High Dimensional Dimension Reduction with r Response Variables

David J. Olive, Kasun G. Pathiranage, and Mohanned S. Alsaudi *
Southern Illinois University

April 2, 2026

Abstract

A common regression and classification technique computes linear combinations $W_i = \hat{\boldsymbol{\gamma}}_i^T \boldsymbol{x}$ of the predictors $\boldsymbol{x} = (x_1, ..., x_p)^T$ for $i = 1, ..., p$ where $W_1, ..., W_p$ are ordered in some way. Then the response variables are regressed on $W_1, ..., W_k$ to produce the $k-$component estimator for $k = 1, ..., M$ with $M \leq p$. Examples include envelopes, sufficient dimension reduction estimators, and variable selection estimators with $W_i = x_{i_j}$. Several methods (including principal component regression, partial least squares, forward selection, lasso, unilasso, and the elastic net) can be used in high dimensions where n/p is small and n is the sample size. Examining some of the properties of $k-$component estimators is useful for unifying these dimension reduction procedures. Some new ordering techniques to obtain the W_i are also given.

KEY WORDS: envelopes, marginal maximum likelihood estimator, PCA, PLS, SDR, variable selection.

1 INTRODUCTION

Some important statistical methods include regression, multivariate statistics, and classification. These methods are useful for machine learning, an important part of artificial intelligence.

High dimensional statistics are used when $n < 5p$ where n is the sample size and p is the number of variables. Such a model is *overfitting*: the model does not have enough data to estimate p parameters accurately. Then n tends to be not large enough for the classical statistical method to be useful. A less general definition of high dimensional statistics is that p is large. Sometimes $p > Jn$ with $J \geq 10$ is called ultrahigh dimensional statistics.

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

In high dimensions, it is very difficult to estimate a $p \times 1$ vector $\boldsymbol{\theta}$. This result is a form of "the curse of dimensionality." If a $\sqrt{n}$ consistent estimator of $\boldsymbol{\theta}$ is available, then the squared norm

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 = \sum_{i=1}^{p} (\hat{\theta}_i - \theta_i)^2 \propto p/n. \tag{1}$$

When p is fixed, $p/n \to 0$ as $n \to \infty$ and $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}$. In high dimensions, often the estimator has not been shown to be consistent, except under very strong regularity conditions.

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. Then $r = 2$ and $p = 3$.

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

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 $\boldsymbol{G} = [\boldsymbol{\gamma}_1 ... \boldsymbol{\gamma}_q]$ be the basis matrix for $\mathcal{L}$. 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 $\boldsymbol{y} \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{P}_q\, \boldsymbol{x}$ is a *dimension reduction subspace* for the regression of $\boldsymbol{y}$ on $\boldsymbol{x}$. Thus the reduced predictors $\boldsymbol{w} = \boldsymbol{P}_q\, \boldsymbol{x}$ 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 $\mathcal{L} = \mathcal{E}_{\boldsymbol{x}}$, suppose a) $\boldsymbol{y} \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{P}_q\, \boldsymbol{x}$, and b) $\text{Cov}(\boldsymbol{P}_q\, \boldsymbol{x}, \boldsymbol{Q}_q\, \boldsymbol{x}) = \boldsymbol{0}$. Notation: $\boldsymbol{w} = \boldsymbol{P}_q\, \boldsymbol{x}$ is *material* for the regression of $\boldsymbol{y}$ on $\boldsymbol{x}$ while $\boldsymbol{u} = \boldsymbol{Q}_q\, \boldsymbol{x}$ is *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 $\text{Cov}(\boldsymbol{P}_q\, \boldsymbol{x}, \boldsymbol{Q}_q\, \boldsymbol{x}) = \boldsymbol{0}$. 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. See Cook and Forzani (2024).

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

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} = \boldsymbol{S}_{\boldsymbol{x}} = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T.$$

Let estimators

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\boldsymbol{y}} = \boldsymbol{S}_{\boldsymbol{x}\boldsymbol{y}} = \text{cov}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{y}_i - \overline{\boldsymbol{y}})^T,$$

and

$$\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{xy}} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{y}_i - \overline{\boldsymbol{y}})^T.$$

Let the population correlation $\rho_{ij} = \rho_{x_i,x_j} = \mathrm{Cor}(x_i, x_j)$ and the sample correlation $r_{ij} = r_{x_i,x_j} = \mathrm{cor}(x_i, x_j)$. Let the population correlation matrices $\mathrm{Cor}(\boldsymbol{x}) = \boldsymbol{\rho}_{\boldsymbol{x}} = (\rho_{ij})$ and $\mathrm{Cor}(\boldsymbol{x}, Y) = \boldsymbol{\rho}_{\boldsymbol{x}Y} = (\rho_{x_1,Y}, ..., \rho_{x_p,Y})^T$. Let the sample correlation matrices be $\boldsymbol{R}_{\boldsymbol{x}} = (r_{ij})$ and $\boldsymbol{r}_{\boldsymbol{x}Y} = (r_{x_1,Y}, ..., r_{x_p,Y})^T$. Then $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}$ and $\boldsymbol{R}$ are dispersion estimators, and $(\overline{\boldsymbol{x}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}})$ is an estimator of multivariate location and dispersion.

For partial least squares (PLS), PLS1 algorithms are for regression with a univariate response Y, while PLS2 algorithms are for a multivariate response $\boldsymbol{y}$, and are also PLS1 algorithms if $r = 1$. PLS, SDR, and envelopes algorithms produce estimated basis vectors $\hat{\boldsymbol{\gamma}}_i$ sequentially, using the response variables.

The NIPALS PLS algorithm uses $\hat{\boldsymbol{\gamma}}_1 =$ 1st eigenvector of $\boldsymbol{S}_{\boldsymbol{xy}}\boldsymbol{S}_{\boldsymbol{xy}}^T$. For $k < q$, the SIMPLS algorithm uses $\hat{\boldsymbol{\gamma}}_1$ and $\hat{\boldsymbol{\gamma}}_{k+1} = argmax_{\boldsymbol{\gamma}\in\mathbb{R}^p}\ \boldsymbol{\gamma}^T\boldsymbol{S}_{\boldsymbol{xy}}\boldsymbol{S}_{\boldsymbol{xy}}^T\boldsymbol{\gamma}$ subject to $\boldsymbol{\gamma}^T\boldsymbol{S}_{\boldsymbol{x}}\hat{\boldsymbol{A}}_k^T = \boldsymbol{0}$ and $\boldsymbol{\gamma}^T\boldsymbol{\gamma} = 1$. See Cook and Forzani (2024) and Wold (1975).

Suppose the positive semidefinite dispersion matrix $\boldsymbol{\Sigma}$ 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 $\hat{\boldsymbol{\Sigma}}$ be $(\hat{\lambda}_1, \hat{\boldsymbol{d}}_1), ..., (\hat{\lambda}_p, \hat{\boldsymbol{d}}_p)$ where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$. These vectors are important quantities for principal component analysis (PCA).

The marginal maximum likelihood estimator (MMLE) is due to Fan and Lv (2008) and Fan and Song (2010). This estimator computes the marginal regression, such as the binary logistic regression, of Y on x_j resulting in the estimator $(\hat{\alpha}_{j,M}, \hat{\beta}_{j,M})$ for $j = 1, ..., p$. Then $\hat{\boldsymbol{\beta}}_{MMLE} = (\hat{\beta}_{1,M}, ..., \hat{\beta}_{p,M})^T$. Let $\hat{\sigma}_i$ be the sample standard deviation of x_i where $\hat{\sigma}_i^2$ is the sample variance of x_i. If the MMLE regresses Y on the predictors $v_i = x_i/\hat{\sigma}_i$ standardized to have unit sample variances, then MMLE variable selection keeps the J predictors corresponding to the largest $|\hat{\beta}_i| = |\hat{\beta}_{i,M}|$. Hence $W_i = x_{i_j}$ and $W_1, ..., W_p$ are ordered by $|\hat{\beta}_{i_1}| \geq |\hat{\beta}_{i_2}| \geq \cdots \geq |\hat{\beta}_{i_p}|$. Similar orderings could be produced using other regression estimators, such as ridge regression applied to all of the predictors v_i. Forward selection also orders the predictors.

Principal components regression (PCR), PLS, SDR, envelopes, and several other dimension reduction methods use p linear combinations $\boldsymbol{\gamma}_1^T\boldsymbol{x}, ..., \boldsymbol{\gamma}_p^T\boldsymbol{x}$. Let the jth column of $\boldsymbol{I}_p$ be $\boldsymbol{c}_j = (0, ..., 0, 1, 0, ..., 0)^T$ where the 1 is in the jth position. The $\boldsymbol{c}_j$ form the standard basis vectors for $\mathbb{R}^p$ and $\boldsymbol{c}_j^T\boldsymbol{x} = x_j$. For variable selection estimators, let $W_i = \hat{\boldsymbol{\gamma}}_i^T\boldsymbol{x} = x_{i_j}$ where $\hat{\boldsymbol{\gamma}}_i = \boldsymbol{c}_{i_j}$. For PCA regression, often $\boldsymbol{\gamma}_i(PCA) = \boldsymbol{d}_i$ and $\hat{\boldsymbol{\gamma}}_i(PCA) = \hat{\boldsymbol{d}}_i$, but other orderings $\hat{\boldsymbol{\gamma}}_i(PCA) = \hat{\boldsymbol{d}}_{i_j}$ may work better.

A common regression and classification technique computes linear combinations $W_i = \hat{\boldsymbol{\gamma}}_i^T\boldsymbol{x}$ of the predictors for $i = 1, ..., p$, where $W_1, ..., W_p$ are ordered in some way. The k-component estimator, e.g. the k-component PLS estimator or the k-component PCR estimator, is obtained by regressing Y or $\boldsymbol{y}$ on $W_1, ..., W_k$ where often a constant or constant vector is in the model. Let $k = 1, ..., M$ where $M \leq p$. The model selection estimator chooses one of the k-component estimators, e.g. using cross validation, and may be denoted by $\hat{\boldsymbol{\beta}}_{MSPLS}$ or $\hat{\boldsymbol{\beta}}_{MSPCR}$.

Remark 1. The following results are useful for several regression and covariance estimators. Let $\boldsymbol{w}_i = \boldsymbol{A}_n \boldsymbol{x}_i$ for $i = 1, ..., n$ where $\boldsymbol{A}_n$ is a full rank $k \times p$ matrix with $1 \leq k \leq p$.

a) Let $\boldsymbol{\Sigma}^*$ be $\hat{\boldsymbol{\Sigma}}$ or $\tilde{\boldsymbol{\Sigma}}$. Then $\boldsymbol{\Sigma}^*_{\boldsymbol{w}} = \boldsymbol{A}_n \boldsymbol{\Sigma}^*_{\boldsymbol{x}} \boldsymbol{A}_n^T$, $\boldsymbol{\Sigma}^*_{\boldsymbol{w}Y} = \boldsymbol{A}_n \boldsymbol{\Sigma}^*_{\boldsymbol{x}Y}$, and $\boldsymbol{\Sigma}^*_{\boldsymbol{w}\boldsymbol{y}} = \boldsymbol{A}_n \boldsymbol{\Sigma}^*_{\boldsymbol{x}\boldsymbol{y}}$.

b) If $\boldsymbol{A}_n$ is a constant matrix, then $\boldsymbol{\Sigma}_{\boldsymbol{w}} = \boldsymbol{A}_n \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{A}_n^T$, $\boldsymbol{\Sigma}_{\boldsymbol{w}Y} = \boldsymbol{A}_n \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$, and $\boldsymbol{\Sigma}_{\boldsymbol{w}\boldsymbol{y}} = \boldsymbol{A}_n \boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{y}}$.

c) Let $r > 1$, and let $\boldsymbol{z}_i = \boldsymbol{D}_n \boldsymbol{y}_i$ where $\boldsymbol{D}_n$ is a full rank $k \times r$ matrix with $1 \leq k \leq r$. Then $\boldsymbol{\Sigma}^*_{\boldsymbol{x}\boldsymbol{z}} = \boldsymbol{\Sigma}^*_{\boldsymbol{x}\boldsymbol{y}} \boldsymbol{D}_n^T$. If $\boldsymbol{D}_n$ is a constant matrix, then $\boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{z}} = \boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{y}} \boldsymbol{D}_n^T$.

d) Let $a > 0$ and $b > 0$. Then the sample correlation $\text{cor}(x, y) = \text{cor}(ax, by) = \text{cor}(-ax, -by) = -\text{cor}(-ax, by) = -\text{cor}(ax, -by)$.

Much of the literature uses multiple linear regression or multivariate linear regression as the regression method. Let the multivariate linear regression model be $\boldsymbol{y} = \boldsymbol{\alpha} + \boldsymbol{B}^T \boldsymbol{x} + \boldsymbol{\epsilon}$ where $\boldsymbol{B} = [\boldsymbol{\beta}_1 \boldsymbol{\beta}_2 \cdots \boldsymbol{\beta}_r]$. The ordinary least squares (OLS) estimator $\hat{\boldsymbol{B}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\boldsymbol{y}}$. 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 $\boldsymbol{D}_k = [\boldsymbol{\theta}_1 \boldsymbol{\theta}_2 \cdots \boldsymbol{\theta}_k]$. Fit the working model

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

with $\hat{\boldsymbol{B}}_k^T = \hat{\boldsymbol{D}}_k^T \hat{\boldsymbol{A}}_k$. Then the OLS estimator $\hat{\boldsymbol{D}}_k = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}\boldsymbol{y}}$, and the $k-$component estimator $\hat{\boldsymbol{B}}_k = \hat{\boldsymbol{A}}_k^T \hat{\boldsymbol{D}}_k = \hat{\boldsymbol{A}}_k^T (\hat{\boldsymbol{A}}_k \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} \hat{\boldsymbol{A}}_k^T)^{-1} \hat{\boldsymbol{A}}_k \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\boldsymbol{y}}$ by Remark 1. Here $k = 1, ..., M$ where $M \leq \min(n-2, p)$ needs to be small enough so that $(\hat{\boldsymbol{A}}_k \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} \hat{\boldsymbol{A}}_k^T)^{-1}$ exists. If $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}$ exists, then $\hat{\boldsymbol{B}}_k = \boldsymbol{P}_{\hat{\boldsymbol{A}}_k^T (\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}})} \hat{\boldsymbol{B}}_{OLS}$. Suppose $k = p$ and both $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}$ and $\hat{\boldsymbol{A}}_p^{-1}$ exist. Then $\hat{\boldsymbol{A}}_p^T (\hat{\boldsymbol{A}}_p \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} \hat{\boldsymbol{A}}_p^T)^{-1} \hat{\boldsymbol{A}}_p \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},\boldsymbol{y}} =$

$$\hat{\boldsymbol{B}}_p = \hat{\boldsymbol{A}}_p^T (\hat{\boldsymbol{A}}_p^T)^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} (\hat{\boldsymbol{A}}_p)^{-1} \hat{\boldsymbol{A}}_p \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},\boldsymbol{y}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},\boldsymbol{y}} = \hat{\boldsymbol{B}}_{OLS}. \tag{2}$$

Replace $\boldsymbol{B}$ by $\boldsymbol{\beta}$ if $r = 1$ so Y is used instead of $\boldsymbol{y}$. Thus for multiple linear regression with OLS, let $Y = \alpha + \boldsymbol{x}^T \boldsymbol{\beta} + e$. Let the working model be $Y = \alpha_k + \boldsymbol{\theta}_k^T \boldsymbol{w} + e$. Then the OLS estimator $\hat{\boldsymbol{\theta}}_k = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}Y}$. By Remark 1, the k-component estimator

$$\hat{\boldsymbol{\beta}}_k = \hat{\boldsymbol{A}}_k^T \hat{\boldsymbol{\theta}}_k = \hat{\boldsymbol{A}}_k^T (\hat{\boldsymbol{A}}_k \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} \hat{\boldsymbol{A}}_k^T)^{-1} \hat{\boldsymbol{A}}_k \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y}.$$

Section 2 considers $r = 1$ with response variable Y, reviews variable selection, and Theorem 5 generalizes Equation (2) to many other regression models. Section 3 generalizes some of the Section 2 results to $r > 1$ response variables $\boldsymbol{y}$. Section 4 considers the OPLS estimator, and Section 5 gives some high dimensional tests.

2 Univariate Response Y

For $r = 1$, let Y be a response variable for regression or classification. Important regression models include generalized linear models (GLMs), nonlinear regression, nonparametric regression, and survival regression models. There are n cases $(Y_i, \boldsymbol{x}_i^T)^T$, and for some important models, Y depends on $\boldsymbol{x}$ through the sufficient predictor $SP = \alpha + \boldsymbol{x}^T \boldsymbol{\beta}$.

Let the estimated sufficient predictor $ESP = \hat{\alpha} + \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ where sometimes $\alpha = \hat{\alpha} = 0$. Some important classification models include binary regression, linear discriminant analysis, and quadratic discriminant analysis. A binary regression model is $Y = Y|SP \sim$ binomial$(1, \rho(\text{SP}))$ where $\rho(SP) = P(Y = 1|SP)$. There are many binary regression models, including binary logistic regression, binary probit regression, and support vector machines (SVMs) (with $Z_i = 2Y_i - 1$). See Cook and Weisberg (1999), Nelder and Wedderburn (1972), and James et al. (2021).

Let the multiple linear regression (MLR) model

$$Y_i = \alpha + x_{i,1}\beta_1 + \cdots + x_{i,p}\beta_p + e_i = \alpha + \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i \tag{3}$$

for $i = 1, ..., n$. In matrix form, this model is $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\phi} + \boldsymbol{e}$, where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ is an $n \times (p+1)$ matrix of predictors, $\boldsymbol{\phi} = (\alpha, \boldsymbol{\beta}^T)^T$ is a $(p+1) \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors. Assume that the e_i are independent and identically distributed (iid) with expected value $E(e_i) = 0$ and variance $V(e_i) = \sigma^2$. A multiple linear regression model with heterogeneity has the zero mean e_i independent with $V(e_i) = \sigma_i^2$.

Then the OLS estimators for model (3) are $\hat{\boldsymbol{\phi}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$, $\hat{\alpha}_{OLS} = \overline{Y} - \hat{\boldsymbol{\beta}}_{OLS}^T\overline{\boldsymbol{x}}$, and

$$\hat{\boldsymbol{\beta}}_{OLS} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\hat{\boldsymbol{\eta}}.$$

For a multiple linear regression model with iid cases, $\hat{\boldsymbol{\beta}}_{OLS}$ is a consistent estimator of $\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ under mild regularity conditions, while $\hat{\alpha}_{OLS}$ is a consistent estimator of $E(Y) - \boldsymbol{\beta}_{OLS}^T E(\boldsymbol{x})$.

