

High Dimensional Dimension Reduction with One Response Variable

David J. Olive, Mohammed S. Alsaudi, and Kasun G. Pathirana^{*}
Southern Illinois University

April 8, 2026

Abstract

A common regression and classification technique computes linear combinations $W_i = \hat{\gamma}_i^T \mathbf{x}$ of the predictors $\mathbf{x} = (x_1, \dots, x_p)^T$ for $i = 1, \dots, p$ where W_1, \dots, W_p are ordered in some way. Then the response variable Y is regressed on W_1, \dots, W_k to produce the k -component estimator for $k = 1, \dots, M$ 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. This section reviews some high dimensional dimension reduction methods.

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.

Some notation for dimension reduction methods is needed. Let Y be the response variable and let $\mathbf{x} = (x_1, \dots, x_p)^T$ be a vector of p predictor variables. For example, predict $Y =$ mussel muscle mass from $x_1 =$ height, $x_2 =$ width, and $x_3 =$ length of the mussel shell. Then $p = 3$.

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

Let \mathcal{L} be a subspace of \mathbb{R}^p . Let $\{\gamma_1, \dots, \gamma_q\}$ be a basis for \mathcal{L} , and let $\{\gamma_1, \dots, \gamma_q, \gamma_{q+1}, \dots, \gamma_p\}$ be a basis for \mathbb{R}^p . Then several dimension reduction methods involve estimating q and $\hat{\gamma}_i$ for $i = 1, \dots, q$. Let $\mathbf{G} = [\gamma_1 \dots \gamma_q]$ be the basis matrix for \mathcal{L} . Let $Y \perp\!\!\!\perp \mathbf{x} | \mathbf{A}\mathbf{x}$ indicate that the response variable Y is independent of the predictors \mathbf{x} given $\mathbf{A}\mathbf{x}$. Then $\mathcal{L} \subseteq \mathbb{R}^p$ that satisfies $Y \perp\!\!\!\perp \mathbf{x} | \mathbf{P}_q \mathbf{x}$ is a *dimension reduction subspace* for the regression of Y on \mathbf{x} . Thus the reduced predictors $\mathbf{w} = \mathbf{P}_q \mathbf{x}$ hold all of the information that \mathbf{x} has about Y . If the intersection of all dimension reduction subspaces is a dimension reduction subspace, then that subspace is called the *central subspace*, denoted by $\mathcal{L}_{Y|\mathbf{x}}$.

For a predictor envelope, suppose a) $Y \perp\!\!\!\perp \mathbf{x} | \mathbf{P}_q \mathbf{x}$, and b) $\mathcal{L} \subseteq \mathbb{R}^p$ reduces $\Sigma \mathbf{x}$ if and only if $\text{Cov}(\mathbf{P}_q \mathbf{x}, \mathbf{Q}_q \mathbf{x}) = \mathbf{0}$. Notation: $\mathbf{w} = \mathbf{P}_q \mathbf{x}$ is *material* for the regression of Y on \mathbf{x} while $\mathbf{u} = \mathbf{Q}_q \mathbf{x}$ is *immaterial* for the regression of Y on \mathbf{x} . Assume $\mathcal{L}_{Y|\mathbf{x}} \subseteq \text{span}(\Sigma \mathbf{x})$. The *predictor envelope* (subspace) $\mathcal{E}\mathbf{x}$ for the regression of Y on \mathbf{x} is the intersection of all dimension reduction subspaces that reduce $\Sigma \mathbf{x}$ and contain $\mathcal{L}_{Y|\mathbf{x}}$. Hence $\mathcal{L}_{Y|\mathbf{x}} \subseteq \mathcal{E}\mathbf{x}$. 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 \mathbf{x} be $\text{Cov}(\mathbf{x}) = \Sigma \mathbf{x} = E[(\mathbf{x} - E(\mathbf{x}))(\mathbf{x} - E(\mathbf{x}))^T]$ and the $p \times 1$ vector $\text{Cov}(\mathbf{x}, Y) = \Sigma \mathbf{x}_Y = E[(\mathbf{x} - E(\mathbf{x}))(Y - E(Y))] = (\text{Cov}(x_1, Y), \dots, \text{Cov}(x_p, Y))^T = \boldsymbol{\eta}$. Let the sample covariance matrix

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

Let estimators

$$\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\eta}}_n = \hat{\Sigma} \mathbf{x}_Y = \mathbf{S} \mathbf{x}_Y = \text{cov}(\mathbf{x}, Y) = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(Y_i - \bar{Y}),$$

$$\text{and } \tilde{\boldsymbol{\eta}} = \tilde{\boldsymbol{\eta}}_n = \tilde{\Sigma} \mathbf{x}_Y = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(Y_i - \bar{Y}).$$

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

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

a) Let Σ^* be $\hat{\Sigma}$ or $\tilde{\Sigma}$. Then $\Sigma^* \mathbf{w} = \mathbf{A}_n \Sigma^* \mathbf{A}_n^T$, $\Sigma^* \mathbf{w}_Y = \mathbf{A}_n \Sigma^* \mathbf{x}_Y$, and $\text{cov}(\mathbf{A}_n \mathbf{x}, \mathbf{D}_n \mathbf{x}) = \mathbf{A}_n \Sigma^* \mathbf{D}_n^T$.

b) If \mathbf{A}_n is a constant matrix, then $\Sigma \mathbf{w} = \mathbf{A}_n \Sigma \mathbf{x} \mathbf{A}_n^T$ and $\Sigma \mathbf{w}_Y = \mathbf{A}_n \Sigma \mathbf{x}_Y$.

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

Partial least squares (PLS), SDR, and envelopes algorithms produce estimated basis vectors $\hat{\gamma}_i$ sequentially, using the response variable Y . Suppose the positive semidefinite dispersion matrix Σ has eigenvalue eigenvector pairs $(\lambda_1, \mathbf{d}_1), \dots, (\lambda_p, \mathbf{d}_p)$ where $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$. Let the eigenvalue eigenvector pairs of $\hat{\Sigma}$ be $(\hat{\lambda}_1, \hat{\mathbf{d}}_1), \dots, (\hat{\lambda}_p, \hat{\mathbf{d}}_p)$ where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_p$. These vectors are important quantities for principal component analysis (PCA) and principal components regression (PCR). If the eigenvectors are orthogonal, $\hat{\Sigma} = \hat{\Sigma} \mathbf{x}$, and $W_i = \hat{\mathbf{d}}_i^T \mathbf{x}$, then $\text{cov}(W_i, W_j) = \hat{\mathbf{d}}_i^T \hat{\Sigma} \mathbf{x} \hat{\mathbf{d}}_j = 0$ for $i \neq j$.

The NIPALS PLS algorithm uses $\hat{\gamma}_1 = \hat{\Sigma} \mathbf{x}_Y / \|\hat{\Sigma} \mathbf{x}_Y\|$, and the $\hat{\gamma}_i$ are orthonormal. Let $\hat{\mathbf{A}}_k$ be the matrix with i th row $\hat{\gamma}_i^T$ for $i = 1, \dots, k$. For $k < q$, the SIMPLS algorithm uses $\hat{\gamma}_1$ and $\hat{\gamma}_{k+1} = \text{argmax}_{\gamma \in \mathbb{R}^p} \gamma^T \mathbf{S} \mathbf{x}_Y \mathbf{S}^T \mathbf{x}_Y \gamma$ subject to $\gamma^T \mathbf{S} \mathbf{x} \hat{\mathbf{A}}_k^T = \mathbf{0}$ and $\gamma^T \gamma = 1$. Note that $\hat{\gamma}_1 = \text{argmax}_{\gamma \in \mathbb{R}^p} \gamma^T \mathbf{S} \mathbf{x}_Y \mathbf{S}^T \mathbf{x}_Y \gamma$ where $\gamma^T \gamma = 1$. If $W_i = \hat{\gamma}_i^T \mathbf{x}$, then $\text{cov}(W_i, W_j) = \hat{\gamma}_i^T \hat{\Sigma} \mathbf{x} \hat{\gamma}_j = 0$ for $i \neq j$. See Cook and Forzani (2024) and Wold (1975).

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 $j = 1, \dots, p$. Then $\hat{\beta}_{MMLE} = (\hat{\beta}_{1,M}, \dots, \hat{\beta}_{p,M})^T$. 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, \dots, W_p are ordered by $|\hat{\beta}_{i_1}| \geq |\hat{\beta}_{i_2}| \geq \dots \geq |\hat{\beta}_{i_p}|$. 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.

A common regression and classification technique computes linear combinations $W_i = \hat{\gamma}_i^T \mathbf{x}$ of the predictors for $i = 1, \dots, p$, where W_1, \dots, W_p are ordered in some way. Let the j th column of \mathbf{I}_p be $\mathbf{c}_j = (0, \dots, 0, 1, 0, \dots, 0)^T$ where the 1 is in the j th position. The \mathbf{c}_j form the standard basis vectors for \mathbb{R}^p and $\mathbf{c}_j^T \mathbf{x} = x_j$. For variable selection estimators, let $W_i = \hat{\gamma}_i^T \mathbf{x} = x_{i_j}$ where $\hat{\gamma}_i = \mathbf{c}_{i_j}$. The k -component estimator, e.g. the k -component PLS estimator or the k -component PCR estimator, is obtained by regressing Y on W_1, \dots, W_k where often a constant is in the model. Let $k = 1, \dots, M$ 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 $\hat{\beta}_{MSPLS}$ or $\hat{\beta}_{MSPCR}$.

