Dimension Reduction in Regression: Comparative Analysis of Principal Component Regression and Partial Least Squares

Khabibullo Ibadullaev, Bekhzod Abulov

February 19, 2025

Winter Semester 2024/2025

Contents

1 Introduction

Regression models are a fundamental tool in statistical modeling, enabling the prediction of a dependent variable based on a set of independent variables. However, in modern applications, datasets often exhibit high-dimensionality, where the number of predictors (p) is significantly larger than the number of observations (n). This scenario, commonly referred to as the curse of dimensionality, poses significant challenges to traditional regression methods, particularly Ordinary Least Squares (OLS) regression.

One of the primary concerns in high-dimensional regression is multicollinearity—a situation where predictor variables are highly correlated. Multicollinearity leads to inflated standard errors, unstable coefficient estimates, and difficulties in interpreting model parameters. Additionally, high-dimensional models tend to overfit the training data, resulting in poor generalization to unseen observations. Computational complexity further exacerbates these challenges, making it difficult to estimate model parameters efficiently.

To address these issues, researchers have developed dimension reduction techniques that aim to transform high-dimensional predictor spaces into lower-dimensional representations while preserving essential information. Two widely used techniques in this context are Principal Component Regression (PCR) and Partial Least Squares (PLS). These methods construct new predictor variables, referred to as components, that capture the most relevant variation in the dataset. The use of PCR and PLS is prevalent in various disciplines, including chemometrics, finance, genomics, and environmental science, where datasets often contain a large number of correlated predictors. While both methods offer solutions to the high-dimensional regression problem, they differ in their approach, underlying assumptions, and effectiveness in different scenarios.

This paper provides a comprehensive analysis of Principal Component Regression (PCR) and Partial Least Squares (PLS), focusing on their theoretical foundations, practical implementations, and empirical performance evaluation across various regression settings. Specifically, in Section 2 we investigate the mathematical principles underlying these methods, highlighting their similarities and differences in handling multicollinearity and dimensionality reduction. A key focus in Section 3 is on understanding the bias-variance trade-off by performing a detailed Mean Squared Error (MSE) decomposition of regression coefficients, as well as evaluating predictive performance through Cross-Validation (CV) MSE. The study systematically compares these methods in different scenarios. To further assess model stability and generalization ability, a Monte Carlo simulation is conducted using synthetic datasets, allowing for an in-depth examination of performance under varying levels of correlation. Additionally, in Section 4 both PCR and PLS are applied to real-world regression problems, including Near-Infrared (NIR) Spectroscopy for Moisture Prediction, to evaluate their practical utility and interpretability. Finally, in Section 5 we discuss the strengths and weaknesses of each method, providing guidance on when to prefer PCR over PLS (or vice versa) and exploring potential hybrid approaches that integrate the advantages of both techniques.

2 Theory

Traditional regression methods, such as OLS, assume that the predictor matrix X has full rank, ensuring a unique solution for the estimated coefficients. However, in many practical applications, datasets exhibit high collinearity among predictors or even contain more predictors than observations ($p \gg n$), making $X^T X$ nearly singular or non-invertible. This leads to large variances in coefficient estimates, making the regression model highly unstable. The instability of OLS is evident in the variance expression:

$$\text{Var}(\hat{\beta}_{\text{OLS}}) = \sigma^2 (X^T X)^{-1} \tag{1}$$

Applying the eigen-decomposition:

$$X^T X = Q \Lambda Q^T \tag{2}$$

we obtain:

$$\text{Var}(\hat{\beta}_{\text{OLS}}) = \sigma^2 Q \Lambda^{-1} Q^T \tag{3}$$

where Λ contains the eigenvalues of $X^T X$, we observe that when multicollinearity is present, some eigenvalues approach zero, making Λ^{-1} contain large values, which significantly inflates the variance of $\hat{\beta}_{\text{OLS}}$.

PCR and PLS overcome this issue by constructing a set of orthogonal predictors, thereby improving numerical stability and reducing overfitting. While, PCR mitigates this issue by removing components corresponding to small eigenvalues, ensuring that the variance of the estimated coefficients remains controlled, PLS focuses on components that maximize the covariance with Y, eliminating high-variance directions that contribute to unstable coefficient estimates, ensuring a more robust regression model.

2.1 Principal Component Regression (PCR)

Principal Component Regression (PCR) is a widely used statistical technique that integrates Principal Component Analysis (PCA) with Ordinary Least Squares (OLS) regression to address issues of high-dimensionality and multicollinearity in regression problems. The fundamental idea behind PCR is to transform the original predictor variables, which may be highly correlated, into a new set of uncorrelated variables called principal components (Figure 1). Regression is then performed using a selected subset of these principal components, effectively reducing the dimensionality of the predictor space while mitigating instability in coefficient estimates.

PCR is an effective tool for dealing with multicollinearity, it is important to note that it is an unsupervised method, meaning that the principal components are chosen solely based on variance in the predictor space, without considering the response variable Y. This characteristic distinguishes it from methods like Partial Least Squares (PLS), which optimize component selection based on predictive relevance to Y.

Mathematical Formulation

The key steps of PCR are as follows:

1. Perform **Principal Component Analysis (PCA)** on X, which involves computing the **Singular Value Decomposition (SVD)**:

$$X = U\Sigma V^T \tag{4}$$

where:

- U is an $n \times n$ orthogonal matrix (left singular vectors),
- Σ is an $n \times p$ diagonal matrix with singular values,
- V is a $p \times p$ orthogonal matrix (right singular vectors), whose columns are the principal component directions.

2. Transform the original predictors into principal components:

$$Z = XV \tag{5}$$

where Z represents the uncorrelated principal components of X.

3. Perform OLS regression on the transformed dataset:

$$Y = Z\gamma + \epsilon \tag{6}$$

where γ is the transformed coefficient vector in the principal component space.