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 algorithms for $k = 1, ..., q$. See Cook and Forzani (2024, p. 96). 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 $W_i = \hat{\boldsymbol{\gamma}}_i^T\boldsymbol{x}$. There are regularity conditions for this result that have been missed in the literature: if $\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ is an eigenvector of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, then $span(\boldsymbol{A}_p^T) = span(\boldsymbol{A}_k^T) = span(\boldsymbol{\Sigma}_{\boldsymbol{x}Y}) = span(\boldsymbol{\beta}_{OLS})$ for HPLS. The literature has noted that the HPLS W_i can have high multicollinearity.

The next result was given by Chun and Keleş (2010) for $r = 1$, but the proof may be new. So PLS is a model free way to get predictors $W_i = \hat{\boldsymbol{\gamma}}_i^T\boldsymbol{x}$ that are fairly highly correlated with the response variable Y, and the absolute correlations tend to decrease quickly. Let $\hat{V}(X)$ be the sample variance of X.

Theorem 1. For $r = 1$ and $1 < k < q$, the SIMPLS algorithm maximizes

$$Q_S(\boldsymbol{\gamma}^T\boldsymbol{x}) = \hat{V}(\boldsymbol{\gamma}^T\boldsymbol{x})[\text{cor}(\boldsymbol{\gamma}^T\boldsymbol{x}, Y)]^2 \tag{4}$$

subject to $\boldsymbol{\gamma}^T\boldsymbol{S}_{\boldsymbol{x}}\hat{\boldsymbol{A}}_k^T = \boldsymbol{0}$ and $\boldsymbol{\gamma}^T\boldsymbol{\gamma} = 1$.

Proof. By Remark 1, $\boldsymbol{\gamma}^T\boldsymbol{S}_{\boldsymbol{x}Y} = \boldsymbol{S}_{\boldsymbol{\gamma}^T\boldsymbol{x},Y} = \text{cov}(\boldsymbol{\gamma}^T\boldsymbol{x}, Y)$, and

$$\boldsymbol{\gamma}^T\boldsymbol{S}_{\boldsymbol{x}Y}\boldsymbol{S}_{\boldsymbol{x}Y}^T\boldsymbol{\gamma} = [\text{cov}(\boldsymbol{\gamma}^T\boldsymbol{x}, Y)]^2 = [\text{cor}(\boldsymbol{\gamma}^T\boldsymbol{x}, Y)]^2\hat{V}(\boldsymbol{\gamma}^T\boldsymbol{x})\hat{V}(Y).$$

Since $\hat{V}(Y)$ is a constant with respect to $\boldsymbol{\gamma}$, the result follows. □

Given $U_j = \hat{\boldsymbol{\gamma}}_j^T \boldsymbol{x}$, the model free variable importance criterion $Q_S(\hat{\boldsymbol{\gamma}}_j^T \boldsymbol{x})$ can be used to order the U_j in importance. Let $W_1, W_2, ..., W_p$ be the variables U_j ordered with respect to a criterion Q, corresponding to $Q(1) \geq Q(2) \geq \cdots \geq Q(p)$. Then a Q scree plot is a plot of i versus $Q(i)$, and scree plot techniques can be used to choose the number $\hat{q}$ of variables to be used in the regression. For example, let

$$R_k = \frac{\sum_{i=1}^{k} Q(i)}{\sum_{i=1}^{p} Q(i)} \tag{5}$$

for $k = 1, ..., p$. Let $D =$ smallest value of k such that $R_k \geq c$, e.g, $c = 0.975$. Then use $\hat{k} = \hat{q} = \min(D, n-2, p)$. This technique can be much faster than using 5-fold cross validation.

Another model free variable importance criterion is an MMLE criterion

$$Q_M(\hat{\boldsymbol{\gamma}}_i^T \boldsymbol{x}) = [\text{cor}(\hat{\boldsymbol{\gamma}}_i^T \boldsymbol{x}, Y)]^2 = r_i^2. \tag{6}$$

If $\boldsymbol{u}_i = \boldsymbol{S}_{\boldsymbol{x}}^{-1/2} \boldsymbol{x}_i$ and $n \geq 5p$, then applying SIMPLS to the $(\boldsymbol{u}_i, Y_i)$ minimizes $[\text{cor}(\boldsymbol{\phi}^T \boldsymbol{u}, Y)]^2$ if $\boldsymbol{S}_{\boldsymbol{x}}$ is nonsingular. If $\hat{\boldsymbol{\phi}}_i^T \boldsymbol{u} = \hat{\boldsymbol{\gamma}}_i^T \boldsymbol{x}$ are the linear combinations, then $\hat{\boldsymbol{\gamma}}_i^T = \hat{\boldsymbol{\phi}}_i^T \boldsymbol{S}_{\boldsymbol{x}}^{-1/2}$.

PCA regression uses $U_j = \hat{\boldsymbol{d}}_j^T \boldsymbol{x}$, and PCR is the special case for multiple linear regression. Using $Q(i) = \hat{\lambda}_i$ is a common, but rather poor, choice for PCA regression. Olive (2025) suggested using $Q(i) = r_i^2$ and the Q_M scree plot = SC scree plot.

To see why (6) is called an MMLE criterion, let $v_i = x_i / \hat{\sigma}_i$ be the standardized x_i. Then the MMLE for multiple linear regression uses the OLS regression of Y on v_i to get $|\hat{\beta}_i| = |\text{cov}(v_i, Y)| / \hat{V}(Y) =$

$$\frac{|\text{cov}(x_i, Y)|}{\sqrt{\hat{\sigma}_i}\ \hat{V}(Y)} = \frac{|\text{cor}(x_i, Y)| \left[\sqrt{\hat{\sigma}_i}\sqrt{\hat{V}(Y)}\right]}{\sqrt{\hat{\sigma}_i}\ \hat{V}(Y)} = \frac{|\text{cor}(x_i, Y)|}{\sqrt{\hat{V}(Y)}}.$$

Thus the largest $|\hat{\beta}_i|$ correspond to the largest $[\text{cor}(x_i, Y)]^2$. See Fan and Lv (2008), who give some pros and cons of MMLE variable selection.

From canonical correlation analysis (CCA), if the $(Y_i, \boldsymbol{x}_i^T)^T$ are iid, then

$$J = \max_{\boldsymbol{\gamma} \neq \boldsymbol{0}} \text{Cor}(\boldsymbol{\gamma}^T \boldsymbol{x}, Y) = \max_{\boldsymbol{\gamma} \neq \boldsymbol{0}} \frac{\boldsymbol{\gamma}^T \boldsymbol{\Sigma}_{\boldsymbol{x}Y}}{\sqrt{V(Y)}\sqrt{\boldsymbol{\gamma}^T \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\gamma}}}.$$

This optimization problem is equivalent to maximizing

$$V(Y) J^2 = \max_{\boldsymbol{\gamma} \neq \boldsymbol{0}} \frac{\boldsymbol{\gamma}^T \boldsymbol{\Sigma}_{\boldsymbol{x}Y} \boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T \boldsymbol{\gamma}}{\boldsymbol{\gamma}^T \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\gamma}}$$

which has a maximum at $\boldsymbol{\gamma} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \boldsymbol{\beta}_{OLS}$. See Mardia, Kent, and Bibby (1979, pp. 168, 282). Hence PLS is a lot like CCA for $(Y_i, \boldsymbol{x}_i^T)^T$ but with more constraints, and PLS can be computed in high dimensions. From the dimension reduction literature, if Y depends on $\boldsymbol{x}$ only through $\alpha + \boldsymbol{\beta}^T \boldsymbol{x}$, then under the assumption of "linearly related

predictors," $\hat{\boldsymbol{\beta}}_{OLS}$ estimates $\boldsymbol{\beta}_{OLS} = c\boldsymbol{\beta}$ for some constant c which is often nonzero. See, for example, Cook and Weisberg (1999, p. 432). Note that in high dimensions, $\hat{\boldsymbol{\gamma}}_1 = \hat{\boldsymbol{\beta}}_{OLS}$ can be replaced by $\hat{\boldsymbol{\gamma}}_1 = \hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is a high dimensional multiple linear regression estimator, such as lasso.

A $k-$component estimator is selected using model selection, which will be variable selection for the $W_i = \hat{\boldsymbol{\gamma}}_i^T \boldsymbol{x}$. Hence the following review of low dimensional variable selection is useful.

2.1 What Is Variable Selection Doing in Low Dimensions?

Variable selection is the search for a subset of predictor variables that can be deleted without important loss of information if n/p is large, and the search for a useful subset of predictors if n/p is not large. *Model selection* generates M models. Then a model is selected from these M models. Variable selection is a special case of model selection.

Ridge regression, lasso, unilasso, and elastic net often fit the model on a grid a λ_i values with $0 \leq \lambda_1 < \lambda_2 < \cdots < \lambda_M$. See Hoerl and Kennard (1970), Tibshirani (1996), Chatterjee, Hastie, and Tibshirani (2025), and Zou and Hastie (2005). The model selection estimator uses $\hat{\lambda}$ chosen from the grid. Several dimension reduction methods, including PLS and PCR, performs the regression of Y on $(W_1, W_2, ..., W_k)$ and a constant where $W_i = \hat{\boldsymbol{\gamma}}^T \boldsymbol{x}$ for $k = 1, ..., M$. Then the model selection estimator uses $\hat{k}$. Model selection and variable selection can also be used for classification models.

2.1.1 The Regularity Conditions for a Variable Selection Estimator to Estimate $\boldsymbol{\beta} = \boldsymbol{\beta}_F$ Are Strong.

Consider a regression model where the response variable Y depends on the predictors $\boldsymbol{x}$ through $SP = SP(F) = \boldsymbol{x}^T\boldsymbol{\beta}$. Such models include MLR, GLMs, and several survival regression models. A *model for variable selection* can be described by

$$SP(F) = \boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S = SP(S) \tag{7}$$

where $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$ is a $p \times 1$ vector of predictors, $\boldsymbol{x}_S$ is a $q \times 1$ vector, and $\boldsymbol{x}_E$ is a $(p-q)\times 1$ vector. Given that $\boldsymbol{x}_S$ is in the model, $\boldsymbol{\beta}_E = \boldsymbol{0}$ and E denotes the subset of terms that can be eliminated given that the subset S is in the model. In Equation (7), there is a "true submodel" S where all of the elements of $\boldsymbol{\beta}_S$ are nonzero, but all of the elements of $\boldsymbol{\beta}$ that are not elements of $\boldsymbol{\beta}_S$ are zero. The full model has $SP = SP(F) = \boldsymbol{x}^T\boldsymbol{\beta}$, the submodel S has $SP = SP(S) = \boldsymbol{x}_S^T\boldsymbol{\beta}_S$, and $SP(F) = SP(S)$. Note that $\boldsymbol{\beta} = \boldsymbol{\beta}_F$. Assume that S is unique. $S = F$ is possible.

Since S is unknown, candidate subsets will be examined. Let $\boldsymbol{x}_I$ be the vector of d_I terms from a candidate subset indexed by I, including a constant, and let $\boldsymbol{x}_O$ be the vector of the remaining predictors (out of the candidate submodel). Then

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I + \boldsymbol{x}_O^T\boldsymbol{\beta}_O.$$

Suppose that S is a subset of I and that model (6) holds. Then

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_{I/S}^T\boldsymbol{\beta}_{(I/S)} + \boldsymbol{x}_O^T\boldsymbol{0} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I \tag{8}$$

where $\boldsymbol{x}_{I/S}$ denotes the predictors in I that are not in S. Since this is true regardless of the values of the predictors, $\boldsymbol{\beta}_O = \boldsymbol{0}$ and the sample correlation $\text{cor}(\boldsymbol{x}^T\boldsymbol{\beta}, \boldsymbol{x}_I^T\boldsymbol{\beta}_I) = \text{cor}(SP(F), SP(I)) = 1$ for the population $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_I$ if $S \subseteq I$. If consistent estimators for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_I$ are used, the full model is useful for prediction, and $S \subseteq I$, then $\text{cor}(ESP(F), ESP(I)) \to 1$ and $\text{cor}(Y - ESP(F), Y - ESP(I)) \to 1$ as $n \to \infty$. See Olive and Hawkins (2005).

Let I_{min} correspond to the set of predictors selected by a variable selection method such as forward selection or lasso variable selection. If $\hat{\boldsymbol{\beta}}_I$ is $d_I \times 1$, use zero padding to form the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I,0}$ from $\hat{\boldsymbol{\beta}}_I$ by adding 0s corresponding to the omitted variables. For example, if $p = 4$ and $\hat{\boldsymbol{\beta}}_{I_{min}} = (\hat{\beta}_1, \hat{\beta}_3)^T$, then the observed variable selection estimator $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_{min},0} = (\hat{\beta}_1, 0, \hat{\beta}_3, 0)^T$. As a statistic, $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities $\pi_{kn} = P(I_{min} = I_k)$ for $k = 1, ..., M$ where there are M subsets, e.g. $M = 2^p - 1$. Note that under model (7), $\boldsymbol{\beta}_{I,0} = \boldsymbol{\beta} = \boldsymbol{\beta}_F$ if $S \subseteq I$.

Assume p is fixed. Suppose model (7) holds, and that if $S \subseteq I_j$ where the dimension of I_j is d_j, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j} - \boldsymbol{\beta}_{I_j}) \xrightarrow{D} N_{d_j}(\boldsymbol{0}, \boldsymbol{V}_j)$ where $\boldsymbol{V}_j$ is the covariance matrix of the asymptotic multivariate normal distribution. Then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V}_{j,0}) \tag{9}$$

where $\boldsymbol{V}_{j,0}$ adds columns and rows of zeros corresponding to the x_i not in I_j, and $\boldsymbol{V}_{j,0}$ is singular unless I_j corresponds to the full model. This large sample theory holds for many models, including multiple linear regression fit by OLS, GLMs fit by maximum likelihood, and Cox (1972) proportional hazards regression fit by maximum partial likelihood. See Pelawa Watagoda and Olive (2021) and Rathnayake and Olive (2023) for references.

Remark 2. If $A_1, A_2, ..., A_k$ are pairwise disjoint and if $\cup_{i=1}^k A_i = S =$ the sample space, then the collection of sets $A_1, A_2, ..., A_k$ is a *partition* of S. Then the *Law of Total Probability* states that if $A_1, A_2, ..., A_k$ form a partition of S such that $P(A_i) > 0$ for $i = 1, ..., k$, then

$$P(B) = \sum_{j=1}^{k} P(B \cap A_j) = \sum_{j=1}^{k} P(B|A_j)P(A_j).$$

Let sets $A_{k+1}, ..., A_m$ satisfy $P(A_i) = 0$ for $i = k+1, ..., m$. Define $P(B|A_j) = 0$ if $P(A_j) = 0$. Then a Generalized Law of Total Probability is

$$P(B) = \sum_{j=1}^{m} P(B \cap A_j) = \sum_{j=1}^{m} P(B|A_j)P(A_j),$$

and will be used in the proof of Theorem 2.

Pötscher (1991) used the conditional distribution of $\hat{\boldsymbol{\beta}}_{VS}|(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0})$ to find the distribution of $\boldsymbol{w}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta})$. Let $\hat{\boldsymbol{\beta}}_{I_k,0}^C$ be a random vector from the conditional distribution $\hat{\boldsymbol{\beta}}_{I_k,0}|(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0})$. Let $\boldsymbol{w}_{kn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_k,0} - \boldsymbol{\beta})|(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}) \sim \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_k,0}^C - \boldsymbol{\beta})$. Denote $F_{\boldsymbol{z}}(\boldsymbol{t}) = P(z_1 \leq t_1, ..., z_p \leq t_p)$ by $P(\boldsymbol{z} \leq \boldsymbol{t})$. Then Pötscher (1991) and Pelawa Watagoda and Olive (2021) show the following result.

Theorem 2. Using the above notation,

$$F_{\boldsymbol{w}_n}(\boldsymbol{t}) = P[n^{1/2}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \le \boldsymbol{t}] = \sum_{k=1}^{J} F_{\boldsymbol{w}_{kn}}(\boldsymbol{t})\pi_{kn}.$$

Hence $\hat{\boldsymbol{\beta}}_{VS}$ has a mixture distribution of the $\hat{\boldsymbol{\beta}}^C_{I_k,0}$ with probabilities π_{kn}, and $\boldsymbol{w}_n$ has a mixture distribution of the $\boldsymbol{w}_{kn}$ with probabilities π_{kn}.

Proof: Let $W = W_{VS} = k$ if $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ where $P(W_{VS} = k) = \pi_{kn}$ for $k = 1, ..., J$. Then $(\hat{\boldsymbol{\beta}}_{VS:n}, W_{VS:n}) = (\hat{\boldsymbol{\beta}}_{VS}, W_{VS})$ has a joint distribution where the sample size n is usually suppressed. Note that $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_W,0}$. Then by Remark 2,

$$F_{\boldsymbol{w}_n}(\boldsymbol{t}) = P[n^{1/2}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \le \boldsymbol{t}] =$$

$$\sum_{k=1}^{J} P[n^{1/2}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \le \boldsymbol{t} | (\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0})] P(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}) =$$

$$\sum_{k=1}^{J} P[n^{1/2}(\hat{\boldsymbol{\beta}}_{I_k,0} - \boldsymbol{\beta}) \le \boldsymbol{t} | (\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0})] \pi_{kn}$$

$$= \sum_{k=1}^{J} P[n^{1/2}(\hat{\boldsymbol{\beta}}^C_{I_k,0} - \boldsymbol{\beta}) \le \boldsymbol{t}] \pi_{kn} = \sum_{k=1}^{J} F_{\boldsymbol{w}_{kn}}(\boldsymbol{t})\pi_{kn}. \quad \square$$

Under the assumptions of Theorem 3, note that $\sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta})$ is selecting from the $\boldsymbol{u}_{kn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_k,0} - \boldsymbol{\beta})$ and asymptotically from the $\boldsymbol{u}_j$ where $\boldsymbol{u}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u}_j \sim N_p(\boldsymbol{0}, \boldsymbol{V}_{j,0})$ where the $\boldsymbol{u}_{jn}$ correspond to the π_j. See Equation (9). Charkhi and Claeskens (2018) showed that $\boldsymbol{w}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}^C_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w}_j$ if $S \subseteq I_j$ for the maximum likelihood estimator (MLE) with AIC, and gave a forward selection example. They claim that $\boldsymbol{w}_j$ is a multivariate truncated normal distribution (where no truncation is possible) that is symmetric about $\boldsymbol{0}$. Hence $E(\boldsymbol{w}_j) = 0$, and $\mathrm{Cov}(\boldsymbol{w}_j) = \boldsymbol{\Sigma}_j$ exits. This claim does not seem to be correct: the selection bias changes the distribution of the selected $\boldsymbol{u}_{jn}$ and $\boldsymbol{u}_j$ to that of $\boldsymbol{w}_{jn}$ and $\boldsymbol{w}_j$ where some of the probabilities are increased and some are decreased, but the probabilities are not driven to 0. The Rathnayake and Olive (2023) Theorem 3 proves that $\boldsymbol{w}$ is a mixture distribution of the $\boldsymbol{w}_j$ with probabilities π_j.

