Abstract

We study and establish a foundation for dimension reduction methods that compress the response and predictor vectors in multivariate regression. While all of the methods studied can perform competitively, depending on the characteristics of the regression, using partial least squares to compress the response and predictor vectors was judged to be the best for prediction and parameter estimation.

Keywords: Envelopes, NIPALS, Sufficient dimension reduction, Two-block method.

2020 MSC: Primary 62H12, Secondary 62F12

1. Introduction

1.1. Background

Partial least squares (PLS) regression algorithms like NIPALS [19, 26, 27] and SIMPLS [16] are at their core dimension reduction methods that can enhance the process of regressing a response vector $Y \in \mathbb{R}^r$ on a vector of predictors $X \in \mathbb{R}^p$ based on training data $(Y_i, X_i), i \in \{1, \ldots, n\}$, that are independent observations on random vectors $Y$ and $X$ with a joint density. They are favored across the applied sciences [20], particularly in chemometrics, because with highly collinear predictors their performance is typically better than that of conventional methods like ordinary least squares, and they are serviceable in high-dimensional regressions where $n$ is not sufficient for conventional methods to yield unambiguous answers. Since the pioneering advances by H. Wold [14] and I. Helland [15], there have appeared a multitude of papers explaining, promoting and extending PLS methods. And yet it seems to us that there is still an air of uncertainty over some operational characteristics of PLS regressions, and that some of the traditional PLS knowledge is akin to folklore.

The standard PLS algorithms perform dimension reduction by calculating a matrix of weights $W \in \mathbb{R}^{p \times q}$, with $q < \min(p, n - 1)$, which is then used to compress the predictor vector $X \mapsto W^T X$, at which point attention turns to fitting the regression of $Y$ on the compressed predictors $W^T X$. For instance, suppose we are entertaining the multivariate linear regression model

$$Y = \alpha + \beta^T X + \varepsilon,$$

where $\alpha \in \mathbb{R}^r, \beta \in \mathbb{R}^{p \times r}, E(X) = \mu_X, \text{var}(X) = \Sigma_X, E(\varepsilon) = 0, \text{var}(\varepsilon) = \Sigma_{\varepsilon}, E(Y) = \mu_Y$ and $\text{cov}(X, Y) = \Sigma_{XY}$. Having reduced from $p$ predictors to $q$ compressed predictors $W^T X$, a PLS algorithm fits the linear model

$$Y_i = \alpha + \eta^T (W^T X_i) + \varepsilon_i, i \in \{1, \ldots, n\},$$

in the compressed predictors $W^T X$ by ordinary least squares, assuming that $q < n$ and hopefully $q \ll n$. The resulting estimators of $\eta$ and $\beta$ are then

$$\hat{\eta} = (W^T S_X W)^{-1} W^T S_{XY}$$

$$\hat{\beta} = W \hat{\eta}. $$

*Corresponding author. Email address: liu1815@gmail.com*
where $S_X$ denotes the sample $p \times p$ variance-covariance matrix of the predictors and $S_{X,Y}$ denotes the $p \times r$ matrix of sample covariances between the predictors and the responses. A prediction rule can be formed by substituting $\hat{\beta}$ into (1), or equivalently $\hat{\eta}$ into (2), and estimating $\alpha$ as $\bar{\eta} = Y - \hat{\beta}^T X$.

Cook et al. [11] showed that, with fixed $p$, the subspace spanned by the columns of $W$ is a $\sqrt{n}$-consistent estimator of the $E_X$-envelope of the subspace spanned by the columns of $\beta$. This was an important advance in understanding PLS regression since it gave for the first time a well-defined population parameter that characterizes the method. Using its envelope foundations, Cook and Forzani [6, 7] studied the asymptotic behavior of PLS estimators in regressions with a univariate response ($r = 1$) as $n$ and $p$ diverge in various alignments. They found that PLS estimators can perform well asymptotically. In particular, when the regression is abundant [19], so nearly all predictors contribute corresponding to the largest eigenvalue of the matrix argument.

**Algorithm 1** NIPALS algorithm adapted from Martens and Naes [19] and Stocchero [24]. The $n \times p$ matrix $X$ contains the centered predictors and the $n \times r$ matrix $Y$ contains the centered responses. $\ell_1(\cdot)$ denotes the eigenvector corresponding to the largest eigenvalue of the matrix argument.

```
(a) Full Version
X_1 = X, Y_1 = Y
Select q ≤ \min(\text{rank}(S_X), n - 1)
for d = 1, \ldots, q do
    S_{X,Y} = n^{-1}X_d Y_d'
    w_d = \ell_1(S_{X,Y}X_d Y_d') // Weights
    s_d = X_d w_d // Scores
    l_d = X_d^T s_d // X loadings
    m_d = Y_d^T s_d // Y loadings
    X_{d+1} = X_d - s_d l_d = Q_d X_d // X deflation
    Y_{d+1} = Y_d - s_d m_d = Q_d Y_d // Y deflation
    W_d = (W_{d-1}, w_d) \in \mathbb{R}^{p\times d}, I_d = (I_{d-1}, l_d) \in \mathbb{R}^{p\times d}
    M_d = (M_{d-1}, m_d) \in \mathbb{R}^{n\times d}
end for
return \hat{\beta}_{npl} = W_q (I_q^T W_q)^{-1} M_q'
```

(b) Bare-bones Version

```
X_1 = X
Select q ≤ \min(\text{rank}(S_X), n - 1)
for d = 1, \ldots, q do
    S_{X,Y} = n^{-1}X_d Y_d'
    w_d = \ell_1(S_{X,Y}X_d Y_d') // X weights
    s_d = X_d w_d // X scores
    l_d = X_d^T s_d // X loadings
    X_{d+1} = X_d - s_d l_d = Q_d X_d // X deflation
    W_d = (W_{d-1}, w_d) \in \mathbb{R}^{p\times d}
end for
Set W = W_q
return \hat{\beta}_{npl} = W (W^T S_X W)^{-1} W^T S_{X,Y}.
```

1.2. Simultaneous response-predictor reduction?

```
Algorithm 1 NIPALS algorithm adapted from Martens and Naes [19] and Stocchero [24]. The $n \times p$ matrix $X$ contains the centered predictors and the $n \times r$ matrix $Y$ contains the centered responses. $\ell_1(\cdot)$ denotes the eigenvector corresponding to the largest eigenvalue of the matrix argument.

(a) Full Version
X_1 = X, Y_1 = Y
Select q ≤ \min(\text{rank}(S_X), n - 1)
for d = 1, \ldots, q do
    S_{X,Y} = n^{-1}X_d Y_d'
    w_d = \ell_1(S_{X,Y}X_d Y_d') // Weights
    s_d = X_d w_d // Scores
    l_d = X_d^T s_d // X loadings
    m_d = Y_d^T s_d // Y loadings
    X_{d+1} = X_d - s_d l_d = Q_d X_d // X deflation
    Y_{d+1} = Y_d - s_d m_d = Q_d Y_d // Y deflation
    W_d = (W_{d-1}, w_d) \in \mathbb{R}^{p\times d}, I_d = (I_{d-1}, l_d) \in \mathbb{R}^{p\times d}
    M_d = (M_{d-1}, m_d) \in \mathbb{R}^{n\times d}
end for
return \hat{\beta}_{npl} = W_q (I_q^T W_q)^{-1} M_q'
```

(b) Bare-bones Version

```
X_1 = X
Select q ≤ \min(\text{rank}(S_X), n - 1)
for d = 1, \ldots, q do
    S_{X,Y} = n^{-1}X_d Y_d'
    w_d = \ell_1(S_{X,Y}X_d Y_d') // X weights
    s_d = X_d w_d // X scores
    l_d = X_d^T s_d // X loadings
    X_{d+1} = X_d - s_d l_d = Q_d X_d // X deflation
    W_d = (W_{d-1}, w_d) \in \mathbb{R}^{p\times d}
end for
Set W = W_q
return \hat{\beta}_{npl} = W (W^T S_X W)^{-1} W^T S_{X,Y}.
```
Algorithm 1(a) shows a standard version of the NIPALS regression algorithm. The ‘loading’ and ‘deflation’ designations appear symmetric in X and Y, evidently leading some to think that standard PLS regression algorithms perform dimension reduction in both X and Y. But that is not so. The X-deflation step is essential for performing predictor reduction, while the Y-deflation step, which gives a type of residual, is mainly for computational efficiency [eg. 21]. Using the mutual orthogonality of the scores, it can be shown that the full algorithm in Algorithm 1(a) gives the same answers as the bare-bones algorithm in Algorithm 1(b) where there is no Y deflation.

Since standard PLS algorithms do not perform simultaneous dimension reduction of the predictors and responses, is methodology available for performing such reductions? Two possibilities come to mind. The first is an established procedure called the two-block method, also known as Wold’s Mode A PLS [28]. We argue in Section 4.4 and illustrate by simulation in Section 5.1 that this method has notables limitations. The second possibility is the envelope-based simultaneous reduction method developed by Cook and Zhang [13]. Being based on a normal likelihood, that methodology will do well when J := (X^T, Y^T)^T is (approximately) normally distributed and the sample size is sufficiently large. It is not serviceable in low sample size regressions.