Unsupervised Nature of PCR

PCR is fundamentally an **unsupervised** technique because it selects principal components based on variance in X without considering Y. This is evident from the PCA decomposition:

$$X^T X = V\Lambda V^T \tag{7}$$

Since V is computed purely from X, it does not incorporate Y. As a result, PCR may discard components that contain important predictive information for Y, making it less efficient for regression tasks compared to supervised alternatives like PLS.
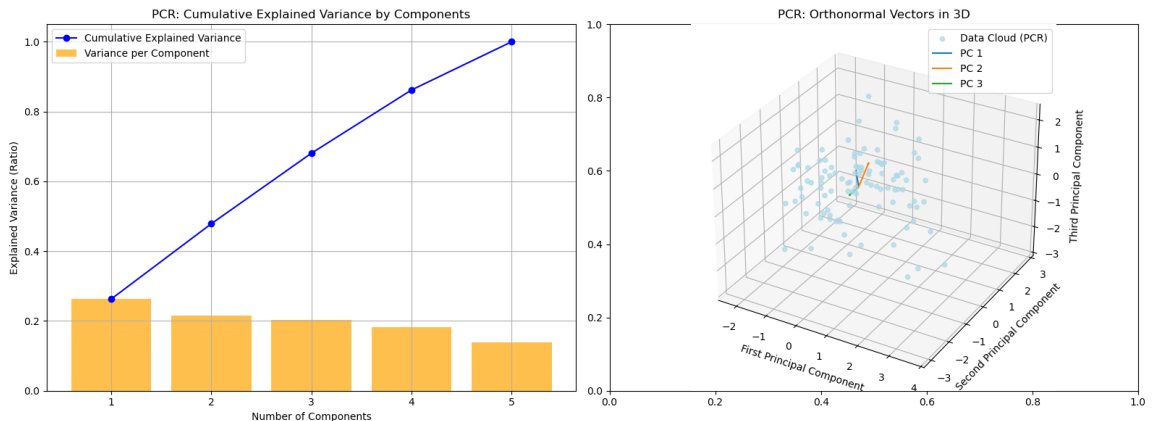

Figure 1: Principal Component Regression (PCR) Analysis Results

Variance Reduction and Bias in PCR

A key feature of PCR is its ability to reduce variance in estimated coefficients by using only first k principal components and eliminating components associated with small eigenvalues:

$$Z_k = XV_k \tag{8}$$

Estimating regression coefficients in the principal component space:

$$\hat{\gamma}_k = (Z_k^T Z_k)^{-1} Z_k^T Y \tag{9}$$

Using the eigen decomposition (7), we obtain:

$$\hat{\gamma}_k = \Lambda_k^{-1} V_k^T X^T Y \tag{10}$$

where Λ_k contains only the top k eigenvalues of $X^T X$, while V_k consists of the first k principal components that capture the most variance in X. The selection of k is often done using cross-validation or based on the cumulative explained variance criterion (Figure 1).

Transforming back to the original space:

$$\hat{\beta}_{\mathrm{PCR}} = V_k \hat{\gamma}_k = V_k \Lambda_k^{-1} V_k^T X^T Y \tag{11}$$

Recall that OLS can be written in terms of PCA decomposition (7). If all principal components are used ($k = p$), then we have:

$$\hat{\beta}_{\mathrm{PCR}} = \hat{\beta}_{\mathrm{OLS}} = V \Lambda^{-1} V^T X^T Y \tag{12}$$

Which confirms that, by discarding small eigenvalues, PCR inherently reduces variance, albeit at the cost of introducing some bias. This trade-off is similar to ridge regression and is crucial in high-dimensional settings. Even though we are selecting a subset of principal components, Eckart-Young theorem[1] guarantees that using the first k principal components in PCR is the best way to approximate and reduce dimensionality while preserving the most important information.

Principal Component Regression provides a robust approach for handling multicollinearity and high-dimensional data. However, its unsupervised nature may lead to suboptimal predictive performance, as it does not consider the response variable in component selection. In the next section, we introduce Partial Least Squares (PLS), which addresses this limitation by incorporating response information into the component selection process.

2.2 Partial Least Squares (PLS)

Partial Least Squares (PLS) is a regression technique specifically designed to handle high-dimensional datasets where multicollinearity and overfitting pose challenges to traditional regression methods. Developed by Herman Wold in the 1960s, PLS integrates elements of dimensionality reduction with regression, making it an effective tool for cases where the number of predictors (p) is large relative to the number of observations (n), or where predictors are highly correlated.

---

[1]Definition and proof are provided in Appendix 6.6

PLS constructs latent variables (LVs) that maximize the covariance between the predictor (X) and response (Y) matrices (Figure 2). Unlike Principal Component Regression (PCR), which selects components based on variance in X without considering Y, PLS selects components that are most relevant for predicting Y. This **supervised** approach makes PLS more effective than PCR when the primary goal is to improve prediction accuracy rather than just reducing dimensionality.

The advantages of PLS make it widely applicable in fields such as chemometrics, finance, genomics, and spectroscopy, where data often exhibit strong collinearities and high dimensionality. By incorporating response information directly into the component selection process, PLS ensures that selected components remain relevant for prediction, reducing the risk of omitting important information.
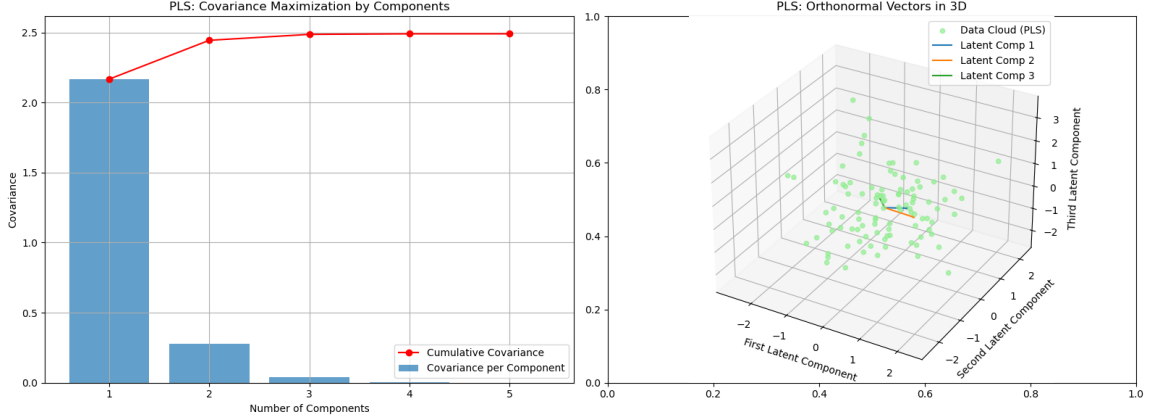

Figure 2: Partial Least Squares (PLS) Analysis Results

Mathematical Formulation

PLS constructs **latent variables (LVs)** as linear combinations of the original predictors:

$$T = XW, \quad U = YQ \tag{13}$$

where:

- T and U are the latent variables (score matrices) for X and Y, respectively,

- W and Q are the weight (loading) matrices that transform X and Y into the latent space.