Much of the literature uses multiple linear regression (MLR) as the regression method. Let the MLR model be

$$Y_i = \alpha + x_{i,1}\beta_1 + \dots + x_{i,p}\beta_p + e_i = \alpha + \mathbf{x}_i^T \boldsymbol{\beta} + e_i \quad (1)$$

for $i = 1, \dots, n$. In matrix form, this model is $\mathbf{Y} = \mathbf{X}\boldsymbol{\phi} + \mathbf{e}$, where \mathbf{Y} is an $n \times 1$ vector of dependent variables, \mathbf{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 \mathbf{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 (1) are $\hat{\boldsymbol{\phi}}_{OLS} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$, $\hat{\alpha}_{OLS} = \bar{Y} - \hat{\boldsymbol{\beta}}_{OLS}^T \bar{\mathbf{x}}$, and

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

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

For the MLR k -component estimator, let $\hat{\mathbf{A}}_k \mathbf{x} = \mathbf{w} = (W_1, \dots, W_k)^T$ where $\hat{\mathbf{A}}_k$ is the matrix with i th row $\hat{\gamma}_i^T$ for $i = 1, \dots, k$. Let $\boldsymbol{\theta}_k = (\theta_1, \theta_2, \dots, \theta_k)^T$. Fit the working model

$$Y = \alpha_k + \boldsymbol{\theta}_k^T \mathbf{w} + \epsilon = \alpha_k + \boldsymbol{\theta}_k^T \hat{\mathbf{A}}_k \mathbf{x} + \epsilon = \alpha_k + \boldsymbol{\beta}_k^T \mathbf{x} + \epsilon$$

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

$$\hat{\boldsymbol{\beta}}_p = \hat{\mathbf{A}}_p^T (\hat{\mathbf{A}}_p^T)^{-1} \hat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{-1} (\hat{\mathbf{A}}_p)^{-1} \hat{\mathbf{A}}_p \hat{\boldsymbol{\Sigma}}_{\mathbf{x},Y} = \hat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{-1} \hat{\boldsymbol{\Sigma}}_{\mathbf{x},Y} = \hat{\boldsymbol{\beta}}_{OLS}. \quad (2)$$

Section 2 considers k -component estimators with response variable Y , reviews variable selection, and Theorem 5 generalizes Equation (2) to many other regression models. Section 3 considers the OPLS estimator, and Section 4 gives some high dimensional tests.

2 Univariate Response Y

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 $(Y_i, \mathbf{x}_i^T)^T$, and for some important models, Y depends on \mathbf{x} through the sufficient predictor $SP = \alpha + \mathbf{x}^T \boldsymbol{\beta}$. Let the estimated sufficient predictor $ESP = \hat{\alpha} + \mathbf{x}^T \hat{\boldsymbol{\beta}}$ 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 $Y = Y|SP \sim \text{binomial}(1, \rho(SP))$ 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).

For a response variable Y , $\text{span}(\mathbf{A}_k^T)$ and $\text{span}(\hat{\mathbf{A}}_k^T)$ are the same for the SIMPLS, NIPALS, and HPLS algorithms for $k = 1, \dots, q$. See Cook and Forzani (2024, p. 96). HPLS uses $\boldsymbol{\gamma}_i = \boldsymbol{\Sigma}_{\mathbf{x}}^{i-1} \boldsymbol{\Sigma}_{\mathbf{x}Y}$ and $\hat{\boldsymbol{\gamma}}_i = \mathbf{S}_{\mathbf{x}}^{i-1} \mathbf{S}_{\mathbf{x}Y}$ with $\mathbf{S}_{\mathbf{x}}^0 = \boldsymbol{\Sigma}_{\mathbf{x}}^0 = \mathbf{I}_p$. Thus $\hat{\boldsymbol{\gamma}}_1 = \mathbf{S}_{\mathbf{x}Y}$ and $W_i = \hat{\boldsymbol{\gamma}}_i^T \mathbf{x}$. There are regularity conditions for this result: if $\boldsymbol{\Sigma}_{\mathbf{x}Y}$ is an eigenvector of $\boldsymbol{\Sigma}_{\mathbf{x}}$, then $\text{span}(\mathbf{A}_p^T) = \text{span}(\hat{\mathbf{A}}_p^T) = \text{span}(\boldsymbol{\Sigma}_{\mathbf{x}Y}) = \text{span}(\boldsymbol{\beta}_{OLS})$ for HPLS. The HPLS W_i can have high multicollinearity.

The next result was given by Chun and Keleş (2010). So PLS is a model free way to get predictors $W_i = \hat{\boldsymbol{\gamma}}_i^T \mathbf{x}$ 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 $1 < k < q$, the SIMPLS algorithm maximizes

$$Q_S(\boldsymbol{\gamma}^T \mathbf{x}) = \hat{V}(\boldsymbol{\gamma}^T \mathbf{x}) [\text{cor}(\boldsymbol{\gamma}^T \mathbf{x}, Y)]^2 \quad (3)$$

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

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

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

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

Given $U_j = \hat{\boldsymbol{\gamma}}_j^T \mathbf{x}$, the model free variable importance criterion $Q_S(\hat{\boldsymbol{\gamma}}_j^T \mathbf{x})$ can be used to order the U_j in importance. Let W_1, W_2, \dots, W_p be the variables U_j ordered with respect to a criterion Q , corresponding to $Q(1) \geq Q(2) \geq \dots \geq Q(p)$. 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

$$R_k = \frac{\sum_{i=1}^k Q(i)}{\sum_{i=1}^p Q(i)} \quad (4)$$

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

Another model free variable importance criterion is an MMLE criterion

$$Q_M(\hat{\boldsymbol{\gamma}}_i^T \mathbf{x}) = [\text{cor}(\hat{\boldsymbol{\gamma}}_i^T \mathbf{x}, Y)]^2 = r_i^2. \quad (5)$$

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

PCA regression uses $U_j = \hat{\mathbf{d}}_j^T \mathbf{x}$, 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 (5) 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 $(Y_i, \mathbf{x}_i^T)^T$ are iid, then

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

This optimization problem is equivalent to maximizing

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

which has a maximum at $\boldsymbol{\gamma} = \boldsymbol{\Sigma} \mathbf{x}^{-1} \boldsymbol{\Sigma} \mathbf{x}_Y = \boldsymbol{\beta}_{OLS}$. See Mardia, Kent, and Bibby (1979, pp. 168, 282). In high dimensions, $\hat{\boldsymbol{\gamma}}_1 = \hat{\boldsymbol{\beta}}_{OLS}$ can be replaced by 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 $W_i = \hat{\boldsymbol{\gamma}}_i^T \mathbf{x}$. 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 $0 \leq \lambda_1 < \lambda_2 < \dots < \lambda_M$. 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, \dots, W_k where $W_i = \hat{\boldsymbol{\gamma}}_i^T \mathbf{x}$ for $k = 1, \dots, M$. 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 \mathbf{x} through $SP = SP(F) = \mathbf{x}^T \boldsymbol{\beta}$. Such models include MLR, GLMs, and several survival regression models. A *model for variable selection* can be described by

$$SP(F) = \mathbf{x}^T \boldsymbol{\beta} = \mathbf{x}_S^T \boldsymbol{\beta}_S + \mathbf{x}_E^T \boldsymbol{\beta}_E = \mathbf{x}_S^T \boldsymbol{\beta}_S = SP(S) \quad (6)$$

where $\mathbf{x} = (\mathbf{x}_S^T, \mathbf{x}_E^T)^T$ is a $p \times 1$ vector of predictors, \mathbf{x}_S is a $q \times 1$ vector, and \mathbf{x}_E is a $(p-q) \times 1$ vector. Given that \mathbf{x}_S is in the model, $\boldsymbol{\beta}_E = \mathbf{0}$ and E denotes the subset of terms that can be eliminated given that the subset S is in the model. In Equation (6), 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 $SP = SP(F) = \mathbf{x}^T \boldsymbol{\beta}$, the submodel S has $SP = SP(S) = \mathbf{x}_S^T \boldsymbol{\beta}_S$, 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 \mathbf{x}_I be the vector of d_I terms from a candidate subset indexed by I , including a constant, and let \mathbf{x}_O be the vector of the remaining predictors (out of the candidate submodel). Then

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

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

$$\mathbf{x}^T \boldsymbol{\beta} = \mathbf{x}_S^T \boldsymbol{\beta}_S = \mathbf{x}_S^T \boldsymbol{\beta}_S + \mathbf{x}_{I/S}^T \boldsymbol{\beta}_{(I/S)} + \mathbf{x}_O^T \mathbf{0} = \mathbf{x}_I^T \boldsymbol{\beta}_I \quad (7)$$

where $\mathbf{x}_{I/S}$ denotes the predictors in I that are not in S . Since this is true regardless of the values of the predictors, $\boldsymbol{\beta}_O = \mathbf{0}$ and the sample correlation $\text{cor}(\mathbf{x}^T \boldsymbol{\beta}, \mathbf{x}_I^T \boldsymbol{\beta}_I) = \text{cor}(SP(F), SP(I)) = 1$ 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 $\text{cor}(ESP(F), ESP(I)) \rightarrow 1$ and $\text{cor}(Y - ESP(F), Y - ESP(I)) \rightarrow 1$ as $n \rightarrow \infty$. 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 $k = 1, \dots, M$ where there are M subsets, e.g. $M = 2^p - 1$. Note that under model (6), $\boldsymbol{\beta}_{I,0} = \boldsymbol{\beta} = \boldsymbol{\beta}_F$ if $S \subseteq I$.

