Qualms about Dimension Reduction Theory: Central Subspace, Envelopes, Variable Selection

David J. Olive*
Southern Illinois University

December 16, 2025

Abstract

Although envelopes, lasso, PLS, SDR, and variable selection are very useful, some qualms about the explanations and theory are given. Balms are needed.

KEY WORDS: Lasso; Model Selection; Outliers; PLS; SDR; Variable Selection

1 INTRODUCTION

Some notation for dimension reduction methods is needed. Let $\boldsymbol{y} = (Y_1, ..., Y_r)^T$ be a vector of r response variables and let $\boldsymbol{x} = (x_1, ..., x_p)^T$ be a vector of p predictor variables. For example, predict Y_1 = mussel muscle mass and Y_2 = mussel shell mass from x_1 = height, x_2 = width, and x_3 = length of the mussel shell. The $r = 2$ and $p = 3$.

Let $\boldsymbol{A}$ be an $k \times c$ matrix and let $\mathcal{A} = span(\boldsymbol{A})$ be the subspace of $\mathbb{R}^k$ spanned by the columns of $\boldsymbol{A}$. For a symmetric $k \times k$ matrix $\boldsymbol{C}$, let $\boldsymbol{C} > 0$ and $\boldsymbol{C} \geq 0$ denote that $\boldsymbol{C}$ is positive definite or positive semidefinite, respectively. Then the projection onto $\mathcal{A}$ in the $\boldsymbol{C}$ inner product is $\boldsymbol{P}_{\mathcal{A}(\boldsymbol{C})} = \boldsymbol{P}_{\boldsymbol{A}(\boldsymbol{C})} = \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{C}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{C}$ provided $\boldsymbol{A}^T\boldsymbol{C}\boldsymbol{A} > 0$. Let $\boldsymbol{Q}_{\mathcal{A}(\boldsymbol{C})} = \boldsymbol{I}_r - \boldsymbol{P}_{\mathcal{A}(\boldsymbol{C})}$ be the orthogonal projection. If $\boldsymbol{C} = \boldsymbol{I}_r$, let $\boldsymbol{P}_{\mathcal{A}} = \boldsymbol{P}_{\boldsymbol{A}} = \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T$.

Let the $p \times p$ covariance matrix of $\boldsymbol{x}$ be $\mathrm{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}} = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))^T]$ and the $p \times 1$ vector $\mathrm{Cov}(\boldsymbol{x}, Y) = \boldsymbol{\Sigma}_{\boldsymbol{x}Y} = E[(\boldsymbol{x} - E(\boldsymbol{x})(Y - E(Y))] = (\mathrm{Cov}(x_1, Y), ..., \mathrm{Cov}(x_p, Y))^T$. Let the estimators be

$$\boldsymbol{S}_{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \text{ and } \boldsymbol{S}_{\boldsymbol{x}\mathrm{Y}} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\boldsymbol{x}_{\mathrm{i}} - \overline{\boldsymbol{x}})(\mathrm{Y_i} - \overline{\mathrm{Y}}).$$

Let $\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{y}}$ be estimated by $\boldsymbol{S}_{\boldsymbol{x}\boldsymbol{y}}$.

*David J. Olive is Professor, School of Mathematical & Statistical Sciences, Southern Illinois University, Carbondale, IL 62901, USA.

