second and the last inequalities are obtained from the fact that $\boldsymbol{\Sigma}_{\mathbf{Y}} \geq \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{L}^T\mathbf{X}}$ and $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{L}^T\mathbf{X}} \geq \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$ for any $\mathbf{L}$. Hence, these two inequalities become equalities if and only if (C2) $\mathrm{span}(\boldsymbol{\Sigma}_{\mathbf{Y}} - \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{L}^T\mathbf{X}}) \subseteq \mathrm{span}(\mathbf{R})$ and (C4) $\mathrm{span}(\boldsymbol{\Sigma}_{\mathbf{XY}}^T) \subseteq \mathrm{span}(\mathbf{L})$. Since d_X is the dimension of the smallest subspace satisfying condition (C1) and (C4), $\mathrm{span}(\mathbf{L}) = \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{L})$. Therefore, condition (C2) and (C3) will be equivalent to (C2′) $\mathrm{span}(\boldsymbol{\Sigma}_{\mathbf{Y}} - \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}) = \mathrm{span}(\boldsymbol{\Sigma}_{\mathbf{XY}}) \subseteq \mathrm{span}(\mathbf{R})$, and (C3′) $\mathrm{span}(\mathbf{R})$ reduces $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}$. Again, because d_Y is the dimension of the smallest subspace satisfying (C2′) and (C3′), we can conclude that $\mathrm{span}(\mathbf{R}) = \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}}(\mathcal{R})$.

The rest of the proof relies on Shapiro's (1986) results on the asymptotics of overparameterized structural models. In order to apply Shapiro's (1986) theory in our context, we define the minimum discrepancy function as

$$\mathrm{F}(\widehat{\boldsymbol{h}}_{\mathrm{full}}, \boldsymbol{h}) = \log|\boldsymbol{\Sigma}_{\mathbf{C}}| + \mathrm{trace}(\boldsymbol{\Sigma}_{\mathbf{C}}^{-1}\mathbf{S}_{\mathbf{C}}) - \log|\mathbf{S}_{\mathbf{C}}| - \mathrm{trace}(\mathbf{S}_{\mathbf{C}}^{-1}\mathbf{S}_{\mathbf{C}}). \tag{G.1}$$

Then $\mathrm{F}(\widehat{\boldsymbol{h}}_{\text{full}}, \boldsymbol{h})$ satisfies: (1) $\mathrm{F}(\widehat{\boldsymbol{h}}_{\text{full}}, \boldsymbol{h}) \geq 0$ for all $\widehat{\boldsymbol{h}}_{\text{full}}$ and $\boldsymbol{h}$; (2) $\mathrm{F}(\widehat{\boldsymbol{h}}_{\text{full}}, \boldsymbol{h}) = 0$ if and only if $\widehat{\boldsymbol{h}}_{\text{full}} = \boldsymbol{h}$; and (3) $\mathrm{F}(\widehat{\boldsymbol{h}}_{\text{full}}, \boldsymbol{h})$ is twice continuously differentiable in $\widehat{\boldsymbol{h}}_{\text{full}}$ and $\boldsymbol{h}$. Recall from Section 5.1 that we use the subscript 0 to emphasize the true parameter: $\boldsymbol{h}_0$ and $\boldsymbol{\phi}_0$ correspond to the true distribution of $\mathbf{C}$. Then $\widehat{\boldsymbol{h}}_{\text{full}}$ is $\sqrt{n}$-consistent for $\boldsymbol{h}_0$. Notice that $\widehat{\boldsymbol{h}}_{\text{full}}$ is a smooth function of the sample covariance matrices which converges in distribution to the population covariance matrices, then by the delta method we know $\sqrt{n}(\widehat{\boldsymbol{h}}_{\text{full}} - \boldsymbol{h}_0) \to N(0, \boldsymbol{\Gamma})$, for some positive definite covariance $\boldsymbol{\Gamma}$. Using Shapiro's (1986) Proposition 3.1 and Proposition 4.1, we will have $\sqrt{n}$-consistency results for $\mathrm{vec}(\widehat{\boldsymbol{h}})$. We next need to compare $\boldsymbol{\Gamma} = \mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{h}}_{\text{full}})$ with $\mathbf{W} = \mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{h}})$.

By noticing that $\mathbf{P}_{\mathbf{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{\Delta}} = \mathbf{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{\Delta}(\boldsymbol{\Delta}^T\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\Delta})^{\dagger}\boldsymbol{\Delta}^T\mathbf{J}_{\boldsymbol{h}}^{1/2}$, we have

$$\begin{aligned}\boldsymbol{\Gamma} - \mathbf{W} &= \boldsymbol{\Gamma} - \mathbf{J}_{\boldsymbol{h}}^{-1/2}\mathbf{P}_{\mathbf{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{\Delta}}\mathbf{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{\Gamma}\mathbf{J}_{\boldsymbol{h}}^{1/2}\mathbf{P}_{\mathbf{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{\Delta}}\mathbf{J}_{\boldsymbol{h}}^{-1/2} \\ &= \mathbf{J}_{\boldsymbol{h}}^{-1/2}\left(\mathbf{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{\Gamma}\mathbf{J}_{\boldsymbol{h}}^{1/2} - \mathbf{P}_{\mathbf{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{\Delta}}\mathbf{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{\Gamma}\mathbf{J}_{\boldsymbol{h}}^{1/2}\mathbf{P}_{\mathbf{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{\Delta}}\right)\mathbf{J}_{\boldsymbol{h}}^{-1/2}.\end{aligned}$$

Then the conclusion follows from the above expression.

H Proof for Proposition 5

Proposition 5 can be seen from Proposition 4 by letting $\boldsymbol{\Gamma} = \mathbf{J}_{\boldsymbol{h}}^{-1}$ under the normality assumption.

I Proof for Proposition 6

I.1 The gradient matrix $\boldsymbol{\Delta} = \frac{\partial \boldsymbol{h}(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}$

The first three columns of the gradient matrix $\boldsymbol{\Delta} \equiv (\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_7)$ are

$$(\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_3) = \begin{pmatrix} \mathbf{R}\otimes\mathbf{L} & \mathbf{R}\boldsymbol{\eta}^T\otimes\mathbf{I}_p & \mathbf{K}_{r,p}\left(\mathbf{L}\boldsymbol{\eta}\otimes\mathbf{I}_r\right) \\ 0 & 2\mathbf{C}_p\left(\mathbf{L}\boldsymbol{\Omega}\otimes\mathbf{I}_p - \mathbf{L}\otimes\mathbf{L}_0\boldsymbol{\Omega}_0\mathbf{L}_0^T\right) & 0 \\ 0 & 0 & 2\mathbf{C}_r\left(\mathbf{R}\boldsymbol{\Phi}\otimes\mathbf{I}_r - \mathbf{R}\otimes\mathbf{R}_0\boldsymbol{\Phi}_0\mathbf{R}_0^T\right) \end{pmatrix},$$

and the last four columns are

$$(\boldsymbol{\delta}_4, \ldots, \boldsymbol{\delta}_7) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \mathbf{C}_p(\mathbf{L}\otimes\mathbf{L})\mathbf{E}_{d_X} & \mathbf{C}_p(\mathbf{L}_0\otimes\mathbf{L}_0)\mathbf{E}_{p-d_X} & 0 & 0 \\ 0 & 0 & \mathbf{C}_r(\mathbf{R}\otimes\mathbf{R})\mathbf{E}_{d_Y} & \mathbf{C}_r(\mathbf{R}_0\otimes\mathbf{R}_0)\mathbf{E}_{r-d_Y} \end{pmatrix},$$

where $\mathbf{C}_p \in \mathbb{R}^{p(p+1)/2 \times p^2}$ is called the contraction matrix and $\mathbf{E}_p \in \mathbb{R}^{p^2 \times p(p+1)/2}$ is called the extraction matrix (Henderson and Searle, 1979). They relate the vec operation and vech operation for any symmetric matrix $\mathbf{M} \in \mathbb{S}^{p \times p}$ as: $\text{vech}(\mathbf{M}) = \mathbf{C}_p \text{vec}(\mathbf{M})$ and $\text{vec}(\mathbf{M}) = \mathbf{E}_p \text{vech}(\mathbf{M})$. More properties about the contraction and extraction matrices can be found from Henderson adn Searle (1979).