Assume p is fixed. Suppose model (6) holds, and that if $S \subseteq I_j$ where the dimension of I_j is d_j , then $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j} - \boldsymbol{\beta}_{I_j}) \xrightarrow{D} N_{d_j}(\mathbf{0}, \mathbf{V}_j)$ where \mathbf{V}_j is the covariance matrix of the asymptotic multivariate normal distribution. Then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \mathbf{V}_{j,0}) \quad (8)$$

where $\mathbf{V}_{j,0}$ adds columns and rows of zeros corresponding to the x_i not in I_j , and $\mathbf{V}_{j,0}$ 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, \dots, 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, \dots, A_k is a *partition* of S . Then the *Law of Total Probability* states that if A_1, A_2, \dots, A_k form a partition of S such that $P(A_i) > 0$ for $i = 1, \dots, k$, 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}, \dots, A_m satisfy $P(A_i) = 0$ for $i = k + 1, \dots, m$. 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 $\mathbf{w}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta})$. 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 $\mathbf{w}_{kn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_k,0} - \boldsymbol{\beta}) | (\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}) \sim \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_k,0}^C - \boldsymbol{\beta})$. Denote $F_{\mathbf{z}}(\mathbf{t}) = P(z_1 \leq t_1, \dots, z_p \leq t_p)$ by $P(\mathbf{z} \leq \mathbf{t})$. Then Pötscher (1991) and Pelawa Watagoda and Olive (2021) show the following result.

Theorem 2. Using the above notation,

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

Hence $\hat{\boldsymbol{\beta}}_{VS}$ has a mixture distribution of the $\hat{\boldsymbol{\beta}}_{I_k,0}^C$ with probabilities π_{kn} , and \mathbf{w}_n has a mixture distribution of the \mathbf{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 $k = 1, \dots, J$. 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,

$$\begin{aligned} F\mathbf{w}_n(\mathbf{t}) &= P[n^{1/2}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \leq \mathbf{t}] = \\ &= \sum_{k=1}^J P[n^{1/2}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \leq \mathbf{t} | (\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0})] P(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}) = \\ &= \sum_{k=1}^J P[n^{1/2}(\hat{\boldsymbol{\beta}}_{I_k,0} - \boldsymbol{\beta}) \leq \mathbf{t} | (\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0})] \pi_{kn} \\ &= \sum_{k=1}^J P[n^{1/2}(\hat{\boldsymbol{\beta}}_{I_k,0}^C - \boldsymbol{\beta}) \leq \mathbf{t}] \pi_{kn} = \sum_{k=1}^J F\mathbf{w}_{kn}(\mathbf{t}) \pi_{kn}. \quad \square \end{aligned}$$

Under the assumptions of Theorem 3, note that $\sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta})$ is selecting from the $\mathbf{u}_{kn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_k,0} - \boldsymbol{\beta})$ and asymptotically from the \mathbf{u}_j where $\mathbf{u}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \mathbf{u}_j \sim N_p(\mathbf{0}, \mathbf{V}_{j,0})$ where the \mathbf{u}_{jn} correspond to the π_j . See Equation (8). Charkhi and Claeskens (2018) showed that $\mathbf{w}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0}^C - \boldsymbol{\beta}) \xrightarrow{D} \mathbf{w}_j$ if $S \subseteq I_j$ for the maximum likelihood estimator (MLE) with AIC, and gave a forward selection example. They claim that \mathbf{w}_j is a multivariate truncated normal distribution (where no truncation is possible) that is symmetric about $\mathbf{0}$. Hence $E(\mathbf{w}_j) = \mathbf{0}$, and $\text{Cov}(\mathbf{w}_j) = \boldsymbol{\Sigma}_j$ exists. This claim does not seem to be correct: the selection bias changes the distribution of the selected \mathbf{u}_{jn} and \mathbf{u}_j to that of \mathbf{w}_{jn} and \mathbf{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 \mathbf{w} is a mixture distribution of the \mathbf{w}_j with probabilities π_j .

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

$$\mathbf{w}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \xrightarrow{D} \mathbf{w} \quad (9)$$

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

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

Remark 3. a) The assumption that $\mathbf{w}_{jn} \xrightarrow{D} \mathbf{w}_j$ may not be mild.
b) If $P(S \subseteq I_{min}) \rightarrow 1$ as $n \rightarrow \infty$, 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 $P(S \subseteq I_{min}) \rightarrow 1$ as $n \rightarrow \infty$ holds for some criterion, such as AIC, BIC, and the C_p criterion for MLR. The assumption $P(S \subseteq I_{min}) \rightarrow 1$ 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

$$Y = \mathbf{x}^T \boldsymbol{\beta} + e \quad (10)$$

where $x_1 = 1$ and x_2, \dots, x_p are the nontrivial predictors. Hence β_1 corresponds to the constant α in the first MLR model (1). Let submodel I have a predictors, including a constant. Let F_I be the partial F statistic for testing whether the $p-a$ predictor variables left out of model I can be deleted. Then for a wide variety of iid error distributions, $F_I \xrightarrow{D} X/(p-a)$ where $X \sim \chi_{p-a}^2$. Then the variable selection criterion

$$C_p(I) = \frac{SSE(I)}{MSE} + 2a - n = (p-a)(F_I - 1) + a \quad (11)$$

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 $\pi_{kn} \rightarrow \pi_k$ as $n \rightarrow \infty$. 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 (10), 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 $X \sim \chi_1^2$. 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 $\pi_S = P(\chi_1^2 + p - 2 < p) = P(\chi_1^2 < 2) = 0.8427$, 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 $n \rightarrow \infty$ because $C_p(I_i) = C_p(S)$ is bounded in probability while $C_p(I_j)$ diverges as $n \rightarrow \infty$ for $j \neq i$. For forward selection with correlated predictors, expect that $\pi_S < P(\chi_1^2 < 2)$, and hence $\pi_F > 1 - P(\chi_1^2 < 2)$ with $\pi_S + \pi_F = 1$.

For the R code below, $\beta = (1, \dots, 1, 0, \dots, 0)^T$ 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, \dots, 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. For GAMs and SVMs, keep $\text{cor}(ESP(I_{\hat{q}}), ESP(I_p))$ high where SP=additive predictor or the weighted sum of kernel functions.

2.1.2 What Are k -Component Estimators Estimating?

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

The following new method is for **any** regression or classification method that depends on \mathbf{x} only through $SP = \alpha + \beta^T \mathbf{x}$. Let the standardized predictors $v_i = x_i / \hat{\sigma}_i$ for $i = 1, \dots, p$. Regress Y on the v_i to obtain the parameter vector $\hat{\theta}$ where the regression on the x_i gives $\hat{\beta}$. 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, and the method is closely related to thresholding 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, \dots, 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 β_F .

Instead of using the response variable Y and the predictors X_1, \dots, X_p , the regression model or classification model can use Y and the predictors W_1, \dots, W_k . Denote this model

by $I_k = \{1, \dots, k\}$. Then the k -component estimator $(\hat{\alpha}_k, \hat{\beta}_k)$ is obtained by fitting the working model

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

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

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

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

Then $\hat{\mathbf{A}}_k \mathbf{x} = \mathbf{w} = (W_1, \dots, W_k)^T$, and $ESP(\mathbf{w}) = \hat{\alpha}_k + \hat{\boldsymbol{\theta}}_k^T \mathbf{w} = \hat{\alpha}_k + \hat{\boldsymbol{\theta}}_k^T \hat{\mathbf{A}}_k \mathbf{x} = \hat{\alpha}_k + \hat{\beta}_k^T \mathbf{x} = ESP(\mathbf{x})$ with $\hat{\beta}_k = \hat{\mathbf{A}}_k^T \hat{\boldsymbol{\theta}}_k$. 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 $\sum_{j=1}^k \hat{\theta}_j \hat{\boldsymbol{\gamma}}_j \xrightarrow{P} \sum_{j=1}^k \theta_j \boldsymbol{\gamma}_j$.

For example, fit a GLM, logistic regression, a SVM, MLR, 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 may be new. The results are the most useful if $\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_k$ are linearly independent and if $\hat{\boldsymbol{\gamma}}_1, \dots, \hat{\boldsymbol{\gamma}}_k$ are linearly independent. The HPLS vectors are a Krylov sequence which tends to be linearly independent for small enough k if $\boldsymbol{\Sigma}_{\mathbf{x}Y}$ is not an eigenvector of $\boldsymbol{\Sigma}_{\mathbf{x}}$. 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 β_{OLS} , assume $\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}$ exists.

