Some Useful Techniques for High Dimensional Statistics

David J. Olive*
Southern Illinois University

April 20, 2025

Abstract

High dimensional statistics are used when $n < 5p$ where n is the sample size and p is the number of predictors. Useful techniques include a) use a sparse fitted model, b) use principal component analysis for dimension reduction, c) use alternative multivariate dispersion estimators instead of the sample covariance matrix, d) eliminate weak predictors, and e) stack low dimensional estimators into a vector. Some variants and theory for these techniques will be given or reviewed.

KEY WORDS: Artificial Intelligence, Lasso, Machine Learning, Model Selection, Outliers, PCA, PLS.

1 Introduction

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 not be large enough for the classical statistical method to be useful. An alternative (but less general) definition of high dimensional statistics is that p is large. Sometimes $p > Kn$ with $K \geq 10$ is called ultrahigh dimensional statistics.

Some important statistical methods include regression, multivariate statistics, and classification. These methods are important for statistical learning $\approx$ machine learning, an important part of artificial intelligence. Let predictor variables for regression or multivariate statistics be $\boldsymbol{x} = (x_1, ..., x_p)^T$. Let Y be a response variable for regression or classification. Important regression models include generalized linear models, nonlinear regression, nonparametric regression, and survival regression models. Inference for multivariate regression where there are m response variables $Y_1, ..., Y_m$ is also of interest. Useful references for the following statistical methods include James et al. (2021) and Cook and Forzani (2024).

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

Let the population covariance matrices

$$\mathrm{Cov}(\boldsymbol{x}) = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))^T] = \boldsymbol{\Sigma}_{\boldsymbol{x}}, \text{ and}$$

$$\mathrm{Cov}(\boldsymbol{x}, Y) = E[(\boldsymbol{x} - E(\boldsymbol{x}))(Y - E(Y))] = \boldsymbol{\Sigma}_{\boldsymbol{x}Y}.$$

Let the sample covariance matrices be

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

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

Suppose the positive semidefinite dispersion matrix $\boldsymbol{\Sigma}$ has eigenvalue eigenvector pairs $(\lambda_1, \boldsymbol{d}_1), ..., (\lambda_p, \boldsymbol{d}_p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. Let the eigenvalue eigenvector pairs of $\hat{\boldsymbol{\Sigma}}$ be $(\hat{\lambda}_1, \hat{\boldsymbol{d}}_1), ..., (\hat{\lambda}_p, \hat{\boldsymbol{d}}_p)$ where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$. These vectors are important quantities for principal component analysis (PCA).

Let the multiple linear regression model

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

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

Principal components regression (PCR), partial least squares (PLS), and several other dimension reduction models use p linear combinations $\boldsymbol{\gamma}_1^T\boldsymbol{x}, ..., \boldsymbol{\gamma}_p^T\boldsymbol{x}$. Estimating the $\boldsymbol{\gamma}_i$ and performing the ordinary least squares (OLS) regression of Y on $(\hat{\boldsymbol{\gamma}}_1^T\boldsymbol{x}, \hat{\boldsymbol{\gamma}}_2^T\boldsymbol{x}, ..., \hat{\boldsymbol{\gamma}}_k^T\boldsymbol{x})$ and a constant gives the k-component estimator, e.g. the k-component PLS estimator or the k-component PCR estimator, for $k = 1, ..., J$ where $J \leq p$ and the p-component estimator is the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS}$. Let $\boldsymbol{\gamma}_i(PCR) = \boldsymbol{d}_i$ and $\boldsymbol{\gamma}_i = \boldsymbol{\gamma}_i(PLS)$. The model selection estimator chooses one of the k-component estimators, e.g. using cross validation, and will be denoted by $\hat{\boldsymbol{\beta}}_{MSPLS}$ or $\hat{\boldsymbol{\beta}}_{MSPCR}$.

Let $\boldsymbol{X} = [\boldsymbol{1}\ \boldsymbol{X}_1]$. Chun and Keleş (2010) noted that one way to formulate PLS is to solve an optimization problem by forming $\boldsymbol{b}_j$ iteratively where

$$\boldsymbol{b}_k = \arg\max_{\boldsymbol{b}}\{[\mathrm{Cor}(\boldsymbol{Y}, \boldsymbol{X}_1\boldsymbol{b})]^2 V(\boldsymbol{X}_1\boldsymbol{b})\} \tag{2}$$

subject to $\boldsymbol{b}^T\boldsymbol{b} = 1$ and $\boldsymbol{b}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{b}_j = 0$ for $j = 1, ..., k-1$. So PLS is a model free way to get predictors $\hat{\boldsymbol{\gamma}}_i^T\boldsymbol{x}$ that are fairly highly correlated with the response, and the absolute correlations tend to decrease quickly.