Theorem 3. Assume $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, and let $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities π_{kn} where $\pi_{kn} \to \pi_k$ as $n \to \infty$. Denote the positive π_k by π_j. Assume $\boldsymbol{w}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}^C_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w}_j$. Then

$$\boldsymbol{w}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w} \tag{10}$$

where the cdf of $\boldsymbol{w}$ is $F_{\boldsymbol{w}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{w}_j}(\boldsymbol{t})$.

Proof. Since $\boldsymbol{w}_n$ has a mixture distribution of the $\boldsymbol{w}_{kn}$ with probabilities π_{kn} by Theorem 2, the cdf of $\boldsymbol{w}_n$ is $F_{\boldsymbol{w}_n}(\boldsymbol{t}) = \sum_k \pi_{kn} F_{\boldsymbol{w}_{kn}}(\boldsymbol{t}) \to F_{\boldsymbol{w}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{w}_j}(\boldsymbol{t})$ at continuity points of the $F_{\boldsymbol{w}_j}(\boldsymbol{t})$ as $n \to \infty$. $\square$

Remark 3. a) The assumption that $\boldsymbol{w}_{jn} \xrightarrow{D} \boldsymbol{w}_j$ may not be mild.
b) If $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, then $\hat{\boldsymbol{\beta}}_{VS}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ since selecting from a finite number M of $\sqrt{n}$ consistent estimators (even on a set that goes to one in probability) results in a $\sqrt{n}$ consistent estimator by Pratt (1959).
c) The assumption $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$ holds for some criterion, such as AIC, BIC, and the C_p criterion for MLR. The assumption $P(S \subseteq I_{min}) \to 1$ also holds for consistent regularized estimators of $\boldsymbol{\beta} = \boldsymbol{\beta}_F$. Thus the lasso variable selection and elastic net variable selection estimators are $\sqrt{n}$ consistent if lasso and elastic net are consistent estimators of $\boldsymbol{\beta}_F$. Lasso is a consistent estimator if $\hat{\lambda}$ is small enough. For MLR it is known how small $\hat{\lambda}$ needs to be, but not for other regression methods such as GLMs.

A second multiple linear regression model is

$$Y = \boldsymbol{x}^T\boldsymbol{\beta} + e \tag{11}$$

where $x_1 = 1$ and $x_2, ..., x_p$ are the nontrivial predictors. Hence β_1 corresponds to the constant α in the first MLR model (3). Let submodel I have a predictors, including a constant. Then for a wide variety of iid error distributions, the Anova F statistic for submodel I satisfies $F_I \xrightarrow{D} X/(p-a)$ where $X \sim \chi^2_{p-a}$. Then the variable selection criterion

$$C_p(I) = \frac{SSE(I)}{MSE} + 2a - n = (p-a)(F_I - 1) + a \tag{12}$$

where MSE is the error mean square for the full model. See Mallows (1973) and Jones (1946).

Example 1. This is an example where the $\pi_{kn} \to \pi_k$ as $n \to \infty$. Assume $S \subseteq I$ where I has a predictors, including a constant. Let F denote the full model with p predictors including a constant from model (11), and let $S = I = I_i$ be the model that deletes predictor x_i with $a = p - 1$. Then $C_p(I) \xrightarrow{D} X + p - 2$ where $X \sim \chi^2_1$. Let F denote the full model and consider all subsets variable selection with C_p. Since only S and F do not underfit, only π_S and π_F are positive. Since $C_p(F) = p$, $I = S$ is selected if $C_p(I) < p$. Hence $\pi_S = P(\chi^2_1 + p - 2 < p) = P(\chi^2_1 < 2) = 0.8427$, and $\pi_F = 1 - \pi_S = 0.1573$. This result also holds for backward elimination since the probability that x_i will be the first predictor deleted goes to 1 as $n \to \infty$ because $C_p(I_i) = C_p(S)$ is bounded in probability while $C_p(I_j)$ diverges as $n \to \infty$ for $j \neq i$. For forward selection with correlated predictors, expect that $\pi_S < P(\chi^2_1 < 2)$, and hence $\pi_F > 1 - P(\chi^2_1 < 2)$ with $\pi_S + \pi_F = 1$

For the R code below, $\boldsymbol{\beta} = (1, ..., 1, 0, ..., 0)^T$ is a $p \times 1$ vector with $k + 1$ ones and $p - k - 1$ zeroes. Hence $k = p - 2$ deletes the predictor x_p. The function `belimsim` generates 1000 data sets, performs backward elimination, and finds the proportion of time the full model was selected, which was $0.158 \approx 0.1573$.

```
belimsim(n=100,p=5,k=3,nruns=1000)
$fullprop
[1] 0.158
```

Remark 4. For $k-$component nonparametric or nonlinear regression, including generalized additive models (GAMs), of Y on $W_1, ..., W_k$ and possibly a constant, if $n \geq 10p$ and the full model using $k = p$ is useful for prediction, then the model selection estimator using $k = \hat{q}$ could be selected such that both $\text{cor}(\hat{Y}_{I_{\hat{q}}}, \hat{Y}_{I_p})$ and $\text{cor}(r_{I_{\hat{q}}}, r_{I_p})$ are very high. The FF and RR plots are useful for checking model $I_{\hat{q}}$.

2.1.2 What Are $k-$Component Estimators Estimating?

For the $k-$component estimators, let $W_i = \hat{\boldsymbol{\gamma}}_i^T \boldsymbol{x}$ for $i = 1, ..., p$. For PLS, let $\hat{\boldsymbol{\gamma}}_i = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{i-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$ with $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^0 = \boldsymbol{\Sigma}_{\boldsymbol{x}}^0 = \boldsymbol{I}_p$. For PCA, let $\hat{\boldsymbol{\gamma}}_i = \hat{\boldsymbol{d}}_i$ be orthogonal eigenvectors of $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}$.

The following new method is for **any** regression or classification method that depends on $\boldsymbol{x}$ only through $SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}$. Let the standardized predictors $v_i = x_i / \hat{\sigma}_i$ for $i = 1, ..., p$. Regress Y on the v_i to obtain the parameter vector $\hat{\boldsymbol{\theta}}$ where the regression on the x_i gives $\hat{\boldsymbol{\beta}}$. Let $W_i = x_{i_j}$ correspond to predictors with the largest $|\hat{\theta}_i|$. If more than one $\hat{\theta}_i = 0$, rank the corresponding predictors by $[\text{cor}(x_i, Y)]^2$. Call the resulting estimator the $k-$component SVM estimator, $k-$component lasso estimator, et cetera.

The above method is useful for doing ridge regression, lasso, or unilasso $k-$component variable selection. The Meinshausen (2007) relaxed lasso estimator with λ_2 computes the lasso estimator with $\hat{\lambda}_1$, then lasso with $\lambda_2 < \hat{\lambda}_1$ is applied to the a predictors with nonzero $\hat{\beta}_i(\hat{\lambda}_1)$. If $\lambda_2 = 0$, then the estimator is often called the relaxed lasso estimator or the variable selection lasso estimator since the estimator applies the regression method, such as OLS or a GLM, to the a predictors. For the new method, the regression of Y on $W_1, ..., W_a$ corresponds to the relaxed lasso estimator if lasso is the regression method used. In low dimensions, if C_p or AIC are used to select the model, then Theorem 3 proves that the $k-$component estimator is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}_F$.

Instead of using the response variable Y and the predictors $X_1, ..., X_p$, the regression model or classification model can use Y and the predictors $W_1, ..., W_k$. Denote this model by $I_k = \{1, ..., k\}$. Then the k-component estimator $(\hat{\alpha}_k, \hat{\boldsymbol{\beta}}_k)$ is obtained by fitting the working model

$$WSP = \alpha_k + \theta_1 W_1 + \cdots + \theta_k W_k = \alpha_k + \boldsymbol{\theta}_k^T \boldsymbol{w}$$

where $\boldsymbol{\theta}_k = (\theta_1, ..., \theta_k)^T$ and $\boldsymbol{w} = \boldsymbol{w}_{I_k} = (W_1, ..., W_k)^T$. Let $ESP(k) = \hat{\alpha}_k + \hat{\boldsymbol{\beta}}_k^T \boldsymbol{x}$ and $SP(k) = \alpha_k + \boldsymbol{\beta}_k^T \boldsymbol{x}$.

Then $ESP(k) \xrightarrow{P} SP(k)$ could happen in several ways. i) For the k-component estimator, we could fix a basis $\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_p$, assume $\hat{\boldsymbol{\theta}}_k \xrightarrow{P} \boldsymbol{\theta}_k$, and assume the $k \times p$ matrix

$$\hat{\boldsymbol{A}}_{k,n} = \hat{\boldsymbol{A}}_k = \begin{pmatrix} \hat{\boldsymbol{\gamma}}_1^T \\ \vdots \\ \hat{\boldsymbol{\gamma}}_k^T \end{pmatrix} \xrightarrow{P} \boldsymbol{A}_k = \begin{pmatrix} \boldsymbol{\gamma}_1^T \\ \vdots \\ \boldsymbol{\gamma}_k^T \end{pmatrix}.$$

Then $\hat{\boldsymbol{A}}_k \boldsymbol{x} = \boldsymbol{w} = (W_1, ..., W_k)^T$, and $ESP(\boldsymbol{w}) = \hat{\alpha}_k + \hat{\boldsymbol{\theta}}_k^T \boldsymbol{w} = \hat{\alpha}_k + \hat{\boldsymbol{\theta}}_k^T \hat{\boldsymbol{A}}_k \boldsymbol{x} = \hat{\alpha}_k + \hat{\boldsymbol{\beta}}_k^T \boldsymbol{x} = ESP(\boldsymbol{x})$ with $\hat{\boldsymbol{\beta}}_k = \hat{\boldsymbol{A}}_k^T \hat{\boldsymbol{\theta}}_k$. Assume $\hat{\boldsymbol{\theta}}_k \xrightarrow{P} \boldsymbol{\theta}_k$.

ii) Assume that $\hat{\theta}_j \hat{\boldsymbol{\gamma}}_j \xrightarrow{P} \theta_j \boldsymbol{\gamma}_j$. For example, if $\hat{\boldsymbol{\gamma}}_j = \hat{\boldsymbol{\eta}}_j$ or $\hat{\boldsymbol{\gamma}}_j = -\hat{\boldsymbol{\eta}}_j$ is an orthonormal eigenvector of some matrix and $|\hat{\theta}_j| \xrightarrow{P} |\theta_j|$ and $\hat{\boldsymbol{\eta}}_j \xrightarrow{P} \boldsymbol{\eta}_j$, then the result holds.

iii) Assume that $\hat{\alpha}_k \xrightarrow{P} \alpha_k$ and $\sum_{j=1}^k \hat{\theta}_j \hat{\boldsymbol{\gamma}} \xrightarrow{P} \sum_{j=1}^k \theta_j \boldsymbol{\gamma}$.

For example, fit a GLM, logistic regression, a support vector machine, multiple linear regression, et cetera. The θ_i depend on the method used to fit the working model. (Also, using θ_{ki} instead of θ_i is more accurate, but suppressing the subscript k is convenient.) Parts e), f), and g) of Theorem 4 and part b) of Theorem 5 are new. The results are the most useful if $\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_k$ are linearly independent and if $\hat{\boldsymbol{\gamma}}_1, ..., \hat{\boldsymbol{\gamma}}_k$ are linearly independent. The HPLS vectors are a Krylov sequence which tends to be linearly independent for small enough k if $\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ is not an eigenvector of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$. See Cook and Forzani (2024, pp. 16-17, 100).

Theorem 4. Consider the above notation. Assume $ESP(k) \xrightarrow{P} SP(k)$. For results with $\boldsymbol{\beta}_{OLS}$, assume $\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}$ exists.

a) $ESP(k) = \hat{\alpha}_k + \hat{\boldsymbol{\beta}}_k^T \boldsymbol{x} = \hat{\alpha}_k + (\sum_{j=1}^k \hat{\theta}_j \hat{\boldsymbol{\gamma}}_j^T)\boldsymbol{x}$.

b) $SP(k) = \alpha_k + \boldsymbol{\beta}_k^T \boldsymbol{x} = \alpha_k + (\sum_{j=1}^k \theta_j \boldsymbol{\gamma}_j^T)\boldsymbol{x}$.

c) $\hat{\boldsymbol{\beta}}_k = \sum_{j=1}^k \hat{\theta}_j \hat{\boldsymbol{\gamma}}_j = \hat{\boldsymbol{A}}_k^T \hat{\boldsymbol{\theta}}_k$.

d) $\boldsymbol{\beta}_k = \sum_{j=1}^k \theta_j \boldsymbol{\gamma}_j = \boldsymbol{A}_k^T \boldsymbol{\theta}_k$.

e) $\hat{\boldsymbol{\beta}}_{kPLS} = (\sum_{j=1}^k \hat{\theta}_j \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{j-1})\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$

f) Under iid cases, $\boldsymbol{\beta}_{kPLS} = (\sum_{j=1}^k \theta_j \boldsymbol{\Sigma}_{\boldsymbol{x}}^{j-1})\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = (\sum_{j=1}^k \theta_j \boldsymbol{\Sigma}_{\boldsymbol{x}}^{j})\boldsymbol{\beta}_{OLS}$.

g) If $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} \xrightarrow{P} \boldsymbol{V}_{\boldsymbol{x}}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} \xrightarrow{P} \boldsymbol{V}_{\boldsymbol{x}Y}$, then $\hat{\boldsymbol{\beta}}_{kPLS} \xrightarrow{P} \boldsymbol{\beta}_{kPLS} = (\sum_{j=1}^k \theta_j \boldsymbol{V}_{\boldsymbol{x}}^{j-1})\boldsymbol{V}_{\boldsymbol{x}Y}$.

Proof. Fit WSP to get the $ESP = \hat{\alpha}_k + \hat{\theta}_1 W_1 + \cdots + \hat{\theta}_k W_k = \hat{\alpha}_k + \hat{\theta}_1 \hat{\boldsymbol{\gamma}}_1^T \boldsymbol{x} + \cdots + \hat{\theta}_k \hat{\boldsymbol{\gamma}}_k^T \boldsymbol{x} = \hat{\alpha}_k + \hat{\boldsymbol{\beta}}_k^T \boldsymbol{x}$. Equating terms gives the result. □

When the cases are not iid, $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$ may be estimating $\boldsymbol{\beta} = \boldsymbol{\beta}_{OLS} \neq \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$. When the errors e_i are iid, a common assumption for OLS MLR theory is

$$n(\boldsymbol{X}^T \boldsymbol{X})^{-1} = \hat{\boldsymbol{V}} = \begin{pmatrix} \hat{\boldsymbol{V}}_{11} & \hat{\boldsymbol{V}}_{12} \\ \hat{\boldsymbol{V}}_{21} & \hat{\boldsymbol{V}}_{22} = n\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}/(n-1) \end{pmatrix} \xrightarrow{P} \boldsymbol{V} = \begin{pmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{pmatrix}.$$

Thus $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \xrightarrow{P} \boldsymbol{V}_{22}$, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} \xrightarrow{P} \boldsymbol{V}_{22}^{-1}$, and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} \xrightarrow{P} \boldsymbol{V}_{22}^{-1}\boldsymbol{\beta}$ since $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} \xrightarrow{P} \boldsymbol{\beta}$.

The following theorem gives a result similar to Equation (2), and shows that for low dimensions, if the p components W_i are plugged into a model that uses maximum likelihood estimation, such as a GLM, then the p-component estimator $\hat{\boldsymbol{\beta}}_p = \hat{\boldsymbol{\beta}}_{\boldsymbol{x}} = \hat{\boldsymbol{\beta}}_F$, the MLE. Hence $\text{cor}(ESP(p), ESP(F)) = 1$, and typically $\hat{\alpha}_p = \hat{\alpha}_F$, so $ESP(p) = ESP(F)$. Similar theory holds for other maximization or minimization problems, such as quasi-likelihood and partial likelihood. The profile likelihood function $L_p(\boldsymbol{\beta}_{\boldsymbol{x}}|\boldsymbol{x}) = L(\boldsymbol{\beta}_{\boldsymbol{x}}, \hat{\boldsymbol{\eta}}|\boldsymbol{x})$ where L is the likelihood function of all of the parameters $(\boldsymbol{\beta}_{\boldsymbol{x}}, \boldsymbol{\eta})$ and $\hat{\boldsymbol{\eta}}$ is the MLE of $\boldsymbol{\eta}$. As above, use $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{\boldsymbol{w}}$ to denote the MLE with $\boldsymbol{w}$ instead of $\hat{\boldsymbol{\beta}}_{\boldsymbol{w}}$.

Theorem 5. Suppose the profile likelihood function $L_p(\boldsymbol{\beta}_{\boldsymbol{x}}|\boldsymbol{x}) = \prod_{i=1}^n f(\boldsymbol{x}_i|\boldsymbol{\beta}_{\boldsymbol{x}}) = \prod_{i=1}^n g(\boldsymbol{x}_i^T \boldsymbol{\beta}_{\boldsymbol{x}})$ depends on $\boldsymbol{x}$ and $\boldsymbol{\beta}_{\boldsymbol{x}}$ only through $\boldsymbol{x}^T \boldsymbol{\beta}_{\boldsymbol{x}}$. a) If the maximum likelihood estimator is computed using $\boldsymbol{w} = \hat{\boldsymbol{A}}_p \boldsymbol{x}$ instead of $\boldsymbol{x}$, then $\hat{\boldsymbol{\beta}}_{\boldsymbol{x}} = \hat{\boldsymbol{A}}_p^T \hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\beta}}_p$ provided that $\hat{\boldsymbol{A}}_p$ is nonsingular. b) Thus $\hat{\boldsymbol{\beta}}_{\boldsymbol{x}} = \hat{\boldsymbol{\beta}}_{pPLS} = (\sum_{j=1}^p \hat{\theta}_j \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{j-1})\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$ if the PLS components are used, and under iid cases, $\boldsymbol{\beta}_{\boldsymbol{x}} = (\sum_{j=1}^p \theta_j \boldsymbol{\Sigma}_{\boldsymbol{x}}^{j-1})\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = (\sum_{j=1}^p \theta_j \boldsymbol{\Sigma}_{\boldsymbol{x}}^{j})\boldsymbol{\beta}_{OLS}$.