PLS selects components by maximizing the covariance between XW and YQ, ensuring that the extracted components remain highly predictive of the response:

$$W = \arg \max_{W} \text{cov}(XW, Y) \tag{14}$$

The final regression model is then expressed in terms of these components:

$$Y = T\delta + \epsilon \tag{15}$$

where δ represents the regression coefficients for the latent variables.

Bias-Variance Tradeoff in PLS

Like PCR, PLS reduces variance by using only the first k components. However, PLS retains predictive power better than PCR because it selects components based on their relevance to Y.

Extracting first k PLS components:

$$T_k = XW_k \tag{16}$$

where W_k contains the first k PLS weight vectors that capture the most covariance between X and Y (Figure 2).

Estimating regression coefficients in the latent space:

$$\hat{\delta}_k = (T_k^T T_k)^{-1} T_k^T Y \tag{17}$$

The final regression coefficient in the original space is:

$$\hat{\beta}_{\text{PCR}} = W_k \hat{\delta}_k \tag{18}$$

After substitution we have:

$$\hat{\beta}_{\text{PCR}} = W_k (W_k^T X^T X W_k)^{-1} W_k^T X^T Y \tag{19}$$

In the case when all p components are used, i.e. $k = p$, we can simplify $\hat{\beta}_{\text{PCR}}$ and we have:

$$\hat{\beta}_{\text{PLS}} = (X^T X)^{-1} X^T Y = \hat{\beta}_{\text{OLS}} \tag{20}$$

where fewer components lead to lower variance but higher bias. This mechanism prevents overfitting, particularly in high-dimensional settings.

Partial Least Squares Regression offers a robust alternative to standard regression techniques by integrating dimensionality reduction with predictive modeling. Compared to PCR, PLS is a supervised approach, ensuring that the extracted components remain relevant for prediction. This makes PLS an effective tool for applications where both dimensionality reduction and accurate predictions are required.

It is important to clarify that Principal Component Regression (PCR) and Partial Least Squares (PLS) are not feature selection techniques in the traditional sense. Unlike methods such as LASSO or stepwise regression, which explicitly remove irrelevant predictors by setting some regression coefficients to zero, both PCR and PLS transform the original predictor space into a new set of latent variables.

Even though PCR and PLS reduce the dimensionality of the predictor space, they do so by forming linear combinations of all original predictors. This means that each principal component or latent variable still contains information from every predictor in the dataset. Thus, PCR and PLS do not eliminate variables but instead project them into a lower-dimensional space where multicollinearity is mitigated, and variance is better controlled.

NIPALS Algorithm

Several algorithms exist for computing PLS regression. One of the common, The **Nonlinear Iterative Partial Least Squares (NIPALS)** algorithm is an iterative approach used to extract PLS components

when the number of predictors is very large.

1. Initialize $u = Y[:, 1]$, the first column of Y.

2. Compute the weight vector:

$$w = \frac{X^T u}{\|X^T u\|} \tag{21}$$

3. Compute the latent score vector:

$$t = Xw \tag{22}$$

4. Compute the loading vector:

$$p = \frac{X^T t}{t^T t} \tag{23}$$

5. Compute the regression coefficient:

$$b = \frac{u^T t}{t^T t} \tag{24}$$

6. Deflate X and Y:

$$X = X - tp^T, \quad Y = Y - btq^T \tag{25}$$

7. Repeat the process until convergence.

We would like to note that the NIPALS algorithm used in Partial Least Squares (PLS) shares structural similarities with the Gram-Schmidt process[2], as both iteratively construct an orthonormal basis through projection and deflation. However, their objectives differ: Gram-Schmidt is an unsupervised linear algebra technique that orthonormalizes column vectors without considering an external target, whereas NIPALS is a supervised method that builds an orthonormal basis of latent variables to maximize covariance with the response variable Y. While both methods ensure orthogonality at each step, NIPALS deflates both X and Y after extracting each latent variable, progressively focusing on predictive directions. This key difference makes PLS more powerful than Principal Component Regression (PCR), as it selects components that are not just orthogonal but also relevant for prediction. In essence, NIPALS can be viewed as a Gram-Schmidt-like process adapted for regression, optimizing feature extraction for predictive accuracy rather than just mathematical convenience.

SIMPLS Algorithm

An alternative to NIPALS is the SIMPLS (Statistically Inspired Modification of PLS) algorithm, which extracts PLS components without iterative deflation, making it computationally more efficient.

1. Compute the covariance matrix:

$$S = X^T Y \tag{26}$$

---

[2]Definition of Gram-Schmidt process is provided in Appendix 6.6

2. Extract the first singular vector r_1 from SVD of S.

3. Compute the score vector:

$$t_1 = Xr_1 \tag{27}$$

4. Compute the loading:

$$p_1 = X^T t_1 / (t_1^T t_1) \tag{28}$$

5. Compute the regression coefficient:

$$b_1 = (t_1^T Y)/(t_1^T t_1) \tag{29}$$

6. Deflate the covariance matrix:

$$S = S - r_1(r_1^T S) \tag{30}$$

7. Repeat for additional components.

SIMPLS is preferred when computational efficiency is a concern, especially for very large datasets. Both methods produce the same final PLS regression coefficients, but their computation methods differ.