We show in this article how PLS can be implemented straightforwardly to carry out dimension reduction in both the response and the predictors. We discuss predictor reductions in Section 2 from a new and fairly unrestrictive perspective that does not relying on a linear model. Those ideas are applied to response reduction in Section 3. The conclusions from Sections 2 and 3 then lead to our proposed methodology in Section 4. Our simulation results in Section 5.1 indicate that our proposal performs well in low sample size regressions, rather like PLS predictor reductions, and is competitive in settings where the likelihood-based methods of [13] are expected to dominate.

1.3. Notation and Supplements

Let \( \mathbb{R}^{r \times c} \) denote the space of all real \( r \times c \) matrices and let \( \mathbb{S}^{r \times r} \) denote the space of all real symmetric \( r \times r \) matrices. For random vectors \( A \in \mathbb{R}^r \) and \( C \in \mathbb{R}^r \) we use \( \Sigma_{A,C} \in \mathbb{R}^{r \times r} \) to denote the matrix of covariances between the elements of A and the elements of C:

\[
\Sigma_{A,C} = E \left[ (A - E(A))(C - E(C))^T \right].
\]

Similarly, \( \Sigma_A \) and \( \Sigma_C \) denote variance-covariance matrices for A and C:

\[
\Sigma_A = E \left[ (A - E(A))(A - E(A))^T \right].
\]

\( \Sigma_{AC} \) indicates the residual covariance matrix from the population linear regression of A on C. Sample versions of these covariances are indicated by replacing \( \Sigma \) with an \( S \). For instance, \( S_A \) and \( S_{AC} \) are the sample versions of \( \Sigma_A \) and \( \Sigma_{AC} \).

Let \( F \in \mathbb{R}^{r \times c} \) and let \( \mathcal{F} = \text{span}(F) \) denote the subspace of \( \mathbb{R}^r \) spanned by the columns of \( F \). Then \( F \) is called a basis matrix for \( \mathcal{F} \) and the projections onto \( \mathcal{F} \) in the usual inner product will be indicated by either \( P_F \) or \( P_{\mathcal{F}} \). A semi-orthogonal basis matrix has \( F^T F = I \). The projection onto \( \mathcal{F} \) in the \( \Delta \in \mathbb{S}^{r \times r} \) inner product will be indicated by either \( P_{F(\Delta)} \) or \( P_{\mathcal{F}(\Delta)} \). We require that \( \Delta \) be positive definite, which is indicated by \( \Delta > 0 \). This projection can be represented in matrix form as

\[
P_{F(\Delta)} = E(F^T \Delta F)^{-1} F \Delta, \quad (4)
\]

provided \( F^T \Delta F > 0 \). Regardless of the inner product, \( Q(\Delta) = I - P(\Delta) \).

Formal proofs of selected results are in the appendix and additional simulation results are available in an online supplement to this article.

2. Envelopes and PLS for predictor reduction

To introduce a general formulation as a foundation for both envelopes and dimension reduction in PLS regressions, recall that \( J = (X^T, Y^T)^T \) and consider the class of regression problems where \( Y \in \mathbb{R}^q \) and \( X \in \mathbb{R}^p \) have a joint density and it is desirable to reduce the dimension of X to enhance the prediction of Y. The requirement that J have a density with finite fourth moments is maintained throughout this article, without necessarily restating it. This broad characterization covers many of the spectrographic regressions encountered in chemometrics, for example. We restrict consideration to linear reductions of X and initially consider subspaces \( S \subseteq \mathbb{R}^p \) so that

\[
Y \perp X | P_S X \quad (5)
\]
which is equivalent to requiring that $Y \perp Q_S X \mid P_S X$. Subspaces that satisfy this condition are called dimension reduction subspaces. When the intersection of all such subspaces is itself a dimension reduction subspace it is called the central subspace and denoted as $S_{\Pi X}$ [13]. A lot is known about $S_{\Pi X}$, including issues like its existence [20] and methods of estimation. Early work that marks the beginning of the area, which is widely referred to as sufficient dimension reduction (SDR), is available from [1]. A good share of the state of the art is described in a recent monograph by Li [17]. The review by Cook [4] covers some additional topics, including method-based approaches.

We know from experience that it is hard to estimate the central subspace when there is collinearity among the predictors because then it can be difficult to distinguish between $P_S X$ and $Q_S X$. We could always mitigate that problem by requiring a large sample size. Or we could require an additional condition that forces a degree of clarity between $P_S X$ and $Q_S X$: 

$$P_S X \perp Q_S X.$$  

(6)

Let $E_{\Pi X}$ denote the intersection of all subspaces that satisfy (5) and (6). Then $S_{\Pi X} \subseteq E_{\Pi X}$, which means we may pay a price of increased dimensionality for imposing clarity via (6). Conditions (5) and (6) are together equivalent to requiring that $Q_S X$ be totally immaterial to the regression in the sense that these conditions are equivalent to the single condition

$$Q_S X \perp (Y, P_S X).$$  

(7)

This condition is hard to use in methodological development, unless $J$ is normally distributed, because we don’t have practicable implementation methods for measuring independence between $Q_S X$ and $(Y, P_S X)$. Methodological development can be facilitated by relaxing (7) and require only that $Q_S X$ and $(Y, P_S X)$ have no linear relationship,

$$\text{cov}(Q_S X, (Y, P_S X)) = 0.$$  

(8)

Condition (8), which still mitigates the effects of collinearity, brings envelopes into the picture, as described in the following proposition. Let $C_{X,Y} = \text{span}(\Sigma_{X,Y})$.

**Proposition 1.** Condition (8) holds if and only if $S$ is a reducing subspace of $\Sigma_X$ that contains $C_{X,Y}$.

It follows from Proposition [1] that the intersection of all subspace satisfying (8) is the $\Sigma_X$-envelope of $C_{X,Y}$; that is, smallest reducing subspace of $\Sigma_X$ that contains $C_{X,Y}$, which we denote as $E_{\Sigma}(C_{X,Y})$.

To facilitate subsequent development, we next review a useful tool in terms of a general envelope. Let $M \in S^{r \times r}$ be positive definite and let $S$ be a subspace of $\mathbb{R}^r$. Then the $M$-envelope of $\text{span}(S)$, denoted $E_M(S)$, is the smallest reducing subspace of $M$ that contains $S$. The tool we need is given in the following lemma [2, 12, Corollary A.2].

**Lemma 1.**

$$E_M(M^{k} S) = E_M(S) \text{ for all real } k,$$

$$E_M(S) = E_M(S) \text{ for all real } k \neq 0.$$  

The population matrix of coefficients from the linear fit of $Y$ on $X$, without necessarily assuming model [1], can be represented as $\beta = \Sigma^{-1}_X \Sigma_{X,Y}$. It follows immediately from Lemma [1] that the envelope $E_{\Sigma}(C_{X,Y})$ arising from Proposition [1] is equal to the $\Sigma_X$-envelope of $B := \text{span}(\beta)$; that is,

$$E_{\Sigma_X}(C_{X,Y}) = E_{\Sigma}(\text{span}(\Sigma^{-1}_X \Sigma_{X,Y})) = E_{\Sigma}(\text{span}(\beta)) = E_{\Sigma}(B).$$

Let $q = \text{dim}(E_{\Sigma}(B))$.

So far we have imposed no restrictions on $J$, other than it have a density in $\mathbb{R}^{\leftarrow p}$. Adding that we have $n$ independent observations on $J$, the subspace spanned by the columns of the $p \times q$ weight matrix produced by the SIMPLS algorithm provides a $\sqrt{n}$-consistent estimator of the matrix that projects onto $E_{\Sigma}(B)$ [11]. It can be shown that the NIPALS algorithm as described in [19, 23] also provides a $\sqrt{n}$-consistent estimator of $E_{\Sigma}(B)$ for fixed $p$. This tells us that the NIPALS and SIMPLS algorithms are both envelope methods and that they estimate the same envelope by using different methods. The weight matrix $W$ that results from either the SIMPLS or NIPALS algorithms is then used to compress the predictor vector $X \mapsto W^T X$. Regressing the response on the compressed predictors results in a PLS estimator of $\beta$ of the form $\hat{\beta}_{\text{pls}} = W(W^T S_X W)^{-1} W^T S_X Y$, as shown in Algorithm [1b]. Let $\hat{\beta}_{\text{pls}}$ and $\hat{\beta}_{\text{opt}}$ denote
the estimators of $\beta$ from using the NIPALS and SIMPLS weight matrices. Then it is known that $\hat{\beta}_{\text{opt}} = \hat{\beta}_{\text{qpl}}$ if the response is univariate ($r = 1$) or if a single component is used $q = 1$ \cite{16}, but otherwise $\hat{\beta}_{\text{opt}} \neq \hat{\beta}_{\text{qpl}}$. We summarize aspects of the previous discussion in the following proposition, which does not require model (1).

**Proposition 2.** Assume that the data $J$, $i \in \{1, \ldots, n\}$, are independent copies of $J$, and also that the envelope dimension $q$ is known. Let $\mathcal{W}_{\text{npl}}$ and $\mathcal{W}_{\text{opt}}$ denote the subspaces spanned by the columns of the weight matrices obtained by applying $q$ iterations of the NIPALS and SIMPLS algorithm with predictors $X$ and responses $Y$. Then, with $p$ fixed, $P_{\mathcal{W}_{\text{npl}}}$ and $P_{\mathcal{W}_{\text{opt}}}$ are $\sqrt{n}$-consistent estimators of $P_{\mathcal{E}_X(\mathcal{B})}$.