Proof. We can write $\boldsymbol{\Delta} = \frac{\partial \boldsymbol{h}(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}$ as $\begin{pmatrix} \partial \boldsymbol{h}_1/\partial \boldsymbol{\phi} \\ \partial \boldsymbol{h}_2/\partial \boldsymbol{\phi} \\ \partial \boldsymbol{h}_3/\partial \boldsymbol{\phi} \end{pmatrix}$ and compute it by rows.

(1) Computation of $\frac{\partial \boldsymbol{h}_1}{\partial \boldsymbol{\phi}}$.

First, $\boldsymbol{h}_1(\boldsymbol{\phi}) = \text{vec}(\boldsymbol{\beta}) = \text{vec}(\mathbf{L}\boldsymbol{\eta}\mathbf{R}^T)$ and thus, $\partial \boldsymbol{h}_1(\boldsymbol{\phi})/\partial \boldsymbol{\phi}_k = 0$, $k = 4, 5, 6, 7$. The rest of the terms are:

$$\begin{aligned}
\frac{\partial \boldsymbol{h}_1(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}_1} &= \frac{\partial \left[\text{vec}(\mathbf{L}\boldsymbol{\eta}\mathbf{R}^T)\right]}{\partial \text{vec}(\boldsymbol{\eta})} = \frac{\partial \left[\{\mathbf{R} \otimes \mathbf{L}\}\text{vec}(\boldsymbol{\eta})\right]}{\partial \text{vec}(\boldsymbol{\eta})} = \mathbf{R} \otimes \mathbf{L} \\
\frac{\partial \boldsymbol{h}_1(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}_2} &= \frac{\partial \left[\text{vec}(\mathbf{L}\boldsymbol{\eta}\mathbf{R}^T)\right]}{\partial \text{vec}(\mathbf{L})} = \frac{\partial \left[\{\mathbf{R}\boldsymbol{\eta}^T \otimes \mathbf{I}_p\}\text{vec}(\mathbf{L})\right]}{\partial \text{vec}(\mathbf{L})} = \mathbf{R}\boldsymbol{\eta}^T \otimes \mathbf{I}_p \\
\frac{\partial \boldsymbol{h}_1(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}_3} &= \frac{\partial \left[\text{vec}(\mathbf{L}\boldsymbol{\eta}\mathbf{R}^T)\right]}{\partial \text{vec}(\mathbf{R})} = \frac{\partial \left[\{\mathbf{I}_r \otimes \mathbf{L}\boldsymbol{\eta}\}\text{vec}(\mathbf{R}^T)\right]}{\partial \text{vec}(\mathbf{R})} = \{\mathbf{I}_r \otimes \mathbf{L}\boldsymbol{\Omega}^{-1}\boldsymbol{\eta}\}\mathbf{K}_{r,m} \\
&= \mathbf{K}_{r,p}\{\mathbf{L}\boldsymbol{\eta} \otimes \mathbf{I}_r\}.
\end{aligned}$$

(2) Computation of $\partial \boldsymbol{h}_2/\partial \boldsymbol{\phi}$.

Since $\boldsymbol{h}_2(\boldsymbol{\phi}) = \text{vech}(\boldsymbol{\Sigma}_{\mathbf{X}}) = \text{vech}(\mathbf{L}\boldsymbol{\Omega}\mathbf{L}^T + \mathbf{L}_0\boldsymbol{\Omega}_0\mathbf{L}_0^T)$, it follows that $\partial \boldsymbol{h}_2(\boldsymbol{\phi})/\partial \boldsymbol{\phi}_k = 0$, $k = 1, 3, 6, 7$. The rest terms can be found in Cook, Li and Chiaromonte (2010). In our notation, they are

$$\begin{aligned}
\frac{\partial \boldsymbol{h}_2(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}_2} &= 2\mathbf{C}_p(\mathbf{L}\boldsymbol{\Omega} \otimes \mathbf{I}_p - \mathbf{L} \otimes \mathbf{L}_0\boldsymbol{\Omega}_0\mathbf{L}_0^T), \\
\frac{\partial \boldsymbol{h}_2(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}_4} &= \mathbf{C}_p(\mathbf{L} \otimes \mathbf{L})\mathbf{E}_{d_X}, \\
\frac{\partial \boldsymbol{h}_2(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}_5} &= \mathbf{C}_p(\mathbf{L}_0 \otimes \mathbf{L}_0)\mathbf{E}_{p-d_X}.
\end{aligned}$$

(3) Computation of $\partial \boldsymbol{h}_3/\partial \boldsymbol{\phi}$.

Since $\boldsymbol{h}_3(\boldsymbol{\phi}) = \mathrm{vech}(\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}) = \mathrm{vech}(\mathbf{R}\boldsymbol{\Phi}\mathbf{R}^T + \mathbf{R}_0\boldsymbol{\Phi}_0\mathbf{R}_0^T)$, we have

$$\begin{aligned}
\frac{\partial \boldsymbol{h}_3(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}_k} &= 0,\ \mathrm{k} = 1, 2, 4, 5, \\
\frac{\partial \boldsymbol{h}_3(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}_3} &= 2\mathbf{C}_r(\mathbf{R}\boldsymbol{\Phi} \otimes \mathbf{I}_r - \mathbf{R} \otimes \mathbf{R}_0\boldsymbol{\Phi}_0\mathbf{R}_0^T), \\
\frac{\partial \boldsymbol{h}_3(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}_6} &= \mathbf{C}_r(\mathbf{R} \otimes \mathbf{R})\mathbf{E}_{d_Y}, \\
\frac{\partial \boldsymbol{h}_3(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}_7} &= \mathbf{C}_r(\mathbf{R}_0 \otimes \mathbf{R}_0)\mathbf{E}_{r-d_Y}.
\end{aligned}$$

□

I.2 Transformation $\boldsymbol{\Delta} = \tilde{\boldsymbol{\Delta}}\mathbf{T}$

We want to find a transformation such that $\mathrm{span}(\boldsymbol{\Delta}) = \mathrm{span}(\tilde{\boldsymbol{\Delta}})$ and $\tilde{\boldsymbol{\Delta}}^T\mathbf{J}_{\boldsymbol{h}}\tilde{\boldsymbol{\Delta}}$ is block diagonal. Partition $\boldsymbol{\Delta}$ into its first three columns and its last four column, and denote them as $\boldsymbol{\Delta} \equiv (\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2)$. It is easy to check that $\boldsymbol{\Delta}_2^T\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\Delta}_2$ is already block diagonal. So we need to find a transformed $\tilde{\boldsymbol{\Delta}}_1$ such that $\tilde{\boldsymbol{\Delta}}_1^T\mathbf{J}_{\boldsymbol{h}}\tilde{\boldsymbol{\Delta}}_1$ is block diagonal and $\tilde{\boldsymbol{\Delta}}_1^T\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\Delta}_2 = 0$.