Proof. a)

$$L_p(\boldsymbol{\theta}|\boldsymbol{w}) = \prod_{i=1}^{n} g(\boldsymbol{w}_i^T\boldsymbol{\theta}) = \prod_{i=1}^{n} g(\boldsymbol{x}_i^T\hat{\boldsymbol{A}}^T\boldsymbol{\theta}) = \prod_{i=1}^{n} g(\boldsymbol{x}_i^T\boldsymbol{\beta}^*).$$

Since the second to last term is maximized by $\hat{\boldsymbol{A}}^T\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\beta}}_p$ and the last term is maximized by $\boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}}_{\boldsymbol{x}}$, it follows that $\hat{\boldsymbol{\beta}}_{\boldsymbol{x}} = \hat{\boldsymbol{A}}^T\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\beta}}_p$, and $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{A}}^T)^{-1}\hat{\boldsymbol{\beta}}_{\boldsymbol{x}}$. Nonsingularity was used so that $\boldsymbol{\beta}^*$ varies through $\mathbb{R}^p$ as $\boldsymbol{\beta}_{\boldsymbol{x}}$ varies through $\mathbb{R}^p$.

b) Plug in $\hat{\boldsymbol{\beta}}_{\boldsymbol{x}} = \hat{\boldsymbol{\beta}}_p = \hat{\boldsymbol{\beta}}_{pPLS}$ from Theorem 4 e). □

2.1.3 "Everything Sensible Works" in Low Dimensions

For real data, an important question in variable selection is whether $\beta_i = 0$ is a reasonable assumption. If $\boldsymbol{X}$ has full rank $p+1$, then having β_i equal to zero for 20 decimal places may not be reasonable. See, for example, Tukey (1991), Nester (1996), and Gelman and Carlin (2017).

Variable selection when some of the $\beta_i = 0$ is interesting, but so is variable selection when none of the $\beta_i = 0$, but some of the β_i are very small in magnitude but nonzero, denoted by $\beta_i = 0^*$. Let I_k be the model that regresses Y on a constant and $W_1, ..., W_k$ for $k = 1, ..., M$. Then the $k-$component model selection estimator becomes a variable selection estimator with respect to $\boldsymbol{\theta} = \boldsymbol{\theta}_p$ and the W_i. For the working model

$$WSP(I_k) = \alpha_k + \theta_1 W_1 + \cdots + \theta_k W_k = \alpha_k + \boldsymbol{\theta}_k^T\boldsymbol{w}_{I_k},$$

consider $WSP(S) = WSP(I_q) = WSP = WSP(I_p) = WSP(F) =$

$$\alpha_q + \theta_1 W_1 + \cdots + \theta_q W_q + 0W_{q+1} + \cdots + 0W_p = \alpha_q + \boldsymbol{\theta}_q^T\boldsymbol{w}_{I_q} = \alpha_p + \boldsymbol{\theta}_p^T\boldsymbol{w}_{I_p}, \quad (13)$$

or

$$WSP = \alpha_p + \theta_1 W_1 + \cdots + \theta_d W_d + (0^*)W_{d+1} + \cdots + (0^*)W_p. \quad (14)$$

Equation (13), which corresponds to Equation (7), is assumed in the envelopes and SDR literature, with a sparsity assumption with respect to the W_i and $\boldsymbol{\theta}_p$. Then the $k-$component parameter vectors $\boldsymbol{\beta}_q = \boldsymbol{\beta}_{q+1} = \cdots = \boldsymbol{\beta}_p$, a result that is analogous to the variable selection result $\boldsymbol{\beta}_{I,0} = \boldsymbol{\beta}_F$ if $S \subseteq I$. Consider Equation (14), perhaps with $0^* = 10^{-6}$ and corresponding observed $|W_j| \leq 10$. Then $I_q = S = F = I_p$ in Equation (13), and the central subspace and the predictor envelope satisfy $\boldsymbol{\mathcal{L}}_{\boldsymbol{y}|\boldsymbol{x}} = \boldsymbol{\mathcal{E}}_{\boldsymbol{x}} = \mathbb{R}^p$. Under Equation (14), the $\boldsymbol{\beta}_k$ are different for $k = 1, ..., p$.

If I is the model selected by the variable selection method, and $P(S \subseteq I) \to 1$ as $n \to \infty$, then the variable selection methods, including SDR and envelope estimators, are asymptotically equivalent to the regression using the full model since $S = F$ is the full model when none of the $\theta_i = 0$ or none of the $\beta_i = 0$.

Remark 5. The above result may seem nice, but in low dimensions, good variable selection estimators do not select the full model with very high probability for moderate n under Equation (14), or if $SP(F) = \alpha + \beta_{i_1}x_{i_1} + \cdots + \beta_{i_d}x_{i_d} + (0^*)x_{i_{d+1}} + \cdots + (0^*)x_{i_p}$. Instead, good variable selection methods divide the predictors W_i or x_i into *wanted*

predictors that are kept in the model I and *unwanted predictors* O that are deleted out of the model such that a) $\text{cor}(ESP(I), ESP(F))$ is high and
b) $\text{cor}(Y - ESP(I), Y - ESP(F))$ is high.

Using the results from Remark 5 a) and b), Olive and Hawkins (2005) suggested 6 plots to compare the submodel I with the full model F. Let a plot of x versus y indicate that y is on the vertical axis. Use the residual plot of ESP versus r for both models, use the response plot of ESP versus Y for both models, make the EE plot of $ESP(I)$ versus $ESP(F)$, and the VV plot of $Y - ESP(I)$ versus $Y - ESP(F)$. If the full model is useful for prediction and if submodel I is good, then the plotted points tend to cluster very tightly about the identity line (with unit slope and zero intercept) for the VV and EE plots. Similar results often hold for an FF plot of the fitted values $\hat{Y}_I$ versus $\hat{Y}_F$ and an RR plot of the residuals r_I versus r_F. See Olive and Hawkins (2005).

For MLR, Remark 5 b) is the correlation of the residuals r_I and $r = r_F$ from the submodel and the full model, and Theorem 6 gives a lower bound on the correlation if the C_p criterion is used.

The "sensible" variable selection methods need to include the full model as one of the models considered, and the full model needs to fit the data "well" (have a good response plot or be useful for prediction). By Theorem 5, the $\boldsymbol{\gamma}_i$ need to be linearly independent and the $\hat{\boldsymbol{\gamma}}_i$ need to be linearly independent for models that use $W_i = \hat{\boldsymbol{\gamma}}_i^T \boldsymbol{x}$ so that $\hat{\boldsymbol{\beta}}_p(\boldsymbol{w}_F) = \hat{\boldsymbol{\beta}}_p(\boldsymbol{x}_F)$ and $\boldsymbol{\beta}_p(\boldsymbol{w}_F) = \boldsymbol{\beta}_p(\boldsymbol{x}_F) = \boldsymbol{\beta}_F$ where $\boldsymbol{w}_F = (W_1, ..., W_p)^T$, $\boldsymbol{x}_F = (x_1, ..., x_p)^T$, and $\boldsymbol{\beta}(\boldsymbol{z})$ means that Y was regressed on $\boldsymbol{z}$. Then $\hat{\boldsymbol{A}}_p$ is nonsingular. $W_i = x_i$ corresponds to $\hat{\boldsymbol{\gamma}}_i = \boldsymbol{\gamma}_i = \boldsymbol{c}_i$ where $\boldsymbol{c}_i$ is the ith column of $\boldsymbol{I}_p$. A good variable selection criterion needs to be used, such as AIC, BIC, $k-$fold cross validation, or C_p for MLR.

For MLR and $W_i = x_i$, consider the Tibshirani (1996) lasso and Chatterjee, Hastie, and Tibshirani (2025) unilasso variable selection estimators that select predictors $\boldsymbol{x}_I$ and a constant to be in the model. Fit OLS to that model, and use model I if $C_p(I) \leq C_p(F) = p+1$. Otherwise use the full model F with $\boldsymbol{\beta}_F = \boldsymbol{\beta} = \boldsymbol{\beta}_{OLS}$. Many variants are possible, including using $k-$component lasso or $k-$component unilasso. Other variable selection criterion can be used for GLMs or the Cox proportional hazards model. The following theorem appeared in Olive (2025) and applies to the W_i. Olive and Hawkins (2005) got the result for the x_i.

Theorem 6. Assume $\boldsymbol{X}$ is full rank so that the OLS full model can be computed. Let r be the residuals from the OLS full model and let r_I be the residuals from OLS submodel I that uses $\hat{\boldsymbol{\beta}}_I$ with k predictors including a constant where $2 \leq k \leq p+1$. If $C_p(I) \leq 2k$, then the sample correlation

$$cor(r, r_I) \geq \sqrt{1 - \frac{p+1}{n}}. \tag{15}$$

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 $\boldsymbol{\beta}$, and $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. 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 iid. OLS also gives consistent estimators for AR(p) and AR(∞) time series, serially correlated errors, et cetera. Note

that the rate at which $P(S \subseteq I_{min}) \to 1$ is not exponentially fast, as is often claimed in the literature.

For moderate sample size n, predictors x_j or W_j will often be omitted as long as $cor(r_I, r)$ stays high, even if $\beta_j \neq 0$ or $\theta_j \neq 0$. The C_p criterion selects wanted predictors I to be in the model, and unwanted predictors O to be out of the model. Under Equation (15), often $\boldsymbol{\beta}_{I,0} \neq \boldsymbol{\beta}_F$, but $\hat{\boldsymbol{\beta}}_I$ is a good estimator of $\boldsymbol{\beta}_I$, and model I fits the data well. Very weak predictors often degrade the full model in that the model is improved when these predictors are omitted. Note that the wanted and unwanted predictors are similar to the material and immaterial predictors for envelopes estimators.

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 $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$.

Theorem 4 shows that $\hat{\boldsymbol{\beta}}_{kPLS} = (\sum_{j=1}^{k} \hat{\theta}_j \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{j-1})\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$ and $\boldsymbol{\beta}_{kPLS} = (\sum_{j=1}^{k} \theta_j \boldsymbol{\Sigma}_{\boldsymbol{x}}^{j-1})\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ (under iid cases). 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 $\theta_j = 0$ 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 $1 \leq m < p$. Heuristically, the model selection result of Equation (15) implies that $\boldsymbol{\beta} \in \mathbb{R}^m$ is approximately true in that $\theta_j \approx 0$ for $k^* + 1 \leq j \leq p$ for some $m = k^* < p$. Hence the "unwanted" predictors W_j are "weak" or "almost immaterial" for $m + 1 \leq j \leq p$. On the other hand, it is possible that $\text{cor}(\hat{\boldsymbol{\beta}}_k^T\boldsymbol{x}, \hat{\boldsymbol{\beta}}^T\boldsymbol{x})$ is very high with $\|\hat{\boldsymbol{\beta}}_k - \hat{\boldsymbol{\beta}}\|$ large.

2.1.4 Variable Selection Summary

For part b) in the following remark, assume that model I is the model selected after variable selection. Then fix I as $n \to \infty$. This result is useful for data splitting. (Model I_{min}, that optimizes a variable selection criterion among models searched, has $P(I_{min} = I_k) = \pi_{kn}$. So condition on model $I = I_k$ selected. It is possible the $P(I_{min} = F) \to 1$ as $n \to \infty$.) In low dimensions, "good regression variable selection estimators" use the full model, and include forward selection, lasso variable selection, $k-$component unilasso, and $k-$component estimators with $k-$fold cross validation, AIC, BIC, or C_p. Using the full model insures that the collection of models considered includes a "good model."

Remark 6. a) In low dimensions where the full model is "good", good variable selection estimators keep $cor(ESP(I), ESP(F))$ high and $cor(Y - ESP(I), Y - ESP(F))$ high, by keeping wanted predictors I in the model and omitting unwanted predictors O.
b) The estimator $\hat{\boldsymbol{\beta}}_I$ estimates $\boldsymbol{\beta}_I$ where often $\boldsymbol{\beta}_{I,0} \neq \boldsymbol{\beta} = \boldsymbol{\beta}_F$.
c) In high dimensions, variable selection is a search for a useful subset I of predictors.

3 r Response Variables $Y_1, ..., Y_r$

The following theorem is a generalization of Theorem 1 for $r > 1$. Thus PLS is a model free way to get predictors $W_i = \hat{\boldsymbol{\gamma}}_i^T\boldsymbol{x}$ that are fairly highly correlated with the response variables $Y_1, ..., Y_r$, and the absolute correlations tend to decrease quickly.

Theorem 7. For $1 < k < q$, the SIMPLS algorithm maximizes

$$Q_S(\boldsymbol{\gamma}^T\boldsymbol{x}) = \hat{V}(\boldsymbol{\gamma}^T\boldsymbol{x}) \sum_{j=1}^{r} [\mathrm{cor}(\boldsymbol{\gamma}^T\boldsymbol{x}, Y_j)]^2 \hat{V}(Y_j) \tag{16}$$

subject to $\boldsymbol{\gamma}^T \boldsymbol{S}_{\boldsymbol{x}} \hat{\boldsymbol{A}}_k^T = \boldsymbol{0}$ and $\boldsymbol{\gamma}^T\boldsymbol{\gamma} = 1$.

Proof. By Remark 1,

$$\boldsymbol{\gamma}^T \boldsymbol{S}_{\boldsymbol{xy}} = \boldsymbol{S}_{\boldsymbol{\gamma}^T\boldsymbol{x},\boldsymbol{y}} = (\mathrm{cov}(\boldsymbol{\gamma}^T\boldsymbol{x}, Y_1), ..., \mathrm{cov}(\boldsymbol{\gamma}^T\boldsymbol{x}, Y_r)),$$

and $Q_S(\boldsymbol{\gamma}^T\boldsymbol{x}) = \boldsymbol{\gamma}^T \boldsymbol{S}_{\boldsymbol{xy}} \boldsymbol{S}_{\boldsymbol{xy}}^T \boldsymbol{\gamma} =$

$$\sum_{j=1}^{r} [\mathrm{cov}(\boldsymbol{\gamma}^T\boldsymbol{x}, Y_j)]^2 = \sum_{j=1}^{r} [\mathrm{cor}(\boldsymbol{\gamma}^T\boldsymbol{x}, Y_j)]^2 \hat{V}(\boldsymbol{\gamma}^T\boldsymbol{x}) \hat{V}(Y_j). \ \square$$

Scaling the response variables to have the same sample variance is often a good idea. Dividing each response variable by its standard deviation is common, and equivalent to using $\boldsymbol{z} = [diag(\boldsymbol{S}_{\boldsymbol{y}})]^{-1/2}\boldsymbol{y} = (Z_1, ..., Z_r)^T$. Then by Remark 1d), the SIMPLS criterion is

$$Q_{\boldsymbol{z}}(\boldsymbol{\gamma}^T\boldsymbol{x}) = \hat{V}(\boldsymbol{\gamma}^T\boldsymbol{x}) \sum_{j=1}^{r} [\mathrm{cor}(\boldsymbol{\gamma}^T\boldsymbol{x}, Z_j)]^2 = \hat{V}(\boldsymbol{\gamma}^T\boldsymbol{x}) \sum_{j=1}^{r} [\mathrm{cor}(\boldsymbol{\gamma}^T\boldsymbol{x}, Y_j)]^2. \tag{17}$$

Hence $Q_{\boldsymbol{z}}(\hat{\boldsymbol{\gamma}}_i^T\boldsymbol{x})$ is a model free variable importance criterion for $W_i = \hat{\boldsymbol{\gamma}}_i^T\boldsymbol{x}$. Note that $V(\boldsymbol{\gamma}^T\boldsymbol{x}) = \mathrm{Cov}(\boldsymbol{\gamma}^T\boldsymbol{x}, \boldsymbol{\gamma}^T\boldsymbol{x}) = \boldsymbol{\gamma}^T \mathrm{Cov}(\boldsymbol{x})\boldsymbol{\gamma}$.

The following, possibly new, model free variable importance criterion is an MMLE criterion

$$Q_M(\hat{\boldsymbol{\gamma}}_i^T\boldsymbol{x}) = \frac{1}{r} \sum_{j=1}^{r} [\mathrm{cor}(\hat{\boldsymbol{\gamma}}_i^T\boldsymbol{x}, Y_j)]^2. \tag{18}$$

Applying SIMPLS to $(\boldsymbol{u}, Z_1, ..., Z_r)$ minimizes $\sum_{j=1}^{r} [\mathrm{cor}(\boldsymbol{\gamma}^T\boldsymbol{u}, Y_j)]^2$ if $\boldsymbol{u}_i = \boldsymbol{S}_{\boldsymbol{x}}^{-1/2}\boldsymbol{x}_i$. If $\hat{\boldsymbol{\phi}}_i^T\boldsymbol{u} = \hat{\boldsymbol{\gamma}}_i^T\boldsymbol{x}$ are the linear combinations, then $\hat{\boldsymbol{\gamma}}_i^T = \hat{\boldsymbol{\phi}}_i^T \boldsymbol{S}_{\boldsymbol{x}}^{-1/2}$.

Given $U_j = \hat{\boldsymbol{\gamma}}_j^T\boldsymbol{x}$, the model free variable importance criterion $Q(\hat{\boldsymbol{\gamma}}_j^T\boldsymbol{x})$ can be used to order the U_j in importance, and scree plot techniques can be used to choose the number $\hat{q}$ of variables to be used in the regression.

To see why (18) is called an MMLE criterion, let $v_i = x_i/\hat{\sigma}_i$ be the standardized x_i. Let $T_i = Y_i/\hat{\sigma}_{Y_i}$ be the standardized Y_i such that $\hat{V}(T_i) = 1$ for $i = 1, ..., r$. Then the MMLE for the OLS multivariate linear regression of $T_1, ..., T_r$ on v_i uses $\hat{\boldsymbol{b}}_i = (\mathrm{cov}(v_i, T_1), \mathrm{cov}(v_i, T_2), ..., \mathrm{cov}(v_i, T_r))^T$ with

$$\|\hat{\boldsymbol{b}}_i\|^2 = \sum_{j=1}^{r} [\mathrm{cov}(v_i, T_j)]^2 = \sum_{j=1}^{r} [\mathrm{cor}(v_i, T_j)]^2 \hat{V}(v_i) \hat{V}(T_j) = \sum_{j=1}^{r} [\mathrm{cor}(x_i, Y_j)]^2.$$

Thus the v_i with the largest $\|\hat{\boldsymbol{b}}_i\|^2$ have the largest $\sum_{j=1}^{r} [\mathrm{cor}(x_i, Y_j)]^2$.

Note that the variable importance criterion (18) does not require p regressions to get p values of $\|\hat{\boldsymbol{b}}_i\|^2$. Criterion (17) can be used in high dimensions to obtain $\hat{\boldsymbol{\gamma}}_i$. Criterion (18) can be used for PCA regression with $U_j = \hat{\boldsymbol{d}}_j^T \boldsymbol{x}$.