We now bring model (1) into our development. Let $\Phi \in \mathbb{R}^{p \times q}$ be a semi-orthogonal basis matrix for $\mathcal{E}_X(\mathcal{B})$, and let $(\Phi, \Phi_0)$ be an orthogonal matrix. Then since $\mathcal{B} \subseteq \mathcal{E}_X(\mathcal{B})$, the matrix of regression coefficients can be expressed as $\hat{\beta} = \Phi h$, where the matrix $h \in \mathbb{R}^{q \times r}$ gives the coordinates of $\hat{\beta}$ relative to basis $\Phi$. Under the multivariate linear model (1), we have the envelope parameterization

\begin{align}
Y &= \alpha + h^T \Phi^T X + \varepsilon, \\
\Sigma_X &= \Phi \Delta \Phi^T + \Phi_0 \Delta_0 \Phi_0^T, \\
\Sigma_{YX} &= 0,
\end{align}

where $\Delta \in \mathbb{R}^{q \times q}$ and $\Delta_0 \in \mathbb{R}^{(p-q) \times (p-q)}$ are positive definite matrices. Requiring that $J$ follow a multivariate normal distribution, Cook et al. \cite{11} derived the maximum likelihood estimators of the parameters in (1). Since these estimators inherit their properties from general likelihood theory, they can be expected to dominate the NIPALS and SIMPLS estimators with sufficiently large sample sizes. Nevertheless the maximum likelihood estimators are not serviceable when $n < p + r + 2$, while the PLS estimators remain so.

Let $W \in \mathbb{R}^{p \times q}$ denote a weight matrix estimated by using a PLS algorithm or maximum likelihood estimation. Then the estimators of the parameters in model (1) via model (7) are given in the following lemma.

**Lemma 2.** After compressing the centered predictor $X \mapsto W^T X$, the estimators of the parameters in models (1) and (7) are $\bar{\alpha} = \bar{Y}, \bar{h} = S^{-1} W^T S\hat{W} X, \bar{\beta} = \bar{W} S^{-1} W^T S\hat{W} X, \bar{\Delta} = S W^T X, \bar{\Delta}_0 = S W^T X$, where $(W, W_0)$ is an orthogonal matrix, $\Sigma_X = \bar{W} \bar{\Delta} \bar{W}^T + \bar{W}_0 \bar{\Delta}_0 \bar{W}_0^T$ and $\Sigma_{YX} = \bar{S} \bar{W} X$.

Cook \cite{5} Definition 5\) proposed a different approach to implementing (7), which is equivalent to the pair of conditions $Y \perp X | P S X$ and $P S X \perp Q S X$. Replacing the second condition with its covariance form, we arrive at the guiding conditions

(a) $Y \perp X | P S X$ and (b) $\text{cov}(P S X, Q S X) = 0$,

which leads to using the central envelope subspace $\mathcal{E}_X(S_{YX})$ as a means of getting a workable version of (6). That approach takes us well beyond PLS and is outside the scope of this article, but it does have promise beyond linear models.

3. Envelopes and PLS for response reduction

As discussed in Section 1\) little is known about using PLS to reduce the response vector with an ultimate goal of improving estimation of $\beta$ in (1). Taking a lesson from how ordinary least squares fits model (1), PLS1 methods fit a separate regression for each response, filling out an estimator of $\beta$ one column at a time. PLS2 methods treat $Y$ as multivariate using an algorithm like that in Algorithm 1\) While neither class of methods deals with response reduction, progress is possible by adapting the development in Section 2\).

Since the novel development of Section 2\) requires only that $Y$ and $X$ have a joint density, we can reverse their roles and, by following the logic that led to (7), end with the following requirement for reducing $Y$:

\[ Q_{\mathcal{R}} Y \perp (X, P_{\mathcal{R}} Y), \]

where $\mathcal{R} \subseteq \mathbb{R}^r$. As argued for (7), to facilitate implementation we consider linear associations only and require that

\[ \text{cov}(Q_{\mathcal{R}} Y, (X, P_{\mathcal{R}} Y)) = 0, \]

which leads to the following dual of Proposition 1. Its proof follows that for Proposition 1 and is omitted. Let $C_{LX} = \text{span}(\Sigma_{LX})$. 

\[ 5 \]
Lemma 3. Equation (11) holds if and only if \( S \) is a reducing subspace of \( \Sigma_Y \) that contains \( C_{YX} \).

It follows from Proposition 3 that the intersection of all subspaces satisfying (11) is the \( \Sigma_Y \)-envelope of \( C_{YX} \); that is, smallest reducing subspace of \( \Sigma_Y \) that contains \( C_{YX} \), which we denote as \( E_{\Sigma_Y}(C_{YX}) \). The discussion after Proposition 1 might suggest that we are now pursuing the regression of \( X \) on \( Y \). However, our goal has not changed: we are still interested in the regression of \( Y \) on \( X \), and in particular in \( \beta = \Sigma_X^{-1}\Sigma_{YX} \). Let \( B' = \text{span}(\beta^T) \). Then

\[
B' = \text{span}(\Sigma_{YX}\Sigma_X^{-1}) = \text{span}(\Sigma_{YX}) = C_{YX},
\]

and so \( E_{\Sigma_Y}(C_{YX}) = E_{\Sigma_X}(B') \). From this calculation we see that response envelopes require enveloping the row space of \( \beta \), while predictor envelopes require enveloping the column space of \( \beta \). Let \( u = \dim(E_{\Sigma_Y}(C_{YX})) \).

The following result [2, 12, Proposition A.5] allows us to write the envelope \( E_{\Sigma_Y}(C_{YX}) \) using the error covariance matrix from model (1).

Lemma 3.

\[
E_{\Sigma_Y}(C_{YX}) = E_{\Sigma_X}(C_{YX}) = E_{\Sigma_Y}(B').
\]

The envelope form \( E_{\Sigma_Y}(B') \) is the one used by [12] in their original treatment of response envelopes. Lemma 3 implies that the weight matrices from applying NIPALS and SIMPLS with \( X \) as the response vector and \( Y \) as the vector of predictors give \( \sqrt{n} \)-consistent estimators of the response envelope \( E_{\Sigma_Y}(B') \). This implies that PLS algorithms can be used for response reduction by simply interchanging the roles of \( X \) and \( Y \) in the algorithms. For clarity we summarize this result in the following proposition

Proposition 4. Assume that the data \( J_i, i \in [1, \ldots, n] \) are independent copies of \( J \), and also that the envelope dimension \( u \) is known. Let \( G_{\text{eqn}} \) and \( G_{\text{eqd}} \) denote the subspaces spanned by the columns of the weight matrices obtained by applying \( u \) iterations of the NIPALS and SIMPLS algorithm with predictor \( Y \) and response \( X \). Then, with \( p \) fixed, \( P_{\text{eqn}} \) and \( P_{\text{eqd}} \) are \( \sqrt{n} \)-consistent estimators of \( P_{E_{\Sigma_Y}(B')} \).

From this we see that the PLS weight matrices for response and predictor reduction are obtained by running a PLS algorithm twice, once with \( Y \) as the response and once with \( X \) as the response. However, the weight matrices are used differently when estimating \( \beta \) in model (1). To find how to use the weight matrices from response reduction, we parameterize the multivariate linear model (1) in terms of the response envelope. Let \( \Gamma \in \mathbb{R}^{r \times u} \) be a semi-orthogonal basis matrix for \( E_{\Sigma_Y}(B') \) and let \( (\Gamma, \Gamma_0) \in \mathbb{R}^{r \times r} \) be an orthogonal matrix. Since \( B' \in E_{\Sigma_Y}(B') \), we can represent \( \beta^T \) in terms of \( \Gamma: \beta^T = \Gamma \eta \), where \( \eta \in \mathbb{R}^{u \times p} \) is a matrix of the coordinates of \( \beta^T \) in terms of the basis \( \Gamma \). The linear model (1) can be expressed in terms of the response envelope as

\[
\begin{align*}
Y &= \alpha + \Gamma \eta X + \varepsilon, \\
\Sigma_{Y|X} &= \Gamma \Omega \Gamma^T + \Gamma_0 \Omega_0 \Gamma_0^T, \\
\Sigma_X &= 0.
\end{align*}
\]

This model satisfies (11) without assuming normality because

\[
\text{cov}(\Gamma_0^T Y, (X, \Gamma^T Y)) = \text{cov}(\Gamma_0^T \varepsilon, (X, \eta^T X + \Gamma^T \varepsilon) = \text{cov}(\Gamma_0^T \varepsilon, \Gamma^T \varepsilon) = \Gamma^T \Sigma_{Y|X} \Gamma_0 = 0.
\]

If \( \Gamma \) is known then we get estimates \( \hat{\eta} = \hat{\Omega} \) of \( \eta \) and \( \hat{\Omega} \) from the ordinary least squares fit of the linear regression of the compressed responses \( \Gamma^T Y \) on \( X \), and \( \hat{\Omega_0} \) can be estimated as \( S_{\Gamma Y} \). From this fit the estimator of \( \beta \) is \( \hat{\beta}^T = \Gamma \hat{\eta} \).