By direct computation, we have

$$\begin{aligned}
\tilde{\boldsymbol{\Delta}}_1 &= \begin{pmatrix} \mathbf{R} \otimes \mathbf{L} & \mathbf{R}\boldsymbol{\eta}^T \otimes \mathbf{L}_0 & \mathbf{K}_{r,p}\,(\mathbf{L}\boldsymbol{\eta} \otimes \mathbf{R}_0) \\ 0 & 2\mathbf{C}_p\,(\mathbf{L}\boldsymbol{\Omega} \otimes \mathbf{L}_0 - \mathbf{L} \otimes \mathbf{L}_0\boldsymbol{\Omega}_0) & 0 \\ 0 & 0 & 2\mathbf{C}_r\,(\mathbf{R}\boldsymbol{\Phi} \otimes \mathbf{R}_0 - \mathbf{R} \otimes \mathbf{R}_0\boldsymbol{\Phi}_0) \end{pmatrix}, \\
&\equiv (\tilde{\boldsymbol{\delta}}_1, \tilde{\boldsymbol{\delta}}_2, \tilde{\boldsymbol{\delta}}_3) \\
\tilde{\boldsymbol{\Delta}}_2 &= \boldsymbol{\Delta}_2.
\end{aligned}$$

And the 7×7 blocks transformation matrix $\mathbf{T}$ can be written as a 7×3 blocks matrix $\mathbf{T}_1$ and a 7×4 block matrix $\mathbf{T}_2 = \begin{pmatrix} \mathbf{0}_{3\times 4} \\ \mathbf{I}_{4\times 4} \end{pmatrix}$ whose blocks conform to those of $\tilde{\boldsymbol{\Delta}}$. Also, $\tilde{\boldsymbol{\Delta}}\mathbf{T}_1 = \boldsymbol{\Delta}_1$ and $\mathbf{T}_1$ is given by

$$\mathbf{T}_1 = \begin{pmatrix} \mathbf{I}_{d_X d_Y} & \boldsymbol{\eta}^T \otimes \mathbf{L}^T & \mathbf{K}_{d_Y,d_X}\,(\boldsymbol{\eta} \otimes \mathbf{R}^T) \\ 0 & \mathbf{I}_{d_X} \otimes \mathbf{L}_0^T & 0 \\ 0 & 0 & \mathbf{I}_{d_Y} \otimes \mathbf{R}_0^T \\ 0 & 2\mathbf{C}_{d_X}(\boldsymbol{\Omega} \otimes \mathbf{L}^T) & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 2\mathbf{C}_{d_Y}(\boldsymbol{\Phi} \otimes \mathbf{R}^T) \\ 0 & 0 & 0 \end{pmatrix}.$$

It can be easily noticed that $\mathbf{T}$ has full row rank, hence, $\mathrm{span}(\boldsymbol{\Delta}) = \mathrm{span}(\tilde{\boldsymbol{\Delta}})$. Therefore, the asymptotic covariance for $\widehat{\boldsymbol{h}}$ under the envelope model is given by

$$\begin{aligned}\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{h}}) &= \boldsymbol{\Delta}\left(\boldsymbol{\Delta}^T\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\Delta}\right)^{\dagger}\boldsymbol{\Delta}^T = \tilde{\boldsymbol{\Delta}}\left(\tilde{\boldsymbol{\Delta}}^T\mathbf{J}_{\boldsymbol{h}}\tilde{\boldsymbol{\Delta}}\right)^{\dagger}\tilde{\boldsymbol{\Delta}}^T \\ &= \sum_{i=1}^{3}\tilde{\boldsymbol{\delta}}_i(\tilde{\boldsymbol{\delta}}_i^T\mathbf{J}_{\boldsymbol{h}}\tilde{\boldsymbol{\delta}}_i)^{\dagger}\tilde{\boldsymbol{\delta}}_i^T + \sum_{j=4}^{7}\boldsymbol{\delta}_j(\boldsymbol{\delta}_j^T\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\delta}_j)^{\dagger}\boldsymbol{\delta}_j^T.\end{aligned}$$

I.3 Asymptotic covariance matrix for $\mathrm{vec}(\widehat{\boldsymbol{\beta}})$

The asymptotic covariance matrix for $\mathrm{vec}(\widehat{\boldsymbol{\beta}})$ is the upper left block of the full covariance matrix for $\widehat{\boldsymbol{h}}$. Under the standard model, it is just $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}\otimes\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}$. From Proposition 2, we recognize this as the asymptotic variance of $\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})$.

$$\mathrm{avar}(\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})) = \boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}\otimes\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}. \tag{I.1}$$

For the envelope model, only $\tilde{\boldsymbol{\delta}}_i$, $i = 1, 2, 3$, have contributions to this block. We use the subscript $\boldsymbol{\beta\beta}$ to denote the upper left block of dimension $pr \times pr$ for a matrix and compute the corresponding three terms.

(1) Contribution from $\tilde{\boldsymbol{\delta}}_1$.

$$\begin{aligned}[\tilde{\boldsymbol{\delta}}_1(\tilde{\boldsymbol{\delta}}_1^T\mathbf{J}_{\boldsymbol{h}}\tilde{\boldsymbol{\delta}}_1)^{\dagger}\tilde{\boldsymbol{\delta}}_1^T]_{\boldsymbol{\beta\beta}} &= (\mathbf{R}\otimes\mathbf{L})\left((\mathbf{R}\otimes\mathbf{L})^T(\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\otimes\boldsymbol{\Sigma}_{\mathbf{X}})(\mathbf{R}\otimes\mathbf{L})\right)^{\dagger}(\mathbf{R}\otimes\mathbf{L})^T \\ &= (\mathbf{R}\otimes\mathbf{L})\left(\boldsymbol{\Phi}^{-1}\otimes\boldsymbol{\Omega}\right)^{\dagger}(\mathbf{R}\otimes\mathbf{L})^T \\ &= \mathbf{R}\boldsymbol{\Phi}\mathbf{R}^T\otimes\mathbf{L}\boldsymbol{\Omega}^{-1}\mathbf{L}^T = \mathrm{avar}(\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathbf{L},\mathbf{R}})).\end{aligned}$$

(2) Contribution from $\tilde{\boldsymbol{\delta}}_2$. Let $\mathbf{A}_1 = \mathbf{E}_p\left(\mathbf{C}_p\left(\mathbf{L}\boldsymbol{\Omega}\otimes\mathbf{L}_0 - \mathbf{L}\otimes\mathbf{L}_0\boldsymbol{\Omega}_0\right)\right)$ then,

$$\begin{aligned}\tilde{\boldsymbol{\delta}}_2^T\mathbf{J}_{\boldsymbol{h}}\tilde{\boldsymbol{\delta}}_2 &= \left(\mathbf{R}\boldsymbol{\eta}^T\otimes\mathbf{L}_0\right)^T(\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\otimes\boldsymbol{\Sigma}_{\mathbf{X}})\left(\mathbf{R}\boldsymbol{\eta}^T\otimes\mathbf{L}_0\right) + 2\mathbf{A}_1^T(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\otimes\boldsymbol{\Sigma}_{\mathbf{X}}^{-1})\mathbf{A}_1 \\ &= \boldsymbol{\eta}\boldsymbol{\Phi}^{-1}\boldsymbol{\eta}^T\otimes\boldsymbol{\Omega}_0 + \boldsymbol{\Omega}\otimes\boldsymbol{\Omega}_0^{-1} + \boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}_0 - 2\mathbf{I}_{d_X}\otimes\mathbf{I}_{p-d_X} \\ &\equiv \mathbf{M}_2.\end{aligned} \tag{I.2}$$