The *response envelope* for the regression of $\boldsymbol{y}$ on $\boldsymbol{x}$ is the predictor envelope for the regression of $\boldsymbol{x}$ on $\boldsymbol{y}$. Let $\mathcal{L}$ be a subspace of $\mathbb{R}^p$. Let $\{\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_q\}$ be a basis for $\mathcal{L}$, and let $\{\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_q, \boldsymbol{\gamma}_{q+1}, ..., \boldsymbol{\gamma}_p\}$ be a basis for $\mathbb{R}^p$. Then several dimension reduction methods involve estimating q and $\hat{\boldsymbol{\gamma}}_i$ for $i = 1, ..., q$. Let $\boldsymbol{S} = [\boldsymbol{\gamma}_1 ... \boldsymbol{\gamma}_q]$ be the basis matrix for $\mathcal{L}$. Let $\boldsymbol{P}_{\boldsymbol{S}} = \boldsymbol{P}_{\mathcal{L}}$ and $\boldsymbol{Q}_{\mathcal{L}} = \boldsymbol{Q}_{\boldsymbol{S}}$. Let $\boldsymbol{y} \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{A}\boldsymbol{x}$ indicate that the response vector $\boldsymbol{y}$ is independent of the predictors $\boldsymbol{x}$ given $\boldsymbol{A}\boldsymbol{x}$. Then a subspace $\mathcal{L} \subseteq \mathbb{R}^p$ that satisfies $\boldsymbol{y} \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{P}_{\mathcal{L}}\ \boldsymbol{x}$ is a *dimension reduction subspace* for the regression of $\boldsymbol{y}$ on $\boldsymbol{x}$. Thus the reduced predictors $\boldsymbol{w} = \boldsymbol{P}_{\mathcal{L}}\ \boldsymbol{x}$ hold all of the information that $\boldsymbol{x}$ has about $\boldsymbol{y}$. If the intersection of all dimension reduction subspaces is a dimension reduction subspace, then that subspace is called the *central subspace*, denoted by $\mathcal{L}_{\boldsymbol{y}|\boldsymbol{x}}$.

For a predictor envelope $\mathcal{L}$, we want a) $\boldsymbol{y} \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{P}_{\mathcal{L}}\ \boldsymbol{x}$, and b) $\boldsymbol{P}_{\mathcal{L}}\ \boldsymbol{x} \perp\!\!\!\perp \boldsymbol{Q}_{\mathcal{L}}\ \boldsymbol{x}$. Hence $\boldsymbol{w} = \boldsymbol{P}_{\mathcal{L}}\ \boldsymbol{x}$ is (wanted or) *material* for the regression of $\boldsymbol{y}$ on $\boldsymbol{x}$ while $\boldsymbol{u} = \boldsymbol{Q}_{\mathcal{L}}\ \boldsymbol{x}$ is (unwanted or) *immaterial* for the regression of $\boldsymbol{y}$ on $\boldsymbol{x}$. Then $\mathcal{L} \subseteq \mathbb{R}^p$ reduces $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ if and only if $\text{Cov}(\boldsymbol{P}_{\boldsymbol{S}}\ \boldsymbol{x}, \boldsymbol{Q}_{\boldsymbol{S}}\ \boldsymbol{x}) = \boldsymbol{0}$. Assume $\mathcal{L}_{\boldsymbol{y}|\boldsymbol{x}} \subseteq span(\boldsymbol{\Sigma}_{\boldsymbol{x}})$. The *predictor envelope* (subspace) $\mathcal{E}_{\boldsymbol{x}}$ for the regression of $\boldsymbol{y}$ on $\boldsymbol{x}$ is the intersection of all dimension reduction subspaces that reduce $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ and contain $\mathcal{L}_{\boldsymbol{y}|\boldsymbol{x}}$. Hence $\mathcal{L}_{\boldsymbol{y}|\boldsymbol{x}} \subseteq \mathcal{E}_{\boldsymbol{x}}$. The envelope subspace may be larger than the central subspace, but tends to handle high predictor collinearity better than sufficient dimension reduction (SDR) estimators of the central subspace.