A model free variable unimportance criterion is

$$Q_U(\hat{\boldsymbol{\gamma}}_i^T \boldsymbol{x}) = \max_{i=1,...,r} [\mathrm{cor}(\hat{\boldsymbol{\gamma}}_i^T \boldsymbol{x}, Y_j)]^2. \tag{19}$$

Delete variables $W_i = \hat{\boldsymbol{\gamma}}_i^T \boldsymbol{x}$ with the smallest values of Q_U. In particular, use $W_i = \boldsymbol{c}_i^T \boldsymbol{x} = x_i$ to delete predictor variables.

3.1 High Dimensional Multivariate Linear Regression

For the predictor and response variables $(x_1, ..., x_p, Y_1, ..., Y_r)$, let the data matrix be

$$\begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} & Y_{1,1} & Y_{1,2} & \dots & Y_{1,r} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} & Y_{2,1} & Y_{2,2} & \dots & Y_{2,r} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} & Y_{n,1} & Y_{n,2} & \dots & Y_{n,r} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^T & \boldsymbol{y}_1^T \\ \boldsymbol{x}_2^T & \boldsymbol{y}_2^T \\ \vdots & \\ \boldsymbol{x}_n^T & \boldsymbol{y}_n^T \end{bmatrix}.$$

Let the multivariate linear regression model be $\boldsymbol{y} = \boldsymbol{\alpha} + \boldsymbol{B}^T \boldsymbol{x} + \boldsymbol{\epsilon}$ where $\boldsymbol{B} = [\boldsymbol{\beta}_1 \boldsymbol{\beta}_2 \cdots \boldsymbol{\beta}_r]$. The OLS estimator $\hat{\boldsymbol{B}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{xy}}$ minimizes the criterion $Q(\boldsymbol{\alpha}, \boldsymbol{B}) = \sum_{i=1}^r Q_i(\alpha_i, \boldsymbol{\beta}_i) = \sum_{i=1}^r [\sum_{j=1}^n r_j^2(\alpha_i, \boldsymbol{\beta}_i)] = \sum_{i=1}^r [\sum_{j=1}^n (Y_{ji} - \alpha_i - \boldsymbol{\beta}_i^T \boldsymbol{x}_j)^2]$ where Q_i is the OLS criterion for the multiple linear regression of the single response variable Y_i on $\boldsymbol{x}$. Hence performing the OLS multivariate linear regression of $\boldsymbol{y}$ on $W_i = \hat{\boldsymbol{\beta}}_i^T \boldsymbol{x}$ for $i = 1, ..., r$ does not change the OLS estimator.

To obtain the $k-$component estimator, 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 $\boldsymbol{D}_k = [\boldsymbol{\theta}_1 \boldsymbol{\theta}_2 \cdots \boldsymbol{\theta}_k]$. Fit the OLS working model

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

with $\hat{\boldsymbol{B}}_k^T = \hat{\boldsymbol{D}}_k^T \hat{\boldsymbol{A}}_k$. Then $\hat{\boldsymbol{D}}_k = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{wy}}$, and $\hat{\boldsymbol{B}}_k = \hat{\boldsymbol{A}}_k^T \hat{\boldsymbol{D}}_k = \hat{\boldsymbol{A}}_k^T (\hat{\boldsymbol{A}}_k \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} \hat{\boldsymbol{A}}_k^T)^{-1} \hat{\boldsymbol{A}}_k \hat{\boldsymbol{\Sigma}}_{\boldsymbol{xy}}$ by Remark 1. Here $k = 1, ..., M$ and $k = \hat{q}$ is of interest.

A common competitor is to fit the r univariate MLR's $Y_j = \alpha_j + \boldsymbol{\eta}_j^T \boldsymbol{x}$ with a method, such as lasso or model selection PLS, 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

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

where $\hat{\boldsymbol{A}}_H \boldsymbol{x} = \boldsymbol{w} = (W_1, ..., W_r)^T$, $W_i = \hat{\boldsymbol{\eta}}_i^T \boldsymbol{x}$, and $\hat{\boldsymbol{\alpha}}_U = (\hat{\alpha}_1, ..., \hat{\alpha}_r)^T$.

A new (high dimensional) method is to regress $\boldsymbol{y}$ on $\hat{\boldsymbol{A}}_H \boldsymbol{x} = \boldsymbol{w}$ with OLS. So let $\boldsymbol{D}_H = [\boldsymbol{\theta}_1 \boldsymbol{\theta}_2 \cdots \boldsymbol{\theta}_r]$, and fit the OLS working model

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

to get $\hat{\boldsymbol{D}}_H = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{wy}}$, and $\hat{\boldsymbol{B}}_H = \hat{\boldsymbol{A}}_H^T \hat{\boldsymbol{D}}_H = \hat{\boldsymbol{A}}_H^T (\hat{\boldsymbol{A}}_H \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} \hat{\boldsymbol{A}}_H^T)^{-1} \hat{\boldsymbol{A}}_H \hat{\boldsymbol{\Sigma}}_{\boldsymbol{xy}}$. Note that the jth column of $\hat{\boldsymbol{B}}_H$ is $\hat{\boldsymbol{\beta}}_j = \sum_{i=1}^r \hat{\theta}_{ij} \hat{\boldsymbol{\eta}}_i$. If r is too large, then randomly select K of the

Y_i, update K of the $\hat{\boldsymbol{\eta}}_i$ by applying $\hat{\boldsymbol{B}}_H$ on the K response variables. Then repeat many times.

The new estimator $\hat{\boldsymbol{B}}_H$ is useful if the $\hat{\boldsymbol{\eta}}_i$ are not obtained from a working model for an OLS multivariate linear regression, and if $\hat{\boldsymbol{D}}_H$ is not close to the identity matrix $\boldsymbol{I}_r$. Univariate lasso estimators could be used to give a competitor for the `glmnet` lasso estimator with "mgaussian." See Qian et al. (2022).

For the $k-$component lasso estimator (with "mgaussian"), let the standardized predictors $v_i = x_i/\hat{\sigma}_i$ for $i = 1, ..., p$. Apply lasso to the model $\boldsymbol{y} = \boldsymbol{\alpha} + \boldsymbol{B}^T\boldsymbol{v} + \boldsymbol{\epsilon}$. Let $\hat{\boldsymbol{b}}_j$ correspond to the jth row of the lasso estimator $\hat{\boldsymbol{B}} = \hat{\boldsymbol{B}}_L$. Let $W_i = x_{i_j}$ correspond to predictors with the largest $\|\hat{\boldsymbol{b}}_i\|$. If more than one $\hat{\boldsymbol{b}}_i = \boldsymbol{0}$, rank the corresponding predictors $x_i = \boldsymbol{c}_i^T\boldsymbol{x}$ using Equation (18) with $\hat{\boldsymbol{\gamma}}_i = \boldsymbol{c}_i$.

The $k-$component PCA estimator using Equation (18) can be used with $r > 1$ for many regression and classification methods. Since the response variables are used to select the PLS W_i, the PLS W_i should work better than the PCA W_i.

3.1.1 Variable Selection with d Linear Combinations

Suppose the regression or classification model has $d \geq 1$ linear combinations. Denote the jth linear combination by $SP_j(I) = \alpha_j + \boldsymbol{\beta}_j^T\boldsymbol{x}_I$ and $ESP_j(I) = \hat{\alpha}_j + \hat{\boldsymbol{\beta}}_j^T\boldsymbol{x}_I$. For multivariate linear regression there are $d = r$ linear combinations, one for each response variable Y_j. For linear discriminant analysis (LDA) there are $d = r$ linear combinations, one for each of the r groups. In low dimensions, Barker and Rayens (2003) proved that PLS-LDA with p components and LDA using $\boldsymbol{x} = (x_1, ..., x_p)^T$ give identical results when $\boldsymbol{S}_{\boldsymbol{x}} > 0$. See Cook and Forzani, (2024, p. 207).

The variable selection model given by Equations (4) and (5) can be extended to d linear combinations as follows. Let $VSP(I) = (SP_1(I), ..., SP_d(I))^T$ and $VESP(I) = (ESP_1(I), ..., ESP_d(I))^T$. Then a model for variable selection can be described by

$$VSP(F) = VSP(S) + VSP(E) = VSP(S) \tag{20}$$

where $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$ is a $p \times 1$ vector of predictors, $\boldsymbol{x}_S$ is an $d_S \times 1$ vector, and $\boldsymbol{x}_E$ is a $(p-d_S) \times 1$ vector. Given that $\boldsymbol{x}_S$ is in the model, $\boldsymbol{\beta}_{jE} = \boldsymbol{0}$ for $j = 1, ..., d$, and E denotes the subset of terms that can be eliminated given that the subset S is in the model.

Since S is unknown, candidate subsets will be examined. Let $\boldsymbol{x}_I$ be the vector of d_I terms from a candidate subset indexed by I, including a constant, and let $\boldsymbol{x}_O$ be the vector of the remaining predictors (out of the candidate submodel). If $S \subseteq I$, then

$$VSP(F) = VSP(S) = VSP(I). \tag{21}$$

Some notation is needed for sample quantities. Let $\text{Cor}(x, y)$ be the population correlation of random variables x and y, and let $\text{cor}(x, y) = \text{cor}(\boldsymbol{x}, \boldsymbol{y})$ be the sample correlation computed from the data vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. Let $\text{cor}(ESP_i(I), ESP_i(F)) = \text{cor}(\boldsymbol{v}_i(I), \boldsymbol{v}_i(F))$, and let $\boldsymbol{v}(I) = (\boldsymbol{v}_1(I)^T, \boldsymbol{v}_2(I)^T, ..., \boldsymbol{v}_d(I)^T)^T$. Let

$$VCor(VSP(I), VSP(F)) = (\text{Cor}(SP_1(I), SP_1(F)), ..., \text{Cor}(SP_d(I), SP_d(F)))^T,$$

and

$$Vcor(VESP(I), VESP(F)) = (\text{cor}(ESP_1(I), ESP_1(F)), ..., \text{cor}(ESP_d(I), ESP_d(F)))^T =$$

$$(\text{cor}(\boldsymbol{v}_1(I), \boldsymbol{v}_1(F)), ..., \text{cor}(\boldsymbol{v}_d(I), \boldsymbol{v}_d(F)))^T.$$

If $S \subseteq I$, then $VCor(VSP(I), VSP(F)) = \mathbf{1} = (1, ..., 1)^T$. Note that $\boldsymbol{v}_i(I) = (\hat{\alpha}_i + \hat{\boldsymbol{\beta}}_i^T \boldsymbol{x}_{1I}, ..., \hat{\alpha}_i + \hat{\boldsymbol{\beta}}_i^T \boldsymbol{x}_{nI})^T$ where the n cases are $(\boldsymbol{x}_1, \boldsymbol{y}_1), ..., (\boldsymbol{x}_1, \boldsymbol{y}_1)$.

In low dimensions, good variable selection estimators keep $Vcor(VESP(I), VESP(F))$ near $\mathbf{1}$ and $\text{cor}(\boldsymbol{v}(I), \boldsymbol{v}(F))$ high, by keeping wanted predictors I in the model and omitting unwanted predictors O. The plotted points in a plot of $\boldsymbol{v}(I)$ versus $\boldsymbol{v}(F)$ should cluster tightly about the identity line. For the multivariate linear regression estimator, $\boldsymbol{v}(I) = (\hat{\boldsymbol{Y}}_1^T(I), ..., \hat{\boldsymbol{Y}}_r^T(I))^T = vec(\boldsymbol{x}_I^T \hat{\boldsymbol{B}}_I)$, which has the nr fitted values from model I stacked into a vector.

Heuristically, if the full model is good in that it gives good predictions, then a submodel I with $Vcor(VESP(I), VESP(F))$ near $\mathbf{1}$ and $\text{cor}(\boldsymbol{v}(I), \boldsymbol{v}(F))$ high, gives good predictions that are nearly identical to those of the full model. Thus "everything sensible works for prediction."

Let $I_j = \{1, ..., j\}$ so that $\boldsymbol{w}_{I_j} = (W_1, W_2, ..., W_j)^T$. Then the $k-$component estimator has $VSP(I_k)$ and $VESP(I_k)$. If the $\hat{\boldsymbol{\gamma}}_i$ come from PCA or if $\hat{\boldsymbol{\gamma}}_i = \boldsymbol{b}_i =$ the ith column of $\boldsymbol{I}_p$, then the $W_i = \hat{\boldsymbol{\gamma}}_i^T \boldsymbol{x}$ can be ordered using, for example, the SIMPLS criterion $Q_{\boldsymbol{Z}}(\hat{\boldsymbol{\gamma}}_i)$.

Suppose that $\hat{q}$ of the W_i are selected. Then even in high dimensions, $\text{cor}(\boldsymbol{v}(I_j), \boldsymbol{v}(I_{\hat{q}}))$ and $Vcor(VESP(I_j), VESP(I_{\hat{q}}))$ can be computed for $i = 1, ..., J$ for some J. For SDR and envelopes, would like the correlations to be high for $j = q+1, ..., J$.

4 The OPLS Estimator

The OPLS estimator is the PLS estimator from Section 2 with $k = 1$. Then the ESP $= \hat{\alpha}_1 + \hat{\theta}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}^T \boldsymbol{x} = \hat{\alpha}_{OPLS} + \hat{\boldsymbol{\beta}}_{OPLS}^T \boldsymbol{x}$ where $\hat{\boldsymbol{\beta}}_{OPLS} = \hat{\theta}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$. Let $\hat{\boldsymbol{\eta}}_{OPLS} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$. Testing $H_0 : \boldsymbol{A}\boldsymbol{\beta}_{OPLS} = \mathbf{0}$ versus $H_1 : \boldsymbol{A}\boldsymbol{\beta}_{OPLS} \neq \mathbf{0}$ is equivalent to testing $H_0 : \boldsymbol{A}\boldsymbol{\eta} = \mathbf{0}$ versus $H_1 : \boldsymbol{A}\boldsymbol{\eta} \neq \mathbf{0}$ where $\boldsymbol{A}$ is a $k \times p$ constant matrix and $\boldsymbol{\eta} = \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$.

For multiple linear regression, Cook, Helland, and Su (2013) and Basa et al. (2024) showed that $\hat{\boldsymbol{\beta}}_{OPLS} = \hat{\theta}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$ estimates $\theta\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \boldsymbol{\beta}_{OPLS}$ where

$$\theta = \frac{\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T \boldsymbol{\Sigma}_{\boldsymbol{x}Y}}{\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\Sigma}_{\boldsymbol{x}Y}} \text{ and } \hat{\theta} = \frac{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\text{Y}}^{\text{T}} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\text{Y}}}{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\text{Y}}^{\text{T}} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\text{Y}}} \tag{22}$$

for $\boldsymbol{\Sigma}_{\boldsymbol{x}Y} \neq \mathbf{0}$. If $\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \mathbf{0}$, then $\boldsymbol{\beta}_{OPLS} = \mathbf{0}$.

Next, some large sample theory is reviewed for $\hat{\boldsymbol{\eta}}_{OPLS} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$ and OPLS for the multiple linear regression model, including some high dimensional tests for low dimensional quantities such as $H_0 : \beta_i = 0$ or $H_0 : \beta_i - \beta_j = 0$. These tests depended on iid cases, but not on linearity or the constant variance assumption. Hence the tests are useful for multiple linear regression with heterogeneity.

The following Olive and Zhang (2025) theorem gives the large sample theory for $\hat{\boldsymbol{\eta}} = \widehat{\text{Cov}}(\boldsymbol{x}, Y)$. Olive et al. (2025) gave alternative proofs. This theory needs $\boldsymbol{\eta} = \boldsymbol{\eta}_{OPLS} =$

$\boldsymbol{\Sigma}_{\boldsymbol{x},Y}$ to exist for $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y}$ to be a consistent estimator of $\boldsymbol{\eta}$. Let $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$ and let $\boldsymbol{w}_i$ and $\boldsymbol{z}_i$ be defined below where

$$\mathrm{Cov}(\boldsymbol{w}_i) = \boldsymbol{\Sigma}_{\boldsymbol{w}} = E[(\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}})(\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}})^T (Y_i - \mu_Y)^2)] - \boldsymbol{\Sigma}_{\boldsymbol{x}Y} \boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T.$$

Then the low order moments are needed for $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}$ to be a consistent estimator of $\boldsymbol{\Sigma}_{\boldsymbol{w}}$.

Theorem 7. Assume the cases $(\boldsymbol{x}_i^T, Y_i)^T$ are iid. Assume $E(x_{ij}^k \; Y_i^m)$ exist for $j = 1, \ldots, p$ and $k, m = 0, 1, 2$. Let $\boldsymbol{\mu}_{\boldsymbol{x}} = E(\boldsymbol{x})$ and $\mu_Y = E(Y)$. Let $\boldsymbol{w}_i = (\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}})(Y_i - \mu_Y)$ with sample mean $\overline{\boldsymbol{w}}_n$. Let $\boldsymbol{\eta} = \boldsymbol{\Sigma}_{\boldsymbol{x},Y}$. Then (a)

$$\sqrt{n}(\overline{\boldsymbol{w}}_n - \boldsymbol{\eta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}}), \;\; \sqrt{n}(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}}), \tag{23}$$

$$and \;\; \sqrt{n}(\tilde{\boldsymbol{\eta}}_n - \boldsymbol{\eta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}}).$$

(b) Let $\boldsymbol{v}_i = (\boldsymbol{x}_i - \overline{\boldsymbol{x}}_n)(Y_i - \overline{Y}_n)$. Then $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{v}} + O_P(n^{-1/2})$. Hence $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{v}} + O_P(n^{-1/2})$.
(c) Let $\boldsymbol{A}$ be a $k \times p$ full rank constant matrix with $k \leq p$, assume $H_0 : \boldsymbol{A}\boldsymbol{\beta}_{OPLS} = \boldsymbol{0}$ is true, and assume $\hat{\theta} \xrightarrow{P} \theta \neq 0$. Then

$$\sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{\beta}}_{OPLS} - \boldsymbol{\beta}_{OPLS}) \xrightarrow{D} N_k(\boldsymbol{0}, \theta^2 \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{w}}\boldsymbol{A}^T). \tag{24}$$

For the following theorem, consider a subset of k distinct elements from $\tilde{\boldsymbol{\Sigma}}$ or from $\hat{\boldsymbol{\Sigma}}$. Stack the elements into a vector, and let each vector have the same ordering. For example, the largest subset of distinct elements corresponds to

$$vech(\tilde{\boldsymbol{\Sigma}}) = (\tilde{\sigma}_{11}, \ldots, \tilde{\sigma}_{1p}, \tilde{\sigma}_{22}, \ldots, \tilde{\sigma}_{2p}, \ldots, \tilde{\sigma}_{p-1,p-1}, \tilde{\sigma}_{p-1,p}, \tilde{\sigma}_{pp})^T = [\tilde{\sigma}_{jk}].$$

For random variables $x_1, \ldots, x_p$, use notation such as $\overline{x}_j$ = the sample mean of the x_j, $\mu_j = E(x_j)$, and $\sigma_{jk} = Cov(x_j, x_k)$. Let

$$n \; vech(\tilde{\boldsymbol{\Sigma}}) = [n \; \tilde{\sigma}_{jk}] = \sum_{i=1}^{n} [(x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k)].$$

For general vectors of elements, the ordering of the vectors will all be the same and be denoted by vectors such as $\hat{\boldsymbol{c}} = [\hat{\sigma}_{jk}]$, $\tilde{\boldsymbol{c}} = [\tilde{\sigma}_{jk}]$, $\boldsymbol{c} = [\sigma_{jk}]$, $\boldsymbol{v}_i = [(x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k)]$, and $\boldsymbol{w}_i = [(x_{ij} - \mu_j)(x_{ik} - \mu_k)]$. Let $\overline{\boldsymbol{w}}_n = \sum_{i=1}^n \boldsymbol{w}_i / n$ be the sample mean of the $\boldsymbol{w}_i$. Assuming that $Cov(\boldsymbol{w}_i) = \boldsymbol{\Sigma}_{\boldsymbol{w}}$ exists, then $E(\boldsymbol{w}_i) = E(\overline{\boldsymbol{w}}_n) = \boldsymbol{c}$.