Thus for the covariance,

$$[\tilde{\boldsymbol{\delta}}_2(\tilde{\boldsymbol{\delta}}_2^T\mathbf{J}_{\boldsymbol{h}}\tilde{\boldsymbol{\delta}}_2)^{\dagger}\tilde{\boldsymbol{\delta}}_2^T]_{\boldsymbol{\beta\beta}} = (\mathbf{R}\boldsymbol{\eta}^T\otimes\mathbf{L}_0)(\mathbf{M}_2)^{\dagger}(\boldsymbol{\eta}\mathbf{R}^T\otimes\mathbf{L}_0^T). \tag{I.3}$$

(3) Contribution from $\tilde{\boldsymbol{\delta}}_3$. Let $\mathbf{A}_2 = \mathbf{E}_r\mathbf{C}_r\left(\mathbf{R}\boldsymbol{\Phi}\otimes\mathbf{R}_0 - \mathbf{R}\otimes\mathbf{R}_0\boldsymbol{\Phi}_0\right)$ then,

$$\begin{aligned}
\tilde{\boldsymbol{\delta}}_3^T\mathbf{J}_{\boldsymbol{h}}\tilde{\boldsymbol{\delta}}_3 &= \left(\mathbf{K}_{r,p}\left(\mathbf{L}\boldsymbol{\eta}\otimes\mathbf{R}_0\right)\right)^T\left(\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\otimes\boldsymbol{\Sigma}_{\mathbf{X}}\right)\left(\mathbf{K}_{r,p}\left(\mathbf{L}\boldsymbol{\eta}\otimes\mathbf{R}_0\right)\right)\\
&+ 2\mathbf{A}^T(\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\otimes\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1})\mathbf{A}\\
&= \boldsymbol{\eta}^T\boldsymbol{\Omega}\boldsymbol{\eta}\otimes\boldsymbol{\Phi}_0^{-1} + \boldsymbol{\Phi}\otimes\boldsymbol{\Phi}_0^{-1} + \boldsymbol{\Phi}^{-1}\otimes\boldsymbol{\Phi}_0 - 2\mathbf{I}_{d_Y}\otimes\mathbf{I}_{r-d_Y}\\
&\equiv \mathbf{M}_3.
\end{aligned}$$

Thus for the covariance,

$$\begin{aligned}
[\tilde{\boldsymbol{\delta}}_3(\tilde{\boldsymbol{\delta}}_3^T\mathbf{J}_{\boldsymbol{h}}\tilde{\boldsymbol{\delta}}_3)^\dagger\tilde{\boldsymbol{\delta}}_3^T]_{\boldsymbol{\beta\beta}} &= \mathbf{K}_{r,p}\left(\mathbf{L}\boldsymbol{\eta}\otimes\mathbf{R}_0\right)\left(\mathbf{M}_3\right)^\dagger\left(\mathbf{K}_{r,p}\left(\mathbf{L}\boldsymbol{\eta}\otimes\mathbf{R}_0\right)\right)^T\\
&= \left(\mathbf{R}_0\otimes\mathbf{L}\boldsymbol{\eta}\right)\mathbf{K}_{r-d_Y,d_Y}\left(\mathbf{M}_3\right)^\dagger\mathbf{K}_{r-d_Y,d_Y}^T\left(\boldsymbol{\eta}^T\mathbf{L}^T\otimes\mathbf{R}_0^T\right)\\
&= \left(\mathbf{R}_0\otimes\mathbf{L}\boldsymbol{\eta}\right)\left(\mathbf{K}_{r-d_Y,d_Y}^T\mathbf{M}_3\mathbf{K}_{r-d_Y,d_Y}\right)^\dagger\left(\boldsymbol{\eta}^T\mathbf{L}^T\otimes\mathbf{R}_0^T\right)
\end{aligned}$$

the last equality is because $\mathbf{K}_{r-d_Y,d_Y}$ is non-singular and $\mathbf{K}_{r-d_Y,d_Y}\mathbf{K}_{r-d_Y,d_Y}^T = \mathbf{I}_{(r-d_Y)d_Y}$.
Therefore,

$$\begin{aligned}
[\tilde{\boldsymbol{\delta}}_3(\tilde{\boldsymbol{\delta}}_3^T\mathbf{J}_{\boldsymbol{h}}\tilde{\boldsymbol{\delta}}_3)^\dagger\tilde{\boldsymbol{\delta}}_3^T]_{\boldsymbol{\beta\beta}} &= \left(\mathbf{R}_0\otimes\mathbf{L}\boldsymbol{\eta}\right)\left(\mathbf{K}_{r-d_Y,d_Y}^T\mathbf{M}_3\mathbf{K}_{r-d_Y,d_Y}\right)^\dagger\left(\boldsymbol{\eta}^T\mathbf{L}^T\otimes\mathbf{R}_0^T\right)\\
&= \left(\mathbf{R}_0\otimes\mathbf{L}\boldsymbol{\eta}\right)\left(\mathbf{M}_4\right)^\dagger\left(\boldsymbol{\eta}^T\mathbf{L}^T\otimes\mathbf{R}_0^T\right),
\end{aligned}$$

where

$$\mathbf{M}_4 = \boldsymbol{\Phi}_0^{-1}\otimes\boldsymbol{\eta}^T\boldsymbol{\Omega}\boldsymbol{\eta} + \boldsymbol{\Phi}_0^{-1}\otimes\boldsymbol{\Phi} + \boldsymbol{\Phi}_0\otimes\boldsymbol{\Phi}^{-1} - 2\mathbf{I}_{r-d_Y}\otimes\mathbf{I}_{d_Y}. \tag{I.4}$$

I.4 Interpretations

The Fisher information matrix for $\boldsymbol{\phi} = (\boldsymbol{\phi}_1^T,\ldots,\boldsymbol{\phi}_7)^T$, $\boldsymbol{\Delta}^T\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\Delta}$, has the form

$$\begin{pmatrix}
\mathbf{J}_{\boldsymbol{\eta}} & \mathbf{J}_{\boldsymbol{\eta}\mathbf{L}} & \mathbf{J}_{\boldsymbol{\eta}\mathbf{R}} & 0 & 0 & 0 & 0\\
\mathbf{J}_{\boldsymbol{\eta}\mathbf{L}}^T & \mathbf{J}_{\mathbf{L}} & \mathbf{J}_{\mathbf{LR}} & \mathbf{J}_{\mathbf{L}\boldsymbol{\Omega}} & 0 & 0 & 0\\
\mathbf{J}_{\boldsymbol{\eta}\mathbf{R}}^T & \mathbf{J}_{\mathbf{LR}}^T & \mathbf{J}_{\mathbf{R}} & 0 & 0 & \mathbf{J}_{\mathbf{R}\boldsymbol{\Phi}} & 0\\
0 & \mathbf{J}_{\mathbf{L}\boldsymbol{\Omega}}^T & 0 & \mathbf{J}_{\boldsymbol{\Omega}} & 0 & 0 & 0\\
0 & 0 & 0 & 0 & \mathbf{J}_{\boldsymbol{\Omega}_0} & 0 & 0\\
0 & 0 & \mathbf{J}_{\mathbf{R}\boldsymbol{\Phi}}^T & 0 & 0 & \mathbf{J}_{\boldsymbol{\Phi}} & 0\\
0 & 0 & 0 & 0 & 0 & 0 & \mathbf{J}_{\boldsymbol{\Phi}_0}
\end{pmatrix}.$$