For partial least squares (PLS), PLS1 algorithms are for regression with a univariate response Y, and while PLS2 algorithms are for a multivariate response $\boldsymbol{y}$, and are also PLS1 algorthms. These algorithms produce estimated basis vectors $\hat{\boldsymbol{\gamma}}_i$ sequentially, using the response variables. Let $\hat{\boldsymbol{A}}_k$ be the matrix with ith row $\hat{\boldsymbol{\gamma}}_i^T$ for $i = 1, ..., k$. Let the estimated coefficient matrix $\hat{\boldsymbol{B}}_k = \hat{\boldsymbol{A}}_k^T (\hat{\boldsymbol{A}}_k \boldsymbol{S}_{\boldsymbol{x}} \hat{\boldsymbol{A}}_k^T)^{-1} \hat{\boldsymbol{A}}_k \boldsymbol{S}_{\boldsymbol{x}\boldsymbol{y}}$ for $k = 1, ..., J$ where $k = q$ is especially important. Then $\hat{\boldsymbol{B}}_k$ can also be obtained by the OLS multivariate linear regression of $\boldsymbol{y}$ on $\boldsymbol{w} = \hat{\boldsymbol{A}}_k \boldsymbol{x} = (W_1, ..., W_k)^T$ where $W_i = \hat{\boldsymbol{\gamma}}_i^T \boldsymbol{x}$. $J \leq \min(n-1, p)$ needs to be small enough so that $(\hat{\boldsymbol{A}}_k \boldsymbol{S}_{\boldsymbol{x}} \hat{\boldsymbol{A}}_k^T)^{-1}$ exists. If $\boldsymbol{S}_{\boldsymbol{x}}^{-1}$ exists, then $\hat{\boldsymbol{B}}_k = \boldsymbol{P}_{\hat{\boldsymbol{A}}_k^T (\boldsymbol{S}_{\boldsymbol{x}})} \hat{\boldsymbol{\beta}}_{OLS}$, and if $\hat{\boldsymbol{A}}_p^{-1}$ exists, then $\hat{\boldsymbol{B}}_p = \hat{\boldsymbol{B}}_{OLS} = \boldsymbol{S}_{\boldsymbol{x}}^{-1} \boldsymbol{S}_{\boldsymbol{x}\boldsymbol{y}}$. Replace $\boldsymbol{B}$ by $\boldsymbol{\beta}$ if $r = 1$ so Y is used instead of $\boldsymbol{y}$.

The NIPALS algorithm uses $\hat{\boldsymbol{\gamma}}_1 =$ 1st eigenvector of $\boldsymbol{S}_{\boldsymbol{x}\boldsymbol{y}} \boldsymbol{S}_{\boldsymbol{x}\boldsymbol{y}}^T$. For $k < q$, the SIMPLS algorithm uses $\hat{\boldsymbol{A}}_k^T$ and $\hat{\boldsymbol{\gamma}}_{k+1} = \max_{\boldsymbol{\gamma} \in \mathbb{R}^p} \boldsymbol{\gamma}^T \boldsymbol{S}_{\boldsymbol{x}\boldsymbol{y}} \boldsymbol{S}_{\boldsymbol{x}\boldsymbol{y}}^T \boldsymbol{\gamma} = \max_{\boldsymbol{\gamma} \in \mathbb{R}^p} (\boldsymbol{\gamma}^T \boldsymbol{S}_{\boldsymbol{x}\boldsymbol{y}})^2$ subject to $\boldsymbol{\gamma}^T \boldsymbol{S}_{\boldsymbol{x}} \hat{\boldsymbol{A}}_k^T = \boldsymbol{0}$ and $\boldsymbol{\gamma}^T \boldsymbol{\gamma} = 1$. From canonical correlation analysis, finding $\max_{\boldsymbol{\gamma} \neq \boldsymbol{0}} Cor(\boldsymbol{\gamma}^T \boldsymbol{x}, Y)$ is equivalent to maximizing

$$\max_{\boldsymbol{\gamma} \neq \boldsymbol{0}} \frac{\boldsymbol{\gamma}^T \boldsymbol{\Sigma}_{\boldsymbol{x}Y} \boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T \boldsymbol{\gamma}}{\boldsymbol{\gamma}^T \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\gamma}}$$

which has a maximum at $\boldsymbol{\gamma} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \boldsymbol{\beta}_{OLS}$. This result suggests that for PLS1 algorithms, the $[Cor(\hat{\boldsymbol{\gamma}}_i^T \boldsymbol{x}, Y)]^2$ are fairly high for i near 1. See Olive (2025).