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

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 = \sum_{i=1}^{p} (\hat{\theta}_i - \theta_i)^2 \propto p/n. \tag{3}$$

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

2 Model Selection Estimators in Low Dimensions

This section explains why "sensible model selection estimators, including variable selection estimators," produce fitted values (predictions) similar to that of the full OLS model when n is much larger than p. The result in Equation 4) that the residuals from the model selection model and the full OLS model are highly correlated was a property of OLS and Mallow's C_p criterion, not of any underlying model, but linearity forces the fitted values to be highly correlated. Hence the result works if OLS is consistent and the population model is linear, so for weighted least squares, AR(p) time series, serially correlated errors, et cetera. In particular, the cases do not need to be iid from some distribution. Since the correlation gets arbitrarily close to 1, the model selection estimator and full OLS estimator are estimating the same population parameter $\boldsymbol{\beta}$, but it is possible that the model selection estimator picks the full OLS model with probability going to one.

Consider the OLS regression of Y on a constant and $\boldsymbol{w} = (W_1, ..., W_p)^T$ where, for example, $W_j = x_j$, $W_j = \hat{\boldsymbol{\gamma}}_j^T \boldsymbol{x}$, or $W_j = \hat{\boldsymbol{d}}_j^T \boldsymbol{x}$. Let I index the variables in the model so $I = \{1, 2, 4\}$ means that $\boldsymbol{w}_I = (W_1, W_2, W_4)^T$ was selected. The full model $I = F$ uses all p predictors and the constant with $\boldsymbol{\beta}_I = \boldsymbol{\beta}_F = \boldsymbol{\beta} = \boldsymbol{\beta}_{OLS}$. Let r be the residuals from the full OLS model and let r_I be the residuals from model I that uses $\hat{\boldsymbol{\beta}}_I$. Suppose model I uses k predictors including a constant with $2 \leq k \leq p+1$. Olive and Hawkins (2005) proved that the model I with k predictors that minimizes Mallows (1973) $C_p(I)$ maximizes $\mathrm{cor}(r, r_I)$, that

$$\mathrm{cor}(r, r_I) = \sqrt{\frac{n - (p+1)}{C_p(I) + n - 2k}},$$

and under linearity, $\mathrm{cor}(r, r_I) \to 1$ forces

$$\mathrm{cor}(\hat{\alpha} + \boldsymbol{w}^{\mathrm{T}} \hat{\boldsymbol{\beta}}, \hat{\alpha}_{\mathrm{I}} + \boldsymbol{w}_{\mathrm{I}}^{\mathrm{T}} \hat{\boldsymbol{\beta}}_{\mathrm{I}}) = \mathrm{cor}(\mathrm{ESP}, \mathrm{ESP(I)}) = \mathrm{cor}(\hat{\mathrm{Y}}, \hat{\mathrm{Y}}_{\mathrm{I}}) \to 1.$$

Thus $C_p(I) \leq 2k$ implies that

$$\mathrm{cor}(\mathrm{r}, \mathrm{r_I}) \geq \sqrt{1 - \frac{\mathrm{p}+1}{\mathrm{n}}}. \tag{4}$$

Let the model I_{min} minimize the C_p criterion among the models considered with $C_p(I) \leq 2k_I$. Then $C_p(I_{min}) \leq C_p(F) = p+1$, and if PLS or PCR is selected using model selection

(on models $I_1, ..., I_p$ with $I_j = \{1, 2, ..., j\}$ corresponding to the j-component regression) with the $C_p(I)$ criterion, and $n \geq 20(p+1)$, then $\text{cor}(r, r_I) \geq \sqrt{19/20} = 0.974$. Hence the correlation of ESP(I) and ESP(F) will typically also be high. (For PCR, the following variant should work better: take $U_j = \hat{\boldsymbol{d}}_j^T \boldsymbol{x}$ and W_1 the U_j with the highest absolute correlation with Y, W_2 the U_j with the second highest absolute correlation, etc.)

Machine learning methods for the multiple linear regression model can be incorporated as follows. Let k be the number of predictors selected by lasso. Standardize the predictors to have unit sample variance, and run the method. Let model I contain the variables corresponding to the $k-1$ predictors variables that have the largest $|\hat{\beta}_i|$. Fit the OLS model I to these predictors and a constant. If $C_p(I) < min(2k, p+1)$ use model I, otherwise use the full OLS model. Many variants are possible. In low dimensions, comparisons between methods like lasso, PCR, PLS, and envelopes might use prediction intervals, the amount of dimension reduction, and standard errors if available.