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

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

c) $\hat{\beta}_k = \sum_{j=1}^k \hat{\theta}_j \hat{\boldsymbol{\gamma}}_j = \hat{\mathbf{A}}_k^T \hat{\boldsymbol{\theta}}_k$. d) $\beta_k = \sum_{j=1}^k \theta_j \boldsymbol{\gamma}_j = \mathbf{A}_k^T \boldsymbol{\theta}_k$. e) $\hat{\beta}_{kPLS} = (\sum_{j=1}^k \hat{\theta}_j \hat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{j-1}) \hat{\boldsymbol{\Sigma}}_{\mathbf{x}Y}$

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

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

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

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

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

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

For discriminant analysis classification, suppose there are two groups, indexed by $Y = 0$ and $Y = 1$. Let the sample mean and covariance matrix for the j th group be $(\hat{\boldsymbol{\mu}}_j(\mathbf{x}) = \bar{\mathbf{x}}_j, \hat{\boldsymbol{\Sigma}}_j(\mathbf{x}))$ for $j = 0, 1$. Let $\hat{\boldsymbol{\Sigma}}_{pool}$ be a pooled covariance matrix estimator. Then the linear discriminant analysis (LDA) rule is allocate \mathbf{x} to group 1 if $ESP(\mathbf{x}) \geq 0$ and to group 0 if $ESP(\mathbf{x}) < 0$ where $ESP(\mathbf{x}) = \hat{\alpha}_{LDA} + \hat{\boldsymbol{\beta}}_{LDA}^T \mathbf{x}$ where $\hat{\boldsymbol{\beta}}_{LDA} = \hat{\boldsymbol{\Sigma}}_{pool}^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)$ and $\hat{\alpha}_{LDA} = 0.5(\hat{\boldsymbol{\mu}}_0^T \hat{\boldsymbol{\Sigma}}_{pool}^{-1} \hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1^T \hat{\boldsymbol{\Sigma}}_{pool}^{-1} \hat{\boldsymbol{\mu}}_1)$. Let the determinant $det(\mathbf{A}) = |\mathbf{A}|$. The quadratic discriminant analysis (QDA) rule is allocate \mathbf{x} to the group with the largest value of

$$Q_j(\mathbf{x}) = \frac{-1}{2} \log(|\hat{\boldsymbol{\Sigma}}_j|) - \frac{1}{2} (\mathbf{x} - \hat{\boldsymbol{\mu}}_j)^T \hat{\boldsymbol{\Sigma}}_j^{-1} (\mathbf{x} - \hat{\boldsymbol{\mu}}_j).$$

The distance discriminant (DD) rule allocates \mathbf{x} to the group with the smallest squared Mahalanobis distance $D_{\mathbf{x}}^2(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) = (\mathbf{x} - \hat{\boldsymbol{\mu}}_j)^T \hat{\boldsymbol{\Sigma}}_j^{-1} (\mathbf{x} - \hat{\boldsymbol{\mu}}_j)$. The K -nearest neighbors (KNN) method identifies the K cases in the training data that are closest to \mathbf{x} . Then allocate \mathbf{x} to the group with the majority of the K cases. Let $R(\mathbf{x})$ be the discriminant analysis rule using $\mathbf{x} = (x_1, \dots, x_p)^T$, and let $R(k)$ be the discriminant analysis rule using $\hat{\mathbf{A}}_k^T \mathbf{x} = (W_1, \dots, W_k)^T$. It is known that the $R(\mathbf{x})$ and $R(p)$ rules are equivalent for some PCA and PLS discriminant analysis methods. See Cook and Forzani (2024, p. 207) and Brereton and Lloyd (2014) for references.

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 $\hat{\boldsymbol{\beta}}_p = \hat{\boldsymbol{\beta}}_{\mathbf{x}} = \hat{\boldsymbol{\beta}}_F$, 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 $L_p(\boldsymbol{\beta}_{\mathbf{x}}|\mathbf{x}) = L(\boldsymbol{\beta}_{\mathbf{x}}, \hat{\boldsymbol{\eta}}|\mathbf{x})$ where L is the likelihood function of all of the parameters $(\boldsymbol{\beta}_{\mathbf{x}}, \boldsymbol{\eta})$ and $\hat{\boldsymbol{\eta}}$ is the MLE of $\boldsymbol{\eta}$. As above, use $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{\mathbf{w}}$ to denote the MLE with \mathbf{w} instead of $\hat{\boldsymbol{\beta}}_{\mathbf{w}}$. The $\hat{\boldsymbol{\gamma}}_i$ from PCA and NIPALS are orthonormal.

Theorem 5. Suppose the profile likelihood function $L_p(\boldsymbol{\beta}_{\mathbf{x}}|\mathbf{x}) = \prod_{i=1}^n f(\mathbf{x}_i|\boldsymbol{\beta}_{\mathbf{x}}) = \prod_{i=1}^n g(\mathbf{x}_i^T \boldsymbol{\beta}_{\mathbf{x}})$ depends on \mathbf{x} and $\boldsymbol{\beta}_{\mathbf{x}}$ only through $\mathbf{x}^T \boldsymbol{\beta}_{\mathbf{x}}$. a) If the maximum likelihood estimator is computed using $\mathbf{w} = \hat{\mathbf{A}}_p \mathbf{x}$ instead of \mathbf{x} , then $\hat{\boldsymbol{\beta}}_{\mathbf{x}} = \hat{\mathbf{A}}_p^T \hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\beta}}_p$ provided that $\hat{\mathbf{A}}_p$ is nonsingular. b) Thus $\hat{\boldsymbol{\beta}}_{\mathbf{x}} = \hat{\boldsymbol{\beta}}_{pPLS} = (\sum_{j=1}^p \hat{\theta}_j \hat{\boldsymbol{\Sigma}}_j^{j-1}) \hat{\boldsymbol{\Sigma}}_{\mathbf{x}Y}$ if the PLS components are used, and under iid cases, $\boldsymbol{\beta}_{\mathbf{x}} = (\sum_{j=1}^p \theta_j \boldsymbol{\Sigma}_{\mathbf{x}}^{j-1}) \boldsymbol{\Sigma}_{\mathbf{x}Y} = (\sum_{j=1}^p \theta_j \boldsymbol{\Sigma}_{\mathbf{x}}^j) \boldsymbol{\beta}_{OLS}$. c) $\hat{\boldsymbol{\Sigma}}_j(\mathbf{w}) = \hat{\mathbf{A}}_p \hat{\boldsymbol{\Sigma}}_j(\mathbf{x}) \hat{\mathbf{A}}_p^T$ and $\hat{\boldsymbol{\Sigma}}_{pool}(\mathbf{w}) = \hat{\mathbf{A}}_p \hat{\boldsymbol{\Sigma}}_{pool}(\mathbf{x}) \hat{\mathbf{A}}_p^T$. d) $\hat{\boldsymbol{\mu}}_j(\mathbf{w}) = \hat{\mathbf{A}}_p \hat{\boldsymbol{\mu}}_j(\mathbf{x})$. e) $D_{\mathbf{x}}^2(\hat{\boldsymbol{\mu}}_j(\mathbf{x}), \hat{\boldsymbol{\Sigma}}_j(\mathbf{x})) = (\mathbf{x} - \hat{\boldsymbol{\mu}}_j(\mathbf{x}))^T \hat{\boldsymbol{\Sigma}}_j^{-1}(\mathbf{x}) (\mathbf{x} - \hat{\boldsymbol{\mu}}_j(\mathbf{x})) = D_{\mathbf{w}}^2(\hat{\boldsymbol{\mu}}_j(\mathbf{w}), \hat{\boldsymbol{\Sigma}}_j(\mathbf{w})) = (\mathbf{w} - \hat{\boldsymbol{\mu}}_j(\mathbf{w}))^T \hat{\boldsymbol{\Sigma}}_j^{-1}(\mathbf{w}) (\mathbf{w} - \hat{\boldsymbol{\mu}}_j(\mathbf{w}))$. Hence the DD(\mathbf{x}) and DD(p) classification rules are equivalent. f) The LDA(\mathbf{x}) and LDA(p) classification rules are equivalent. g) If the $\hat{\boldsymbol{\gamma}}_i$ are orthonormal, then the QDA(\mathbf{x}) and QDA(p) classification rules are equivalent. h) If the $\hat{\boldsymbol{\gamma}}_i$ are orthonormal, then the KNN(\mathbf{x}) and KNN(p) classification rules are equivalent.