The following Olive et al. (2025) theorem provides large sample theory for $\hat{\boldsymbol{c}}$ and $\tilde{\boldsymbol{c}}$. We use $Cov(\boldsymbol{w}_i) = \boldsymbol{\Sigma}_{\boldsymbol{d}}$ to avoid confusion with the $\boldsymbol{\Sigma}_{\boldsymbol{w}}$ used in Theorem 3. Note that $\boldsymbol{x}_i$ are dummy variables and could be replaced by $\boldsymbol{u}_i = (Y_{i1}, \ldots, Y_{im}, x_{i1}, \ldots, x_{ip})^T$ to get information about m response variables $Y_1, \ldots, Y_m$.

Theorem 8. Assume the cases $\boldsymbol{x}_i$ are iid and that $Cov(\boldsymbol{w}_i) = \boldsymbol{\Sigma}_{\boldsymbol{d}}$ exists. Using the above notation with $\boldsymbol{c}$ a $k \times 1$ vector,

(i) $\sqrt{n}(\tilde{\boldsymbol{c}} - \boldsymbol{c}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{d}})$.

(ii) $\sqrt{n}(\hat{\boldsymbol{c}} - \boldsymbol{c}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{d}})$.

(iii) $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{v}} + O_P(n^{-1/2})$ and $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{d}} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{v}} + O_P(n^{-1/2})$.

5 Large Sample Theory and Testing

Suppose the classification or regression model has a response variable Y that depends on the predictors $\boldsymbol{x}$ through $SP = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$. In low dimensions, important tests include a) $H_0 : \beta_i = 0$ (the Wald tests for MLR), b) $H_0 : \boldsymbol{\beta} = \mathbf{0}$ (the Anova F test for MLR), and c) $H_0 : (\beta_{i_1}, \ldots, \beta_{i_k})^T = \mathbf{0}$ (the partial F test for MLR).

This section will derive some high dimensional analogs of the above tests.

5.1 Testing $H_0 : \boldsymbol{\beta} = \mathbf{0}$

An Omnibus or Universal Test

This subsection follows Abid, Quaye, and Olive (2025) closely. Consider classification and regression models where the response variable Y only depends on the $p \times 1$ vector of predictors $\boldsymbol{x} = (x_1, ..., x_p)^T$ through the sufficient predictor $SP = \alpha + \boldsymbol{x}^T\boldsymbol{\beta}$. Let the covariance vector $\mathrm{Cov}(\boldsymbol{x}, Y) = \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$. Assume the cases $(\boldsymbol{x}_i^T, Y_i)^T$ are iid random vectors for $i = 1, ..., n$. Then for many such regression models, $\boldsymbol{\beta} = \mathbf{0}$ if and only if $\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \mathbf{0}$ where $\mathbf{0} = (0, ..., 0)^T$ is the $p \times 1$ vector of zeroes.

The test of $H_0 : \boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \mathbf{0}$ versus $H_1 : \boldsymbol{\Sigma}_{\boldsymbol{x}Y} \neq \mathbf{0}$ is equivalent to the high dimensional one sample test $H_0 : \boldsymbol{\mu} = \mathbf{0}$ versus $H_A : \boldsymbol{\mu} \neq \mathbf{0}$ applied to $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$ where $\boldsymbol{w}_i = (\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}})(Y_i - \mu_Y)$ and the expected values $E(\boldsymbol{x}) = \boldsymbol{\mu}_{\boldsymbol{x}}$ and $E(Y) = \mu_Y$. Since $\boldsymbol{\mu}_{\boldsymbol{x}}$ and μ_Y are unknown, the test of $H_0 : \boldsymbol{\beta} = \mathbf{0}$ versus $H_1 : \boldsymbol{\beta} \neq \mathbf{0}$ is implemented by applying the one sample test to $\boldsymbol{v}_i = (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(Y_i - \overline{Y})$ for $i = 1, ..., n$.

Zhao et al. (2024) have an interesting result for the multiple linear regression model (2). Assume that the cases $(\boldsymbol{x}_i^T, Y_i)^T$ are iid with $E(Y) = \mu_Y$, $E(\boldsymbol{x}) = \boldsymbol{\mu}_{\boldsymbol{x}}$ and nonsingular $Cov(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$. Let $\boldsymbol{\beta} = \boldsymbol{\beta}_{OLS}$. Then testing $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ versus $H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0$ is equivalent to testing $H_0 : \boldsymbol{\mu} = \mathbf{0}$ versus $H_1 : \boldsymbol{\mu} \neq \mathbf{0}$ with $\boldsymbol{\mu} = E(\boldsymbol{w}_i) = \boldsymbol{\Sigma}_{\boldsymbol{x}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$ where $\boldsymbol{w}_i = (\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}})(Y_i - \mu_Y - (\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}})^T\boldsymbol{\beta}_0)$, and a one sample test can be applied to $\boldsymbol{v}_i = (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(Y_i - \overline{Y} - (\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T\boldsymbol{\beta}_0)$.

Abid, Quaye, and Olive (2025) used the above test for $\boldsymbol{\beta}_0 = \mathbf{0}$. The resulting test can be used for many regression models, not just multiple linear regression. Suppose $\boldsymbol{\beta}_D = \boldsymbol{D}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ where $\boldsymbol{D}$ is a $p \times p$ nonsingular matrix. Then $\boldsymbol{\beta}_D = \mathbf{0}$ if and only if $\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \mathbf{0}$. Then $\boldsymbol{D}^{-1} = \theta\boldsymbol{I}$ for OPLS, $\boldsymbol{D}^{-1} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}$ for OLS, and $\boldsymbol{D}^{-1} = [diag(\boldsymbol{\Sigma}_{\boldsymbol{x}})]^{-1}$ for the MMLE for multiple linear regression (MLR). By Theorem 1e), $\boldsymbol{\beta}_{kPLS} = \mathbf{0}$ if $\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \mathbf{0}$. Thus if the cases $(\boldsymbol{x}_i^T, Y_i)^T$ are iid, then using $\boldsymbol{\beta}_0 = \mathbf{0}$ gives tests for $H_0 : \boldsymbol{\beta} = \mathbf{0}$, $H_0 : \boldsymbol{\beta}_{MMLE} = \mathbf{0}$ (for MLR), $H_0 : \boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \mathbf{0}$, $H_0 : \boldsymbol{\beta}_{OPLS} = \mathbf{0}$, and $H_0 : \boldsymbol{\beta}_{kPLS} = \mathbf{0}$. For multiple linear regression with heterogeneity, $\hat{\boldsymbol{\beta}}_{OLS}$ is still a consistent estimator of $\boldsymbol{\beta} = \boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$. Hence the test can be used when the constant variance assumption is violated.

Assume the cases $(\boldsymbol{x}_i^T, Y_i)^T$ are iid. For a generalized linear model and several other regression models that depend on the predictors $\boldsymbol{x}$ only through $SP = \alpha + \boldsymbol{x}^T\boldsymbol{\beta}$, if $\boldsymbol{\beta} = \mathbf{0}$, then the Y_i are iid and do not depend on $\boldsymbol{x}$, and thus satisfy a multiple linear regression model with $\boldsymbol{\beta}_{OLS} = \mathbf{0}$. Typically, if $\boldsymbol{\beta} \neq \mathbf{0}$, then $\boldsymbol{\Sigma}_{\boldsymbol{x}Y} \neq 0$. Also see Theorem 2 b). An exception is when there is a lot of symmetry which rarely occurs with real data. For example, suppose $Y = m(SP) + e$ where the iid errors $e_i \sim N(0, \sigma_1^2)$ are

independent of the predictors, $SP \sim N(0, \sigma_2^2)$, and the function m is symmetric about 0, e.g. $m(SP) = (SP)^2$. Then $\boldsymbol{\beta}_{OLS} = 0$ and $\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = 0$ even if $\boldsymbol{\beta} \neq \mathbf{0}$.

If $\boldsymbol{\beta}_0 = \mathbf{0}$, then $\boldsymbol{w}_i = (\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}})(Y_i - \mu_Y)$, and $E(\boldsymbol{w}_i) = E(\boldsymbol{u}_i) = E[\boldsymbol{x}_i(Y_i - \mu_Y)] = \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$. Then apply a high dimensional one sample test on the $\boldsymbol{v}_i = (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(Y_i - \overline{Y})$. Note that the sample mean $\overline{\boldsymbol{v}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$.

Suppose $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid random vectors with $E(\boldsymbol{x}) = \boldsymbol{\mu}$ and covariance matrix $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}$. Then the test $H_0 : \boldsymbol{\mu} = \mathbf{0}$ versus $H_1 : \boldsymbol{\mu} \neq \mathbf{0}$ is equivalent to the test $H_0 : \boldsymbol{\mu}^T\boldsymbol{\mu} = 0$ versus $H_1 : \boldsymbol{\mu}^T\boldsymbol{\mu} \neq 0$. Let $\boldsymbol{S} = \hat{\boldsymbol{\Sigma}}$. A U-statistic for estimating $\boldsymbol{\mu}^T\boldsymbol{\mu}$ is

$$T_n = T_n(\boldsymbol{x}) = \frac{1}{n(n-1)} \sum_{i \neq j} \boldsymbol{x}_i^T \boldsymbol{x}_j = \frac{n\overline{\boldsymbol{x}}^T\overline{\boldsymbol{x}} - tr(\boldsymbol{S})}{n} \tag{25}$$

where tr() is the trace function. See, for example, Abid, Quaye, and Olive (2025).

Let the variance $V(W) = V(W_{ij}) = V(\boldsymbol{x}_i^T\boldsymbol{x}_j) = \sigma_W^2$ for $i \neq j$. Let $m = \text{floor}(\text{n}/2) = \lfloor \text{n}/2 \rfloor$ be the integer part of $n/2$. So $\text{floor}(100/2) = \text{floor}(101/2) = 50$. Let the iid random variables $W_i = \boldsymbol{x}_{2i-1}^T\boldsymbol{x}_{2i}$ for $i = 1, ..., m$. Hence $W_1, W_2, ..., W_m = \boldsymbol{x}_1^T\boldsymbol{x}_2, \boldsymbol{x}_3^T\boldsymbol{x}_4, ..., \boldsymbol{x}_{2m-1}^T\boldsymbol{x}_{2m}$. Note that $E(W_i) = \boldsymbol{\mu}^T\boldsymbol{\mu}$ and $V(W_i) = \sigma_W^2$. Let S_W^2 be the sample variance of the W_i:

$$S_W^2 = \frac{1}{m-1} \sum_{i=1}^{m} (W_i - \overline{W})^2. \tag{26}$$

Zhao et al. (2024, p. 2024) showed that $\sigma_W^2 = tr(\boldsymbol{\Sigma}^2) + 2\boldsymbol{\mu}^T\boldsymbol{\Sigma}\boldsymbol{\mu}$.

The following Abid, Quaye, and Olive (2025) theorem derived the variance $V(T_n)$ under simpler regularity conditions than those in the literature. The second formula in Theorem 5a) was obtained by Chen and Qin (2010).

Theorem 9. Assume $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid, $E(\boldsymbol{x}_i) = \boldsymbol{\mu}$, and the variance $V(\boldsymbol{x}_i^T\boldsymbol{x}_j) = \sigma_W^2$ for $i \neq j$. Let $W_{ij} = \boldsymbol{x}_i^T\boldsymbol{x}_j$ for $i \neq j$. Let $\theta = Cov(W_{ij}, W_{id}) = \boldsymbol{\mu}^T\boldsymbol{\Sigma}\boldsymbol{\mu}$ where $j \neq d$, $i < j$, and $i < d$. Then

$$a)\ V(T_n) = \frac{2\sigma_W^2}{n(n-1)} + \frac{4(n-2)\theta}{n(n-1)} = \frac{2}{n(n-1)} tr(\boldsymbol{\Sigma}^2) + \frac{4\boldsymbol{\mu}^T\boldsymbol{\Sigma}\boldsymbol{\mu}}{n}.$$

b) If $H_0 : \boldsymbol{\mu} = \mathbf{0}$ is true, then $\theta = 0$ and

$$V_0 = V(T_n) = \frac{2\sigma_W^2}{n(n-1)} = \frac{2tr(\boldsymbol{\Sigma}^2)}{n(n-1)} = \frac{2\sigma_W^2 - 4\theta}{n(n-1)}.$$

Let $\hat{V}(T_n)$ and $\hat{V}_0(T_n)$ be consistent estimators of $V(T_n)$ and $V_0(T_n)$, respectively. Then Srivastava and Du (2008), Bai and Saranadasa (1996), Chen and Qin (2010), Li (2023), and others proved that under mild regularity conditions when H_0 is true,

$$T_n / \sqrt{\hat{V}(T_n)} = T_n / \sqrt{\hat{V}_0(T_n)} \xrightarrow{D} N(0, 1).$$

Under regularity conditions when H_0 is true, Li (2023) proved that $T_n / \sqrt{\hat{V}_0(T_n)} \xrightarrow{D} t_k$ as $p \to \infty$ for fixed $n \geq 3$ where $k = 0.5n(n-1) - 1$.

A consistent estimator of $V_0(T_n)$ needs a consistent estimator of $\sigma_W^2 = 0.5n(n-1)$ $V_0(T_n)$. Let $s_n^2 = \hat{V}_0(T_n)$. Then one estimator is $0.5n(n-1)s_n^2 = S_W^2$ from Equation (8). An estimator nearly the same as the one used by Li (2023) is

$$0.5n(n-1)s_n^2 = \hat{\sigma}_W^2 = \frac{1}{n(n-1)} \sum \sum_{i \neq j} (\boldsymbol{x}_i^T \boldsymbol{x}_j - T_n)^2 = \frac{1}{n(n-1)} \sum \sum_{i \neq j} (W_{ij} - T_n)^2.$$

A New Competing Test

If the parametric distribution D is known, then the iid cases assumption can be changed to independent cases. Assume $Y_i | \boldsymbol{x}_i^T \boldsymbol{\beta} \sim D(\tau(\alpha + \boldsymbol{x}_i^T \boldsymbol{\beta}), \boldsymbol{\theta})$. If $\boldsymbol{\beta} = \mathbf{0}$, then the iid $Y_i \sim D(\tau(\alpha), \boldsymbol{\theta})$. Hence testing $H_0 : \boldsymbol{\beta} = \mathbf{0}$ vs. $H_1 : \boldsymbol{\beta} \neq \mathbf{0}$ is equivalent to testing whether the Y_i are a random sample from the $D(\tau(\alpha), \boldsymbol{\theta})$ distribution. Such a test can be done with the Kolmogorov-Smirnov test, the chi-square test, the Anderson-Darling test, the Cramér-von Mises test, et cetera. For specific distributions, there are often tests. For example, the Lilliefors test can be used to test if the Y_i are iid from a $N(\mu, \sigma^2)$ distribution where μ and σ^2 are unknown. See, for example, Kellison and London (2011, pp. 455-465), Conover (1971, pp. 295-308), Zheng, Lai, and Gould (2023), and Zheng et al. (2025).

This test has great level and extreme dimension reduction since the test does not depend on the predictors $\boldsymbol{x}$. The power can be sometimes be very poor if the cases are iid. a) If the $(Y_i, \boldsymbol{x}_i^T)^T$ are iid from a multivariate normal distribution, then the Y_i are iid $N(\mu_Y, \sigma_Y^2)$ regardless of whether $\boldsymbol{\beta} = \mathbf{0}$ or $\boldsymbol{\beta} \neq \mathbf{0}$ for the multiple linear regression model $Y | (\alpha + \boldsymbol{x}^T \boldsymbol{\beta}) \sim N(\alpha + \boldsymbol{x}^T \boldsymbol{\beta}, \sigma^2)$. b) If the $(Y_i, \boldsymbol{x}_i^T)^T$ are iid from some distribution where the $Y_i \in \{0, 1\}$ are binary, then the Y_i are iid $bin(n = 1, \rho_Y)$ regardless of whether $\boldsymbol{\beta} = \mathbf{0}$ or $\boldsymbol{\beta} \neq \mathbf{0}$ for the binary regression model $Y | (\alpha + \boldsymbol{x}^T \boldsymbol{\beta}) \sim bin(n = 1, \rho(\alpha + \boldsymbol{x}^T \boldsymbol{\beta}))$.

The test does not depend on $\boldsymbol{x}$, and can thus be done after variable selection. Also, all of the predictors can have outliers and missing values.

A Test for Binary Regression or Classification

Olive (2017, pp. 396-397) gave the result for a binary response variable $Y \in \{0, 1\}$.

Theorem 10. Let $\pi_j = P(Y = j)$ for $j = 0, 1$. Let $\boldsymbol{\mu}_j = E(\boldsymbol{x} | Y = j)$ for $j = 0, 1$. Then a) $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \hat{\pi}_1 \hat{\pi}_0 (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)$, and b) $\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \pi_1 \pi_0 (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$.