If the above procedure is used, then model selection estimators, such as $\hat{\boldsymbol{\beta}}_{MSPLS}$, produce predictions that are similar to those of the OLS full model if $n \geq 20(p+1)$. Empirically, variable selection estimators and model selection estimators often do not select the full model. Equation 4) suggests that "weak" predictors will often be omitted, as long as $\text{cor}(r, r_I)$ stays high. (If the predictors are not orthogonal, 'weak" might mean the predictor is not very useful given that the other predictors are in the model.)

It is common in the model selection literature to assume, for the full model, that there is a model S such that $\beta_i \neq 0$ for $i \in S$, and $\beta_i = 0$ for $i \notin S$. Then model I underfits unless $S \subseteq I$. If $S \not\subseteq I$, then an "important" predictor has been left out of the model. Under the model $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S$, $\text{cor}(r, r_I)$ will not converge to 1 as $n \to \infty$, and for large enough n, $[\text{cor}(r, r_I)]^2 \leq \gamma < 1$. Thus $C_p(I) \to \infty$ as $n \to \infty$. Hence $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. Thus the probability that the model selection estimator underfits goes to zero as $n \to \infty$ if p is fixed, the full model is one of the models considered, and the C_p criterion is used, as noted by Rathnayake and Olive (2023).

For real data, an important question in variable selection is whether $\beta_i = 0$ is a reasonable assumption. If $\boldsymbol{X}$ has full rank $p+1$, then having β_i equal to zero for 20 decimal places may not be reasonable. See, for example, Tukey (1991), Nester (1996), and Gelman and Carlin (2017). Then the probability that the variable selection estimator chooses the full model goes to one if the probability of underfitting goes to 0 as $n \to \infty$. 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 $a \times 1$, 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 $k = 1, ..., J$ where there are J subsets, e.g. $J = 2^p - 1$.

3 Sparse Fitted Models

A fitted or population regression model is sparse if a of the predictors are active (have nonzero $\hat{\beta}_i$ or β_i) where $n \geq Ja$ with $J \geq 10$. Otherwise the model is nonsparse. A

high dimensional population full regression model is abundant or dense if the regression information is spread out among the p predictors (nearly all of the predictors are active). Hence an abundant model is a nonsparse model. Under the above definitions, most classical low dimensional models use sparse fitted models, and statisticians have over one hundred years of experience with such models.

The literature for high dimensional sparse regression models often assumes that i) $\boldsymbol{\beta}_{I,0} = \boldsymbol{\beta} = \boldsymbol{\beta}_F$, that ii) $S \subseteq I$ where I uses k predictors including a constant, and that iii) $n \geq 10k$. When these assumptions hold, the population model is sparse, the fitted model is sparse, and Equation 3) becomes $\|\hat{\boldsymbol{\beta}}_{I,0} - \boldsymbol{\beta}\|^2$, which can be small. Getting rid of assumption i) and the assumption that $S \subseteq I$ greatly increases the applicability of variable selection estimators, such as forward selection, lasso, and the elastic net, for high dimensional data, even if $\|\hat{\boldsymbol{\beta}}_{I,0} - \boldsymbol{\beta}\|^2$ is huge. As argued in the following paragraphs, the sparse fitted model often fits the data well, and often $\hat{\boldsymbol{\beta}}_I$ is a good estimator of $\boldsymbol{\beta}_I$.

A sparse fitted model transforms a high dimensional problem into a low dimensional problem, and the sparse fitted model can be checked with the goodness of fit diagnostics available for that low dimensional model. If the predictors used by the sparse fitted regression model are $\boldsymbol{x}_I$, and if the regression model depends on $\boldsymbol{x}_I$ only through the sufficient predictor $SP = \alpha_I + \boldsymbol{x}_I^T \boldsymbol{\beta}_I$, then a useful diagnostic is the response plot of $ESP(I) = \hat{\alpha}_I + \boldsymbol{x}_I^T \hat{\boldsymbol{\beta}}_I$ versus the response Y on the vertical axis. If there is goodness of fit, then $\hat{\boldsymbol{\beta}}_I$ tends to estimate $\boldsymbol{\beta}_I$ regardless of whether the population model is sparse or nonsparse. Data splitting may be needed for valid inference such as hypothesis testing.

Suppose the cases $(\boldsymbol{x}_i^T, Y_i)^T$ are iid for $i = 1, ..., n$. Then $Y_1, ..., Y_n$ are iid, resulting in a valid sparse fitted model regardless of whether the population model is sparse or nonsparse. This *null model* omits all of the predictors. For high dimensional data, a reasonable goal is to find a model that greatly outperforms the null model.