Proof. a)

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

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

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

c) Use Remark 1. d) Use $\mathbf{w} = \hat{\mathbf{A}}_p \mathbf{x}$. e) $(\mathbf{w} - \hat{\boldsymbol{\mu}}_j(\mathbf{w}))^T \hat{\boldsymbol{\Sigma}}_j^{-1}(\mathbf{w})(\mathbf{w} - \hat{\boldsymbol{\mu}}_j(\mathbf{w})) = [\hat{\mathbf{A}}_p(\mathbf{x} - \hat{\boldsymbol{\mu}}_j(\mathbf{x}))]^T [\hat{\mathbf{A}}_p \hat{\boldsymbol{\Sigma}}_{pool}(\mathbf{x}) \hat{\mathbf{A}}_p^T]^{-1} [\hat{\mathbf{A}}_p(\mathbf{x} - \hat{\boldsymbol{\mu}}_j(\mathbf{x}))]^T = D_{\mathbf{x}}^2(\hat{\boldsymbol{\mu}}_j(\mathbf{x}), \hat{\boldsymbol{\Sigma}}_j(\mathbf{x}))$.

f) $[\hat{\boldsymbol{\beta}}_{LDA}(\mathbf{w})]^T \mathbf{w} = [\hat{\boldsymbol{\Sigma}}_{pool}^{-1}(\mathbf{w})(\hat{\boldsymbol{\mu}}_1(\mathbf{w}) - \hat{\boldsymbol{\mu}}_0(\mathbf{w}))]^T \mathbf{w} = [(\hat{\mathbf{A}}_p \hat{\boldsymbol{\Sigma}}_{pool}(\mathbf{x}) \hat{\mathbf{A}}_p^T)^{-1} (\hat{\mathbf{A}}_p(\hat{\boldsymbol{\mu}}_1(\mathbf{x}) - \hat{\boldsymbol{\mu}}_0(\mathbf{x})))]^T \hat{\mathbf{A}}_p \mathbf{x} = [(\hat{\mathbf{A}}_p^T)^{-1} \hat{\boldsymbol{\Sigma}}_{pool}^{-1}(\mathbf{x})(\hat{\boldsymbol{\mu}}_1(\mathbf{x}) - \hat{\boldsymbol{\mu}}_0(\mathbf{x}))]^T \hat{\mathbf{A}}_p \mathbf{x} = (\hat{\boldsymbol{\mu}}_1(\mathbf{x}) - \hat{\boldsymbol{\mu}}_0(\mathbf{x}))^T [\hat{\boldsymbol{\Sigma}}_{pool}^{-1}(\mathbf{x})] \mathbf{x} = [\hat{\boldsymbol{\beta}}_{LDA}(\mathbf{x})]^T \mathbf{x}$. Similarly, $\hat{\alpha}_{LDA}(\mathbf{w}) = \hat{\alpha}_{LDA}(\mathbf{x})$.

g) The result follows by an argument similar to e) if

$$|\hat{\boldsymbol{\Sigma}}_j(\mathbf{w})| = |\hat{\mathbf{A}}_p \hat{\boldsymbol{\Sigma}}_j(\mathbf{x}) \hat{\mathbf{A}}_p^T| = |\hat{\mathbf{A}}_p^T \hat{\mathbf{A}}_p| |\hat{\boldsymbol{\Sigma}}_j(\mathbf{x})| = |\hat{\boldsymbol{\Sigma}}_j(\mathbf{x})|,$$

which holds under orthonormality since then $|\hat{\mathbf{A}}_p^T \hat{\mathbf{A}}_p| = |\mathbf{I}_p| = 1$.

h) Let \mathbf{w}_i and \mathbf{w}_j be two cases. Then $(\mathbf{w}_i - \mathbf{w}_j)^T (\mathbf{w}_i - \mathbf{w}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T \hat{\mathbf{A}}_p^T \hat{\mathbf{A}}_p (\mathbf{x}_i - \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T \hat{\mathbf{A}}_p (\mathbf{x}_i - \mathbf{x}_j)$. Hence the Euclidean distances are the same. \square

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 \mathbf{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 W_1, \dots, W_k for $k = 1, \dots, M$. 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

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

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

$$\alpha_q + \theta_1 W_1 + \dots + \theta_q W_q + 0W_{q+1} + \dots + 0W_p = \alpha_q + \boldsymbol{\theta}_q^T \mathbf{w}_{I_q} = \alpha_p + \boldsymbol{\theta}_p^T \mathbf{w}_{I_p}, \quad (12)$$

or

$$WSP = \alpha_p + \theta_1 W_1 + \dots + \theta_d W_d + (0^*) W_{d+1} + \dots + (0^*) W_p. \quad (13)$$

Equation (12), which corresponds to Equation (6), 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 $\boldsymbol{\beta}_q = \boldsymbol{\beta}_{q+1} = \dots = \boldsymbol{\beta}_p$, a result that is analogous to the variable selection result $\boldsymbol{\beta}_{I,0} = \boldsymbol{\beta}_F$ if $S \subseteq I$. Consider Equation (13), perhaps with $0^* = 10^{-6}$ and corresponding observed $|W_j| \leq 10$. Then $I_q = S = F = I_p$ in Equation (12), and the central subspace and the predictor envelope satisfy $\mathcal{L}_{\mathbf{y}|\mathbf{x}} = \mathcal{E}_{\mathbf{x}} = \mathbb{R}^p$. Under Equation (13), the $\boldsymbol{\beta}_k$ are different for $k = 1, \dots, p$.

If I is the model selected by the variable selection method, and $P(S \subseteq I) \rightarrow 1$ as $n \rightarrow \infty$, 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 (13), 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 . Use these plots to check Remark 4 models.

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 γ_i need to be linearly independent and the $\hat{\gamma}_i$ need to be linearly independent for models that use $W_i = \hat{\gamma}_i^T \mathbf{x}$ so that $\hat{\beta}_p(\mathbf{w}_F) = \hat{\beta}_p(\mathbf{x}_F)$ and $\beta_p(\mathbf{w}_F) = \beta_p(\mathbf{x}_F) = \beta_F$ where $\mathbf{w}_F = (W_1, \dots, W_p)^T$, $\mathbf{x}_F = (x_1, \dots, x_p)^T$, and $\beta(\mathbf{z})$ means that Y was regressed on \mathbf{z} . Then $\hat{\mathbf{A}}_p$ is nonsingular. $W_i = x_i$ corresponds to $\hat{\gamma}_i = \gamma_i = \mathbf{c}_i = i$ th column of \mathbf{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, 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.

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 \mathbf{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 $\beta_F = \beta = \beta_{OLS}$. 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 \mathbf{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 $\hat{\beta}_I$ with k predictors including a constant where $2 \leq k \leq p + 1$. If

$C_p(I) \leq 2k$, then the sample correlation

$$\text{cor}(r, r_I) \geq \sqrt{1 - \frac{p+1}{n}}. \quad (14)$$

Since the correlation gets arbitrarily close to 1 as $n \rightarrow \infty$, the model selection estimator and full OLS estimator are estimating the same population parameter β , and $P(S \subseteq I_{min}) \rightarrow 1$ as $n \rightarrow \infty$. This result holds if $\hat{\beta}_{OLS}$ is a consistent estimator of β : 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 $P(S \subseteq I_{min}) \rightarrow 1$ is not exponentially fast.

For moderate sample size n , predictors x_j or W_j will often be omitted as long as $\text{cor}(r_I, r)$ 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 (14), often $\beta_{I,0} \neq \beta_F$, but $\hat{\beta}_I$ is a good estimator of β_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): $Y|\mathbf{x} = \alpha + \mathbf{x}^T \beta_{kPLS} + e$ for some k . If $Y|\mathbf{x} = \alpha + \mathbf{x}^T \beta + e$, then under mild regularity conditions, $\beta = \beta_{OLS}$. Hence assumption (a1) forces $\beta_{kPLS} = \beta_{OLS}$. For $k = 1$, (a1) forces $\Sigma \mathbf{x}_Y$ and $\beta_{OLS} = \beta_{1PLS}$ to be eigenvectors of the covariance matrix $\text{Cov}(\mathbf{x}) = \Sigma \mathbf{x}$.

Theorem 4 shows that $\hat{\beta}_{kPLS} = (\sum_{j=1}^k \hat{\theta}_j \hat{\Sigma} \mathbf{x}^{j-1}) \hat{\Sigma} \mathbf{x}_Y$ and $\beta_{kPLS} = (\sum_{j=1}^k \theta_j \Sigma \mathbf{x}^{j-1}) \Sigma \mathbf{x}_Y$ (under iid cases). This result suggests that the β_{kPLS} are typically different for each $k = 1, \dots, p$, but $\beta_{kPLS} = \beta_{qPLS}$ for $k \leq q \leq p$ if $\theta_j = 0$ for $k+1 \leq j \leq p$. Note that $\beta \in \mathbb{R}^p$ is a much weaker assumption than $\beta \in \mathbb{R}^m$ where $1 \leq m < p$.

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 $n \rightarrow \infty$. This result is useful for data splitting. 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 $\text{cor}(ESP(I), ESP(F))$ high and $\text{cor}(Y - ESP(I), Y - ESP(F))$ high, by keeping wanted predictors I in the model and omitting unwanted predictors O .

b) The estimator $\hat{\beta}_I$ estimates β_I where often $\beta_{I,0} \neq \beta = \beta_F$.

c) In high dimensions, variable selection is a search for a useful subset I of predictors.

d) The k -component regression is a sequential variable selection procedure. An alternative method is to apply alternative variable selection methods on the W_i , for example lasso. See, for example, Song and Zou (2026). This technique ignores the ordering of the variables.