2.3 Selection of the Optimal Number of Components in PCR and PLS

A crucial aspect of applying PCR and PLS effectively is determining the appropriate number of components (k) to retain (Figure 2). Selecting too few components may result in underfitting, where important predictive information is lost. Conversely, retaining too many components can lead to overfitting, reducing the generalization ability of the model. In this section we will consider several methods for selecting the optimal number of components in PCR and PLS.

Elbow Method

The Elbow Method is widely used for selecting the optimal number of components in both Principal Component Regression (PCR) and Partial Least Squares (PLS). The method involves plotting a performance metric—such as cumulative explained variance for PCR or cumulative explained covariance for PLS—against the number of components (Figures 1 and 2). The optimal number of components is chosen at the "elbow point", where the marginal gain in variance or covariance explained diminishes significantly.

In Principal Component Regression (PCR), component selection is based on Principal Component Analysis (PCA), which transforms the original predictor space into a set of orthogonal components that explain the variance in X. Since PCR does not consider the response variable Y in component selection, it relies on the cumulative explained variance of the predictors.

Mathematically, the proportion of variance explained by the first k principal components is given by:

$$\text{Explained Variance}(k) = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \tag{31}$$

where λ_i are the eigenvalues of the covariance matrix $X^T X$, which correspond to the variances captured by each principal component.

The selection of k is determined by identifying the "elbow point" on the cumulative variance plot. This is the point where adding more components results in a negligible increase in explained variance. In practical applications, a commonly used threshold is:

$$\sum_{i=1}^{k} \lambda_i \geq 0.95 \sum_{i=1}^{p} \lambda_i \tag{32}$$

indicating that at least 95% of the total variance in X has been retained.

Unlike PCR, which selects components based solely on variance in X, Partial Least Squares (PLS) extracts components that maximize the covariance between X and Y. Therefore, the Elbow Method in PLS is based on the cumulative explained covariance rather than variance.

For PLS, we define the cumulative explained covariance as:

$$\text{Explained Covariance}(k) = \frac{\sum_{i=1}^{k} \text{Cov}(t_i, y)^2}{\sum_{i=1}^{p} \text{Cov}(t_i, y)^2} \tag{33}$$

where t_i are the latent variables obtained from PLS, which are linear combinations of the original predictors.

Similar to PCR, the "elbow point" is identified on the cumulative covariance plot, marking the point beyond which additional components contribute minimally to predictive power. The typical threshold for PLS component selection is:

$$\sum_{i=1}^{k} \text{Cov}(t_i, y)^2 \geq 0.90 \sum_{i=1}^{p} \text{Cov}(t_i, y)^2 \tag{34}$$

indicating that at least 90% of the total covariance between X and Y is captured by the first k components.

The Elbow Method provides a simple yet effective means of selecting k, but its effectiveness depends on the dataset. If the "elbow" is not clearly visible, alternative methods such as the **Needle Algorithm** or **Cross-Validation (CV)** should be used to refine component selection.

Needle Algorithm

The Needle Algorithm is an alternative method for determining the optimal number of components, which provides a formalized way of detecting the point beyond which additional components contribute negligibly to the model. It is particularly useful when the variance or covariance decay is gradual, making the elbow point ambiguous.

In PCR, The Needle Algorithm identifies k by examining the second derivative of the cumulative explained variance function. Mathematically, the proportion of variance explained by the first k components is:

$$\text{Explained Variance}(k) = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \tag{35}$$

where λ_i are the eigenvalues of $X^T X$.

The **Needle Criterion** is defined as:

$$\Delta_k = \lambda_k - \lambda_{k+1} \tag{36}$$

$$\text{Relative Gain}(k) = \frac{\Delta_k}{\sum_{i=1}^{k} \lambda_i} \tag{37}$$

The optimal k is determined when the relative gain in explained variance falls below a predefined threshold ϵ, typically around 1%:

$$\frac{\lambda_k - \lambda_{k+1}}{\sum_{i=1}^{k} \lambda_i} < \epsilon \tag{38}$$

This ensures that additional components do not contribute meaningfully to variance retention.

For PLS, the Needle Algorithm follows the same principle but is applied to the explained covariance between X and Y instead of variance in X. The proportion of covariance explained by the first k components is given by:

$$\text{Explained Covariance}(k) = \frac{\sum_{i=1}^{k} \text{Cov}(t_i, y)^2}{\sum_{i=1}^{p} \text{Cov}(t_i, y)^2} \tag{39}$$

where t_i are the latent variables extracted by PLS.

The Needle Criterion for PLS is:

$$\Delta_k = \text{Cov}(t_k, y)^2 - \text{Cov}(t_{k+1}, y)^2 \tag{40}$$

$$\text{Relative Gain}(k) = \frac{\Delta_k}{\sum_{i=1}^{k} \text{Cov}(t_i, y)^2} \tag{41}$$

As in PCR, the optimal k is found when the relative gain in explained covariance drops below a predefined threshold ϵ:

$$\frac{\text{Cov}(t_k, y)^2 - \text{Cov}(t_{k+1}, y)^2}{\sum_{i=1}^{k} \text{Cov}(t_i, y)^2} < \epsilon \tag{42}$$

Typically, ϵ is set to 1% to 2%, ensuring that only significant components contributing to the prediction of Y are retained.

The Needle Algorithm provides a more formal alternative to the Elbow Method, particularly in cases where variance or covariance decays smoothly without a clear "elbow point".

Cross-Validation

Cross-Validation (CV) is a data-driven approach to selecting the optimal number of components. Unlike the Elbow Method or the Needle Algorithm, which focus on variance or covariance retention, CV directly evaluates the predictive performance of the model by minimizing the generalization error. This ensures that the selected number of components leads to the best tradeoff between bias and variance.

In PCR, the key challenge is to determine how many principal components (k) should be retained to ensure strong predictive performance. Since PCR does not use the response variable Y when selecting components, cross-validation is crucial in identifying the point where including additional components no longer improves prediction accuracy.