The sparse fitted model using $(Y, \boldsymbol{x}_I)$ is often useful when there are one or more strong predictors. The following Olive and Zhang (2025) theorem gives two more situations where a sparse fitted model can greatly outperform the null model. The population models in Theorem 1 can be sparse or nonsparse. The high dimensional multiple linear regression literature often assumes that the cases are iid from a multivariate normal distribution, and that the population model is sparse. Let $\Sigma_Y = \sigma_Y^2$. For multiple linear regression, note that $\sigma_O^2 < \sigma_Y^2 = \Sigma_Y$ unless $\boldsymbol{\eta}^T \boldsymbol{\Sigma}_{\boldsymbol{x}Y} = 0$.

Theorem 1 *Suppose the cases $(Y_i, \boldsymbol{x}_i^T)^T$ are iid from some distribution.*
a) If the joint distribution of $(Y, \boldsymbol{x}^T)^T$ is multivariate normal,

$$\begin{pmatrix} Y \\ \boldsymbol{x} \end{pmatrix} \sim N_{p+1} \left(\begin{pmatrix} \mu_Y \\ \boldsymbol{\mu}_{\boldsymbol{x}} \end{pmatrix}, \begin{pmatrix} \Sigma_Y & \boldsymbol{\Sigma}_{Y\boldsymbol{x}} \\ \boldsymbol{\Sigma}_{\boldsymbol{x}Y} & \boldsymbol{\Sigma}_{\boldsymbol{x}} \end{pmatrix} \right),$$

then $Y|\boldsymbol{x} \sim Y|(\alpha_{OLS} + \boldsymbol{\beta}_{OLS}^T \boldsymbol{x}) \sim N(\alpha_{OLS} + \boldsymbol{\beta}_{OLS}^T \boldsymbol{x}, \sigma^2)$ follows a multiple linear regression model, but so does $Y|\boldsymbol{\eta}^T \boldsymbol{x} \sim N(\alpha_O + \boldsymbol{\beta}_O^T \boldsymbol{x}, \sigma_O^2)$ where $\alpha_O = \mu_Y - \boldsymbol{\beta}_O^T \boldsymbol{\mu}_{\boldsymbol{x}}$, $\boldsymbol{\beta}_O = \lambda \boldsymbol{\eta}$, $\sigma_O^2 = \Sigma_Y - \boldsymbol{\beta}_O^T \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$, and

$$\lambda = \frac{\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T \boldsymbol{\eta}}{\boldsymbol{\eta}^T \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\eta}}.$$

b) If the response Y is binary, then $Y|(\alpha_O + \boldsymbol{\beta}_O^T \boldsymbol{x}) \sim$ binomial$(m = 1, \rho(\alpha_O + \boldsymbol{\beta}_O^T \boldsymbol{x}))$ where $E[Y|(\alpha_O + \boldsymbol{\beta}_O^T \boldsymbol{x})] = \rho(\alpha_O + \boldsymbol{\beta}_O^T \boldsymbol{x}) = P[Y = 1|(\alpha_O + \boldsymbol{\beta}_O^T \boldsymbol{x})]$. Hence every linear

combination of the predictors satisfies a binary regression model.

4 PCA-PLS

Another technique is to use PCA for dimension reduction. Let $U_1, ..., U_p$ be the PCA linear combinations ($U_i = \hat{\boldsymbol{\gamma}}_i^T \boldsymbol{x}$) ordered with respect to the largest eigenvalues. Then use $U_1, ..., U_k$ in the regression or classification model where k is chosen in some manner. This method can be used for models with m response variables $Y_1, ..., Y_m$.

Consider a low or high dimensional regression or classification method with a univariate response variable Y. Let $W_1, ..., W_p$ be the linear combinations ordered with respect to the highest squared correlations $r_1^2, ..., r_p^2$ where the sample correlation $r_i = \text{cor}(x_i, Y)$. From a model selection viewpoint, using $W_1, ..., W_k$ should work much better than using $U_1, ..., U_k$. Also, the PLS components W_i should be used instead of the PCA W_i, since the PLS components are chosen to be fairly highly correlated with Y. See Equation 2). Brown (1993, pp. 71-72) shows that an equivalent way to compute the k-component PLS estimator is to maximize $\hat{\boldsymbol{\gamma}}^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$ under some constraints. If the predictors are standardized to have unit sample variance, then this method becomes a correlation vector optimization problem.