If $\mathbf{L}$ and $\mathbf{R}$ are known, then the asymptotic variance of the MLE of $\mathrm{vec}(\boldsymbol{\eta})$, denoted

by $\mathrm{vec}(\widehat{\boldsymbol{\eta}}_{\mathbf{L},\mathbf{R}})$, is simply $\mathbf{J}_{\boldsymbol{\eta}}^{-1} = \boldsymbol{\Phi} \otimes \boldsymbol{\Omega}^{-1}$. As we discussed before,

$$\begin{aligned}
\mathrm{avar}(\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathbf{L},\mathbf{R}})) &= \mathrm{avar}(\sqrt{n}\mathrm{vec}(\mathbf{L}\widehat{\boldsymbol{\eta}}_{\mathbf{L},\mathbf{R}}\mathbf{R}^T)) \\
&= (\mathbf{R} \otimes \mathbf{L})\mathrm{avar}(\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\eta}}_{\mathbf{L},\mathbf{R}}))(\mathbf{R}^T \otimes \mathbf{L}^T) \\
&= \mathbf{R}\boldsymbol{\Phi}\mathbf{R}^T \otimes \mathbf{L}\boldsymbol{\Omega}^{-1}\mathbf{L}^T.
\end{aligned}$$

I.5 If $\boldsymbol{\eta}$ and $\mathbf{R}$ are known,

then the asymptotic variance of the MLE of $\mathrm{vec}(\mathbf{L})$, denoted by $\mathrm{vec}(\widehat{\mathbf{L}}_{\boldsymbol{\eta},\mathbf{R}})$, is $\left(\mathbf{J}_{\mathbf{L}} - \mathbf{J}_{\mathbf{L}\boldsymbol{\Omega}}\mathbf{J}_{\boldsymbol{\Omega}}^{-1}\mathbf{J}_{\mathbf{L}\boldsymbol{\Omega}}^T\right)^{-1}$. Notice that

$$\begin{aligned}
\mathbf{J}_{\mathbf{L}} - \mathbf{J}_{\mathbf{L}\boldsymbol{\Omega}}\mathbf{J}_{\boldsymbol{\Omega}}^{-1}\mathbf{J}_{\mathbf{L}\boldsymbol{\Omega}}^T &= \boldsymbol{\delta}_2^T\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\delta}_2 - \boldsymbol{\delta}_2^T\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\delta}_4\left(\boldsymbol{\delta}_4^T\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\delta}_4\right)^{\dagger}\boldsymbol{\delta}_4^T\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\delta}_2 \\
&= \boldsymbol{\delta}_2^T\mathbf{J}_{\boldsymbol{h}}^{1/2}\left(\mathbf{I} - \mathbf{P}_{\mathbf{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{\delta}_4}\right)\mathbf{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{\delta}_2 \\
&= \boldsymbol{\delta}_2^T\left(\mathbf{J}_{\boldsymbol{h}} - \mathbf{J}_{\boldsymbol{h}}\boldsymbol{\delta}_4\left(\boldsymbol{\delta}_4^T\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\delta}_4\right)^{\dagger}\boldsymbol{\delta}_4^T\mathbf{J}_{\boldsymbol{h}}\right)\boldsymbol{\delta}_2 \\
&\equiv \boldsymbol{\delta}_2^T\left(\mathbf{J}_{\boldsymbol{h}} - \mathbf{J}_0\right)\boldsymbol{\delta}_2.
\end{aligned}$$

In the second equality above, $\mathbf{P}_{\mathbf{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{\delta}_4}$ is the projection matrix onto $\mathbf{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{\delta}_4$. But it not obvious how the projection might facilitate the calculation here. We look into the last equality, recall that $\boldsymbol{\delta}_4^T = (0, (\mathbf{C}_p(\mathbf{L} \otimes \mathbf{L})\mathbf{E}_{d_X})^T, 0)$ and that $\mathbf{J}_{\boldsymbol{h}}$ was given in (5.1). Therefore,

$$\begin{aligned}
\mathbf{J}_0 &= \mathbf{J}_{\boldsymbol{h}}\boldsymbol{\delta}_4\left(\boldsymbol{\delta}_4^T\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\delta}_4\right)^{\dagger}\boldsymbol{\delta}_4^T\mathbf{J}_{\boldsymbol{h}} \\
&= \mathbf{J}_{\boldsymbol{h}}\boldsymbol{\delta}_4\left(\frac{1}{2}\mathbf{E}_{d_X}^T(\mathbf{L}^T \otimes \mathbf{L}^T)\mathbf{P}_{\mathbf{E}_p}(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{\mathbf{X}}^{-1})\mathbf{P}_{\mathbf{E}_{d_X}}(\mathbf{L} \otimes \mathbf{L})\mathbf{E}_p\right)^{\dagger}\boldsymbol{\delta}_4^T\mathbf{J}_{\boldsymbol{h}} \\
&= \mathbf{J}_{\boldsymbol{h}}\boldsymbol{\delta}_4\left(\frac{1}{2}\mathbf{E}_{d_X}^T(\mathbf{L}^T \otimes \mathbf{L}^T)(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{\mathbf{X}}^{-1})(\mathbf{L} \otimes \mathbf{L})\mathbf{E}_{d_X}\right)^{\dagger}\boldsymbol{\delta}_4^T\mathbf{J}_{\boldsymbol{h}} \\
&= \mathbf{J}_{\boldsymbol{h}}\boldsymbol{\delta}_4\left(\frac{1}{2}\mathbf{E}_{d_X}^T(\boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Omega}^{-1})\mathbf{E}_{d_X}\right)^{\dagger}\boldsymbol{\delta}_4^T\mathbf{J}_{\boldsymbol{h}}.
\end{aligned}$$

The matrix $\boldsymbol{\delta}_4\left(\frac{1}{2}\mathbf{E}_{d_X}^T(\boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Omega}^{-1})\mathbf{E}_{d_X}\right)^{\dagger}\boldsymbol{\delta}_4^T$ has 3×3 blocks. Only its middle $[\cdot]_{22}$ block is nonzero, other 8 blocks are all zeros. Then we compute the middle block as following,

where we let $\mathbf{A}_3 = \mathbf{C}_p(\mathbf{L}\otimes\mathbf{L})\mathbf{E}_{d_X}$.

$$\begin{aligned}
&\left[\boldsymbol{\delta}_4\left(\frac{1}{2}\mathbf{E}_{d_X}^T(\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}^{-1})\mathbf{E}_{d_X}\right)^\dagger\boldsymbol{\delta}_4^T\right]_{22} && \text{(I.5)}\\
=\;& \mathbf{A}_3\left(\frac{1}{2}\mathbf{E}_{d_X}^T(\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}^{-1})\mathbf{E}_{d_X}\right)^\dagger\mathbf{E}_{d_X}^T\mathbf{A}_3^T\\
=\;& 2\mathbf{A}_3\mathbf{E}_{d_X}\left(\mathbf{E}_{d_X}^T(\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}^{-1})\mathbf{E}_{d_X}\right)^\dagger\mathbf{E}_{d_X}^T\mathbf{A}_3^T\\
=\;& 2\mathbf{A}_3\mathbf{P}_{\mathbf{E}_{d_X}}(\boldsymbol{\Omega}\otimes\boldsymbol{\Omega})\mathbf{P}_{\mathbf{E}_{d_X}}\mathbf{A}_3^T && \text{(I.6)}\\
=\;& 2\mathbf{C}_p(\mathbf{L}\boldsymbol{\Omega}\mathbf{L}^T\otimes\mathbf{L}\boldsymbol{\Omega}\mathbf{L}^T)\mathbf{C}_p^T.
\end{aligned}$$