For a univariate response Y, $span(\boldsymbol{A}_k^T)$ and $span(\hat{\boldsymbol{A}}_k^T)$ are the same for the SIMPLS, NIPALS, and HPLS PLS1 algorithms for $k = 1, ..., q$. See Cook and Forzani (2024, p. 96). The Helland algorithm (HPLS) uses $\boldsymbol{\gamma}_i = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{i-1} \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ and $\hat{\boldsymbol{\gamma}}_i = \boldsymbol{S}_{\boldsymbol{x}}^{i-1} \boldsymbol{S}_{\boldsymbol{x}Y}$ with $\boldsymbol{S}_{\boldsymbol{x}}^0 = \boldsymbol{\Sigma}_{\boldsymbol{x}}^0 = \boldsymbol{I}_p$. Thus $\hat{\boldsymbol{\gamma}}_1 = \boldsymbol{S}_{\boldsymbol{x}Y}$ and $W_i = \hat{\boldsymbol{\gamma}}_i^T \boldsymbol{x}$.

Suppose $\boldsymbol{\Sigma_x}$ has eigenvalue eigenvector pairs $(\lambda_1, \boldsymbol{d}_1), ..., (\lambda_p, \boldsymbol{d}_p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. Let the eigenvalue eigenvector pairs of $\boldsymbol{S_x}$ be $(\hat{\lambda}_1, \hat{\boldsymbol{d}}_1), ..., (\hat{\lambda}_p, \hat{\boldsymbol{d}}_p)$ where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$. Then a common principal component analysis (PCA) uses $\hat{\boldsymbol{\gamma}}_i = \hat{\boldsymbol{d}}_i$.

Although envelopes, lasso, PLS, SDR, and variable selection are very useful, there are some qualms about the explanations and theory.

2 Qualms

2.1 Predictor Envelope – PCA Relationship

A very interesting property of the predictor envelope is that if the eigenvalues of $\boldsymbol{\Sigma_x}$ are unique, and if $dim(\mathcal{E}_{\boldsymbol{x}}) = q$, then $\mathcal{E}_{\boldsymbol{x}}$ is spanned by q eigenvectors of $\boldsymbol{\Sigma_x}$. See Cook and Forzani (2024, p. 38). Hence the predictor envelope and principal component analysis (PCA) have some similarities. Then PCA replaces the variables $x_1, ..., x_p$ by $w_1, ..., w_J$ where $w_i = \hat{\boldsymbol{d}}_i^T \boldsymbol{x}$ for $i = 1, ..., J$ where J is chosen using a scree plot or some other method. The estimated predictor envelope replaces the variables $x_1, ..., x_p$ by $w_1, ..., w_q$ where $w_i = \hat{\boldsymbol{d}}_j^T \boldsymbol{x}$ for $i = 1, ..., q$ where q is chosen in some way using the response variables $Y_1, ..., Y_r$. Hence q is often much smaller than J. In particular, let $r = 1$ so there is a single response variable Y. If $\boldsymbol{\Sigma_x} = diag(1, ..., p)$, then the eigenvectors of $\boldsymbol{\Sigma_x}$ are the columns of the identity matrix $\boldsymbol{I}_p$ (up to spans). Suppose the MLR model $Y = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} + e$ holds with $\boldsymbol{\beta} = (1, ..., 1)^T = \boldsymbol{1}$ where the zero mean errors are iid with variance $V(e_i) = \sigma^2$. Then the central subspace $= span(\boldsymbol{\beta})$, but PCA needs all p eigenvectors. This result appears to imply that $q = p$ for the predictor envelope. For PCA, the assumption that $\boldsymbol{\beta} \in \mathbb{R}^p$ is a much weaker assumption than $\boldsymbol{\beta} \in \mathbb{R}^q$ where $1 \leq q < p$. As r increases, the regularity conditions get stronger.

Intuitively, since PCA does not use the response variables to form the $\hat{\boldsymbol{\gamma}}_i = \hat{\boldsymbol{d}}_i$, then SDR methods, response envelopes, and PLS should outperform PCA at least if $n \geq 10p$ and r is small.

Let $W_1, ..., W_p$ be the linear combinations $\hat{\boldsymbol{d}}_k^T \boldsymbol{x}$ ordered with respect to the highest squared correlations $r_1^2 \geq r_2^2 \geq \cdots \geq r_p^2$ where the sample correlation $r_{i,Y} = \text{cor}(x_i, Y)$. Then for the predictor envelope, does $w_i = W_i$ for $i = 1, ..., q$?