3 The OPLS Estimator

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

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

$$\theta = \frac{\Sigma_{\mathbf{x}_Y}^T \Sigma_{\mathbf{x}_Y}}{\Sigma_{\mathbf{x}_Y}^T \Sigma_{\mathbf{x}} \Sigma_{\mathbf{x}_Y}} \quad \text{and} \quad \hat{\theta} = \frac{\hat{\Sigma}_{\mathbf{x}_Y}^T \hat{\Sigma}_{\mathbf{x}_Y}}{\hat{\Sigma}_{\mathbf{x}_Y}^T \hat{\Sigma}_{\mathbf{x}} \hat{\Sigma}_{\mathbf{x}_Y}} \quad (15)$$

for $\Sigma_{\mathbf{x}_Y} \neq \mathbf{0}$. If $\Sigma_{\mathbf{x}_Y} = \mathbf{0}$, then $\beta_{OPLS} = \mathbf{0}$. Remark 1 can be used to prove (15).

Next, some large sample theory is reviewed for $\hat{\boldsymbol{\eta}}_{OPLS} = \hat{\Sigma}_{\mathbf{x}_Y}$ 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 $\hat{\boldsymbol{\eta}} = \widehat{\text{Cov}}(\mathbf{x}, Y)$. Olive et al. (2026) gave alternative proofs. This theory needs $\boldsymbol{\eta} = \boldsymbol{\eta}_{OPLS} = \Sigma_{\mathbf{x}_Y}$ to exist for $\hat{\boldsymbol{\eta}} = \hat{\Sigma}_{\mathbf{x}_Y}$ to be a consistent estimator of $\boldsymbol{\eta}$. Let $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^T$ and let \mathbf{w}_i and \mathbf{z}_i be defined below where

$$\text{Cov}(\mathbf{w}_i) = \Sigma_{\mathbf{w}} = E[(\mathbf{x}_i - \boldsymbol{\mu}_{\mathbf{x}})(\mathbf{x}_i - \boldsymbol{\mu}_{\mathbf{x}})^T (Y_i - \mu_Y)^2] - \Sigma_{\mathbf{x}_Y} \Sigma_{\mathbf{x}_Y}^T.$$

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

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

$$\sqrt{n}(\bar{\mathbf{w}}_n - \boldsymbol{\eta}) \xrightarrow{D} N_p(\mathbf{0}, \Sigma_{\mathbf{w}}), \quad \sqrt{n}(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}) \xrightarrow{D} N_p(\mathbf{0}, \Sigma_{\mathbf{w}}), \quad (16)$$

$$\text{and} \quad \sqrt{n}(\tilde{\boldsymbol{\eta}}_n - \boldsymbol{\eta}) \xrightarrow{D} N_p(\mathbf{0}, \Sigma_{\mathbf{w}}).$$

b) Let $\mathbf{v}_i = (\mathbf{x}_i - \bar{\mathbf{x}}_n)(Y_i - \bar{Y}_n)$. Then $\hat{\Sigma}_{\mathbf{w}} = \hat{\Sigma}_{\mathbf{v}} + O_P(n^{-1/2})$. Hence $\tilde{\Sigma}_{\mathbf{w}} = \tilde{\Sigma}_{\mathbf{v}} + O_P(n^{-1/2})$.

c) Let \mathbf{A} be a $k \times p$ full rank constant matrix with $k \leq p$, assume $H_0 : \mathbf{A} \hat{\beta}_{OPLS} = \mathbf{0}$ is true, and assume $\hat{\theta} \xrightarrow{P} \theta \neq 0$. Then

$$\sqrt{n} \mathbf{A} (\hat{\beta}_{OPLS} - \beta_{OPLS}) \xrightarrow{D} N_k(\mathbf{0}, \theta^2 \mathbf{A} \Sigma_{\mathbf{w}} \mathbf{A}^T). \quad (17)$$

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

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

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

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

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

The following Olive et al. (2026) theorem provides large sample theory for $\hat{\mathbf{c}}$ and $\tilde{\mathbf{c}}$. We use $Cov(\mathbf{w}_i) = \Sigma_{\mathbf{d}}$ to avoid confusion with the $\Sigma_{\mathbf{w}}$ used in Theorem 7. Note that \mathbf{x}_i are dummy variables and could be replaced by $\mathbf{u}_i = (Y_{i1}, \dots, Y_{im}, x_{i1}, \dots, x_{ip})^T$ to get information about m response variables Y_1, \dots, Y_m .

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

- (i) $\sqrt{n}(\tilde{\mathbf{c}} - \mathbf{c}) \xrightarrow{D} N_k(\mathbf{0}, \Sigma_{\mathbf{d}})$.
- (ii) $\sqrt{n}(\hat{\mathbf{c}} - \mathbf{c}) \xrightarrow{D} N_k(\mathbf{0}, \Sigma_{\mathbf{d}})$.
- (iii) $\hat{\Sigma}_{\mathbf{d}} = \hat{\Sigma}_{\mathbf{v}} + O_P(n^{-1/2})$ and $\tilde{\Sigma}_{\mathbf{d}} = \tilde{\Sigma}_{\mathbf{v}} + O_P(n^{-1/2})$.

4 Large Sample Theory and Testing

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

4.1 Testing $H_0 : \beta = \mathbf{0}$

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 $\mathbf{x} = (x_1, \dots, x_p)^T$ through the sufficient predictor $SP = \alpha + \mathbf{x}^T \beta$. Assume the cases $(\mathbf{x}_i^T, Y_i)^T$ are iid random vectors for $i = 1, \dots, n$. Then for many such regression models, $\beta = \mathbf{0}$ if and only if $\Sigma_{\mathbf{x}Y} = \mathbf{0}$ where $\mathbf{0} = (0, \dots, 0)^T$ is the $p \times 1$ vector of zeroes. By Theorem 5b), this result holds for $\beta_{\mathbf{x}} = \beta_{pPLS}$ provided that $\Sigma_{\mathbf{x}Y}$ is not in the null space of the matrix $\sum_{j=1}^p \theta_j \Sigma_{\mathbf{x}}^{j-1}$. Suppose $\beta_D = \mathbf{D}^{-1} \Sigma_{\mathbf{x}Y}$ where \mathbf{D} is a $p \times p$ nonsingular matrix. Then $\beta_D = \mathbf{0}$ if and only if $\Sigma_{\mathbf{x}Y} = \mathbf{0}$. Then $\mathbf{D}^{-1} = \theta \mathbf{I}_p$ for OPLS, $\mathbf{D}^{-1} = \Sigma_{\mathbf{x}}^{-1}$ for OLS, and $\mathbf{D}^{-1} = [\text{diag}(\Sigma_{\mathbf{x}})]^{-1}$ for the MMLE for multiple linear regression.

Zhao et al. (2024) have an interesting result for the multiple linear regression model (1). Assume that the cases $(\mathbf{x}_i^T, Y_i)^T$ are iid with $E(Y) = \mu_Y$, $E(\mathbf{x}) = \mu_{\mathbf{x}}$ and nonsingular $Cov(\mathbf{x}) = \Sigma_{\mathbf{x}}$. Let $\beta = \beta_{OLS}$. Then testing $H_0 : \beta = \beta_0$ versus $H_1 : \beta \neq \beta_0$ is equivalent to testing $H_0 : \mu = \mathbf{0}$ versus $H_1 : \mu \neq \mathbf{0}$ with $\mu = E(\mathbf{w}_i) = \Sigma_{\mathbf{x}}(\beta - \beta_0)$

where $\mathbf{w}_i = (\mathbf{x}_i - \boldsymbol{\mu}_X)(Y_i - \mu_Y - (\mathbf{x}_i - \boldsymbol{\mu}_X)^T \boldsymbol{\beta}_0)$, and a one sample test can be applied to $\mathbf{v}_i = (\mathbf{x}_i - \bar{\mathbf{x}})(Y_i - \bar{Y} - (\mathbf{x}_i - \bar{\mathbf{x}})^T \boldsymbol{\beta}_0)$.

Abid, Quaye, and Olive (2025) used the above test for $\boldsymbol{\beta}_0 = \mathbf{0}$. If $\boldsymbol{\beta} = \mathbf{0}$, then the Y_i are iid and do not depend on \mathbf{x} , and thus satisfy a multiple linear regression model with $\boldsymbol{\beta}_{OLS} = \mathbf{0}$. Then $\boldsymbol{\mu} = E(\mathbf{w}_i) = \boldsymbol{\Sigma} \mathbf{x}_Y$, and a high dimensional one sample test is applied to $\mathbf{v}_i = (\mathbf{x}_i - \bar{\mathbf{x}})(Y_i - \bar{Y})$. Note that the sample mean $\bar{\mathbf{v}} = \tilde{\boldsymbol{\Sigma}} \mathbf{x}_Y$. This test of $H_0 : \boldsymbol{\Sigma} \mathbf{x}_Y = \mathbf{0}$ versus $H_1 : \boldsymbol{\Sigma} \mathbf{x}_Y \neq \mathbf{0}$ is equivalent to the test of $H_0 : \boldsymbol{\beta} = \mathbf{0}$ versus $H_1 : \boldsymbol{\beta} \neq \mathbf{0}$. For multiple linear regression with heterogeneity, $\hat{\boldsymbol{\beta}}_{OLS}$ is still a consistent estimator of $\boldsymbol{\beta} = \boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma} \mathbf{x}_Y$. Hence the test can be used when the constant variance assumption is violated.