Where the equation (I.6) is obtained using Corollary E.1. in CLC (2010; supplementary material): $\mathbf{E}_{d_X}\in\mathbb{R}^{d_X^2\times d_X(d_X+1)/2}$ is nonsingular and $\mathbf{P}_{\mathbf{E}_l}$ commute with $(\boldsymbol{\Omega}\otimes\boldsymbol{\Omega})$, then $\mathbf{E}_{d_X}\left(\mathbf{E}_{d_X}^T(\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}^{-1})\mathbf{E}_{d_X}\right)^{-1}\mathbf{E}_{d_X}^T=\mathbf{P}_{\mathbf{E}_{d_X}}\left[(\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}^{-1})\right]^{-1}\mathbf{P}_{\mathbf{E}_{d_X}}$.

Now, $\mathbf{J}_0=\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\delta}_4\left(\boldsymbol{\delta}_4^T\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\delta}_4\right)^\dagger\boldsymbol{\delta}_4^T\mathbf{J}_{\boldsymbol{h}}$ only has its $[\cdot]_{22}$ block nonzero. For notational convenience, let $\mathbf{A}=\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\otimes\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}$ and $\mathbf{B}=\mathbf{L}\boldsymbol{\Omega}\mathbf{L}^T\otimes\mathbf{L}\boldsymbol{\Omega}\mathbf{L}^T$. Then

$$\begin{aligned}
[\mathbf{J}_0]_{22} &= \left[\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\delta}_4\left(\boldsymbol{\delta}_4^T\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\delta}_4\right)^\dagger\boldsymbol{\delta}_4^T\mathbf{J}_{\boldsymbol{h}}\right]_{22}=2^{-1}\mathbf{E}_p^T\mathbf{A}\mathbf{E}_p\times2\mathbf{C}_p\mathbf{B}\mathbf{C}_p^T\times2^{-1}\mathbf{E}_p^T\mathbf{A}\mathbf{E}_p\\
&= 2^{-1}\mathbf{E}_p^T\mathbf{A}\mathbf{P}_{\mathbf{E}_p}\mathbf{B}\mathbf{P}_{\mathbf{E}_p}\mathbf{A}\mathbf{E}_p=2^{-1}\mathbf{E}_p^T\mathbf{A}\mathbf{B}\mathbf{A}\mathbf{E}_p=2^{-1}\mathbf{E}_p^T\mathbf{B}\mathbf{E}_p.
\end{aligned}$$

Then, recall that $\boldsymbol{\delta}_2=\begin{pmatrix}\mathbf{R}\boldsymbol{\eta}^T\otimes\mathbf{I}_p\\ 2\mathbf{C}_p\left(\mathbf{L}\boldsymbol{\Omega}\otimes\mathbf{I}_p-\mathbf{L}\otimes\mathbf{L}_0\boldsymbol{\Omega}_0\mathbf{L}_0^T\right)\\ 0\end{pmatrix}$ and,

$$\begin{aligned}
\mathbf{J}_{\mathbf{L}}-\mathbf{J}_{\mathbf{L}\boldsymbol{\Omega}}\mathbf{J}_{\boldsymbol{\Omega}}^{-1}\mathbf{J}_{\mathbf{L}\boldsymbol{\Omega}}^T &= \boldsymbol{\delta}_2^T\mathbf{J}_{\boldsymbol{h}}\boldsymbol{\delta}_2-\boldsymbol{\delta}_2^T\mathbf{J}_0\boldsymbol{\delta}_2\\
&= \left(\boldsymbol{\eta}\mathbf{R}^T\otimes\mathbf{I}_p\right)\left(\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}\otimes\boldsymbol{\Sigma}_{\mathbf{X}}\right)\left(\mathbf{R}\boldsymbol{\eta}^T\otimes\mathbf{I}_p\right)\\
&\quad+2\left(\boldsymbol{\Omega}\mathbf{L}^T\otimes\mathbf{I}_p-\mathbf{L}^T\otimes\mathbf{L}_0\boldsymbol{\Omega}_0\mathbf{L}_0^T\right)\mathbf{P}_{\mathbf{E}_p}\{(\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\otimes\boldsymbol{\Sigma}_{\mathbf{X}}^{-1})-\\
&\quad(\mathbf{L}\boldsymbol{\Omega}^{-1}\mathbf{L}^T\otimes\mathbf{L}\boldsymbol{\Omega}^{-1}\mathbf{L}^T)\}\mathbf{P}_{\mathbf{E}_p}\left(\mathbf{L}\boldsymbol{\Omega}\otimes\mathbf{I}_p-\mathbf{L}\otimes\mathbf{L}_0\boldsymbol{\Omega}_0\mathbf{L}_0^T\right)\\
&= \boldsymbol{\eta}\boldsymbol{\Phi}^{-1}\boldsymbol{\eta}^T\otimes\boldsymbol{\Sigma}_{\mathbf{X}}+2\mathbf{A}^T\mathbf{P}_{\mathbf{E}_p}\mathbf{B}\mathbf{P}_{\mathbf{E}_p}\mathbf{A}.
\end{aligned}$$

The second term $2\mathbf{A}^T\mathbf{P}_{\mathbf{E}_p}\mathbf{B}\mathbf{P}_{\mathbf{E}_p}\mathbf{A}$ is defined by $\mathbf{A}=\mathbf{L}\boldsymbol{\Omega}\otimes\mathbf{I}_p-\mathbf{L}\otimes\mathbf{L}_0\boldsymbol{\Omega}_0\mathbf{L}_0^T$, and

$$\begin{aligned}
\mathbf{B} &= (\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\otimes\boldsymbol{\Sigma}_{\mathbf{X}}^{-1})-(\mathbf{L}\boldsymbol{\Omega}^{-1}\mathbf{L}^T\otimes\mathbf{L}\boldsymbol{\Omega}^{-1}\mathbf{L}^T)\\
&= \mathbf{L}\boldsymbol{\Omega}^{-1}\mathbf{L}^T\otimes\mathbf{L}_0\boldsymbol{\Omega}_0^{-1}\mathbf{L}_0^T+\mathbf{L}_0\boldsymbol{\Omega}_0^{-1}\mathbf{L}_0^T\otimes\mathbf{L}\boldsymbol{\Omega}^{-1}\mathbf{L}^T+\mathbf{L}_0\boldsymbol{\Omega}_0^{-1}\mathbf{L}_0^T\otimes\mathbf{L}_0\boldsymbol{\Omega}_0^{-1}\mathbf{L}_0^T.
\end{aligned}$$

Then we compute this term directly by noticing $\mathbf{P}_{\mathbf{E}_p}=\frac{1}{2}(\mathbf{I}_{p^2}+\mathbf{K}_{p,p})$ and $\mathbf{B}$ commutes with $\mathbf{P}_{\mathbf{E}_p}$, $2\mathbf{A}^T\mathbf{P}_{\mathbf{E}_p}\mathbf{B}\mathbf{P}_{\mathbf{E}_p}\mathbf{A}=2\mathbf{A}^T\mathbf{B}\mathbf{P}_{\mathbf{E}_p}\mathbf{A}=\mathbf{A}^T\mathbf{B}(\mathbf{I}_{p^2}+\mathbf{K}_{p,p})\mathbf{A}$. Then, notice that in