A basis \( \Gamma \) can be estimated by (a) using a PLS algorithm or (b) by assuming normality for \( J \) and using the maximum likelihood estimation [12]. In either case the estimated semi-orthogonal basis matrix \( G \in \mathbb{R}^{r \times u} \) can be substituted for \( \Gamma \) in the \( \Gamma \)-known case, resulting in a fully estimated model based on \( u \) components. We summarize the estimators in the following lemma.

Lemma 4. After compressing the response \( Y \mapsto \Gamma^T Y \) and assuming that the predictors are centered, the estimators of the parameters in model (12) are

\[
\hat{\alpha} = \hat{Y}, \hat{\eta} = \hat{S}_{G^T Y} S_X^{-1} \hat{\beta} = S_X^{-1} S_{XG^T} \hat{G}^T, \hat{\Omega} = S_{G^T Y|X}, \hat{\Omega_0} = S_{G^T 0}, \text{ where } (G, G_0) \text{ is an orthogonal matrix, } \hat{\Sigma}_X = S_X \text{ and } \hat{\Sigma}_{Y|X} = GS_{G^T Y} G^T + G_0 S_{G^T G_0} G_0^T.
\]
4. Envelopes and PLS for simultaneous predictor-response reduction

4.4. Foundations

For simultaneous reduction of predictors and responses, we first combine (7) and (10), as described in the following proposition.

**Proposition 5.** Define the subspaces \( S \subseteq \mathbb{R}^p \) and \( R \subseteq \mathbb{R}^q \). Then the two conditions

(a) \( Q_S X \perp (Y, P_S X) \) and (b) \( Q_R Y \perp (X, P_R Y) \)

hold if and only if the following two condition hold:

(I) \( Q_R Y \perp Q_S X \) and (II) \( (P_R Y, P_S X) \perp (Q_R Y, Q_S X) \).

To implement Proposition 5 in practice, we replace conditions (I) and (II) with corresponding zero covariance conditions:

\[
(I') \text{ cov}(Q_R Y, Q_S X) = 0 \quad \text{and} \quad (II') \text{ cov}((P_R Y, P_S X), (Q_R Y, Q_S X)) = 0. \tag{13}
\]

These covariance conditions can be deduced also from (8) and (11) which imply that \( Q_S \Sigma_{XY} = \Sigma_{XY} Q_R = 0 \). But this alone would miss the independence underpinning. The covariance conditions in (13) imply straightforwardly that

\[
\mathbf{var} \begin{pmatrix} P_R Y \\ P_S X \\ Q_R Y \\ Q_S X \end{pmatrix} = \begin{pmatrix} P_R \Sigma_Y P_R & P_R \Sigma_{XY} P_S & 0 & 0 \\ P_S \Sigma_{XY} P_R & P_S \Sigma_{XY} P_S & 0 & 0 \\ 0 & 0 & Q_S \Sigma_Y Q_R & 0 \\ 0 & 0 & 0 & Q_S \Sigma_X Q_S \end{pmatrix}. \tag{14}
\]

Then using (14) we get the covariance matrix of \( J = (X^T, Y^T)^T \),

\[
\Sigma_J = \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{XY} & \Sigma_Y \end{pmatrix} = \begin{pmatrix} P_S \Sigma_X P_S + Q_S \Sigma_{XY} Q_S & P_S \Sigma_{XY} P_R \\ P_S \Sigma_{XY} P_R & P_R \Sigma_Y P_R + Q_R \Sigma_Y Q_R \end{pmatrix}, \tag{15}
\]

which is satisfied by \( R = \mathcal{E}_{S_X}(G_{XY}) = \mathcal{E}_{S_Y}(G_Y') \) and \( S = \mathcal{E}_{S_X}(G_{XY}) = \mathcal{E}_{S_Y}(G_Y) \). Assuming that linear model (1) holds we can now express it in terms of our parameterizations in Sections 2 and 3.

\[
Y = \sigma + \Gamma_\eta^T X + \epsilon, \tag{16}
\]

\[
\Sigma_J = \Gamma \Omega \Gamma^T + \Gamma_0 \Omega_0 \Gamma_0^T, \quad \Sigma_X = \Phi \Delta \Phi^T + \Phi_0 \Delta_0 \Phi_0^T,
\]

where \( \beta = \Phi_0 \eta^T \) and \( \eta \) contains the coordinates of \( \beta \) relative to bases \( \Gamma \) and \( \Phi \). In the following sections we discuss estimation under model (16) by maximum likelihood, PLS and a related two-block method.

Let \( G \) and \( W \) denote semi-orthogonal matrices that denote estimated versions of \( \Gamma \) and \( \Phi \). These may come from maximum likelihood estimation, PLS estimation or some other methods. Estimators of the remaining parameters in (16) are given in the following lemma.

**Lemma 5.** After compressing the centered predictor \( X \) \( \mapsto \) \( W^T X \) and the response \( Y \) \( \mapsto \) \( G^T Y \), estimators of the parameters in model (1) via model (16) are \( \hat{\alpha} = \bar{Y}, \hat{\eta} = S_{w^TX}^{-1} S_{w^T GX} G^T, \hat{\beta} = W S_{w^T GX}^{-1} S_{w^T GX} G^T, \hat{\Delta} = S_{w^TX}, \hat{\Delta}_0 = S_{w^TX} W_0, \bar{\Omega} = S_{G^T Y w^T X}, \bar{\Omega}_0 = S_{G^T Y W_0}, \Sigma_{YX} = G \bar{\Omega} G^T + G_0 \bar{\Omega}_0 G_0^T, \) and \( \Sigma_X = W \Delta W^T + W_0 \Delta_0 W_0^T \), where \( (W, W_0) \) and \( (G, G_0) \) are orthogonal matrices.

Using the estimators from Lemma 5 the covariance matrix of the residuals from model (16) is \( S_{res} = n^{-1} \sum_{i=1}^n [(Y_i - \bar{Y}) - G \hat{\eta}^T w^T X_i] [(Y_i - \bar{Y}) - G \hat{\eta}^T w^T X_i]^T = \Sigma_{YX} + P_G S_{Y w^T X} Q_G + Q_G S_{Y w^T X} P_G \). From this we see that \( S_{res} = \Sigma_{YX} \) if and only if \( G \) is a reducing subspace of \( S_{Y w^T X} \). As this holds in the population, it fits well with the basic theory leading to model (16). It also suggests that lack-of-fit diagnostics might be developed from the term \( P_G S_{Y w^T X} Q_G \).
4.2. Likelihood-based estimation

Assuming that \( J \) is normally distributed, Cook and Zhang \cite{13} developed maximum likelihood estimation under model \( (16) \). They also studied asymptotic properties under different scenarios and connections with principal component analysis, canonical correlation analysis and reduced rank regression. They proposed two algorithms for getting estimators \( G \) and \( W \) of the semi-orthogonal bases \( \Gamma \) and \( \Phi \) in model \( (16) \). One is a likelihood-based algorithm that alternates between iterations on two non-convex objective functions for predictor and response reductions. That algorithm is generally reliable, but can be slow and can get caught in a local optimum. Their second algorithm optimizes over one basis vector at a time. It is generally faster and nearly as efficient as the likelihood-based algorithm. Once suitable semi-orthogonal bases have been obtained, the estimators of the remaining parameters are as given in Lemma \[5\].

They also prove that if the data \( J_i, i \in \{1, \ldots, n\} \), are independent observations on \( J \) which has finite fourth moments then jointly the vectorized maximum likelihood estimators of \( \beta, \Sigma_X \) and \( \Sigma_{Y|X} \) are asymptotically distributed as a normal random vector. If \( J \) is normally distributed that asymptotic distribution simplifies considerably.

4.3. PLS estimation

The Cook-Zhang normal likelihood analysis under the joint reduction model \( (16) \) is not serviceable unless the sample size is sufficiently large. This is where PLS comes in. We have see in Proposition \[5\] that at the population level, the requirements for simultaneous reduction of \( X \) and \( Y \) are obtained by combining the requirements for separate reductions. This suggests that we proceed in an analogous fashion using PLS: Obtain the weight matrix \( W \) for predictor compression by running a PLS algorithm, NIPALS or SIMPLS, as usual. Then interchange the roles \( X \) and \( Y \) and run the algorithm again to obtain the weight matrix \( G \) for response compression. Since we are running two separate applications of the NIPALS or SIMPLS algorithms, basic asymptotic behavior of the reductions is as described in Propositions \[2\] and \[4\].

The weight matrices \( W \) and \( G \) can be used in Lemma \[5\] to estimate the remaining parameters. For clarity, NIPALS-type algorithms for constructing \( W \) and \( G \) is shown in Algorithm \[2\]. Algorithm \[2\](a) is the same as the bare-bones version in Algorithm \[1\].