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

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

This optimization problem is equivalent to maximizing

$$\boldsymbol{\Sigma}_Y M^2 = \max_{\boldsymbol{\gamma} \neq \boldsymbol{0}} \frac{\boldsymbol{\gamma}^T \boldsymbol{\Sigma}_{\boldsymbol{x}Y} \boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T \boldsymbol{\gamma}}{\boldsymbol{\gamma}^T \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\gamma}}$$

which has a maximum at $\boldsymbol{\gamma} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \boldsymbol{\beta}_{OLS}$. See Mardia, Kent, and Bibby (1979, pp. 168, 282). Hence PLS is a lot like CCA but with more constraints, and PLS can be computed in high dimensions. From the dimension reduction literature, if Y depends on $\boldsymbol{x}$ only through $\alpha + \boldsymbol{\beta}^T \boldsymbol{x}$, then under the assumption of "linearly related predictors" $\hat{\boldsymbol{\beta}}_{OLS}$ estimates $\boldsymbol{\beta}_{OLS} = c\boldsymbol{\beta}$ for some constant c which is often nonzero. See, for example Cook and Weisberg (1999, p. 432).

The above results suggest computing the lasso for multiple linear regression, find the number of predictors k chosen by lasso, and take k linear combinations. An SC scree plot of i versus r_i^2 behaves like a scree plot of i versus the eigenvalues. Hence quantities like $\sum_{i=1}^j r_i^2 / \sum_{i=1}^p r_i^2$ are of interest for $j = 1, ..., p$, and scree plot techniques could be adapted to choose k. Many other possibilities exist, and there are many possibilities for models with m response variables $Y_1, ..., Y_m$.

Another useful technique is to eliminate weak predictors before finding $W_1, ..., W_k$. By Equation 3), $\hat{\boldsymbol{\gamma}}_i$ may not be close to $\boldsymbol{\gamma}_i$ in high dimensions, e.g. $p = n^6$. For example, the sample eigenvectors $\hat{\boldsymbol{d}}_i$ tend to be poor estimators of the population eigenvectors $\boldsymbol{d}_i$ of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$. An exception is when the correlation $\text{Cor}(x_i, x_j) = \rho$ for $i \neq j$ where ρ is close to 1. See Jung and Marron (2009). One possibility is to take the j predictors x_i with the

highest squared correlations with Y. The SC scree plot is useful. Then do lasso (meant for the multiple linear regression model) to further reduce the number of x_i. Here j should be proportional to n, for example $j = \min(Kn, p)$, where $K = 1$ is an interesting choice.

5 Stack Low Dimensional Estimators into a Vector

Another technique is to stack low dimensional estimators into a vector. The MMLE, one component PLS estimator, $\overline{\boldsymbol{x}_2} - \overline{\boldsymbol{x}_1}$, and elements from an estimated covariance matrix such as $\boldsymbol{c} = vech(\boldsymbol{\Sigma}_{\boldsymbol{z}})$. Using $\boldsymbol{z} = (Y_1, ..., Y_m, x_1, ..., x_p)^T$ can give information about a multivariate regression. Then tests for low dimensional quantities such as $Cov(x_i, Y)$ or $Cov(x_i, x_i) = Var(x_i)$ can be done for $i = 1, ..., p$. Theory for several of these estimators appears in Olive et al. (2025).

6 Alternative Dispersion Estimators

Let $\hat{\boldsymbol{\Sigma}}$ be a $p \times p$ symmetric positive semidefinite matrix such as $\boldsymbol{R}, \boldsymbol{R}^{-1}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}, \boldsymbol{X}^T\boldsymbol{X}$ or $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. When $\hat{\boldsymbol{\Sigma}}$ is singular or ill conditioned, some common techniques are to replace $\hat{\boldsymbol{\Sigma}}$ with a symmetric positive definite matrix $\hat{\boldsymbol{D}}$ such as $\hat{\boldsymbol{D}} = diag(\hat{\boldsymbol{\Sigma}}), \boldsymbol{D} = (\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I}_p)$ where the constant $\lambda > 0$, or $\hat{\boldsymbol{D}} = \boldsymbol{D} = \boldsymbol{I}_p$. Regularized estimators are also used.

For $\delta \geq 0$, a simple way to regularize a $p \times p$ correlation matrix $\boldsymbol{R} = (r_{ij})$ is to use

$$\boldsymbol{R}_\delta = \frac{1}{1+\delta}(\boldsymbol{R} + \delta \boldsymbol{I}_p) = (t_{ij}) \tag{5}$$

where $t_{ii} = 1$ and

$$t_{ij} = \frac{r_{ij}}{1+\delta}$$

for $i \neq j$. Note that each correlation r_{ij} is divided by the same factor $1 + \delta$. If λ_i is the ith eigenvalue of $\boldsymbol{R}$, then $(\lambda_i + \delta)/(1 + \delta)$ is the ith eigenvalue of $\boldsymbol{R}_\delta$. The eigenvectors of $\boldsymbol{R}$ and $\boldsymbol{R}_\delta$ are the same since if $\boldsymbol{R}\ \boldsymbol{x} = \lambda_i\ \boldsymbol{x}$, then

$$\boldsymbol{R}_\delta\ \boldsymbol{x} = \frac{1}{1+\delta}(\boldsymbol{R} + \delta \boldsymbol{I}_p)\ \boldsymbol{x} = \frac{1}{1+\delta}(\lambda_i + \delta)\ \boldsymbol{x}.$$