$\mathbf{A}^T\mathbf{B}$ only the first term in $\mathbf{B}$ will not vanish. Hence,

$$\begin{aligned}\mathbf{A}^T\mathbf{B} &= (\mathbf{L}\boldsymbol{\Omega}\otimes\mathbf{I}_p - \mathbf{L}\otimes\mathbf{L}_0\boldsymbol{\Omega}_0\mathbf{L}_0^T)^T\mathbf{L}\boldsymbol{\Omega}^{-1}\mathbf{L}^T\otimes\mathbf{L}_0\boldsymbol{\Omega}_0^{-1}\mathbf{L}_0^T \\ &= \mathbf{L}^T\otimes\mathbf{L}_0\boldsymbol{\Omega}_0^{-1}\mathbf{L}_0^T - \boldsymbol{\Omega}^{-1}\mathbf{L}^T\otimes\mathbf{L}_0\mathbf{L}_0^T.\end{aligned}$$

And,

$$\begin{aligned}\mathbf{A}^T\mathbf{B}(\mathbf{I}_{p^2}+\mathbf{K}_{p,p})\mathbf{A} &= \mathbf{A}^T\mathbf{B}\left(\mathbf{L}\boldsymbol{\Omega}\otimes\mathbf{I}_p - \mathbf{L}\otimes\mathbf{L}_0\boldsymbol{\Omega}_0\mathbf{L}_0^T\right) \\ &\quad +\mathbf{A}^T\mathbf{B}\left(\mathbf{I}_p\otimes\mathbf{L}\boldsymbol{\Omega} - \mathbf{L}_0\boldsymbol{\Omega}_0\mathbf{L}_0^T\otimes\mathbf{L}\right)\mathbf{K}_{l,p} \\ &= \mathbf{A}^T\mathbf{B}\left(\mathbf{L}\boldsymbol{\Omega}\otimes\mathbf{I}_p - \mathbf{L}\otimes\mathbf{L}_0\boldsymbol{\Omega}_0\mathbf{L}_0^T\right) \\ &= \boldsymbol{\Omega}\otimes\mathbf{L}_0\boldsymbol{\Omega}_0^{-1}\mathbf{L}_0^T - \mathbf{I}_{d_X}\otimes\mathbf{L}_0\mathbf{L}_0^T - \mathbf{I}_{d_X}\otimes\mathbf{L}_0\mathbf{L}_0^T + \boldsymbol{\Omega}^{-1}\otimes\mathbf{L}_0\boldsymbol{\Omega}_0\mathbf{L}_0^T \\ &= \boldsymbol{\Omega}\otimes\mathbf{L}_0\boldsymbol{\Omega}_0^{-1}\mathbf{L}_0^T - 2\mathbf{I}_{d_X}\otimes\mathbf{L}_0\mathbf{L}_0^T + \boldsymbol{\Omega}^{-1}\otimes\mathbf{L}_0\boldsymbol{\Omega}_0\mathbf{L}_0^T.\end{aligned}$$

So far, we have found,

$$\begin{aligned}\mathbf{J}_{\mathbf{L}} - \mathbf{J}_{\mathbf{L}\boldsymbol{\Omega}}\mathbf{J}_{\boldsymbol{\Omega}}^{-1}\mathbf{J}_{\mathbf{L}\boldsymbol{\Omega}}^T &= \boldsymbol{\eta}\boldsymbol{\Phi}^{-1}\boldsymbol{\eta}^T\otimes\boldsymbol{\Sigma}_{\mathbf{X}} + 2\mathbf{A}^T\mathbf{P}_{\mathbf{E}_p}\mathbf{B}\mathbf{P}_{\mathbf{E}_p}\mathbf{A} \\ &= \boldsymbol{\eta}\boldsymbol{\Phi}^{-1}\boldsymbol{\eta}^T\otimes\boldsymbol{\Sigma}_{\mathbf{X}} + \mathbf{A}^T\mathbf{B}(\mathbf{I}_{p^2}+\mathbf{K}_{p,p})\mathbf{A} \\ &= \boldsymbol{\eta}\boldsymbol{\Phi}^{-1}\boldsymbol{\eta}^T\otimes\boldsymbol{\Sigma}_{\mathbf{X}} + \boldsymbol{\Omega}\otimes\mathbf{L}_0\boldsymbol{\Omega}_0^{-1}\mathbf{L}_0^T - 2\mathbf{I}_{d_X}\otimes\mathbf{L}_0\mathbf{L}_0^T + \boldsymbol{\Omega}^{-1}\otimes\mathbf{L}_0\boldsymbol{\Omega}_0\mathbf{L}_0^T.\end{aligned}$$

Comparing this expression with (I.2) and (I.3), we have

$$\begin{aligned}\tilde{\boldsymbol{\delta}}_2^T\mathbf{J}_{\boldsymbol{h}}\tilde{\boldsymbol{\delta}}_2 &= (\mathbf{I}_{d_X}\otimes\mathbf{L}_0^T)\left(\mathbf{J}_{\mathbf{L}} - \mathbf{J}_{\mathbf{L}\boldsymbol{\Omega}}\mathbf{J}_{\boldsymbol{\Omega}}^{-1}\mathbf{J}_{\mathbf{L}\boldsymbol{\Omega}}^T\right)(\mathbf{I}_{d_X}\otimes\mathbf{L}_0) \\ &= (\mathbf{I}_{d_X}\otimes\mathbf{L}_0^T)\left(\mathrm{avar}(\sqrt{n}\mathrm{vec}(\widehat{\mathbf{L}}_{\boldsymbol{\eta},\mathbf{R}}))\right)^{\dagger}(\mathbf{I}_{d_X}\otimes\mathbf{L}_0).\end{aligned}$$

Consequently,

$$\begin{aligned}&[\tilde{\boldsymbol{\delta}}_2(\tilde{\boldsymbol{\delta}}_2^T\mathbf{J}_{\boldsymbol{h}}\tilde{\boldsymbol{\delta}}_2)^{\dagger}\tilde{\boldsymbol{\delta}}_2^T]_{\boldsymbol{\beta}\boldsymbol{\beta}} \\ &= (\mathbf{R}\boldsymbol{\eta}^T\otimes\mathbf{L}_0)\left((\mathbf{I}_{d_X}\otimes\mathbf{L}_0^T)\left(\mathrm{avar}(\sqrt{n}\mathrm{vec}(\widehat{\mathbf{L}}_{\boldsymbol{\eta},\mathbf{R}}))\right)^{\dagger}(\mathbf{I}_{d_X}\otimes\mathbf{L}_0)\right)^{\dagger}(\boldsymbol{\eta}\mathbf{R}^T\otimes\mathbf{L}_0^T) \\ &= (\mathbf{R}\boldsymbol{\eta}^T\otimes\mathbf{L}_0\mathbf{L}_0^T)\left(\mathrm{avar}(\sqrt{n}\mathrm{vec}(\widehat{\mathbf{L}}_{\boldsymbol{\eta},\mathbf{R}}))\right)(\boldsymbol{\eta}\mathbf{R}^T\otimes\mathbf{L}_0\mathbf{L}_0^T) \qquad \text{(I.7)} \\ &= \mathrm{avar}\left(\sqrt{n}\mathrm{vec}(\mathbf{Q}_{\mathbf{L}}\widehat{\mathbf{L}}_{\boldsymbol{\eta},\mathbf{R}}\boldsymbol{\eta}\mathbf{R}^T)\right) = \mathrm{avar}\left(\sqrt{n}\mathrm{vec}(\mathbf{Q}_{\mathbf{L}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\eta},\mathbf{R}})\right).\end{aligned}$$