The above ordering is marginal maximum likelihood estimator (MMLE) variable selection. See Fan and Lv (2008) and Fan and Song (2010), who give some drawbacks of the method.

For $r > 1$, let $W_1, ..., W_p$ be the linear combinations $\hat{\boldsymbol{\gamma}}_k^T \boldsymbol{x}$ ordered with respect $r_1^2 \geq r_2^2 \geq \cdots \geq r_p^2$ where $r_{i,\boldsymbol{y}}^2 = \frac{1}{r} \sum_{j=1}^{r} [\text{cor}(x_i, Y_j)]^2$ for $i = 1, ..., p$. Let $U_1, ..., U_p$ be the linear combinations $\hat{\boldsymbol{\gamma}}_k^T \boldsymbol{x}$ ordered with respect to $m_1^2 \geq m_2^2 \geq \cdots \geq m_p^2$ where $m_{i,\boldsymbol{y}}^2 = \max_{j=1,...,r} [\text{cor}(x_i, Y_j)]^2$. Hence $m_i^2 = m_{(i),\boldsymbol{y}}^2$, the order statistics of the $m_{i,\boldsymbol{y}}^2$. These two methods of ordering variables appear to be new, although the r_i^2 can be motivated by marginal maximum likelihood estimator (MMLE) variable selection. The Olive (2025) SC scree plot of i versus r_i^2 behaves like a scree plot of i versus the eigenvalues. Hence quantities like $\sum_{i=1}^{j} r_i^2 / \sum_{i=1}^{p} r_i^2$ are of interest for $j = 1, ..., p$, and scree plot techniques could be adapted to choose q for PCA.

2.2 Envelopes and PLS

Cook and Forzani (2024, pp. 83,92,111): to establish the link between envelopes and PLS, and for the theory for $\hat{\boldsymbol{\beta}}_{1PLS}$, a very strong regularity condition for a univariate response is that $\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ is an eigenvector of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, e.g., $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \tau^2 \boldsymbol{I}_p$. The Cook and Forzani (2024, equation (4.1)) model is $Y|\boldsymbol{x} = \alpha_1 + \boldsymbol{\beta}_{1PLS}^T \boldsymbol{x} + e$. OLS is a consistent estimator for such models, hence (4.1) forces $\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \lambda \boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \boldsymbol{\beta}_{1PLS}$. If $\boldsymbol{\Sigma}_{\boldsymbol{x}} = diag(1, ..., p)$, then the eigenvectors of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ are the columns of the identity matrix $\boldsymbol{I}_p$ (up to spans). Hence $\boldsymbol{\beta}_{OLS}$ and $\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ have $p-1$ zero elements and one nonzero element under (4.1).

Note that if $\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ is an eigenvector of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, then for HPLS, $span(\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_p) = span(\boldsymbol{\Sigma}_{\boldsymbol{x}Y})$ which is equal to $span(\boldsymbol{\beta}_{OLS})$ if $\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}$ exists. For $r = 1$, SIMPLS, NIPALS, and HPLS, to get linearly independent population basis vectors, a necessary condition is that $\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ is not an eigenvector of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$.

2.3 What Is Variable Selection Actually Doing?

Currently, this discussion is from Olive (2025b).

Some simple results hold for model selection estimators when the number of predictors p is fixed. Several methods, including PLS, use p linear combinations $\boldsymbol{\eta}_1^T \boldsymbol{x}, ..., \boldsymbol{\eta}_p^T \boldsymbol{x}$. Performing the ordinary least squares (OLS) regression of Y on $(\hat{\boldsymbol{\eta}}_1^T \boldsymbol{x}, \hat{\boldsymbol{\eta}}_2^T \boldsymbol{x}, ..., \hat{\boldsymbol{\eta}}_k^T \boldsymbol{x})$ and a constant gives the k-component estimator.