4.4. Two-block estimation

The two-block algorithm is apparently the only PLS method that simultaneously reduces the predictors and the response. A version of the two-block algorithm is given in Algorithm \[3\]. In the simultaneous NIPALS algorithm of Algorithm \[2\] the reduction weights are eigenvectors of the covariance matrices between deflated predictors and the response, and between deflated responses and the predictors, the deflations serving to remove by projection the weight vectors already identified. The two-block algorithm is similar, except response and predictors are both deflated together based on singular vectors. This may be clearer in Proposition \[6\] which gives the population version of the two-block algorithm. Plugging in the sample versions of \( \Sigma_{X,Y}, \Sigma_X \) and \( \Sigma_Y \) in Proposition \[6\] gives the sample version in Algorithm \[3\]. Correspondingly, plugging in the population versions of \( S_{X,Y}, S_X \) and \( S_Y \) in Algorithm \[3\] gives the population version in Proposition \[6\]. This may be clear from the proof of the proposition.

**Proposition 6.** The population version of the two-block algorithm in Algorithm \[3\] Initialize as

\[
(w_1, g_1) = \arg \max_{w^T g = 1} w^T \Sigma_{X,Y} g. \tag{17}
\]

and \( W_1 = (w_1) \) and \( G_1 = (g_1) \). And for \( d \in \{1, 2 \ldots\} \),

\[
(w_{d+1}, g_{d+1}) = \arg \max_{w^T g = 1} w^T Q_{W_d(S_X)}^T \Sigma_{X,Y} Q_{G_d(S_Y)} g, \tag{18}
\]

where \( W_d = (w_1, \ldots, w_d) \) and \( G_d = (g_1, \ldots, g_d) \). Terminate when

\[
w^T Q_{W_d(S_X)}^T \Sigma_{X,Y} Q_{G_d(S_Y)} g = 0
\]

for all \( w \) and \( g \). The value of \( d \) at termination is the number of components, \( m \).
Algorithm 2 NIPALS algorithm for computing the weight matrices $W$ and $G$ for compressing $X$ and $Y$. The $n \times p$ matrix $X$ contains the centered predictors and the $n \times r$ matrix $Y$ contains the centered responses. $\ell_1(\cdot)$ denotes the eigenvector corresponding to the largest eigenvalue of the matrix argument.

(a) Predictor reduction, $W$

\begin{align*}
X_1 &= X \quad \text{// Initialize} \\
\text{Select } q &\leq \min\{\text{rank}(S_X), n - 1\} \\
\text{for } d = 1, \ldots, q &\text{ do} \\
S_{X_d,Y} &= n^{-1}X_d^T Y \\
w_d &= \ell_1(S_{X_d,Y}S_{X_d,Y}^T) \quad \text{// X weights} \\
s_d &= \mathcal{Y}_d w_d \quad \text{// X scores} \\
l_d &= \mathcal{Y}_d^T s_d \mathcal{Y}_d^T s_d \quad \text{// X loadings} \\
X_{d+1} &= X_d - s_d l_d^T = Q_s X_d = X_d Q_{w_d(S_{X_d})} \quad \text{// X deflation} \\
W_d &= (W_{d-1}, w_d) \in \mathbb{R}^{p \times d} \quad \text{// Append}\end{align*}

end for

Set $W = W_q$

(b) Response reduction, $G$

\begin{align*}
Y_1 &= Y \quad \text{// Initialize} \\
\text{Select } u &\leq \min\{\text{rank}(S_Y), n - 1\} \\
\text{for } d = 1, \ldots, q &\text{ do} \\
S_{Y_d,X} &= n^{-1}Y_d^T X \\
g_d &= \ell_1(S_{Y_d,X}S_{Y_d,X}^T) \quad \text{// Y weights} \\
c_d &= \mathcal{Y}_d g_d \quad \text{// Y scores} \\
h_d &= \mathcal{Y}_d^T c_d \mathcal{Y}_d^T c_d \quad \text{// Y loadings} \\
Y_{d+1} &= Y_d - c_d h_d^T = Q_c Y_d = Y_d Q_{g_d(S_{Y_d})} \quad \text{// Y deflation} \\
G_d &= (G_{d-1}, g_d) \in \mathbb{R}^{p \times d} \quad \text{// Append}\end{align*}

end for

Set $G = G_u$

return $\hat{\beta} = WS_{Y_d,X}^{-1}S_{W^T X, Y}G^T Y$ \quad // Compute regression coefficients
Algorithm 3 Two-block algorithm – also known as Wold’s Mode A PLS – for computing the weight matrices $W$ and $G$ for compressing $X$ and $Y$; adapted from Wegelin [25] via [28]. The $n \times p$ matrix $\mathcal{X}$ contains the centered predictors and the $n \times r$ matrix $\mathcal{Y}$ contains the centered responses. $\ell_L(\cdot)$ and $\ell_R(\cdot)$ denote the left and right singular vectors corresponding to the largest singular value of the matrix argument.

$$\mathcal{X}_1 = \mathcal{X}, \mathcal{Y}_1 = \mathcal{Y}$$  
Select $m \leq \min(\text{rank}(\mathcal{X}), \text{rank}(\mathcal{Y}), n-1)$  

for $d = 1, \ldots, m$ do

- $S_{\mathcal{X}_d, \mathcal{Y}_d} = n^{-1} \mathcal{X}_d^T \mathcal{Y}_d$  
  // Sample covariance matrix

X iteration

- $w_d = \ell_L(S_{\mathcal{X}_d, \mathcal{Y}_d})$  
  // $X$ weights
- $s_d = \mathcal{X}_d w_d$  
  // $X$ scores
- $l_d = \mathcal{X}_d^T s_d / s_d^T s_d$  
  // $X$ loadings
- $\mathcal{X}_{d+1} = \mathcal{X}_d - s_d l_d^T = \mathcal{Q}_d \mathcal{X}_d = \mathcal{X}_d \mathcal{Q}_d(S_{\mathcal{X}_d})$  
  // $X$ deflation
- $W_d = (W_{d-1}, w_d) \in \mathbb{R}^{p \times d}$  
  // Append

Y iteration

- $g_d = \ell_R(S_{\mathcal{X}_d, \mathcal{Y}_d})$  
  // $Y$ weights
- $c_d = \mathcal{Y}_d g_d$  
  // $Y$ scores
- $h_d = \mathcal{Y}_d^T g_d / g_d^T g_d$  
  // $Y$ loadings
- $\mathcal{Y}_{d+1} = \mathcal{Y}_d - c_d h_d^T = \mathcal{Q}_d \mathcal{Y}_d = \mathcal{Y}_d \mathcal{Q}_d(S_{\mathcal{Y}_d})$  
  // $Y$ deflation
- $G_d = (G_{d-1}, g_d) \in \mathbb{R}^{p \times d}$  
  // Append

end for

Set $W = W_m, G = G_m$

return $W$ and $G$

The following corollary follows immediately from Proposition [6] but its conclusions warrant separate statements since they highlight the differences with the simultaneous PLS reduction in Algorithm [2].

Corollary 1. In the two-block algorithm the number $m$ of response components is equal to the number $m$ of predictor components and $m = \min(q, u) \leq \min(p, r)$.

This corollary pinpoints one of the main disadvantages of the two-block algorithm; namely, the number of components must be the same for both the responses and predictors, while the number of response and predictor components are not linked in the PLS algorithm of Algorithm [2]. For instance, suppose that $\Gamma = (1, 1, \ldots, 1)^T / \sqrt{r}$ so that only $u = 1$ response component is needed. Without loss of regression information we can reduce model (16) to

$$\Gamma^T Y = \Gamma^T \alpha + \eta^T \Phi \Gamma X + \Gamma^T \varepsilon,$$

$$\text{var} (\Gamma^T \varepsilon) = \Omega,$$

$$\Sigma_X = \Phi \Delta \Phi^T + \Phi_0 \Delta_0 \Phi_0^T.$$

This is a version of model (7) for predictor reduction with one response and the number of components $1 \leq q \leq p$. In contrast, the population two-block algorithm must produce only one $q = 1$ predictor component, potentially missing substantial predictor information.

The two-block algorithm has a connection with envelopes. As shown in the next proposition the spans of the predictor and response weight matrices $W$ and $G$ are contained in the corresponding envelopes, $\text{span}(W) \subseteq E_{\mathcal{E}_X}(\mathcal{B})$ and $\text{span}(G) \subseteq E_{\mathcal{E}_Y}(\mathcal{B}^r)$.

Proposition 7. The subspaces $\text{span}(W)$ and $\text{span}(G)$ found by population version of the two-block algorithm are contained in their respective envelopes:

$$\text{span}(W) \subseteq E_{\mathcal{E}_X}(\mathcal{B}) \text{ and } \text{span}(G) \subseteq E_{\mathcal{E}_Y}(\mathcal{B}^r).$$
5. Empirical observations

5.1. Simulations

In this section we present representative simulation results for comparing five methods for predicting $Y$ and estimating $\beta$ in model (1): (1) ordinary least squares (OLS), (2) the two block method (Section 4.4), (3) the envelope-based likelihood reduction method for simultaneously reducing $X$ and $Y$ (XY-ENV, Section 4.2), (4) PLS for predictor reduction only (X-PLS, Algorithm 1) and (5) the newly proposed PLS method for reducing both $X$ and $Y$ (XY-PLS, Section 4.3).

Extensive simulation results are available in the supplement to this article. We present here only representative results to support our general conclusions:

• Any of the five methods can perform competitively, depending on the regression. For instance, although the two block method does not do well overall, it can outperform the rest when the sample size is modest and the gains from the predictor and response reductions are small, as we observed in Supplement Tables 6 and 7.