where (I.7) can be obtained using Corollary E.1. in CLC (2010; supplementary material): $\mathbf{I}_{d_X}\otimes\mathbf{L}_0$ has full column rank, and its projection $\mathbf{I}_{d_X}\otimes\mathbf{L}_0\mathbf{L}_0^T$ commute with $\mathbf{C}\equiv\left(\mathrm{avar}(\sqrt{n}\mathrm{vec}(\widehat{\mathbf{L}}_{\boldsymbol{\eta},\mathbf{R}}))\right)^{\dagger} = \left(\mathbf{J}_{\mathbf{L}} - \mathbf{J}_{\mathbf{L}\boldsymbol{\Omega}}\mathbf{J}_{\boldsymbol{\Omega}}^{-1}\mathbf{J}_{\mathbf{L}\boldsymbol{\Omega}}^T\right)$, then

$$\mathbf{I}_{d_X}\otimes\mathbf{L}_0\left(\left(\mathbf{I}_{d_X}\otimes\mathbf{L}_0^T\right)\mathbf{C}\left(\mathbf{I}_{d_X}\otimes\mathbf{L}_0\right)\right)^{-1}\mathbf{I}_{d_X}\otimes\mathbf{L}_0^T = \left(\mathbf{I}_{d_X}\otimes\mathbf{L}_0\mathbf{L}_0^T\right)\mathbf{C}^{-1}\left(\mathbf{I}_{d_X}\otimes\mathbf{L}_0\mathbf{L}_0^T\right).$$

I.6 If $\boldsymbol{\eta}$ and $\mathbf{L}$ are known,

then the asymptotic variance of the MLE of $\mathrm{vec}(\mathbf{R})$, denoted by $\mathrm{vec}(\widehat{\mathbf{R}}_{\boldsymbol{\eta},\mathbf{L}})$, is $\left(\mathbf{J}_{\mathbf{R}} - \mathbf{J}_{\mathbf{R}\boldsymbol{\Phi}}\mathbf{J}_{\boldsymbol{\Phi}}^{-1}\mathbf{J}_{\mathbf{R}\boldsymbol{\Phi}}^{T}\right)^{-1}$. Either by similar computation or by symmetry, we have

$$\begin{aligned}[\tilde{\boldsymbol{\delta}}_3(\tilde{\boldsymbol{\delta}}_3^T\mathbf{J}_{\boldsymbol{h}}\tilde{\boldsymbol{\delta}}_3)^{\dagger}\tilde{\boldsymbol{\delta}}_3^T]_{\boldsymbol{\beta}\boldsymbol{\beta}} &= \mathrm{avar}\left(\sqrt{n}\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\eta},\mathbf{L}}\mathbf{Q}_{\mathbf{R}})\right) = \mathrm{avar}\left(\sqrt{n}\mathrm{vec}(\mathbf{L}\boldsymbol{\eta}\widehat{\mathbf{R}}_{\boldsymbol{\eta},\mathbf{L}}^T\mathbf{Q}_{\mathbf{R}})\right)\\ &= (\mathbf{R}_0\mathbf{R}_0^T \otimes \mathbf{L}\boldsymbol{\eta})\mathrm{avar}\left(\sqrt{n}\mathrm{vec}(\widehat{\mathbf{R}}_{\boldsymbol{\eta},\mathbf{L}}^T)\right)(\mathbf{R}_0\mathbf{R}_0^T \otimes \boldsymbol{\eta}^T\mathbf{L}^T).\end{aligned}$$

J Additional simulations

We used the same models as in Section 7.1. We simulated 1000 data sets for each setting and used 500 for training and the others for testing. We used the known dimension for each method: the true ranks of $\boldsymbol{\beta}$ were used for CCA and RRR, the $\mathbf{X}$-envelope dimensions d_X were used for PLS, and the true envelope dimensions were used for envelope methods. The Frobenius norm $||\boldsymbol{\alpha}||_F$ was used as a measure of overall prediction error. And we also used the Frobenius norm $||\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}||_F$ as a measure of overall parameter estimation accuracy.

Simulation results were summarized in Table 4. We tried different dimensions and different sample sizes for all these models. As expected, the simultaneous envelope estimators always had the best performance among all estimators and always performed substantially better than OLS. For each of the other methods, it could perform as well as the simultaneous envelope in some cases but could have much worse performances in other cases. We did not include the model (R2) because the true envelope dimensions were p and r, which gave no reductions. Interpretations of the results were similar as in Section 7.

References

[1] AMEMIYA, T. (1985), Advanced Econometrics, *Harvard University Press*.

Prediction Criterion $\|\|\boldsymbol{\alpha}\|\|_F$									
(d_X, d_Y, N) or (d, N)		OLS	PLS	CCA	RRR	$\mathbf{X}$-env.	$\mathbf{Y}$-env.	S. env.	S. E.$\leq$
E1	$(3, 3, 100)$	3.2	1.3	0.75	2.3	1.3	**0.57**	**0.48**	0.02
	$(3, 3, 400)$	0.70	0.32	0.14	0.37	0.31	**0.11**	**0.10**	0.004
	$(3, 3, 900)$	0.30	0.14	0.06	0.15	0.13	**0.05**	**0.04**	0.002
E2	$(2, 5, 200)$	1.6	0.55	2.8	1.6	**0.51**	0.76	**0.28**	0.03
	$(5, 3, 200)$	1.9	1.2	11	1.9	1.2	**0.12**	**0.10**	0.04*
	$(7, 4, 200)$	1.9	1.8	44	1.9	1.5	**0.29**	**0.23**	0.04*
R1	$(1, 100)$	3.8	**1.5**	1.83	2.3	2.2	1.7	**1.4**	0.03
	$(3, 100)$	3.8	**2.5**	3.9	2.7	2.8	2.7	**2.6**	0.03
	$(5, 100)$	3.8	**3.0**	4.7	3.3	3.3	3.2	**3.1**	0.04
Estimation Criterion $\|\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|\|_F$									
(d_X, d_Y, N) or (d, N)		OLS	PLS	CCA	RRR	$\mathbf{X}$-env.	$\mathbf{Y}$-env.	S. env.	S. E.$\leq$
E1	$(3, 3, 100)$	253	28	31	142	28	**25**	**11**	2.6
	$(3, 3, 400)$	14	2	**1**	3	2	**1**	**1**	0.1
	$(3, 3, 900)$	2.6	0.3	**0.2**	0.4	0.3	**0.2**	**0.1**	0.02
E2	$(2, 5, 200)$	451	**23**	116	425	127	197	**59**	11
	$(5, 3, 200)$	994	94	80	946	493	**45**	**26**	26
	$(7, 4, 200)$	1152	599	225	1112	876	**154**	**104**	36
R1	$(1, 100)$	73	**30**	36	46	42	34	**27**	0.7
	$(3, 100)$	73	**49**	76	53	54	52	**50**	0.8
	$(5, 100)$	73	**58**	90	64	64	62	**60**	0.8

Table 4: Prediction performances measured by $||\boldsymbol{\alpha}||_F$ and estimation performances measured by $||\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}||_F$, where * in the last column means the corresponding S.E. upper bound were computed without the S.E. of CCA. The best two methods for each setting were emphasized with boldface.