Consider the OLS regression of Y on a constant and $\boldsymbol{w} = (W_1, ..., W_p)^T$ where, e.g., $W_j = x_j$ or $W_j = \hat{\boldsymbol{\eta}}_j^T \boldsymbol{x}$. Let I index the variables in the model so $I = \{1, 2, 4\}$ means that $\boldsymbol{w}_I = (W_1, W_2, W_4)^T$ was selected. The full model $I = F$ uses all p predictors and the constant with $\boldsymbol{\beta}_F = \boldsymbol{\beta} = \boldsymbol{\beta}_{OLS}$. Let r be the residuals from the full OLS model and let r_I be the residuals from model I that uses $\hat{\boldsymbol{\beta}}_I$ with k predictors including a constant where $2 \leq k \leq p+1$. Olive (2025a) noted that if the Mallows (1973) criterion $C_p(I) \leq 2k$, then the sample correlation

$$cor(r, r_I) \geq \sqrt{1 - \frac{p+1}{n}}. \tag{1.1}$$

Since the correlation gets arbitrarily close to 1 as $n \to \infty$, the model selection estimator and full OLS estimator are estimating the same population parameter $\boldsymbol{\beta}$. This result holds if $\hat{\boldsymbol{\beta}}_{OLS}$ is a consistent estimator of $\boldsymbol{\beta}$: heterogeneity is allowed and the cases do not need to be independent and identically distributed (iid). For moderate sample size n, "weak" predictors W_j will often be omitted as long as $cor(r_I, r)$ stays high.

The PLS literature often assumes (a1): $Y|\boldsymbol{x} = \alpha + \boldsymbol{x}^T \boldsymbol{\beta}_{kPLS} + e$ for some k. If $Y|\boldsymbol{x} = \alpha + \boldsymbol{x}^T \boldsymbol{\beta} + e$, then under mild regularity conditions, $\boldsymbol{\beta} = \boldsymbol{\beta}_{OLS}$. Hence assumption (a1) forces $\boldsymbol{\beta}_{kPLS} = \boldsymbol{\beta}_{OLS}$. For $k = 1$, (a1) forces $\boldsymbol{\beta}_{OLS} = \boldsymbol{\beta}_{1PLS} =$ an eigenvector of the covariance matrix $Cov(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$. Assume instead that the cases $(\boldsymbol{x}_i, Y_i)$ are iid with $E(Y) = \mu_Y$ and $E(\boldsymbol{x}) = \boldsymbol{\mu}_{\boldsymbol{x}}$. Let $Cov(\boldsymbol{x}, Y) = \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$. Then $\boldsymbol{\beta} = \boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$, $\boldsymbol{\beta}_{1PLS} = \theta \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ where θ is a constant, and $\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\beta}$, even when heterogeneity is present. Since $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ is positive definite, $\boldsymbol{\beta} = \boldsymbol{0}$ iff $\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \boldsymbol{0}$. This hypothesis can be tested by applying a one sample test to $\boldsymbol{v}_i = (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(Y_i - \overline{Y})$ for $i = 1, ..., n$. See Olive and Zhang (2025), Olive et al. (2025), and Abid, Quaye, and Olive (2025).

Let an OLS working model for $\hat{\boldsymbol{\beta}}_{kPLS}$ be $Y_i = \alpha_k + \theta_{k1}W_{1i} + \cdots + \theta_{kk}W_{ki} + e_{ki}$ where $W_{ji} = \boldsymbol{x}_i^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{j-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$ with $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^0 = \boldsymbol{\Sigma}_{\boldsymbol{x}}^0 = \boldsymbol{I}_p$. Then $\hat{\boldsymbol{\beta}}_{kPLS} = (\sum_{j=1}^k \hat{\theta}_{kj} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{j-1}) \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$ and $\boldsymbol{\beta}_{kPLS} = (\sum_{j=1}^k \theta_{kj} \boldsymbol{\Sigma}_{\boldsymbol{x}}^{j-1}) \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ (under iid cases) using $Y = \alpha_{kPLS} + \boldsymbol{x}^T \boldsymbol{\beta}_{kPLS} + e_k$. This result suggests that the $\boldsymbol{\beta}_{kPLS}$ are typically different for each $k = 1, ..., p$, but $\boldsymbol{\beta}_{kPLS} = \boldsymbol{\beta}_{qPLS}$ for $k \leq q \leq p$ if $\theta_{pj} = 0$ for $k+1 \leq j \leq p$.