The procedure for selecting k via CV follows these steps:

1. Split the dataset into K **folds** (typically $K = 5$ or $K = 10$).

2. For each fold, fit the **PCR model** using the first k principal components and compute the prediction error on the held-out data.

3. Compute the **Cross-Validation Mean Squared Error (CV-MSE)** for each k:

$$\text{CV-MSE}(k) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_{-i,k})^2 \tag{43}$$

   where $\hat{y}_{-i,k}$ is the prediction for the i-th observation using a model trained without that observation and with k components.

4. Select k^* that minimizes CV-MSE(k):

$$k^* = \arg\min_k \text{CV-MSE}(k) \tag{44}$$

Since PCR selects principal components without considering Y, CV ensures that the selected components contribute to predictive accuracy rather than just maximizing variance in X.

Unlike PCR, PLS constructs components that maximize the covariance between X and Y. Thus, CV in PLS plays a slightly different role: it helps determine the number of latent variables that maximize predictive accuracy.

The CV procedure for PLS follows the same general steps as in PCR:

1. Split the dataset into K **folds**.

2. For each fold, fit the **PLS model** using the first k latent variables and compute the prediction error on the held-out data.

3. Compute the **CV-MSE** for each k:

$$\text{CV-MSE}(k) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_{-i,k})^2 \tag{45}$$

4. Select k^* that minimizes CV-MSE(k):

$$k^* = \arg\min_k \text{CV-MSE}(k) \tag{46}$$

Since PLS incorporates Y in selecting components, it is expected to achieve lower optimal k^* and lower CV-MSE compared to PCR for a given dataset, particularly when X and Y are highly correlated.

Cross-validation is the most rigorous method for selecting k, but it comes with some practical considerations. One major concern is computational cost, as it requires training multiple models

for different values of k across several folds, making the process resource-intensive. Additionally, cross-validation plays a crucial role in managing the bias-variance tradeoff. Smaller values of k tend to introduce higher bias, whereas larger values lead to increased variance. The goal is to find an optimal balance between these two factors. In practice, common choices for k are 5 or 10. While opting for a smaller k can reduce computation time, a larger k generally provides a more reliable error estimate.

The choice of k should be guided by a combination of these methods, balancing the need for variance retention and predictive accuracy. In empirical applications, cross-validation is typically preferred due to its ability to optimize for prediction error directly, therefore CV is applied in our simulations and empirical estimation.

3   Simulation Study

The goal of this simulation study is to evaluate the performance of Principal Component Regression (PCR) and Partial Least Squares (PLS) across different data-generating scenarios, focusing on their predictive accuracy, robustness, and bias-variance trade-offs in high-dimensional settings. Given the common challenges posed by multicollinearity and high-dimensional predictor spaces, we employ a Monte Carlo simulation framework to systematically analyze the behavior of these methods under varying feature dependencies. Simulation study is structured around three distinct setups:

1. **Classical Data (No Multicollinearity)**: Predictors are independently sampled from a normal distribution, ensuring no multicollinearity.

2. **Moderate Multicollinearity**: The predictor matrix consists of a set of base features along with additional highly correlated features, introducing a controlled level of dependence.

3. **Severe Multicollinearity (Low Observations)**: The predictors exhibit strong correlations, with only a subset contributing to the response variable. This setup mimics scenarios where multicollinearity is extreme and sample size is limited, testing the methods' ability to extract meaningful structure from highly dependent data.

To assess performance, we first apply PCR by performing Principal Component Analysis (PCA) on the predictor matrix and selecting the optimal number of components via k-fold cross-validation, minimizing the Root Mean Squared Error (RMSE) of prediction. Using these selected components, an Ordinary Least Squares (OLS) regression model is fitted, and predictive accuracy is evaluated through RMSE and Mean Squared Error (MSE). Additionally, we analyze the bias-variance trade-off of the coefficient estimates, comparing them to standard OLS results to quantify the extent of regularization.

Similarly, for PLS, we employ the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm to extract latent components, selecting the optimal number via cross-validation to balance model complexity and predictive power. A PLS regression model is then fitted to the transformed predictor matrix, with performance assessed through RMSE, MSE, and coefficient stability.

To ensure the asymptotic validity of our findings, we conduct a Monte Carlo simulation, generating 1000 datasets for each three setups to evaluate the stability of coefficient estimates, their bias-variance decomposition, and how these methods perform in large-sample scenarios. By systematically comparing

PCR, PLS, and OLS across these conditions, we gain deeper insights into their relative advantages and limitations in high-dimensional regression problems.

3.1 Data Description

The dataset for each scenario is generated as follows:

Classical Data (No Multicollinearity)

- The predictor matrix X consists of $n = 200$ observations and $p = 10$ predictors, generated independently from a standard normal distribution:

$$X \sim \mathcal{N}(0, I_p)$$

- The response variable follows:

$$Y = X\beta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 0.1)$$

- All 10 predictors are actively used to generate Y.

Moderate Multicollinearity

- The dataset contains $n = 200$ observations and $p = 10$ predictors.

- A **base predictor matrix** X_{base} is generated as:

$$X_{\text{base}} \sim \mathcal{N}(0, I)$$

- Additional features are constructed as:

$$X_{\text{extra}} = X_{\text{base}} + \eta, \quad \eta \sim \mathcal{N}(0, 0.1)$$

  where η introduces correlation between the predictors.

- The final predictor matrix is formed by concatenation:

$$X = [X_{\text{base}}, X_{\text{extra}}]$$

- The response variable is generated as:

$$Y = X\beta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1.0)$$

- All 10 predictors are actively used to generate Y.

Severe Multicollinearity (Low Observations)

- The dataset contains only $n = 40$ observations and $p = 10$ predictors, creating a scenario where the number of observations is significantly lower than in previous setups.

- A base predictor matrix is generated as:

$$X_{\text{gen}} \sim \mathcal{N}(0, I)$$

- Additional predictors are formed via a linear transformation:

$$X_{\text{additional}} = X_{\text{gen}}W + \mathcal{N}(0, 0.1)$$

where W represents a randomly generated weight matrix, introducing strong collinearity among predictors.