Note that $\boldsymbol{R}_\delta = \kappa \boldsymbol{R} + (1-\kappa)\boldsymbol{I}_p$ where $\kappa = 1/(1+\delta) \in (0, 1]$. See Ledoit and Wolf (2004) and Warton (2008).

Following Datta (1995, pp. 250-254), the condition number of a symmetric positive definite $p \times p$ matrix $\boldsymbol{A}$ is $cond(\boldsymbol{A}) = \lambda_1(\boldsymbol{A})/\lambda_p(\boldsymbol{A})$ where $\lambda_1(\boldsymbol{A}) \geq \lambda_2(\boldsymbol{A}) \geq \cdots \geq \lambda_p(\boldsymbol{A}) > 0$ are the eigenvalues of $\boldsymbol{A}$. Note that $cond(\boldsymbol{A}) \geq 1$. A well conditioned matrix has condition number $cond(\boldsymbol{A}) \leq c$ for some number c such as 50 or 500. Hence $\boldsymbol{R}_\delta$ is nonsingular for $\delta > 0$ and well conditioned if

$$cond(\boldsymbol{R}_\delta) = \frac{\lambda_1 + \delta}{\lambda_p + \delta} \leq c,$$

or

$$\delta = \max\left(0, \frac{\lambda_1 - c\lambda_p}{c-1}\right) \tag{6}$$

if $1 < c \leq 500$. Taking $c = 50$ suggests using

$$\delta = \max\left(0, \frac{\lambda_1 - 50\lambda_p}{49}\right).$$

The matrix can be further regularized by setting $t_{ij} = 0$ if $|t_{ij}| \leq \tau$ where $\tau \in [0, 1)$ should be less than 0.5. Denote the resulting matrix by $\boldsymbol{R}(\delta, \tau)$. We suggest using $\tau = 0.05$. Note that $\boldsymbol{R}_\delta = \boldsymbol{R}(\delta, 0)$. Using τ is known as *thresholding*. We recommend computing $\boldsymbol{I}_p, \boldsymbol{R}(\delta, 0)$ and $\boldsymbol{R}(\delta, 0.05)$ for $c = 50$, 100, 200, 300, 400, and 500. Compute $\boldsymbol{R}$ if it is nonsingular. Note that a regularized covariance matrix can be found using

$$\boldsymbol{S}(\delta, \tau) = \boldsymbol{D}_S \ \ \boldsymbol{R}(\delta, \tau) \ \ \boldsymbol{D}_S \tag{7}$$

where $\boldsymbol{S} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}$ and $\boldsymbol{D}_S = \text{diag}(\sqrt{\text{S}_{11}}, ..., \sqrt{\text{S}_{\text{pp}}})$.

A common type of regularization of a covariance matrix $\boldsymbol{S}$ is to use $\boldsymbol{S}_D = diag(\boldsymbol{S})$ where the ijth element of $\boldsymbol{S}_D = 0$ and $\boldsymbol{S}_D(i,i) = \boldsymbol{S}(i,i)$. The corresponding correlation matrix is the identity matrix, and Mahalanobis distances using the identity matrix correspond to Euclidean distances. These estimators tend to use too much regularization, and underfit. Note that as $\delta \to \infty$, $\boldsymbol{R}_\delta \to \boldsymbol{I}_p$, and $\boldsymbol{I}_p$ corresponds to $c = 1$. Note that $\boldsymbol{S}_D$ corresponds to using $\boldsymbol{R}(\delta = \infty, 0) = \boldsymbol{I}_p$ in Equation (7).

For the population correlation matrix $\boldsymbol{\rho}_{\boldsymbol{x}}$ and the population precision matrix $\boldsymbol{\rho}_{\boldsymbol{x}}^{-1}$, the literature often claims that most of the population correlations $\rho_{ij} = 0$, so that the population matrix is sparse, and that $\hat{\boldsymbol{D}}$ is a good estimator of the population matrix. Assume that $\hat{\boldsymbol{D}}$ estimates a population dispersion matrix $\boldsymbol{D}$. Note that this assumption always holds when $\hat{\boldsymbol{D}} = \boldsymbol{I}_p = \boldsymbol{D}$. Note that $diag(\boldsymbol{S})$ estimates $diag(\boldsymbol{\Sigma}_{\boldsymbol{x}})$ since $(\hat{\sigma}_1^2, ..., \hat{\sigma}_p^2)^T$ estimates $(\sigma_1^2, ..., \sigma_p^2)^T$ where $\sigma_i^2 = V(x_i)$ for $i = 1, ..., p$. However, by Equation 3), the estimator tends not to be good in high dimensions.

Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ where a $g \times 1$ statistic T_n satisfies $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u} \sim N_g(\boldsymbol{0}, \boldsymbol{\Sigma})$. If $\hat{\boldsymbol{\Sigma}}^{-1} \xrightarrow{P} \boldsymbol{\Sigma}^{-1}$ and H_0 is true, then

$$D_n^2 = D_{\boldsymbol{\theta}_0}^2(T_n, \hat{\boldsymbol{\Sigma}}/n) = n(T_n - \boldsymbol{\theta}_0)^T \hat{\boldsymbol{\Sigma}}^{-1}(T_n - \boldsymbol{\theta}_0) \xrightarrow{D} \boldsymbol{u}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{u} \sim \chi_g^2$$

as $n \to \infty$. Then a Wald type test rejects H_0 if $D_n^2 > \chi_{g,1-\delta}^2$ where $P(X \leq \chi_{g,1-\delta}^2) = 1 - \delta$ if $X \sim \chi_g^2$, a chi-square distribution with g degrees of freedom. Note that $D_{\boldsymbol{\theta}_0}^2(T_n, \hat{\boldsymbol{\Sigma}}/n)$ is a squared Mahalanobis distance.

It is common to implement a Wald type test using

$$D_n^2 = D_{\boldsymbol{\theta}_0}^2(T_n, \boldsymbol{C}_n/n) = n(T_n - \boldsymbol{\theta}_0)^T \boldsymbol{C}_n^{-1}(T_n - \boldsymbol{\theta}_0) \xrightarrow{D} \boldsymbol{u}^T \boldsymbol{C}^{-1} \boldsymbol{u}$$

as $n \to \infty$ if H_0 is true, where the $g \times g$ symmetric positive definite matrix $\hat{\boldsymbol{D}} = \boldsymbol{C}_n \xrightarrow{P} \boldsymbol{C} \neq \boldsymbol{\Sigma}$. Hence $\boldsymbol{C}_n$ is the wrong dispersion matrix, and $\boldsymbol{u}^T \boldsymbol{C}^{-1} \boldsymbol{u}$ does not have a χ_g^2 distribution when H_0 is true. Rajapaksha and Olive (2024) showed how to bootstrap

Wald tests with the wrong dispersion matrix. When $\boldsymbol{C}_n = \boldsymbol{I}_g$, the bootstrap tests often became conservative as g increased to n. For some methods, better high dimensional tests are reviewed by Hu and Bai (2015). For some of these tests, the m out of n bootstrap, which draws a sample of size m without replacement from the n, works better than the nonparametric bootstrap. Sampling without replacement is also known as subsampling and the delete d jackknife. See Abid and Olive (2025).

Using a high dimensional dispersion estimator with considerable outlier resistance is another useful technique. Let $\boldsymbol{W}$ be a data matrix, where the rows $\boldsymbol{w}_i$ correspond to cases. For example, $\boldsymbol{w}_i = \boldsymbol{x}_i$ or $\boldsymbol{w}_i = \boldsymbol{z}_i = (Y_{i1}, ..., Y_{im}, x_{i1}, ..., x_{ip})^T$. One of the simplest outlier detection methods uses the Euclidean distances of the $\boldsymbol{x}_i$ from the coordinatewise median $D_i = D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$. Concentration type steps compute the weighted median MED_j: the coordinatewise median computed from the "half set" of cases $\boldsymbol{x}_i$ with $D_i^2 \leq \text{MED}(D_i^2(\text{MED}_{j-1}, \boldsymbol{I}_p))$ where $\text{MED}_0 = \text{MED}(\boldsymbol{W})$. We often used $j = 0$ (no concentration type steps) or $j = 9$. Let $D_i = D_i(\text{MED}_j, \boldsymbol{I}_p)$. Let $W_i = 1$ if $D_i \leq \text{MED}(D_1, ..., D_n) + k\text{MAD}(D_1, ..., D_n)$ 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 $\text{MAD}(n) = \text{MED}(|D_i - \text{MED}(n)|,\ i = 1, \ldots, n)$ where $\text{MED}(n) = \text{MED}(D_1, ..., D_n)$ is the sample median of $D_1, ..., D_n$.