• With sufficiently large sample size, XY-ENV performs the best. It is not serviceable when the sample size is smaller than $r + p$.

• The newly proposed method XY-PLS performs the best over all. Its performance is competitive with XY-ENV in large samples and it does not suffer appreciably with relatively small sample sizes. If one wishes to choose a serviceable simultaneous reduction method without fretting over the characteristics of the regression, this is the method to choose from among those studied.

The performance of a method can depend on the response and predictor dimensions $r$ and $p$, the dimensions of the response and predictor envelopes $u$ and $q$, the components $\Delta$, $\Delta_0$, $\Omega$ and $\Omega_0$ of the covariance matrices $\Sigma_{XY}$ and $\Sigma_X$ given at (16), $\beta = \Phi \eta \Gamma^T$ and the distribution of $J$. We felt that a comprehensive study varying all of these systematically was a too tall order, so we focused on the dimensions and the covariance matrices. Given all dimensions, the parameters $\Phi$, $\eta$ and $\Gamma$ were filled out using observations on independent uniform $(0, 1)$ random variables. The distribution of $J$ was taken to be multivariate normal. This gives an edge to the XY-ENV method so we felt comfortable thinking of it as the gold standard in large samples.

Following [13] we set

$$
\Delta = a I_q, \quad \Delta_0 = I_{p-q},
$$

$$
\Omega = b I_u, \quad \Omega_0 = 10 I_{r-u},
$$

where $a$ and $b$ were selected constants. Generally, the effectiveness of predictor envelopes increases with $a$, while the effectiveness of response envelopes increases as $b$ decreases [2, 13]. We compared methods based on the average root mean squared prediction error per response

$$
predMSE = \frac{1}{m} \sum_{i=1}^{m} \left( \frac{1}{r} \sum_{j=1}^{r} (Y_{ij} - \hat{Y}_{ij})^2 \right),
$$

where $\hat{Y}_{ij}$ denotes a fitted value and $Y_{ij}$ is an observations from an independent testing sample of size $m = 1000$ generated independently with the same parameters as the data for the fitted model. To focus on estimation, we also computed the average root mean squared estimation error per response

$$
betaMSE = \frac{1}{p} \sum_{i=1}^{p} \left( \frac{1}{r} \sum_{j=1}^{r} (\beta_{ij} - \hat{\beta}_{ij})^2 \right).
$$

The true dimensions were used for all the methods. Representative results are given in Tables 1–4. Some cells in these tables are missing data because a method ran into rank issues and could not be implement cleanly.

The results in Table 1 are for $p = 50$ predictors with $q = 10$ components and $r = 4$ responses with $u = 3$ components. At $n = 1000$ the XY-ENV had the best performance, as expected, with XY-PLS a close second. The
performance of all methods deteriorated as the sample size decreased. After \( n = 57 \), the rank requirements for OLS and XY-ENV were no longer met and so computing ceased for these methods. At \( n = 50 \), the two-block method deteriorated badly, while XY-PLS continued to be serviceable. The difference between the performances of X-PLS and XY-PLS attests to the potential advantages of response reduction in addition to predictor reduction.

<table>
<thead>
<tr>
<th>( n )</th>
<th>OLS</th>
<th>Two Block</th>
<th>XY-ENV</th>
<th>X-PLS</th>
<th>XY-PLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>2.64</td>
<td>3.52</td>
<td>2.51</td>
<td>2.63</td>
<td>2.53</td>
</tr>
<tr>
<td>100</td>
<td>0.07</td>
<td>0.04</td>
<td>0.00</td>
<td>0.07</td>
<td>0.01</td>
</tr>
<tr>
<td>57</td>
<td>5.17</td>
<td>11.74</td>
<td>2.55</td>
<td>3.39</td>
<td>2.67</td>
</tr>
<tr>
<td>50</td>
<td>0.33</td>
<td>0.11</td>
<td>0.02</td>
<td>0.16</td>
<td>0.04</td>
</tr>
<tr>
<td>25</td>
<td>26.99</td>
<td>20.20</td>
<td>2.68</td>
<td>3.93</td>
<td>2.87</td>
</tr>
<tr>
<td></td>
<td>0.93</td>
<td>0.15</td>
<td>0.07</td>
<td>0.16</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>—</td>
<td>21.12</td>
<td>—</td>
<td>4.16</td>
<td>2.94</td>
</tr>
<tr>
<td></td>
<td>—</td>
<td>0.15</td>
<td>—</td>
<td>0.16</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>—</td>
<td>36.93</td>
<td>—</td>
<td>6.74</td>
<td>4.37</td>
</tr>
<tr>
<td></td>
<td>—</td>
<td>0.21</td>
<td>—</td>
<td>0.17</td>
<td>0.10</td>
</tr>
</tbody>
</table>

The results in Table 2 are for \( p = 30 \) predictors with \( q = 20 \) components and \( r = 50 \) responses with \( u = 2 \) components. At \( n = 1000 \) the XY-ENV again had the best performance, as expected, with XY-PLS a close second. The two-block method does relatively better in this case because there is substantial response reduction possible, while X-PLS does relatively worse because it neglects response reduction. We again conclude that with sufficiently large sample size, XY-ENV performs the best, while XY-PLS is serviceable overall.

The settings for the simulation results in Table 2 are the same as those in Table 1, except \( a \) has been increased from 50 to 500, so predictor reduction becomes more effective, and \( b \) has been increased from 0.01 to 0.05, so response reduction becomes less effective. Now the two-block method does terribly because the advantages of response reduction are less clear to it. The XY-ENV method is still the best for the larger sample sizes and we still judge the XY-ENV method to be the best overall. The performance of X-PLS is relatively better because there is not as much to gain from response reduction. Nevertheless, the advantages of response reduction are still apparent when comparing X-PLS and XY-PLS at \( n = 25 \).

The simulation settings for Table 3 are the same as those for Table 2 except the sample size was fixed at \( n = 85 \) and \( a \) was varied. The XY-ENV method is still judged to be the best over the table, particularly when estimation error is taken into account. The XY-PLS method does better than XY-ENV for the two larger values of \( a \), but does worse for the three smaller values of \( a \). Again, we judge XY-PLS to be a serviceable method. The performance of the two-block method improves as \( a \) decreases.

5.2. Three illustrative data analyses

The slump flow of concrete depends on the components of the concrete: cement, fly ash, slag, water, super plasticizer, coarse aggregate, and fine aggregate, which were measured in kilograms per cubic meter of concrete and comprise the \( p = 7 \) predictors for this regression. The \( r = 3 \) response variables were slump, flow, and 28-day compressive strength. The data set [29] contains 103 multivariate observations, which are comprised of 78 original records and 25 new records that were added later. We used the first 78 records as a training set and the new 25 records as the testing set. Predictions were evaluated against the testing set using predMSE. The dimensions were determined via ten fold cross validation for the two-blocks, XY-PLS and X-PLS. BIC was used to determine the dimensions for
Table 2: Mean square errors of prediction (predMSE) and estimation (betaMSE) for various sample sizes \((n)\) using five fitting methods under the simulation setting \(a = 50, b = 0.01, p = 30, r = 50, q = 20, u = 2\). Dashes mean the method is not serviceable under the indicated sample size.

<table>
<thead>
<tr>
<th></th>
<th>OLS</th>
<th>Two Block</th>
<th>XY-ENV</th>
<th>X-PLS</th>
<th>XY-PLS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>n = 1000</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>predMSE</td>
<td>9.91</td>
<td>9.89</td>
<td>9.62</td>
<td>9.81</td>
<td>9.64</td>
</tr>
<tr>
<td>betaMSE</td>
<td>0.41</td>
<td>0.08</td>
<td>0.00</td>
<td>0.08</td>
<td>0.03</td>
</tr>
<tr>
<td><strong>n = 100</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>predMSE</td>
<td>13.97</td>
<td>12.19</td>
<td>9.72</td>
<td>12.18</td>
<td>9.95</td>
</tr>
<tr>
<td>betaMSE</td>
<td>1.54</td>
<td>0.25</td>
<td>0.03</td>
<td>0.29</td>
<td>0.09</td>
</tr>
<tr>
<td><strong>n = 85</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>predMSE</td>
<td>15.25</td>
<td>12.63</td>
<td>9.84</td>
<td>12.83</td>
<td>10.01</td>
</tr>
<tr>
<td>betaMSE</td>
<td>1.74</td>
<td>0.27</td>
<td>0.06</td>
<td>0.32</td>
<td>0.10</td>
</tr>
<tr>
<td><strong>n = 50</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>predMSE</td>
<td>26.08</td>
<td>14.27</td>
<td>—</td>
<td>16.66</td>
<td>10.33</td>
</tr>
<tr>
<td>betaMSE</td>
<td>2.95</td>
<td>0.34</td>
<td>—</td>
<td>0.47</td>
<td>0.13</td>
</tr>
<tr>
<td><strong>n = 25</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>predMSE</td>
<td>—</td>
<td>16.99</td>
<td>—</td>
<td>44.08</td>
<td>11.68</td>
</tr>
<tr>
<td>betaMSE</td>
<td>—</td>
<td>0.42</td>
<td>—</td>
<td>1.20</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Table 3: Mean square errors of prediction (predMSE) and estimation (betaMSE) for various sample sizes \((n)\) using five fitting methods under the simulation setting \(a = 500, b = 0.05, p = 50, r = 4, q = 10, u = 3\). Dashes mean the method is not serviceable under the indicated sample size.