Note that $\boldsymbol{\beta} \in \mathbb{R}^p$ is a much weaker assumption than $\boldsymbol{\beta} \in \mathbb{R}^m$ where $1 \leq m < p$. The model selection result of Equation (1) implies that $\boldsymbol{\beta} \in \mathbb{R}^m$ approximately true in that $\theta_{pj} \approx 0$ for $k^* + 1 \leq j \leq p$ for some $m = k^* < p$. Hence the predictors W_j are "weak" or "almost immaterial" for $m + 1 \leq j \leq p$.

2.4 Oracle Property

2.5 Bigger Model

2.6 MLR with Heterogeneity

2.7 High Dimensional Multivariate Linear Regression

Let the multivariate linear regression model be $\boldsymbol{y} = \boldsymbol{\alpha} + \boldsymbol{B}^T \boldsymbol{x} + \boldsymbol{\epsilon}$ where $\boldsymbol{B} = [\boldsymbol{\beta}_1 \boldsymbol{\beta}_2 \cdots \boldsymbol{\beta}_r]$. As before, let $\hat{\boldsymbol{A}}_k \boldsymbol{x} = \boldsymbol{w} = (W_1, ..., W_k)^T$ where $\hat{\boldsymbol{A}}_k$ is the matrix with ith row $\hat{\boldsymbol{\gamma}}_i^T$ for $i = 1, ..., k$. Let $\boldsymbol{D}_k = [\boldsymbol{\theta}_1 \boldsymbol{\theta}_2 \cdots \boldsymbol{\theta}_k]$. Let, for example, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} = n\boldsymbol{S}_{\boldsymbol{x}}/(n-1)$. Fit the working model

$$\boldsymbol{y} = \boldsymbol{\alpha} + \boldsymbol{D}_k^T \boldsymbol{w} + \boldsymbol{\epsilon} = \boldsymbol{\alpha} + \boldsymbol{D}_k^T \hat{\boldsymbol{A}}_k \boldsymbol{x} + \boldsymbol{\epsilon}$$

with $\hat{\boldsymbol{B}}_k^T = \hat{\boldsymbol{D}}_k^T \hat{\boldsymbol{A}}_k$. Then $\hat{\boldsymbol{D}}_k = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}\boldsymbol{y}}$, and $\hat{\boldsymbol{B}}_k = \hat{\boldsymbol{A}}_k^T \hat{\boldsymbol{D}}_k = \hat{\boldsymbol{A}}_k^T (\hat{\boldsymbol{A}}_k \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} \hat{\boldsymbol{A}}_k^T)^{-1} \hat{\boldsymbol{A}}_k \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\boldsymbol{y}}$. PLS2 algorithms can be used to get $\hat{\boldsymbol{B}}_k$.

Alternatively, let r univariate MLR's be fit to get $\hat{\boldsymbol{B}}_U = \hat{\boldsymbol{A}}_H^T = [\hat{\boldsymbol{\eta}}_1 \hat{\boldsymbol{\eta}}_2 \cdots \hat{\boldsymbol{\eta}}_r]$ and

$$\hat{\boldsymbol{y}} = \hat{\boldsymbol{\alpha}} + \hat{\boldsymbol{B}}_U^T \boldsymbol{x}.$$

Then $\hat{\boldsymbol{A}}_H \boldsymbol{x} = \boldsymbol{w} = (W_1, ..., W_r)^T$ where $\hat{\boldsymbol{A}}_H$ is the matrix with ith row $\hat{\boldsymbol{\eta}}_i^T$ for $i = 1, ..., r$. Let $\boldsymbol{D}_H = [\boldsymbol{\theta}_1 \boldsymbol{\theta}_2 \cdots \boldsymbol{\theta}_r]$, and fit the working model

$$\boldsymbol{y} = \boldsymbol{\alpha} + \boldsymbol{D}_H^T \boldsymbol{w} + \boldsymbol{\epsilon} = \boldsymbol{\alpha} + \boldsymbol{D}_H^T \hat{\boldsymbol{A}}_H \boldsymbol{x} + \boldsymbol{\epsilon}$$

to get $\hat{\boldsymbol{D}}_H = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}\boldsymbol{y}}$, and $\hat{\boldsymbol{B}}_H = \hat{\boldsymbol{A}}_H^T \hat{\boldsymbol{D}}_H = \hat{\boldsymbol{A}}_H^T (\hat{\boldsymbol{A}}_H \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} \hat{\boldsymbol{A}}_H^T)^{-1} \hat{\boldsymbol{A}}_H \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\boldsymbol{y}}$. PLS1 algorithms could be used to get the $\hat{\boldsymbol{\eta}}_i$.