- The response variable follows:

$$Y = X\beta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1.0)$$

- Importantly, only **three** of the 10 predictors actively contribute to generating Y, while the remaining seven are purely collinear noise variables.

In the severe multicollinearity scenario, the reduction in the number of observations relative to the number of predictors amplifies the risk of overfitting and instability in regression models. Furthermore, since only three predictors actively contribute to the response variable while the remaining seven are collinear noise, methods like PCR and PLS must effectively distinguish between relevant and redundant information.

PCR, which selects components based on variance, may retain principal components influenced by multicollinear features, leading to potential inefficiencies in prediction. Conversely, PLS, which maximizes covariance with Y, is expected to perform better in identifying relevant predictors, thus mitigating the negative impact of multicollinearity.

3.2 Simulation Results: Classical Dataset (No Multicollinearity)

In this scenario, the predictors are generated independently, ensuring there is no correlation between them. This setup creates the ideal conditions for Ordinary Least Squares (OLS) regression, which serves as the benchmark for evaluating Principal Component Regression (PCR) and Partial Least Squares (PLS). Our goal is to analyze how these methods perform when no multicollinearity is present and how their behavior changes as the number of components increases.

As expected, OLS performs the best, achieving the lowest Mean Squared Error (MSE) and producing coefficient estimates that closely match the true values. Because there is no multicollinearity, OLS remains unbiased and efficient, making it the optimal estimator in this setting. However, when we apply PCR and PLS, we see a clear difference in how they handle feature selection and regularization (Figure 5).

Since PCR selects components based purely on variance, it struggles when using only a few components. For instance, with PCR(1) (one principal component), the model captures only a limited portion of

the predictor space, leading to high bias. However, as we increase the number of components (PCR(5), PCR(10)), the model recovers more information, reducing bias and approaching OLS-level accuracy. This improvement demonstrates the classic bias-variance trade-off: fewer components introduce higher bias but lower variance, while more components reduce bias at the cost of increasing variance (Figure 3).

In contrast, PLS finds a balance between bias and variance more efficiently than PCR. Unlike PCR, which selects components based solely on predictor variance, PLS chooses components based on their relationship with the response variable. This allows PLS to extract relevant predictive features even with a small number of components. As a result, PLS with fewer components (e.g., PLS(1), PLS(5)) already provides better coefficient estimates and lower MSE compared to PCR with the same number of components. This suggests that PLS is a more efficient dimensionality reduction method when the goal is to improve prediction accuracy (Figure 6).

The impact of component selection is further reflected in the quality of predictions. OLS predictions align almost perfectly with real values, reinforcing its reliability in this scenario. On the other hand, PCR with very few components, such as PCR(1), struggles to capture the full structure of the data, leading to poor predictions and a weaker fit. However, as the number of components increases, the predictions improve significantly, with PCR(10) closely matching the performance of OLS. In comparison, PLS consistently performs well even with fewer components, confirming its ability to extract meaningful features efficiently (Figure 7).

The stability of coefficient estimates also reflects these patterns. OLS estimates remain centered around the true values with minimal variation, confirming that it is an unbiased estimator. PCR, on the other hand, shows greater variability when fewer components are used, and its estimates are further from the true values. However, when enough components are included, PCR estimates stabilize and align with OLS. PLS estimates remain more stable than PCR, even with fewer components, highlighting its advantage in capturing relevant information effectively (Figure 8).

(a) Classic Data Analysis   (b) Moderate Multicollinearity   (c) Severe Multicollinearity
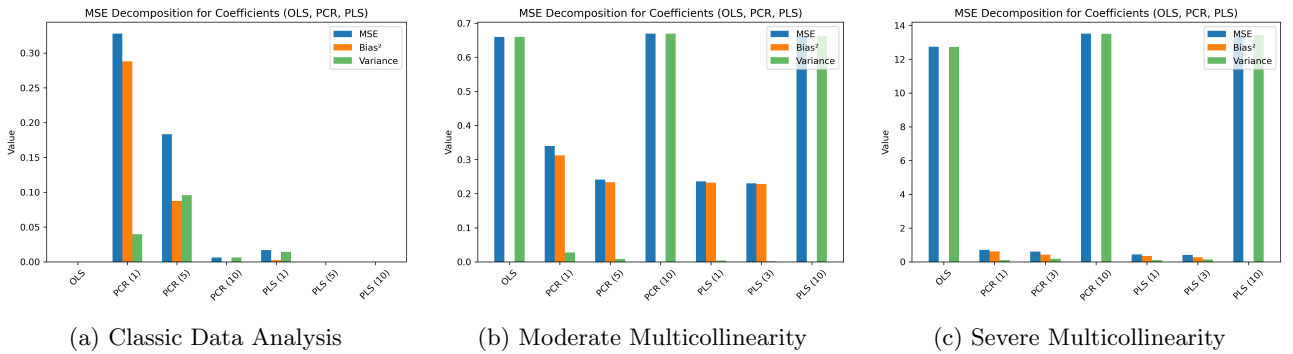
Figure 3: Bias-Variance trade-off: OLS, PCR, and PLS for different cases

OLS is the best option when there is no multicollinearity, as it provides the most accurate and stable coefficient estimates. PCR requires a sufficient number of components to recover the full predictor space. Using too few components results in high bias and poor predictions, but increasing the number of components improves performance. PLS is more efficient than PCR in selecting relevant features, achieving strong predictive performance with fewer components. The bias-variance trade-off is evident—fewer components lead to high bias (underfitting), while using too many components increases variance (potential overfitting).

This classical dataset serves as a baseline reference, showing how PCR and PLS behave under ideal conditions. The next step is to explore how these methods perform when moderate multicollinearity is introduced, where predictors become correlated and the assumptions of OLS begin to break down.

3.3   Simulation Results: Moderate Multicollinearity Dataset

In second scenario, the predictor variables are moderately correlated, introducing a controlled level of multicollinearity. This setup challenges the assumptions of Ordinary Least Squares (OLS) regression and allows us to assess how Principal Component Regression (PCR) and Partial Least Squares (PLS) handle collinearity and improve prediction performance.