<table>
<thead>
<tr>
<th></th>
<th>OLS</th>
<th>Two Block</th>
<th>XY-ENV</th>
<th>X-PLS</th>
<th>XY-PLS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>n = 1000</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>predMSE</td>
<td>2.67</td>
<td>12.44</td>
<td>2.55</td>
<td>2.59</td>
<td>2.55</td>
</tr>
<tr>
<td>betaMSE</td>
<td>0.08</td>
<td>0.03</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td><strong>n = 100</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>predMSE</td>
<td>5.23</td>
<td>92.89</td>
<td>2.60</td>
<td>2.82</td>
<td>2.66</td>
</tr>
<tr>
<td>betaMSE</td>
<td>0.33</td>
<td>0.10</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td><strong>n = 57</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>predMSE</td>
<td>27.49</td>
<td>175.82</td>
<td>2.66</td>
<td>3.21</td>
<td>2.78</td>
</tr>
<tr>
<td>betaMSE</td>
<td>0.95</td>
<td>0.14</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td><strong>n = 50</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>predMSE</td>
<td>—</td>
<td>184.20</td>
<td>—</td>
<td>3.31</td>
<td>2.81</td>
</tr>
<tr>
<td>betaMSE</td>
<td>—</td>
<td>0.14</td>
<td>—</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td><strong>n = 25</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>predMSE</td>
<td>—</td>
<td>338.29</td>
<td>—</td>
<td>4.86</td>
<td>3.29</td>
</tr>
<tr>
<td>betaMSE</td>
<td>—</td>
<td>0.20</td>
<td>—</td>
<td>0.04</td>
<td>0.03</td>
</tr>
</tbody>
</table>

XY-ENV. For completeness, we also included the PLS1 method with dimension determined by cross validation. The results shown in Table 3 conform to the general conclusions given at the outset of Section 5.1. The XY-PLS method performs the best with XY-ENV coming in second.

As a second illustration we use data from a study of a low density tubular polyethylene reactor [23]. There are \(n = 56\) multivariate observations, each with \(p = 22\) predictors and \(r = 6\) responses. The predictor variables consist of 20 temperatures measured at equal distances along the reactor together with the wall temperature of the reactor and the feed rate. The responses are the output characteristics of the polymers produced. Because the distributions of the values of all the response variables are highly skewed to the right, the responses were transformed to the log scale and then standardized to have unit variance. Dimensions were determined as described in the concrete example.
Table 4: Mean square errors of prediction (predMSE) and estimation (betaMSE) for various values of \(a\), the common variance of the compressed predictors, using five fitting methods under the simulation setting \(n = 85, b = 0.01, p = 30, r = 50, q = 20, u = 2\).

<table>
<thead>
<tr>
<th></th>
<th>OLS</th>
<th>Two Block</th>
<th>XY-ENV</th>
<th>X-PLS</th>
<th>XY-PLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a = 1000)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>predMSE</td>
<td>15.25</td>
<td>63.33</td>
<td>10.41</td>
<td>12.82</td>
<td>10.00</td>
</tr>
<tr>
<td>betaMSE</td>
<td>1.71</td>
<td>0.26</td>
<td>0.04</td>
<td>0.07</td>
<td>0.02</td>
</tr>
<tr>
<td>(a = 500)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>predMSE</td>
<td>15.25</td>
<td>36.64</td>
<td>10.34</td>
<td>12.82</td>
<td>10.00</td>
</tr>
<tr>
<td>betaMSE</td>
<td>1.71</td>
<td>0.26</td>
<td>0.05</td>
<td>0.10</td>
<td>0.03</td>
</tr>
<tr>
<td>(a = 50)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>predMSE</td>
<td>15.25</td>
<td>12.63</td>
<td>9.84</td>
<td>12.83</td>
<td>10.01</td>
</tr>
<tr>
<td>betaMSE</td>
<td>1.74</td>
<td>0.27</td>
<td>0.06</td>
<td>0.32</td>
<td>0.10</td>
</tr>
<tr>
<td>(a = 5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>predMSE</td>
<td>15.16</td>
<td>10.28</td>
<td>9.73</td>
<td>12.84</td>
<td>10.11</td>
</tr>
<tr>
<td>betaMSE</td>
<td>2.02</td>
<td>0.41</td>
<td>0.07</td>
<td>1.12</td>
<td>0.35</td>
</tr>
<tr>
<td>(a = 0.5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>predMSE</td>
<td>15.16</td>
<td>10.10</td>
<td>9.72</td>
<td>12.86</td>
<td>10.22</td>
</tr>
<tr>
<td>betaMSE</td>
<td>3.83</td>
<td>0.90</td>
<td>0.09</td>
<td>2.81</td>
<td>1.05</td>
</tr>
</tbody>
</table>

Table 5: Prediction errors (predMSE) from five methods determined by cross validation or a holdout sample for three datasets. Cases that were not serviceable are indicated by dashes.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>OLS</th>
<th>Two Block</th>
<th>XY-ENV</th>
<th>X-PLS</th>
<th>XY-PLS</th>
<th>X-PLS1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concrete</td>
<td>223.7</td>
<td>101.5</td>
<td>71.1</td>
<td>81.6</td>
<td>61.8</td>
<td>174.1</td>
</tr>
<tr>
<td>Reactor</td>
<td>2.22</td>
<td>1.14</td>
<td>1.86</td>
<td>0.94</td>
<td>0.93</td>
<td>—</td>
</tr>
<tr>
<td>Biscuit</td>
<td>—</td>
<td>2.18</td>
<td>—</td>
<td>1.73</td>
<td>0.95</td>
<td>1.44</td>
</tr>
</tbody>
</table>

The prediction error predMSE was determined using the leave-one-out method to compare with their results. The results shown in Table 5 suggest that for these data there is little advantage in compressing the responses, although we estimated \(q = 3\). We did not run X-PLS1 for these data.

Our third illustration comes from a near-infrared spectroscopy study on the composition of biscuit dough [22]. The original data has \(n = 39\) samples in a training dataset and 31 samples in a testing set. The two sets were created on separate occasions and are not the result of a random split of a larger dataset. Each sample consists of \(r = 4\) responses, the percentages of fat, sucrose, flour and water, and spectral readings at 700 wavelengths. Cook and Zhang [13] used a subset of these data to illustrate application of XY-ENV. They constructed the data subset by reducing the spectral range to restrict the number of predictors to 20 from a potential of 700. This allowed them to avoid ‘\(n < p\)’ issues and again dimension was chosen by cross-validation. In a separate study, Li et al. [18] reasoned that the leading and trailing wavelengths contain little information, which motivated them to use middle wavelengths and end with \(p = 64\) predictors. Using these 64 predictors we applied the four methods that do not require \(n > p\), which gave the results in the last row of Table 5. We see that XY-PLS again performed the best. Since its performance was better than that of X-PLS, we again see an advantage to compressing the responses.

6. Discussion

It is worth emphasizing one of our conclusions from Section 5.1: XY-PLS is a serviceable method for compressing the response and predictor vectors in regression. Although other methods exhibited somewhat better performance in particular settings, XY-PLS did well overall and was judged to be serviceable across the range of regressions studied. We did not give special emphasis to sparse regressions because in areas like chemometrics where PLS has been used for decades, sparsity is apparently not a notable concern [8, Section 2].
We know a lot about the asymptotic behavior as $n, p \to \infty$ of estimators from X-PLS fits of regressions with a real response ($r = 1$). In abundant regressions where many predictors contribute information on the response, X-PLS estimators can converge at the $\sqrt{n}$-rate without regard to the relationship between $n$ and $p$ \cite{15, 16}. On the other hand, in sparse regressions where very few predictors contribute information on the response, $p/n$ must converge to 0 for consistency.

As far as we know, there is no published work on the asymptotic behavior of estimators from X-PLS fits of regressions with multiple responses ($r > 1$). We conjecture that, with $r$ fixed, estimators from X-PLS fits may behave asymptotically as $n, p \to \infty$ in a manner similar to that from regressions with a real response. In particular, if the regression of each real response on $X$ is abundant then estimators from an X-PLS fit may again converge at the $\sqrt{n}$-rate without regard to the relationship between $n$ and $p$. If some of the individual regressions are sparse, the task of assessing convergence becomes more complicated, as it does if we allow also $r \to \infty$.

Since estimators from $XY$-PLS fits utilize the duality between $X$-PLS and $Y$-PLS compressions (Section 4.3), intuition on the behavior of $X$-PLS estimators may serve as a rough guide to the asymptotic behavior of $XY$-PLS estimators. We conjecture that estimators from an $XY$-PLS fit of an abundant regression will perform well asymptotically, but that their performance will degrade in the presence of sparsity. A detail study along these lines is nontrivial, but is needed for definitive conclusions.

Acknowledgement

Lan Liu’s research is partially sponsored by NSF DMS 1916013 and NIH R01HL155417-01.

References

Proof of Proposition 5. Equation (8) is equivalent to the pair of conditions

(i) \( \text{cov}(Q_S X, Y) = Q_S \Sigma_{XY} = 0 \),

(ii) \( \text{cov}(Q_S X, P_S X) = Q_S \Sigma_{XX} P_S = 0 \).