Let the *covmb2 set* B of at least $n/2$ cases correspond to the cases with weight $W_i = 1$. Then the Olive (2017, p. 120) *covmb2* estimator $(T, \boldsymbol{C})$ is the sample mean and sample covariance matrix applied to the cases in set B. Hence

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

This estimator was built for speed, applications, and outlier resistance. In low dimensions, the population dispersion matrix is the population covariance matrix of a spherically truncated distribution. In high dimensions, spherical truncation is still used, but the sample weighted median varies about the population weighted median by Equation 3).

A useful application is to apply high (and low) dimensional methods to the cases that get weight 1. If the ith case $\boldsymbol{w}_i = (\boldsymbol{y}_i^T, \boldsymbol{x}_i^T)^T$ where $\boldsymbol{y} = (Y_1, ..., Y_m)^T$, then this application can be used if all of the variables are continuous. For a variant, let the continuous predictors from $\boldsymbol{x}_i$ be denoted by $\boldsymbol{u}_i$ for $i = 1, ..., n$. Apply the `covmb2` estimator to the $\boldsymbol{u}_i$, and then run the method on the m cases $\boldsymbol{w}_i$ corresponding to the `covmb2` set B indices $i_1, ..., i_m$, where $m \geq n/2$. If the estimator has large sample theory "conditional" on the predictors $\boldsymbol{x}$, then typically the same theory applies for the "robust estimator" since the response variables were not used to select the cases in B. These two applications can be used for regression, classification, neural networks, et cetera.

7 Conclusions

The `covmb2` estimator attempts to give a robust dispersion estimator that reduces the bias by using a big ball about MED_j instead of a ball that contains half of the cases. The weighting is the default method, but you can also plot the squared Euclidean distances and estimate the number $m \geq n/2$ of cases with the smallest distances to be used. The *median ball* is the hypersphere centered at the coordinatewise median with radius $r = \text{MED}(D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p),\ i = 1, ..., n)$ that tends to contain $(n+1)/2$ of the cases if n is odd. The *slpack* function `medout` makes the plot, and the *slpack* function `getB` gives the set B of cases that got weight 1 along with the index `indx` of the case numbers that got weight 1.

The function `ddplot5` plots the Euclidean distances from the coordinatewise median versus the Euclidean distances from the `covmb2` location estimator. Typically the plotted points in this DD plot cluster about the identity line, and outliers appear in the upper right corner of the plot with a gap between the bulk of the data and the outliers.

8 References

Abid, A.M. and Olive, D.J. (2025), "Some Simple High Dimensional One and Two Sample Tests," is at (http://parker.ad.siu.edu/Olive/pphd1samp.pdf).

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

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

Cook, R.D., and Forzani, L. (2018), "Big Data and Partial Least Squares Prediction," *The Canadian Journal of Statistics*, 46, 62-78.

Cook, R.D., and Forzani, L. (2019), "Partial Least Squares Prediction in High-Dimensional Regression," *The Annals of Statistics,* 47, 884-908.

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.

Datta, B.N. (1995), *Numerical Linear Algebra and Applications,* Brooks/Cole Publishing Company, Pacific Grove, CA.

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

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

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer, New York, NY.

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

Hu, J., and Bai, Z. (2015), "A Review of 20 Years of Naive Tests of Significance for High-Dimensional Mean Vectors and Covariance Matrices," *Science China Mathematics*, 55, online.

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.

Jung, S. and Marron, J.S. (2012), "PCA Consistency in High Dimension Low Sample Size Context," *The Annals of Statistics*, 37, 4104-4130.

Ledoit, O., and Wolf, M. (2004), "A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices," *Journal of Multivariate Analysis*, 88, 365-411.

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.

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).

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

Olive, D.J. (2017), *Robust Multivariate Analysis*, Springer, New York, NY.

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

Olive, D.J., Alshammari, A.A., Pathiranage, K.G., and Hettige, L.A.W. (2025), "Testing with the One Component Partial Least Squares and the Marginal Maximum Likelihood Estimators." See (http://parker.ad.siu.edu/Olive/pphdwls.pdf).

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.

Rajapaksha, K.W.G.D.H., and Olive, D.J. (2022), "Wald Type Tests with the Wrong Dispersion Matrix," *Communications in Statistics: Theory and Methods*, 53, 2236-2251.

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.

Seber, G.A.F., and Lee, A.J. (2003), *Linear Regression Analysis,* 2nd ed., Wiley, New York, NY.

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.

Warton, D.I. (2008), "Penalized Normal Likelihood and Ridge Regularization of Correlation and Covariance Matrices," *Journal of the American Statistical Association*, 103, 340-349.