When predictors exhibit moderate multicollinearity, OLS starts to struggle as the variance of coefficient estimates increases. This is evident in the Mean Squared Error (MSE) decomposition, where OLS has a significantly high variance component. While OLS remains unbiased, its estimates become unstable due to the correlated predictors.

PCR mitigates this issue by reducing the predictor space to uncorrelated principal components. However, with too few components (e.g., PCR(1)), the model captures only limited variance, resulting in high bias. As the number of components increases (PCR(5), PCR(10)), bias decreases, and the model performance improves. In contrast, PLS achieves a better balance between bias and variance with fewer components compared to PCR. Even with a small number of components (e.g., PLS(3)), it outperforms PCR with the same number of components, as it better identifies the underlying structure in the data (Figure 3 and Table 2).

The impact of multicollinearity on model predictions becomes evident when comparing actual vs. predicted values. OLS predictions are less stable due to inflated coefficient variance, leading to fluctuations in its predictive performance. PCR with very few components underfits the data, failing to capture enough variance for accurate predictions. However, as the number of components increases, PCR's predictions align more closely with actual values, though it still lags behind PLS (Figure 11).

PLS consistently provides a stronger predictive fit, even with fewer components. Its predictions are closer to the true values, demonstrating that selecting components based on their relevance to the response variable improves generalization. This makes PLS a more robust choice when multicollinearity is present (Table 1).

Coefficient estimates in OLS show high variability, confirming its sensitivity to multicollinearity. PCR, when using too few components, exhibits significant bias, whereas with more components, its estimates stabilize. PLS remains more stable than PCR throughout, highlighting its effectiveness in handling multicollinearity (Figure 12).

OLS struggles in the presence of moderate multicollinearity, with high variance in coefficient estimates reducing prediction stability. PCR mitigates multicollinearity but requires enough components to recover essential variance. Selecting too few components results in high bias, while increasing components gradually improves predictions. PLS effectively balances bias and variance, leveraging response-driven component selection to achieve strong predictive performance with fewer components. PLS outperforms PCR in handling multicollinearity, as it selects components that are directly relevant to the response variable rather than relying solely on variance. The bias-variance trade-off remains crucial—too few components lead to underfitting (high bias), whereas too many components can introduce instability.

This dataset highlights the limitations of OLS in correlated predictor settings and demonstrates the advantages of PCR and PLS as alternatives. The next step is to explore how these methods perform under severe multicollinearity, where predictor correlations are extreme, and the number of observations is limited.

3.4 Simulation Results: Severe Multicollinearity with Low Observations

In our last scenario, the dataset contains only 40 observations with 10 predictors, where the predictors exhibit strong multicollinearity. The additional predictors are generated via a linear transformation of a base predictor matrix, leading to extreme correlations. Furthermore, only three predictors contribute to the response variable, while the remaining predictors act as redundant noise. This setup presents an extreme case of multicollinearity, making it a challenging setting for regression models.

With severe multicollinearity and a low number of observations, OLS struggles the most. The variance of coefficient estimates becomes excessively high, leading to unstable and unreliable predictions. The MSE decomposition shows that OLS has an extremely high variance component, making it unsuitable for this setting. Even though OLS remains unbiased, its high variance leads to large fluctuations in predictions, reducing its overall reliability.

PCR and PLS attempt to mitigate this issue by selecting fewer components. However, using too few components (PCR(1) and PLS(1)), the model oversimplifies the predictor space, leading to high bias. As reaching optimal number of components, the model performance improves significantly. It is worth noting that, PLS needs the same number of components to reach a similar level of accuracy as PCR (Table 1).

At their optimal points, both PCR and PLS still exhibit some level of bias, as the component selection inherently introduces regularization. However, this bias is outweighed by the substantial reduction in variance, leading to a much lower overall MSE compared to OLS (Figure 3). This highlights the key advantage of dimension-reduction methods: while they sacrifice some degree of unbiasedness, they compensate by stabilizing predictions and significantly improving generalization performance.

The stability of coefficient estimates also supports these findings. OLS estimates show extreme variability, confirming its inefficiency in this setting. PCR estimates stabilize as more components are included but still suffer from inefficiencies due to noise retention when too few components are used. PLS maintains stable coefficient estimates even with fewer components due to its response-driven selection. However, when the appropriate number of components is selected, both methods achieve comparable stability (Figure 16).

Overall, OLS fails to provide reliable estimates in the presence of severe multicollinearity with low observations due to excessive variance in coefficient estimates. PCR helps reduce multicollinearity effects, but selecting too few components results in high bias, while selecting too many components may introduce noise from redundant predictors. PLS, leveraging its supervised nature, achieves strong predictive performance with fewer components. However, when PCR is given enough components, it reaches a similar level of predictive accuracy, making both methods viable solutions depending on the data and modeling constraints (Figure 14 and Table 2).

At their optimal component selections, PCR and PLS remain slightly biased but achieve significantly lower variance compared to OLS, leading to a much lower overall MSE. This highlighted by the

effectiveness of both methods in controlling variance and improving model stability, particularly in high-dimensional, low-sample-size settings (Table 1).

Results of all three scenarios highlight the extreme limitations of OLS and demonstrates the importance of dimensionality reduction in challenging regression problems. Both PCR and PLS emerge as effective methods for handling highly collinear predictors in small datasets, achieving similar results through different mechanisms—PLS using fewer components due to its supervised nature, while PCR requiring slightly more components to reach similar performance. To further evaluate the real-world applicability of PCR and PLS, the next section applies these methods to actual spectral data, assessing their predictive performance beyond the controlled scenarios examined so far.