Condition (i) holds if and only if \( S \subseteq C_{XY} \) and condition (ii) holds if and only if \( S \) can be used to partition \( \Sigma_X \) as

\[ \Sigma_X = P_S \Sigma_X P_S + Q_S \Sigma_X Q_S; \]

that is, if and only if \( S \) is a reducing subspace of \( \Sigma_X \) [2, 12, Proposition A.1]. □

Proof of Proposition 5. For notational convenience, let \( M = (P_R Y, P_S X) \). We first show that conditions (a) and (b) are equivalent to the conditions

\[ (I') \ Q_R Y \perp Q_S X \mid (P_R Y, P_S X) \text{ and } (II) \ (P_R Y, P_S X) \perp (Q_R Y, Q_S X). \]

Assume that conditions (a) and (b) hold. Let \( R \in \mathbb{R}^{p \times m} \) and \( R_0 \in \mathbb{R}^{p \times u} \) be semi-orthogonal basis matrices for \( R \) and its orthogonal complement \( R^\perp \). Then \( (R, R_0) \) is an orthogonal matrix and condition (a) holds if and only if \( Q_S X \perp (R^T Y, R_0^T Y, P_S X) \). Consequently, condition (a) implies that

\[ Q_S X \perp (P_R Y, Q_R Y, P_S X) \Rightarrow Q_S X \perp (M, Q_R Y) \Rightarrow (a1) \ Q_S X \perp Q_R Y \mid M \text{ and } (a2) \ Q_S X \perp M \Rightarrow (a3) \ Q_S X \perp M \mid Q_R Y \text{ and } (a4) \ Q_S X \perp Q_R Y. \]

Similarly, condition (b) implies \( Q_R Y \perp (P_S X, Q_S X, P_R Y) \Rightarrow Q_R Y \perp (M, Q_S X) \Rightarrow (b1) \ Q_S X \perp Q_R Y \mid M \text{ and } (b2) \ Q_R Y \perp M \Rightarrow (b3) \ Q_R Y \perp M \mid Q_S X \text{ and } (b4) \ Q_S X \perp Q_R Y. \) Condition (I') follows immediately from either condition (a1) or condition (b1). Condition (II) is implied by (a3) and (b2). Thus condition (a) and (b) imply conditions (I') and (II).

Assume that conditions (I') and (II) hold. Then these imply that

\[ (I') \ Q_R Y \perp Q_S X \mid M, \ (II1) \ M \perp Q_R Y \text{ and } (II2) \ M \perp Q_S X. \]

Conditions (I') and (II1) imply that \( Q_R Y \perp (M, Q_S X) \), while conditions (I') and (II2) imply that \( Q_S X \perp (M, Q_R Y) \). Replacing \( M \) with its definition, these implications give

\[ Q_R Y \perp (P_S X + Q_S X, P_R Y), \quad Q_S X \perp (P_R Y + Q_R Y, P_S X). \]

The required condition (a) and (b) follow from here since if \( A \) is independent of \( B \) then \( A \) is independent of any non-stochastic function of \( B \). That is, these statements imply that

\[ Q_R Y \perp (P_S X + Q_S X, P_R Y), \quad Q_S X \perp (P_R Y + Q_R Y, P_S X). \]

This established that conditions (I') and (II) are equivalent to conditions (a) and (b).
To establish (a) and (b) with (I) and (II), first assume conditions (a) and (b). Then (I') and (II) hold. But (II) implies that (I) holds if and only if (I') holds. Next, assume that (I') and (II) hold. Then (a) and (b) hold and, again, (II) implies that (I) holds if and only if (I') holds. \( \square \)

**Proof of Proposition 8**

The conclusion holds for \( d = 1 \) since (17) gives the first left and right singular vectors of \( \Sigma_{XY} \), the population version of \( S_{XY} \) in the algorithm. For an arbitrary \( d + 1 \leq m \) we have directly from Algorithm 2 \( S_{X_{d+1}Y_{d+1}} = Q_{w(d)}^T(S_{X_d})S_{X_d}Q_{w(d)}(S_{Y_d}) \). For \( d = 1 \) this implies immediately that \( S_{X_1Y_1} = Q_{w(1)}^T(S_{X_1})S_{X_1}Q_{w(1)}(S_{Y_1}) \). Replacing the sample statistics with their population versions, this confirms (18) holds for \( d = 1 \).

For \( d = 2 \) we explicitly switch to the population version \( \Sigma_{X_2} = Q_{w(2)}^T(S_{X_1})\Sigma_{X_1}Q_{w(2)}(S_{X_1}), = Q_{w(2)}^T(S_{X_2})\Sigma_{X_2}Q_{w(2)}(S_{X_2}), \Sigma_{X_2} = Q_{w(2)}^T(S_{X_2})\Sigma_{X_2}Q_{w(2)}(S_{X_2}), \Sigma_{X_2} = Q_{w(2)}^T(S_{X_2})\Sigma_{X_2}Q_{w(2)}(S_{X_2}). \) It follows that (18) holds for \( d = 1 \). The general result can be demonstrated by induction. \( \square \)

**Proof of Proposition 7**

We show only that \( \text{span}(W) \subseteq \text{span}(\Phi) \). The proof for the response envelope follows similarly.

We first show that \( w_1 \) is the first eigenvector of \( \Sigma_{X_Y} \). Let \( t = \text{rank}(\Sigma_{X_Y}) \leq \min(p, r) \) and let \( \Sigma_{XY} = UDV^T \) denote the compact singular value decomposition of \( \Sigma_{XY} \), where \( U \in \mathbb{R}^{p \times t} \) and \( V \in \mathbb{R}^{t \times r} \) are semi-orthogonal matrices, and \( D \in \mathbb{R}^{t \times t} \) is a positive definite diagonal matrix of positive singular values ordered from largest to smallest. Then

\[
(w_1, g_1) = \arg \max_{w \in \mathbb{R}^p, g \in \mathbb{R}^t} w^T UDV^T g.
\]

It follows that without loss of generality we can constrain \( w \) and \( g \) to be in the column spaces of \( U \) and \( V \): \( w = Uw_s \) and \( v = Vg_s \). Then \((w_1, g_1) = (Uw_s, Vg_s)\), where

\[
(w'_1, g'_1) = \arg \max_{w \in \mathbb{R}^p, g \in \mathbb{R}^t} w^T Dg = (D)_{11}.
\]

It follows that \( w_1 \) is the first left singular vector of \( \Sigma_{XY} \) and \( g_1 \) is the first right singular vector of \( \Sigma_{XY} \). Equivalently, \( w_1 \) and \( g_1 \) are the first eigenvectors of \( \Sigma_{X_Y} \) and \( \Sigma_{X_Y} \).

Because \( w_1 \) is the first eigenvector of \( \Sigma_{X_Y} \), we have \( w_1 \in \text{span}(\Phi) \). Let \( w_i \) be the weight vector for the \( i \)-th step, that is,

\[
w_i = \arg \max_{w \in \mathbb{R}^p} \frac{w^T Q_{W_{w_i}(S_{X_{d+1}})}^T \Sigma_{X_Y} Q_{G_{w_i}(S_{X_{d+1}})}^T} {w^T Q_{W_{w_i}(S_{X_{d+1}})}^T \Sigma_{X_Y} Q_{W_{w_i}(S_{X_{d+1}})}^T} w_i.
\]

Suppose that we have \( \text{span}(W_{w_i}) \subseteq \text{span}(\Phi) \) but \( w_i = \Phi A + \Phi_0 A_0 \), where \( i \leq m, A \in \mathbb{R}^d, A_0 \in \mathbb{R}^{p-t} \) with \( A_0 \neq 0 \), and thus \( w_i \notin \text{span}(\Phi) \). We prove by contradiction that \( \Phi A \sqrt{A} \) achieves a larger value of (19) than \( w_i \). We first show that \( \Phi_0^T Q_{W_{w_i}(S_{X_{d+1}})}^T \Sigma_{X_Y} = 0 \). Because \( \text{span}(W_{w_i}) \subseteq \text{span}(\Phi) \), we have \( \Phi_0^T \Sigma_X W_{w_i} = \Phi_0^T \Sigma_X M = 0 \). Moreover, because \( \Sigma_{XY} = \Phi M \), we have

\[
\Phi_0^T Q_{W_{w_i}(S_{X_{d+1}})}^T \Sigma_{XY} = \Phi_0^T \Sigma_X - \Phi_0^T \Sigma_X W_{w_i} (W_{w_i}^T \Sigma_X W_{w_i})^{-1} W_{w_i}^T \Sigma_X Y = 0.
\]

Because \( w_i^TW_{w_i} = 1 \) and \( A_0 \neq 0 \), we have \( A^T A < 1 \). Let \( V_{w_i} = Q_{W_{w_i}(S_{X_{d+1}})}^T \Sigma_{XY} Q_{G_{w_i}(S_{X_{d+1}})}^T \). Then \( \tilde{w}_i^T V_{w_i} V_{w_i}^T \tilde{w}_i = \frac{1}{A^T A} w_i^T V_{w_i} V_{w_i}^T w_i > w_i^T V_{w_i} V_{w_i}^T w_i \), contradicting the definition of \( w_i \) given at (19) and therefore proving the result.