Suppose $\mathbf{x}_1, \dots, \mathbf{x}_n$ are iid random vectors with $E(\mathbf{x}) = \boldsymbol{\mu}$ and covariance matrix $\text{Cov}(\mathbf{x}) = \boldsymbol{\Sigma}$. 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 $\mathbf{S} = \hat{\boldsymbol{\Sigma}}$. A U-statistic for estimating $\boldsymbol{\mu}^T \boldsymbol{\mu}$ is

$$T_n = T_n(\mathbf{x}) = \frac{1}{n(n-1)} \sum_{i \neq j} \mathbf{x}_i^T \mathbf{x}_j = \frac{n \bar{\mathbf{x}}^T \bar{\mathbf{x}} - \text{tr}(\mathbf{S})}{n} \quad (18)$$

where $\text{tr}()$ is the trace function. See, for example, Abid, Quaye, and Olive (2025).

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

$$S_W^2 = \frac{1}{m-1} \sum_{i=1}^m (W_i - \bar{W})^2. \quad (19)$$

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

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 9a) was obtained by Chen and Qin (2010).

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

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

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

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

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 $p \rightarrow \infty$ 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 (19). An estimator nearly the same as the one used by Li (2023) is

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

A New Competing Test

If the parametric distribution D is known, then the iid cases assumption can be changed to independent cases. Assume $Y_i | \mathbf{x}_i^T \boldsymbol{\beta} \sim D(\tau(\alpha + \mathbf{x}_i^T \boldsymbol{\beta}), \boldsymbol{\theta})$. 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 \mathbf{x} . The test does not depend on \mathbf{x} , and can thus be done after variable selection. Also, all of the predictors can have outliers and missing values. The power can be sometimes be very poor if the cases are iid. a) If the $(Y_i, \mathbf{x}_i^T)^T$ 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 $Y | (\alpha + \mathbf{x}^T \boldsymbol{\beta}) \sim N(\alpha + \mathbf{x}^T \boldsymbol{\beta}, \sigma^2)$. b) If the $(Y_i, \mathbf{x}_i^T)^T$ 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 $Y | (\alpha + \mathbf{x}^T \boldsymbol{\beta}) \sim bin(n=1, \rho(\alpha + \mathbf{x}^T \boldsymbol{\beta}))$.

A Test for Binary Regression or Classification

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

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

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

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

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

$$\begin{aligned} \tilde{\boldsymbol{\Sigma}}_{\mathbf{x}Y} &= \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i Y_i - \bar{\mathbf{x}} \bar{Y} = \frac{1}{n} \left[\sum_{j: Y_j=1} \mathbf{x}_j (1) + \sum_{j: Y_j=0} \mathbf{x}_j (0) \right] - \bar{\mathbf{x}} \hat{\pi}_1 = \\ &= \frac{1}{n} (N_1 \hat{\boldsymbol{\mu}}_1) - \frac{1}{n} (N_1 \hat{\boldsymbol{\mu}}_1 + N_0 \hat{\boldsymbol{\mu}}_0) \hat{\pi}_1 = \hat{\pi}_1 \hat{\boldsymbol{\mu}}_1 - \hat{\pi}_1^2 \hat{\boldsymbol{\mu}}_1 - \hat{\pi}_1 \hat{\pi}_0 \hat{\boldsymbol{\mu}}_0 = \end{aligned}$$

$$\hat{\pi}_1(1 - \hat{\pi}_1)\hat{\boldsymbol{\mu}}_1 - \hat{\pi}_1\hat{\pi}_0\hat{\boldsymbol{\mu}}_0 = \hat{\pi}_1\hat{\pi}_0(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0).$$

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

This result means $\boldsymbol{\eta} = \boldsymbol{\Sigma}_{\mathbf{x},Y} = \pi_1\pi_0(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$ and $\boldsymbol{\phi} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$ are quantities of interest for binary regression. Note that $\mathbf{x} = (w_1, \dots, w_k, w_1w_2, \dots, w_1w_k, \dots, w_{k-1}w_k)^T$ could be used to include pairwise interactions of the w_i . Theorem 5b) suggests that often the binary regression $\hat{\boldsymbol{\beta}} = \hat{\mathbf{C}}\hat{\boldsymbol{\Sigma}}\mathbf{x}_Y$. If the cases $(Y_i, \mathbf{x}_i^T)^T$ are iid, then $H_0 : \boldsymbol{\beta} = \mathbf{0}$ can be tested with the omnibus test for $H_0 : \boldsymbol{\Sigma}_{\mathbf{x}Y} = \mathbf{0}$ or with a high dimensional two sample test for $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_0$. If the cases within each group are iid, then $\boldsymbol{\beta}_{LDA} = \boldsymbol{\Sigma}_{pool}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$. Hence a high dimensional two sample test for $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_0$ can also be used to test $H_0 : \boldsymbol{\beta}_{LDA} = \mathbf{0}$.

Let $\mathbf{w} = \mathbf{w}(k) = (W_1, \dots, W_k)^T$. Let (k) denote that the estimator was computed using \mathbf{w} . Bing and Wegkamp (2023) use iid cases and $\hat{\boldsymbol{\beta}}(k) = \hat{\pi}_1\hat{\pi}_0\hat{\boldsymbol{\Sigma}}_{\mathbf{w}}^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0) = \hat{\boldsymbol{\beta}}_{OLS}(\mathbf{w}, Y)$ by Theorem 10, where $\hat{\boldsymbol{\beta}}(k)$ is found from the OLS regression of Y on $\mathbf{w}(k)$, and $\hat{\alpha}(k) = -0.5[\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_0]^T\hat{\boldsymbol{\beta}}(k) + [1 - (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)^T\hat{\boldsymbol{\beta}}(k)] \log(\hat{\pi}_1/\hat{\pi}_0)$. The W_i were found using PCA. Using (5) to order the W_i or using PLS W_i may greatly improve this method.

4.2 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.

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

For OPLS, testing $H_0 : \boldsymbol{\beta}_I = \mathbf{0}$ versus $H_A : \boldsymbol{\beta}_I \neq \mathbf{0}$ is equivalent to testing $H_0 : [Cov(x_{i_1}, Y), \dots, Cov(x_{i_k}, Y)]^T = \mathbf{0}$. 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 $H_0 : \mathbf{A}\boldsymbol{\beta}_{BR} = \mathbf{0}$ versus $H_1 : \mathbf{A}\boldsymbol{\beta}_{BR} \neq \mathbf{0}$ is equivalent to testing $H_0 : \mathbf{A}\boldsymbol{\eta} = \mathbf{0}$ versus $H_1 : \mathbf{A}\boldsymbol{\eta} \neq \mathbf{0}$ where \mathbf{A} is a $k \times p$ constant matrix. Let $Cov(\hat{\boldsymbol{\eta}}) = \boldsymbol{\Sigma}_{\mathbf{w}}$ be the asymptotic covariance matrix of $\hat{\boldsymbol{\eta}}$. In high dimensions where $n < 5p$, we can't get a good nonsingular estimator of $Cov(\hat{\boldsymbol{\eta}})$, but we can get good nonsingular estimators of $Cov((\hat{\eta}_{i_1}, \dots, \hat{\eta}_{i_k})^T)$ with $\mathbf{u} = (x_{i_1}, \dots, x_{i_k})^T$ 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 is near 0 or 1.) Simply use the sample covariance matrix with \mathbf{u} replacing \mathbf{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.

5 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.

5.1 Everyone is Trying to Estimate β , and Nearly Nothing Works in High Dimensions

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

$$\|\hat{\theta} - \theta\|^2 = \sum_{i=1}^p (\hat{\theta}_i - \theta_i)^2 \propto p/n. \quad (20)$$

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

Here $\theta = \beta$ or $\theta = \gamma$ are possible. For example, the sample eigenvectors \hat{d}_i tend to be poor estimators of the population eigenvectors d_i of $\Sigma_{\mathbf{x}}$. 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).

Often it is useful to eliminate weak predictors before finding the $\hat{\gamma}_i$. A possible exception is when the predictor variables come in G groups where $n \geq 10G$, say, and the predictors in each group are highly correlated. This scenario often happens for chemometrics data and when Y and the x_i are time series. Then $W = \hat{\gamma}^T \mathbf{x}$ is roughly approximated by $(c_1 \bar{x}_{w1}, \dots, c_G \bar{x}_{wG})^T$ where \bar{x}_{wj} is a weighted average of the predictors in the j th group. Then W may outperform selecting one predictor from each of the G groups.

5.1.1 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) $\beta = \beta_F = \beta_{I,0}$.
- c) The number of variables a in I is small compared to n so that $\hat{\beta}_{I,0}$ is a good estimator of β_F . Hence $\hat{\beta}_I$ is a good estimator of β_I .

A useful bigger model is $\hat{\beta}_I$ is a good estimator of β_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 β_I , not for model $\beta_{I,0} = \beta_F$.