The new estimator $\hat{\boldsymbol{B}}_H$ may be competitive with $\hat{\boldsymbol{B}}_q$ if $n \geq Jr$ with $J \geq 5$, $r \geq 2$, and if the $\hat{\boldsymbol{\eta}}_i$ are not obtained from a working model for a multivariate linear regression. J much larger than 5 may be needed. Univariate lasso estimators could be used to give a competitor for the `glmnet` lasso estimator with "mgaussian."

For a comparison of using r MLR's (e.g. PLS1) and multivariate linear regression of $\boldsymbol{y}$ on $\boldsymbol{w}$ using $\hat{\gamma}_i \boldsymbol{x} = W_i$ (possibly with $\hat{\gamma}_i = \gamma_j = j$th column of the identity matrix $\boldsymbol{I}_p$ so that $W_i = x_j$), see, for example, Cook and Forzani (2024, pp. 106-107).

2.8 Outlier Resistance

To make outlier resistant analogs for many statistical techniques, including envelopes, lasso, PCA, and PLS, suppose the cases are $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$. Find the Euclidean distances D_i of the $\boldsymbol{w}_i$ from the coordinatewise median of the n cases. Find the n_R cases that satisfy $D_j \leq \text{MED}(D_1, ..., D_n) + 5\text{MAD}(D_1, ..., D_n)$, and apply the statistical technique on the n_R cases. For variants, R code, and more explanation, see Olive (2025).

3 CONCLUSIONS

Outliers are important, and outlier resistant methods should be used for low and high dimensional statistics.

References

Abid, A.M., Quaye, P.A., and Olive, D.J. (2025), "A High Dimensional Omnibus Regression Test," *Stats*, 8, 107.

Basa, J., Cook, R.D., Forzani, L., and Marcos, M. (2024), "Asymptotic Distribution of One-Component Partial Least Squares Regression Estimators in High Dimensions," *The Canadian Journal of Statistics*, 52, 118-130.

Cook, R.D., and Forzani, L. (2024), *Partial Least Squares Regression: and Related Dimension Reduction Methods*, Chapman and Hall/CRC, Boca Raton, FL.

Cook, R.D., Helland, I.S., and Su, Z. (2013), "Envelopes and Partial Least Squares Regression," *Journal of the Royal Statistical Society, B*, 75, 851-877.

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space," *Journal of the Royal Statistical Society, B*, 70, 849-911.

Fan, J., and Song, R. (2010), "Sure Independence Screening in Generalized Linear Models with np-Dimensionality," *The Annals of Statistics*, 38, 3217-3841.

Mallows, C. (1973), "Some Comments on C_p," *Technometrics,* 15, 661-676.

Olive, D.J. (2025a), "Some Useful Techniques for High Dimensional Statistics," *Stats,* 8, 60.

Olive, D.J. (2025b), "David J. Olive's Contribution to the Discussion of "On Optimal Linear Prediction" by I. Helland," *Scandinavian Journal of Statistics*, to appear. http://doi.org/10.1111/sjos.70037

Olive, D.J., Alshammari, A.A., Pathiranage, K.G., and Hettige, L.A.W. (2025), "Testing with the One Component Partial Least Squares and the Marginal Maximum Likelihood Estimators," *Communications in Statistics: Theory and Methods*, to appear.

Olive, D.J., and Zhang, L. (2025), "One Component Partial Least Squares, High Dimensional Regression, Data Splitting, and the Multitude of Models," *Communications in Statistics: Theory and Methods*, 54, 130-145.