Proof. Let N_i be the number of Ys that are equal to i for $i = 0, 1$ with $n = N_1 + N_2$. Then

$$\hat{\boldsymbol{\mu}}_i = \frac{1}{N_i} \sum_{j:Y_j = i} \boldsymbol{x}_j$$

for $i = 0, 1$ while $\hat{\pi}_i = N_i / n$ and $\hat{\pi}_1 = 1 - \hat{\pi}_0$. Hence $\hat{\boldsymbol{\mu}}_i = \overline{\boldsymbol{x}}_i$ is the sample mean of the $\boldsymbol{x}_k$ corresponding to $Y_k = j$ for $j = 0, 1$. Then

$$\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i Y_i - \overline{\boldsymbol{x}}\, \overline{Y}.$$

$$\text{Thus } \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}\mathrm{Y}} = \frac{1}{\mathrm{n}} \left[\sum_{\mathrm{j:Y_j=1}} \boldsymbol{x}_{\mathrm{j}}(1) + \sum_{\mathrm{j:Y_j=0}} \boldsymbol{x}_{\mathrm{j}}(0) \right] - \overline{\boldsymbol{x}}\, \hat{\pi}_1 =$$

$$\frac{1}{n}(N_1\hat{\boldsymbol{\mu}}_1) - \frac{1}{n}(N_1\hat{\boldsymbol{\mu}}_1 + N_0\hat{\boldsymbol{\mu}}_0)\hat{\pi}_1 = \hat{\pi}_1\hat{\boldsymbol{\mu}}_1 - \hat{\pi}_1^2\hat{\boldsymbol{\mu}}_1 - \hat{\pi}_1\hat{\pi}_0\hat{\boldsymbol{\mu}}_0 =$$
$$\hat{\pi}_1(1-\hat{\pi}_1)\hat{\boldsymbol{\mu}}_1 - \hat{\pi}_1\hat{\pi}_0\hat{\boldsymbol{\mu}}_0 = \hat{\pi}_1\hat{\pi}_0(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0).$$

Thus $\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \pi_1\pi_0(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$. $\square$

This result means $\boldsymbol{\eta} = \boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \pi_1\pi_0(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$ and $\boldsymbol{\phi} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$ are quantities of interest for binary regression. Note that $\boldsymbol{x} = (w_1, ..., w_k, w_1w_2, ..., w_1w_k, ..., w_{k-1}w_k)^T$ could be used to include pairwise interactions of the w_i.

Theorem 2b) suggests that typically the binary regression $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{C}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$. If the cases $(Y_i, \boldsymbol{x}_i^T)^T$ are iid, then $H_0 : \boldsymbol{\beta} = \boldsymbol{0}$ can be tested with the omnibus test for $H_0 : \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$. If the cases within each group are iid, if the two groups are independent, and if $N_1/(N_1+N_2) \to \pi_1$, then $\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \pi_1\pi_0(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$ by Theorem 6b). Thus $H_0 : \boldsymbol{\beta} = \boldsymbol{0}$ can be tested with a high dimensional two sample test for $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_0$.

5.2 OPLS: Testing $H_0 : \beta_i = 0$

For OPLS, testing $H_0 : \beta_i = 0$ versus $H_A : \beta_i \neq 0$is equivalent to testing $H_0 : \eta_i = 0$ or $H_0 : Cov(x_i, Y) = 0$. Theorem 7 or Theorem 9 can be used.

5.3 OPLS: Testing $H_0 : \boldsymbol{\beta}_I = (\beta_{i_1}, ..., \beta_{i_k})^T = \boldsymbol{0}$

For OPLS, testing $H_0 : \boldsymbol{\beta}_I = \boldsymbol{0}$ versus $H_A : \boldsymbol{\beta}_I \neq \boldsymbol{0}$is equivalent to testing $H_0 : [Cov(x_{i_1}, Y), ..., Cov(x_{i_k}, Y)]^T = \boldsymbol{0}$. Theorem 9 can be used, or Theorem 7 can be used if $n \geq Jk$ with $J \geq 5$ where sometimes J needs to be much larger than 5.

High Dimensional Tests

Some tests when n/p is not large are simple. Testing $H_0 : \boldsymbol{A}\boldsymbol{\beta}_{BR} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{A}\boldsymbol{\beta}_{BR} \neq \boldsymbol{0}$ is equivalent to testing $H_0 : \boldsymbol{A}\boldsymbol{\eta} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{A}\boldsymbol{\eta} \neq \boldsymbol{0}$ where $\boldsymbol{A}$ is a $k \times p$ constant matrix. Let $\text{Cov}(\hat{\boldsymbol{\eta}}) = \boldsymbol{\Sigma}_{\boldsymbol{W}}$ be the asymptotic covariance matrix of $\hat{\boldsymbol{\eta}}$. In high dimensions where $n < 5p$, we can't get a good nonsingular estimator of $\text{Cov}(\hat{\boldsymbol{\eta}})$, but we can get good nonsingular estimators of $\text{Cov}((\hat{\eta}_{i1}, ..., \hat{\eta}_{ik})^T)$ with $\boldsymbol{u} = (x_{i1}, ..., x_{ik})^T$ where $n \geq Jk$ with $J \geq 10$. (Values of J much larger than 10 may be needed if some of the k predictors are skewed or if a π_i in near 0 or 1.) Simply use the sample covariance matrix with $\boldsymbol{u}$ replacing $\boldsymbol{x}$. Hence we can test hypotheses like $H_0 : \beta_i - \beta_j = 0$. In particular, testing $H_0 : \beta_i = 0$ is equivalent to testing $H_0 : \eta_i = 0$.

Data splitting uses model selection (variable selection is a special case) to reduce the high dimensional problem to a low dimensional problem. The above procedure also reduces the high dimensional problem to a low dimensional problem.

6 Bigger Model

Often there is a smaller constrained model and a bigger model without the constraints. Sometimes the bigger model greatly increases the applicability of the model.

6.1 Everyone is Trying to Estimate $\boldsymbol{\beta}$, and Nearly Nothing Works in High Dimensions

In high dimensions, it is very difficult to estimate a $p \times 1$ vector $\boldsymbol{\theta}$. This result is a form of "the curse of dimensionality." If a $\sqrt{n}$ consistent estimator of $\boldsymbol{\theta}$ is available, then the squared norm

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 = \sum_{i=1}^{p} (\hat{\theta}_i - \theta_i)^2 \propto p/n. \tag{27}$$

When p is fixed, $p/n \to 0$ as $n \to \infty$ and $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}$. In high dimensions, often the estimator has not been shown to be consistent, except under very strong regularity conditions.

Here $\boldsymbol{\theta} = \boldsymbol{\beta}$ or $\boldsymbol{\theta} = \boldsymbol{\gamma}$ are possible. For example, the sample eigenvectors $\hat{\boldsymbol{d}}_i$ tend to be poor estimators of the population eigenvectors $\boldsymbol{d}_i$ of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$. An exception is when the correlation $\text{Cor}(x_i, x_j) = \rho$ for $i \neq j$ where ρ is close to 1. See Jung and Marron (2009).

6.1.1 Bet on Sparsity Principle

The "don't bet on sparsity principle" is $S = F$ because $\beta_i = 0$ is not reasonable for any i. Note that with $r > 1$ response variables, the sparsity assumption becomes much stronger than with $r = 1$, since there is an $r \times (p - q)$ matrix of 0s.

The "bet on sparsity principle" makes the following assumption. Let I be the model selected, e.g., by lasso or elastic net.
a) $S \subseteq I$.
b) $\boldsymbol{\beta} = \boldsymbol{\beta}_F = \boldsymbol{\beta}_{I,0}$.
c) The number of variables a in I is small compared to n so that $\hat{\boldsymbol{\beta}}_{I,0}$ is a good estimator of $\boldsymbol{\beta}_F$. Hence $\hat{\boldsymbol{\beta}}_I$ is a good estimator of $\boldsymbol{\beta}_I$.

A useful bigger model is $\hat{\boldsymbol{\beta}}_I$ is a good estimator of $\boldsymbol{\beta}_I$. This assumption can be checked if $n \geq Ja$ where $J \geq 5$. As always, sometimes J much larger than 5 is needed. The bigger model greatly increases the applicability of lasso since assumptions a) and b) are not needed. The response plot of $ESP(I)$ versus Y is useful to check the model. See, for example, Olive (2013). Note that data splitting is for model $\boldsymbol{\beta}_I$, not for model $\boldsymbol{\beta}_{I,0} = \boldsymbol{\beta}_F$.

The "variable selection principle" is that in low dimensions, sensible variable selection estimators keep the sample correlation $\text{cor}(ESP(I), ESP(F))$ high. This principle corresponds to the bigger model, and can be checked.

Lasso selects no more than n predictors and a constant to be in the model. Lasso uses a grid $\lambda_1 < \lambda_2 < \cdots < \lambda_M$. When p is fixed, $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau$ does not do variable selection well. For variable selection, want $\hat{\lambda}_{1,n}/\sqrt{n} \to \infty$, but $\hat{\lambda}_{1,n}/n \to 0$. See Fan and Li (2001). Let $\lambda_1 = 2n\lambda$. Guan and Tibshirani (2020) (and likely `glmnet`) use $\lambda < Cn^{-1/4}$ for some large constant C. Hence $\lambda_{1,n} = \lambda_1 \propto n^{3/4}$, and the consistency rate of the lasso algorithm is as best $n^{1/4}$, but variable selection lasso has the $\sqrt{n}$ rate (if the OLS full model is added as one of the models considered).

Adding $\lambda_0 = C\sqrt{n}/\log(n)$ or $\lambda_0 = n^{0.49}$ to the grid does not seem to be a good idea, because $k-$fold cross validation chooses λ_0 too often, and lasso with the modified grid does not do variable selection well: lasso selects the full model too often compared

to forward selection with C_p. Heuristically, in low dimensions, lasso with λ_0 is $\sqrt{n}$ consistent while lasso with λ_1 is $n^{1/4}$ consistent, and $k-$fold cross validation prefers the $\sqrt{n}$ consistent estimator. Using λ_1 makes the lasso variable selection estimator more like forward selection with C_p in low dimensions. Getting rid of weak (but population active) predictors that degrade the performance of the full model is more important than having $\sqrt{n}$ consistency or minimizing lasso without the λ_1 constraint. Empirically, researchers have decades of experience showing that sparse fitted models from variable selection often work better than the full model. Variable selection is an important widely used technique.

The unilasso estimator finds the marginal estimators $(\hat{\alpha}_i, \hat{\eta}_i)$ from regressing Y on x_i, like the MMLE. Then the leave one out estimators $(\hat{\alpha}_i^{-i}, \hat{\eta}_i^{-i})$ are computed. Let $\boldsymbol{\theta} = (\theta_0, \theta_1, ..., \theta_p)^T$. Fit lasso with an intercept, no standardization, and nonnegativity constraints:

$$\min_{\boldsymbol{\theta}} \left\{ \frac{1}{n} \sum_{i=1}^{n} [Y_i - \theta_0 - \sum_{j=1}^{p} (\hat{\alpha}_i^{-i} + \hat{\eta}_i^{-i} x_{ij}) \theta_j]^2 + \lambda \sum_{j=1}^{p} \theta_j \right\} \text{ with } \theta_{\rm j} \geq 0 \ \ \forall {\rm j} \geq 1.$$

Perform the minimization over a grid of λ values and select $\hat{\lambda}$ using $k-$fold cross validation. Let $I = \{i_1, ..., i_k\}$ correspond to the k predictors with nonzero $\hat{\beta}_i$ where the unilasso estimator $\hat{\boldsymbol{\beta}}_I = (\hat{\beta}_{i_1}, ..., \hat{\beta}_{i_k})^T$ and $\hat{\boldsymbol{\beta}}_U = \hat{\boldsymbol{\beta}}_{I,0} = (\hat{\beta}_1, ..., \hat{\beta}_p)^T$. Then $ESP(I) = \hat{\alpha} + \hat{\boldsymbol{\beta}}_I^T \boldsymbol{x}_I = \hat{\alpha} + \hat{\boldsymbol{\beta}}_U^T \boldsymbol{x} = \hat{\alpha} + \sum_{i=1}^{p} \hat{\beta}_i x_i = \hat{\alpha} + \sum_{j=1}^{k} \hat{\beta}_{i_j} x_{i_j}$ where $\hat{\beta}_i = \hat{\theta}_i \hat{\eta}_i$ and $\hat{\alpha} = \hat{\theta}_0 + \sum_{j=1}^{p} \hat{\alpha}_j \hat{\theta}_j$. As λ decreases to zero, the limiting estimator is called the unireg estimator $\hat{\boldsymbol{\beta}}_{UR}$, and still has nonnegativity constraints.

The unilasso estimator has some interesting properties. Let $ESP(F) = \hat{\alpha}_F + \hat{\boldsymbol{\beta}}_F^T \boldsymbol{x}$ for a full model that depends on $\boldsymbol{x}$ through $SP = \alpha_F + \boldsymbol{\beta}_F^T \boldsymbol{x}$. Such models include MLR, the Nelder and Wedderburn (1972) GLMs, and the Cox (1972) proportional hazards regression model. The predictors x_i are replaced by $W_i = \hat{\alpha}_i^{-i} + \hat{\eta}_i^{-i} x_i$ for $i = 1, ..., p$ in the lasso type criterion. The full model $\boldsymbol{\beta}_F$ often changes the sign of the marginal estimator to get a better fit. The unilasso estimator does not allow sign changes from the marginal model because of the nonnegativity constraints. Hence the unilasso $\hat{\beta}_j$ may become 0 because of lasso type regularization or to avoid the sign change. Then the unilasso k is often less than the lasso a = number of nonzero lasso coefficients. In general, $\boldsymbol{\beta}_U \neq \boldsymbol{\beta}_F$, and unilasso does not satisfy the bet on sparsity principle for the population model. Unilasso variable selection (where the full model is added as one of the models considered) is useful in low dimensions by Subsection 2.3.3. In high dimensions, sparse fitted models are useful if they can be checked or have a better value for $k-$fold cross validation than competing models.

7 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).

8 Examples

The following example demonstrates the Theorem 5 result that $\hat{\boldsymbol{\beta}}_{MLE} = \hat{\boldsymbol{\beta}}_{\boldsymbol{x}} = \hat{\boldsymbol{\beta}}_p$ if $\hat{\boldsymbol{A}}_p$ is nonsingular.

Example 1. The species data is from Cook and Weisberg (1999, pp. 285-286) and Johnson and Raven (1973). The response variable is the total number of species recorded on each of 29 islands in the Galápagos Archipelago. Predictors include area of island, *areanear* = the area of the closest island, the distance to the closest island, the elevation, and *endem* = the number of endemic species (those that were not introduced from elsewhere). The R output below used predictors $log(endem)$ and $log(areanear)$. For both Poisson regression and negative binomial regression, HPLS $\hat{\boldsymbol{\beta}}_p = \hat{\boldsymbol{\beta}}_{\boldsymbol{x}}$.

```
source("http://parker.ad.siu.edu/Olive/sldata.txt")
Y<-species[,1]
endem<-species[,2]
lnendem <- log(endem)
areanear <- species[,7]
lnareanear <- log(areanear)
#use HPLS for the p-component estimator
out1 <- glm(Y~log(endem)+log(areanear),family=poisson)
ESP1 <- predict(out1)
x <- cbind(lnendem,lnareanear)
covxy <- cov(x,Y)
gam1hat<- covxy
w1 <- x%*%covxy
gam2hat <- cov(x) %*% covxy
w2 <-  x%*%gam2hat
out2 <- glm(Y~w1+w2,family=poisson)
out2$coef #PR thetahat = (0.01421,-0.001482)'
 (Intercept)           w1           w2
-0.044393306  0.014205369 -0.001481973
ESP2 <- predict(out2)
plot(ESP2,ESP1)
abline(0, 1)
AhatTrans <- cbind(gam1hat,gam2hat)
AhatTrans%*%out2$coef[c(2,3)]
                  [,1]  #PR betahat_p
lnendem     1.32751783
lnareanear -0.02241533
out1$coef   #PR betahat = (1.3257,-0.02242)'
  (Intercept)    log(endem) log(areanear)
```

```
  -0.04439331    1.32751783   -0.02241533
#above was Poisson regression, now do negative binomial regression
library(MASS) #theta=37 needs to be specified
out3 <- glm(Y~log(endem)+log(areanear),family=negative.binomial(37))
ESP3 <- predict(out3)
out4 <- glm(Y~w1+w2,family=negative.binomial(37))
ESP4 <- predict(out4)
plot(ESP4,ESP3)
abline(0, 1)
AhatTrans%*%out4$coef[c(2,3)]
                  [,1] #NBR betahat_p
lnendem     1.32370941
lnareanear -0.02328905
out3$coef  #NBR betahat = (1.3237,-0.2329)'
(Intercept)    log(endem) log(areanear)
  -0.02914982    1.32370941   -0.02328905
```

Example 2. The Hebbler (1847) data was collected from $n = 26$ districts in Prussia in 1843. Let Y_1 = the *number of women married to civilians* in the district, Y_2 = the *number of women married to husbands in the military* in the district with a constant and predictors x_1 = the *population of the district in 1843*, x_2 = the *number of married civilian men* in the district, x_3 = the *number of married men in the military* in the district. Sometimes the person conducting the survey would not count a spouse if the spouse was not at home. Hence Y_1 and x_2 are highly correlated but not equal. Similarly, Y_2 and x_3 are highly correlated but not equal. The predictor x_1 was highly correlated with Y_1 and x_2. For the multivariate OLS and lasso estimators,

$$\hat{\boldsymbol{B}}_{OLS} = \begin{bmatrix} 0.00035 & 0.00005 \\ 0.9996 & 0.00047 \\ 0.0871 & 0.9509 \end{bmatrix}, \ \hat{\boldsymbol{B}}_{mlasso} = \begin{bmatrix} 0.00078 & 0.00001 \\ 0.9700 & 0.0056 \\ 0 & 0 \end{bmatrix}.$$

PCR and PLS chose the p-component estimator, whether regressed on $\boldsymbol{y}$ or Y_1 and then Y_2, and were thus equivalent to multivariate OLS for this data set. Performing lasso on Y_1 and then Y_2, and then finding $\hat{\boldsymbol{B}}_H$ gave

$$\hat{\boldsymbol{B}}_{lasso} = \begin{bmatrix} 0.0018 & 0 \\ 0.9618 & 0.0013 \\ 0 & 0.9464 \end{bmatrix}, \hat{\boldsymbol{B}}_H = \begin{bmatrix} 0.0019 & 0+ \\ 0.9913 & 0.00017 \\ -0.1091 & 0.9501 \end{bmatrix}.$$

Here $b = 0+$ indicates that $|b| < 0.00001$. Performing relaxed lasso on Y_1 and then Y_2, and then finding $\hat{\boldsymbol{B}}_H$ gave

$$\hat{\boldsymbol{B}}_{RL} = \begin{bmatrix} 0.00007 & 0 \\ 1.0006 & 0.00017 \\ 0 & 0.9501 \end{bmatrix}, \hat{\boldsymbol{B}}_H = \begin{bmatrix} 0.00007 & 0+ \\ 1.0011 & 0.00013 \\ -0.0871 & 0.9464 \end{bmatrix}.$$

The uniLasso estimator, applied to Y_1 and then Y_2, worked well since each response variable is best represented by a simple linear regression estimator. Forward selection,

on Y_1 and then Y_2, also worked well. This estimator would also be the relaxed uniLasso estimator.