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 $\lambda_1 < \lambda_2 < \dots < \lambda_M$. When p is fixed, $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau$ does not do variable selection well. For variable selection, want $\hat{\lambda}_{1,n}/\sqrt{n} \rightarrow \infty$, but $\hat{\lambda}_{1,n}/n \rightarrow 0$. 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 $\boldsymbol{\theta} = (\theta_0, \theta_1, \dots, \theta_p)^T$. Fit lasso with an intercept, no standardization, and nonnegativity constraints:

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

Perform the minimization over a grid of λ values and select $\hat{\lambda}$ using k -fold cross validation. Let $I = \{i_1, \dots, i_k\}$ correspond to the k predictors with nonzero $\hat{\beta}_i$ where the unilasso estimator $\hat{\boldsymbol{\beta}}_I = (\hat{\beta}_{i_1}, \dots, \hat{\beta}_{i_k})^T$ and $\hat{\boldsymbol{\beta}}_U = \hat{\boldsymbol{\beta}}_{I,0} = (\hat{\beta}_1, \dots, \hat{\beta}_p)^T$. Then $ESP(I) = \hat{\alpha} + \hat{\boldsymbol{\beta}}_I^T \mathbf{x}_I = \hat{\alpha} + \hat{\boldsymbol{\beta}}_U^T \mathbf{x} = \hat{\alpha} + \sum_{i=1}^p \hat{\beta}_i x_i = \hat{\alpha} + \sum_{j=1}^k \hat{\beta}_{i_j} x_{i_j}$ 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 $ESP(F) = \hat{\alpha}_F + \hat{\boldsymbol{\beta}}_F^T \mathbf{x}$ for a full model that depends on \mathbf{x} through $SP = \alpha_F + \boldsymbol{\beta}_F^T \mathbf{x}$. 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 $i = 1, \dots, p$ 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.

5.2 Envelope Theory

When the cases are iid with 4th moments, the assumptions $Y|\mathbf{x} = Y|\mathbf{Ax}$ and $Y \perp\!\!\!\perp \mathbf{x}|\mathbf{Ax}$ tend to be very strong compared to these assumptions when the cases are iid from a multivariate normal distribution. The $Y|\mathbf{x}$ assumption assumes that every relevant transformation of the predictors $t_j(x_1, \dots, x_p)$ is correct and in the model, including interactions.

To establish the link between envelopes and PLS, and for the theory for $\hat{\beta}_{1PLS}$, Cook and Forzani (2024, pp. 83, 92, 111) make the very strong assumption that $\Sigma_{\mathbf{x}Y}$ is an eigenvector of $\Sigma_{\mathbf{x}}$.

For the predictor envelope, Cook and Forzani (2024, p. 38) state that if the eigenvalues of $\Sigma_{\mathbf{x}}$ are unique, and if $\dim(\mathcal{E}_{\mathbf{x}}) = q$, then $\mathcal{E}_{\mathbf{x}}$ is spanned by q eigenvectors of $\Sigma_{\mathbf{x}}$. Intuitively, the PLS components should outperform PCA components since Y is used for PLS. When the above result is correct, using (5) to order the PLS components should result in an envelopes estimator that is fast to compute, and the technique can be used in high dimensions.

6 Outlier Resistance

To make outlier resistant analogs for many statistical techniques, including envelopes, lasso, PCA, and PLS, let \mathbf{U} be a data matrix, where the rows \mathbf{u}_i correspond to cases. For example, $\mathbf{u}_i = \mathbf{x}_i$ or $\mathbf{u}_i = (Y_{i1}, x_{i1}, \dots, x_{ip})^T$. One of the simplest outlier detection methods uses the Euclidean distances of the \mathbf{u}_i from the coordinatewise median $D_i = D_i(\text{MED}(\mathbf{U}), \mathbf{I}_p)$. Concentration type steps compute the weighted median MED_j : the coordinatewise median computed from the cases \mathbf{u}_i with $D_i^2 \leq \text{MED}(D_i^2(\text{MED}_{j-1}, \mathbf{I}_p))$ where $\text{MED}_0 = \text{MED}(\mathbf{U})$. We often used $j = 0$ (no concentration type steps) or $j = 9$. Let $D_i = D_i(\text{MED}_j, \mathbf{I}_p)$. Let $W_i = 1$ if $D_i \leq \text{MED}(D_1, \dots, D_n) + k\text{MAD}(D_1, \dots, D_n)$ where $k \geq 0$ and $k = 5$ is the default choice. Let $W_i = 0$, otherwise. Using $k \geq 0$ insures that at least half of the cases get weight 1. This weighting corresponds to the weighting that would be used in a one sided metrically trimmed mean (Huber type skipped mean) of the distances. Here, the sample median absolute deviation is $\text{MAD}(n) = \text{MAD}(D_1, \dots, D_n) = \text{MED}(|D_i - \text{MED}(n)|, i = 1, \dots, n)$ where $\text{MED}(n) = \text{MED}(D_1, \dots, D_n)$ is the sample median of D_1, \dots, D_n .

Let the *covmb2* set B of $n_R \geq n/2$ cases correspond to the cases with weight $W_i = 1$, and apply the statistical technique on the n_R cases. For variants, R code, and more explanation, see Olive (2017b, 2025). The Olive (2017b, p. 120) *covmb2* estimator (T, \mathbf{C}) is the sample mean and sample covariance matrix applied to the cases in set B . If $\mathbf{u}_i = \mathbf{x}_i$, then

$$T = \frac{\sum_{i=1}^n W_i \mathbf{x}_i}{\sum_{i=1}^n W_i} \quad \text{and} \quad \mathbf{C} = \frac{\sum_{i=1}^n W_i (\mathbf{x}_i - T)(\mathbf{x}_i - T)^T}{\sum_{i=1}^n W_i - 1}.$$

This estimator was built for speed, applications, and outlier resistance. The estimator can be used in high dimensions, and in low dimensions, the population dispersion matrix is the population covariance matrix of a spherically truncated distribution.

7 Example

The following example demonstrates the Theorem 5 result that $\hat{\beta}_{MLE} = \hat{\beta}_x = \hat{\beta}_p$ if \hat{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 $\log(\textit{endem})$ and $\log(\textit{areanear})$. For both Poisson regression and negative binomial regression, HPLS $\hat{\beta}_p = \hat{\beta}_x$.

```
source("http://parker.ad.siu.edu/Olive/slldata.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$coef #PR thetihat = (0.01421,-0.001482)'
  (Intercept)          w1          w2
-0.044393306  0.014205369 -0.001481973
ESP2 <- predict(out2)
plot(ESP2,ESP1)
abline(0, 1)
AhatTrans <- cbind(gam1hat,gam2hat)
AhatTrans%*%out2$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

```

8 CONCLUSIONS

A useful high dimensional technique is to use PCA for dimension reduction. Let U_1, \dots, U_p be the PCA linear combinations ($U_i = \hat{\mathbf{d}}_i^T \mathbf{x}$) ordered with respect to the largest eigenvalues. Then use U_1, \dots, 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, \dots, W_k should work much better than using U_1, \dots, 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 \mathbf{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, \dots, W_p be ordered with respect to the highest squared correlations $r_1^2 \geq r_2^2 \geq \dots \geq r_p^2$ where the sample correlation $Q(i) = r_i = r_{i,Y} = \text{cor}(x_i, Y)$. See Olive (2025).

Software

The R software was used in the simulations. See R Core Team (2024). Programs will be added to the collection 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 `p1s` 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.
- Brereton, R.G., and Lloyd, G.R. (2014), “Partial Least Squares Discriminant Analysis: Taking the Magic Away,” *Journal of Chemometrics*, 28, 213-225.
- Bing, X., and Wegkamp, M. (2023), “Optimal Discriminant Analysis in High-Dimensional Latent Factor Models,” *The Annals of Statistics*, 51, 1232-1257.
- Charkhi, A., and Claeskens, G. (2018), “Asymptotic Post-Selection Inference for the Akaike Information Criterion,” *Biometrika*, 105, 645-664.
- 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 Li, R. (2001), “Variable Selection via Noncave Penalized Likelihood and Its Oracle Properties,” *Journal of the American Statistical Association*, 96, 1348-1360.
- 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 Problem and Why They Wont Work,” *Journal of the American Statistical Association*, 112, 899-901.
- Guan, L., and Tibshirani, R. (2020), “Post Model-Fitting Exploration via a “Next-Door” Analysis,” *Canadian Journal of Statistics*, 48, 447-470.
- 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>).
- 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.
- Jin, Y., and Olive, D.J. (2026), “Large Sample Theory for Some Ridge-Type Regression Estimators,” is at (<http://parker.ad.siu.edu/Olive/ppridgetype.pdf>).
- 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.
- Jung, S. and Marron, J.S. (2012), “PCA Consistency in High Dimension Low Sample Size Context,” *The Annals of Statistics*, 37, 4104-4130.
- 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. (2013), “Plots for Generalized Additive Models,” *Communications in Statistics: Theory and Methods*, 42, 2610-2628.
- Olive, D.J. (2017a), *Linear Regression*, Springer, New York, NY.
- Olive, D.J. (2017b), *Robust Multivariate Analysis*, 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.
- Pötscher, B. (1991), “Effects of Model Selection on Inference,” *Econometric Theory*, 7, 163-185.
- 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).
- Rathnayake, R.C., and Olive, D.J. (2023), “Bootstrapping Some GLM and Survival Regression Variable Selection Estimators,” *Communications in Statistics: Theory and Methods*, 52, 2625-2645.
- Song, Y., and Zou, H. (2026), “Another Look at High-Dimensional Regression in Principal Components Space and The Blessing of Dimensionality,” *Journal of the American Statistical Association*, to appear.
- 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.
- 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.