$$\hat{\boldsymbol{B}}_{uL} = \begin{bmatrix} 0 & 0 \\ 0.9716 & 0 \\ 0 & 0.9415 \end{bmatrix}, \hat{\boldsymbol{B}}_{FS} = \begin{bmatrix} 0 & 0 \\ 1.0010 & 0 \\ 0 & 0.9585 \end{bmatrix}.$$

Some R code is below.

```
x <- marry[,-c(3,5)]
Y1 <- marry[,3]
Y2 <- marry[,5]
y <- cbind(Y1,Y2)
mols <- lsfit(x,y)
                     Y1            Y2
Intercept  2.398139e+02 -1.682749e+01
pop        3.467458e-04 -5.368749e-05
mmen       9.995575e-01  4.662285e-04
mmilmen   -8.713382e-02  9.508827e-01
library(glmnet)
outlasso <- cv.glmnet(x,y,family="mgaussian")
lam <- outlasso$lambda.min
lcoef <- predict(outlasso,type="coefficients",s=lam)
lcoef
$Y1
4 x 1 sparse Matrix of class "dgCMatrix"
              s=1066.251
(Intercept) 2.800322e+03
pop         7.857383e-04
mmen        9.700227e-01
mmilmen     .
$Y2
4 x 1 sparse Matrix of class "dgCMatrix"
              s=1066.251
(Intercept) 1.750867e+02
pop         1.081750e-05
mmen        5.560175e-03
mmilmen     .
out1 <- cv.glmnet(x,Y1)
lam <- out1$lambda.min
lcoef <- predict(out1,type="coefficients",s=lam)
lcoef
4 x 1 sparse Matrix of class "dgCMatrix"
              s=1170.189
(Intercept) 3.000397e+03
pop         1.800342e-03
mmen        9.618035e-01
```

```
mmilmen     .
out2 <- cv.glmnet(x,Y2)
lam <- out2$lambda.min
lcoef <- predict(out2,type="coefficients",s=lam)
lcoef
pop          .
mmen         1.315187e-04
mmilmen      9.463703e-01
w1 <- x%*%c(1.800342e-03,9.618035e-01,0)
w2 <- x%*%c(0,1.315187e-04,9.463703e-01)
w<- cbind(w1,w2)
Dhat <- lsfit(w,y)$coef[-1,]
                    Y1            Y2
X1           1.0307024  4.222075e-05
X2          -0.1153572  1.003938e+00
eta1hat <- c(1.800342e-03,9.618035e-01,0)
eta2hat <- c(0,1.315187e-04,9.463703e-01)
AhatTrans <- cbind(eta1hat,eta2hat)
BhatH <- AhatTrans%*%Dhat
                Y1           Y2
[1,]  0.001855617 7.601178e-08
[2,]  0.991318031 1.726447e-04
[3,] -0.109170643 9.500975e-01
#relaxed lasso
rout1 <- lsfit(x[,-3],Y1)$coef
rout2 <- lsfit(x[,-1],Y2)$coef
rout1
   Intercept          pop         mmen
2.380912e+02 6.556895e-05 1.000603e+00
rout2
    Intercept          mmen       mmilmen
-1.950581e+01  1.731124e-04  9.500960e-01
w1 <- x%*%c(6.556895e-05, 1.000603e+00,0)
w2 <- x%*%c(0,1.731124e-04,  9.500960e-01)
w<- cbind(w1,w2)
Dhat <- lsfit(w,y)$coef[-1,]
eta1hat <- c(6.556895e-05, 1.000603e+00,0)
eta2hat <- c(0,1.315187e-04,9.463703e-01)
AhatTrans <- cbind(eta1hat,eta2hat)
BhatH <- AhatTrans%*%Dhat
                 Y1            Y2
[1,]  6.560201e-05 -1.245153e-13
[2,]  1.001096e+00  1.315168e-04
[3,] -8.268911e-02  9.463704e-01
##uniLasso
```

```
library(uniLasso)
out1<-cv.uniLasso(x,Y1)
lam <- out1$lambda.min
lcoef <- predict(out1,type="coefficients",s=lam)
lcoef
4 x 1 sparse Matrix of class "dgCMatrix"
              s=46983463
(Intercept) 3109.5982108
pop            .
mmen           0.9715993
mmilmen        .
out2<-cv.uniLasso(x,Y2)
lam <- out2$lambda.min
lcoef <- predict(out2,type="coefficients",s=lam)
lcoef
            s=3278.221
(Intercept)  4.3077246
pop          0.0000000
mmen         0.0000000
mmilmen      0.9415391
#forward selection with Cp
library(leaps)
temp<-regsubsets(x,Y1,method="forward")
out<-summary(temp)
vars <- as.vector(1:3)
mincp <- out$which[out$cp==min(out$cp),]
vin <- vars[mincp[-1]]
vin
2
lsfit(x[,vin],Y1)$coef
 Intercept          X
241.544496   1.000967
temp<-regsubsets(x,Y2,method="forward")
out<-summary(temp)
vars <- as.vector(1:3)
mincp <- out$which[out$cp==min(out$cp),]
vin <- vars[mincp[-1]]
vin
3
lsfit(x[,vin],Y2)$coef
 Intercept          X
-9.0176168  0.9585048
##PLS
library(pls)
z <- as.data.frame(cbind(y,x))
```

```
out <- plsr(y~pop+mmen+mmilmen,data=z,method="simpls",validation="CV")
bhat<- coef(out,intercept=TRUE)
bhat              Y1            Y2 #same as OLS
(Intercept)  2.398139e+02 -1.682749e+01
pop          3.467459e-04 -5.368749e-05
mmen         9.995574e-01  4.662285e-04
mmilmen     -8.713384e-02  9.508827e-01
out <- plsr(y~pop+mmen+mmilmen,data=z,method="simpls",scale=TRUE,validation="CV")
bhat<- coef(out,intercept=TRUE)
                     Y1        Y2  #this method did not work
(Intercept)   239.81391 -16.82749
pop            79.15141 -12.25520
mmen        40880.32880  19.06801
mmilmen       -39.25344 428.36886
out1 <- plsr(Y1~pop+mmen+mmilmen,data=z,method="simpls",validation="CV")
bhat<- coef(out1,intercept=TRUE)
bhat              Y1  #same as OLS
(Intercept)  2.398139e+02
pop          3.467411e-04
mmen         9.995575e-01
mmilmen     -8.713236e-02
out2 <- plsr(Y2~pop+mmen+mmilmen,data=z,method="simpls",validation="CV")
bhat<- coef(out2,intercept=TRUE)
bhat
                   Y2 #same as OLS
(Intercept) -1.682749e+01
pop         -5.368749e-05
mmen         4.662285e-04
mmilmen      9.508827e-01  #below is the same for PLS and PCR
w1 <- x%*%c(3.467411e-04, 9.995575e-01, -8.713236e-02)
w2 <- x%*%c(-5.368749e-05, 4.662285e-04, 9.508827e-01)
w<- cbind(w1,w2)
Dhat <- lsfit(w,y)$coef[-1,]
eta1hat <- c(3.467411e-04, 9.995575e-01, -8.713236e-02)
eta2hat <- c(-5.368749e-05, 4.662285e-04, 9.508827e-01)
AhatTrans <- cbind(eta1hat,eta2hat)
BhatH <- AhatTrans%*%Dhat
BhatH  #same as OLS
                Y1            Y2
[1,]  0.0003467412 -5.368749e-05
[2,]  0.9995574755  4.662285e-04
[3,] -0.0871337489  9.508827e-01
##PCR
library(pls)
x <- marry[,-c(3,5)]
```

```
Y1 <- marry[,3]
Y2 <- marry[,5]
y <- cbind(Y1,Y2)
z <- as.data.frame(cbind(y,x))
out <- pcr(y~pop+mmen+mmilmen,data=z,validation="CV")
bhat<- coef(out,intercept=TRUE)
bhat   #same as OLS
                        Y1            Y2
(Intercept)  2.398139e+02 -1.682749e+01
pop          3.467458e-04 -5.368749e-05
mmen         9.995575e-01  4.662285e-04
mmilmen     -8.713382e-02  9.508827e-01
out1 <- pcr(Y1~pop+mmen+mmilmen,data=z,validation="CV")
bhat<- coef(out1,intercept=TRUE)
bhat   #same as OLS
                        Y1
(Intercept)  2.398139e+02
pop          3.467458e-04
mmen         9.995575e-01
mmilmen     -8.713382e-02
out2 <- pcr(Y2~pop+mmen+mmilmen,data=z,validation="CV")
bhat<- coef(out2,intercept=TRUE)
bhat #same as OLS
                        Y2
(Intercept) -1.682749e+01
pop         -5.368749e-05
mmen         4.662285e-04
mmilmen      9.508827e-01
```

9 CONCLUSIONS

A useful high dimensional technique is to use PCA for dimension reduction. Let $U_1, ..., U_p$ be the PCA linear combinations ($U_i = \hat{\boldsymbol{d}}_i^T \boldsymbol{x}$) ordered with respect to the largest eigenvalues. Then use $U_1, ..., U_k$ in the regression or classification model where k is chosen in some manner. For example, use Equation (5) with $Q(i) = \hat{\lambda}_i$. The problem with this idea is that principal components are used to explain the structure of the dispersion matrix of the data, not to be linear combinations of the data that are good for classification. See, for example, Artigue and Smith (2019), Cook (2007, 2018), and Zhang and Chen (2020). Cook and Forzani (2021) used the PLS components as predictors for nonlinear regression.

From a model selection viewpoint, using $W_1, ..., W_k$ should work much better than using $U_1, ..., U_k$. Also, the PLS components W_i should be used instead of the PCA W_i, since the PLS components are chosen to be fairly highly correlated with $\boldsymbol{y}$. One method

to select k is to find D such that

$$\frac{\sum_{i=1}^{D} Q(i)}{\sum_{i=1}^{p} Q(i)} \geq 0.975.$$

Then use the $\hat{k} = \min(D, n-2, p)$ W_i as the variables. The Q scree plot can also be used. For $r = 1$ with a univariate response variable Y, let $W_1, ..., W_p$ be 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 $Q(i) = r_i = r_{i,Y} = \text{cor}(x_i, Y)$. See Olive (2025).

Binary regression is closely related to two sample tests. Note that $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2$ can use other multivariate location estimators than sample means. For example, sample coordinatewise medians, sample coordinatewise trimmed means, and the Olive (2017b) T_{RMVN} estimator have large sample theory given by Rupasinghe Arachchige Don and Olive (2019) and Rupasinghe Arachchige Don and Pelawa Watagoda (2018).

Some papers on binary regression include Cai, Guo, and Ma (2023), Candès and Sur (2020), Mukherjee, Pillai, and Lin (2015), Sur and Candès (2019), Sur, Chen, and Candès (2019), and Tang and Ye (2020). Empirically, often $\boldsymbol{\beta}_{LR} \approx d\ \boldsymbol{\beta}_{OLS}$. Haggstrom (1983) suggests that d is not far from 1/MSE for logistic regression.

These binary regression estimators also give new ways to compare multivariate location estimators from two groups. The tests using k predictors can be performed. High dimensional tests for means from two groups can also be used. The tests that make very strong assumptions, such as multivariate normality or equal covariance matrices for the two groups, should be avoided. See Feng and Sun (2015), Gregory et al. (2015), Hu and Bai (2015), Rajapaksha and Olive (2024), and Xue and Yao (2020).

Yee (2015) considers many models with $r > 1$, and has R software. The OLS multivariate linear regression techniques are also useful for vector autoregressive $VAR(p)$ time series. See Samadi and Herath (2024) for references.

Software

The R software was used in the simulations. See R Core Team (2024). Programs will be added to the Olive (2025) collections of R functions *slpack.txt*, available from (http://parker.ad.siu.edu/Olive/slpack.txt). Some R packages used include `glmnet` Friedman et al. (2015), `uniLasso` Hastie, Tibshirani, and Chatterjee (2026), `leaps` Lumley (2009), and `pls` Mevik et al. (2015).

References

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

Artigue, H., and Smith, G. (2019), "The Principal Problem with Principal Components Regression," *Cogent Mathematics & Statistics*, 6, 1622190.

Bai, Z.D., and Saranadasa, H. (1996), "Effects of High Dimension: by an Example of a Two Sample Problem," *Statistica Sinica*, 6, 311-329.

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.

Brown, P.J. (1993), *Measurement, Regression, and Calibration*, Oxford University Press, New York, NY.

Chatterjee, S., Hastie, T., and Tibshirani, R. (2025), "Univariate-Guided Sparse Regression," *Harvard Data Science Review*, 7(3). https://doi.org/10.1162/99608f92.c79ff6db

Chen, S.X., and Qin, Y.L. (2010), "A Two Sample Test for High-dimensional Data with Applications to Gene-Set Testing," *The Annals of Statistics*, 38, 808-835.

Chun, H., and Keleş, S. (2010), "Sparse Partial Least Squares Regression for Simultaneous Dimension Reduction and Predictor Selection," *Journal of the Royal Statistical Society, B,* 72, 3-25.

Conover, W.J. (1971), *Practical Nonparametric Statistics*, Wiley, New York, NY.

Cook, R.D. (2007), "Fisher Lecture: Dimension Reduction in Regression," *Statistical Science*, (with discussion), 22, 1-26.

Cook, R.D. (2018), "Principal Components, Sufficient Dimension Reduction, and Envelopes," *Annual Review of Statistics and Its Application*, 5, 533-559.

Cook, R.D., and Forzani, L. (2021), "PLS Regression Algorithms in the Presence of Nonlinearity," *Chemometrics and Intelligent Laboratory Systems*, 213, 104307.

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.

Cook, R.D., and Weisberg, S. (1999), *Applied Regression Including Computing and Graphics,* Wiley, New York, NY.

Cox, D.R. (1972), "Regression Models and Life-Tables," *Journal of the Royal Statistical Society, B,* 34, 187-220.

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.

Friedman, J., Hastie, T., Simon, N., and Tibshirani, R. (2015), *glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models, R* Package version 2.0, (http://cran.r-project.org/package=glmnet).

Gelman, A., and Carlin, J. (2017), "Some Natural Solutions to the p-Value Communication Problemand Why They Wont Work," *Journal of the American Statistical Association*, 112, 899-901.

Hastie, T., Tibshirani, R., and Chatterjee, S. (2026), *uniLasso: Univariate-Guided Sparse Regression, R* Package version 2.11, (https://CRAN.R-project.org/package=uniLasso).

Hebbler, B. (1847), "Statistics of Prussia," *Journal of the Royal Statistical Society, A*, 10, 154-186.

Helland, I.S. (1990), "Partial Least Squares Regression and Statistical Models," *Scandanavian Journal of Statistics*, 17, 97-114.

Hoerl, A.E., and Kennard, R. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12, 55-67.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021), *An Introduction to Statistical Learning with Applications in R*, 2nd ed., Springer, New York, NY.

Johnson, M.P., and Raven, P.H. (1973), "Species Number and Endemism, the Galápagos Archipelago Revisited," *Science*, 179, 893-895.

Jones, H.L. (1946), "Linear Regression Functions with Neglected Variables," *Journal of the American Statistical Association,* 41, 356-369.

Kellison, S.G. and London, R.L. (2011), *Risk Models and Their Estimation*, ACTEX Publications, Winsted, CT.

Li, J. (2023), "Finite Sample t-Tests for High-dimensional Means," *Journal of Multivariate Analysis*, 196, 105183.

Lumley, T. (2009) (using Fortran code by Alan Miller), *leaps: Regression Subset Selection, R* package version 2.9, (https://CRAN.R-project.org/package =leaps).

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

Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979), *Multivariate Analysis,* Academic Press, London, UK.

Meinshausen, N. (2007), "Relaxed Lasso," *Computational Statistics & Data Analysis*, 52, 374-393.

Mevik, B.–H., Wehrens, R., and Liland, K.H. (2015), *pls: Partial Least Squares and Principal Component Regression*, *R* package version 2.5-0, (https: //CRAN.R-project.org/package=pls).

Nelder, J.A., and Wedderburn, R.W.M. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society, A*, 135, 370-384.

Nester, M.R. (1996), "An Applied Statistician's Creed," *Journal of the Royal Statistical Society, Series C*, 45, 401-410.

Olive, D.J. (2017), *Linear Regression*, Springer, New York, NY.

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

Olive, D.J., Alshammari, A.A., Pathiranage, K.G., and Hettige, L.A.W. (2026), "Testing with the One Component Partial Least Squares and the Marginal Maximum Likelihood Estimators," *Communications in Statistics: Theory and Methods*, 55, 1492-1507.

Olive, D.J., and Hawkins, D.M. (2005), "Variable Selection for 1D Regression Models," *Technometrics*, 47, 43-50.

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.

Pelawa Watagoda, L.C.R., and Olive, D.J. (2021), "Comparing Six Shrinkage Estimators with Large Sample Theory and Asymptotically Optimal Prediction Intervals," *Statistical Papers,* 62, 2407-2431.

Pratt, J.W. (1959), "On a General Concept of "in Probability"," *The Annals of Mathematical Statistics,* 30, 549-558.

R Core Team (2024), "R: a Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria, (www.R-project.org).

Samadi, S.Y., and Herath, H.M.W.B. (2024), "Reduced-Rank Envelope Vector Autoregressive Model," *Journal of Business & Economic Statistics,* 42, 918-932.

Srivastava, M.S., and Du, M. (2008), "A Test for the Mean Vector with Fewer Observations Than the Dimension," *Journal of Multivariate Analysis*, 99, 386-402.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, B*, 58, 267-288.

Tukey, J.W. (1991), "The Philosophy of Multiple Comparisons," *Statistical Science*, 6, 100-116.

Wold, H. (1975), "Soft Modelling by Latent Variables: the Non-Linear Partial Least Squares (NIPALS) Approach," *Journal of Applied Probability*, 12, 117-142.

Yee, T. (2015), *Vector Generalized Linear and Additive Models*, Springer, New York, NY.

Zhang, J., and Chen, X. (2020), "Principal Envelope Model," *Journal of Statistical Planning and Inference*, 206, 249-262.

Zhao, A., Li, C., Li, R., and Zhang, Z. (2024), "Testing High-Dimensional Regression Coefficients in Linear Models," *The Annals of Statistics*, 52, 2034-2058.

Zheng, W., Lai, D., and Gould, K. L. (2023), "A Simulation Study of a Class of Nonparametric Test Statistics: a Close Look of Empirical Distribution Function-Based Tests," *Communications in Statistics - Simulation and Computation*, 52, 1132-1148.

Zheng, W., Zhu, H., Lance Gould, K., and Lai, D. (2025), "Comparing Heart PET Scans: an Adjustment of Kolmogorov-Smirnov Test under Spatial Autocorrelation," *Journal of Applied Statistics*, 52, 253-269.

Zou, H., and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society Series, B,* 67, 301-320.