4 Analysis of NIR Data: Corn Moisture Prediction Using PLS and PCR

Dataset description

The dataset used in this study consists of NIR spectra collected from 80 corn samples, measured using the mp5 NIR spectrometer. Each spectrum contains 700 data points recorded in the wavelength range of 2498–1100 nm at 2 nm intervals. The target variable, moisture content, serves as the dependent variable. This dataset and methodology align closely with the results reported in the referenced paper, where similar approaches were applied to corn data. The high-dimensional nature of the spectral data makes this a challenging regression problem, particularly due to the presence of collinearity among the predictor variables (Figure 4).
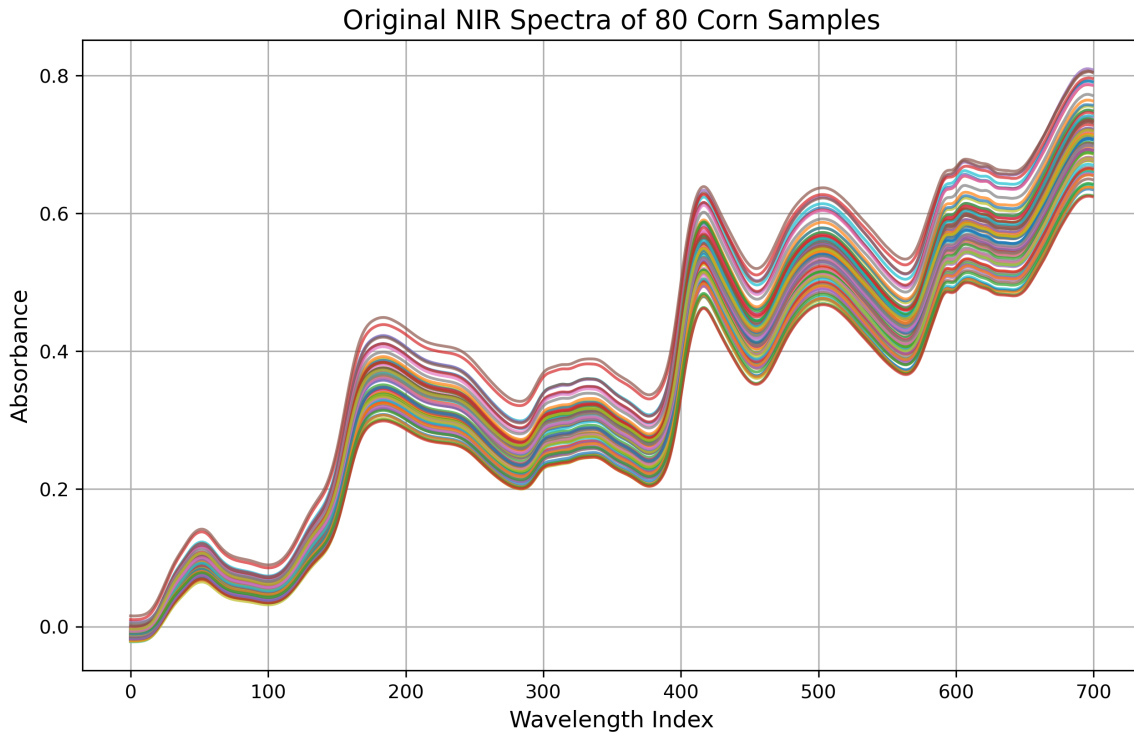

Figure 4: NIR absorption spectra from the Corn Data Set

Results and Analysis

In our empirical part cross-validation is used to determine the optimal number of components for PCR and PLS. The results show a sharp decrease in MSE as the number of components increases, after which the performance stabilizes. In our training data, PLS achieves its optimal performance with fewer components, while PCR requires slightly more components to reach the same level (Figure 17). While in test data, we can observe that beyond a certain threshold, increasing the number of components leads to overfitting in PLS, causing MSE to rise. This reinforces the importance of selecting the appropriate number of components to balance bias and variance. Meanwhile, PCR is less affected by overfitting due to its unsupervised nature. After reaching its optimal level, PCR performance starts fluctuating instead of consistently increasing with the addition of more components, as it does not rely directly on maximizing covariance with the response variable (Figure 18).

Examining predictive accuracy of PCR and PLS predictions, we can see that both methods are close to the actual values, demonstrating their ability to capture meaningful patterns in the data. Since PLS selects components in a way that directly relates to the response variable, it achieves a good predictive fit with fewer components. However, when PCR is allowed to utilize a sufficient number of components, it reaches a similar performance level (Figure 19).

Summary and Conclusions

Overall, this study validates that PCR and PLS provide significantly more reliable estimates for high-dimensional spectral data. Both methods ultimately achieve similar predictive accuracy, but PLS requires fewer components due to its supervised approach, whereas PCR needs slightly more components to reach the same results (Table 1). The stability of coefficient estimates and predictive performance confirm that both PCR and PLS are highly effective in handling multicollinearity in NIR spectroscopy data. Cross-validation results further reinforce their advantages, showing that regularization through dimensionality reduction is essential for improving model stability and predictive performance. The findings align with previous research demonstrating the bias-variance trade-off in PLS, where increasing the number of latent variables initially reduces bias but eventually leads to overfitting, as evidenced by increasing cross-validation error. In contrast, PCR, being unsupervised, is less prone to overfitting; instead, after reaching its optimal performance, its MSE fluctuates rather than consistently increasing.

Additionally, standard regression techniques such as Ordinary Least Squares (OLS) cannot be directly applied in this setting due to singularity issues, as the number of predictors is much larger than the number of observations. While modifications such as applying a pseudo-inverse could technically be used to solve the regression problem, it would not help mitigate the noise amplification caused by collinearity. Another approach would be to use Tikhonov regularization, which effectively transforms the OLS formulation into a Ridge regression model by adding a penalty term. However, to avoid additional complexity and ensure comparability across methods, we exclude OLS and its modified forms from the analysis, focusing instead on PCR and PLS as more robust alternatives for handling high-dimensional NIR data.

5 Limitations and Possible Extensions

Limited Comparison with Other Regression Techniques

This study focuses on Principal Component Regression (PCR) and Partial Least Squares (PLS) but does not compare them with Ridge, Lasso, or Elastic Net Regression, which also address multicollinearity. Ridge uses an L2 penalty for coefficient shrinkage, Lasso enforces sparsity through an L1 penalty, and Elastic Net combines both. Unlike PCR and PLS, which reduce dimensionality discretely, these shrinkage methods optimize continuously. Future research could compare these approaches to provide a broader perspective on their trade-offs.

Assumptions on Data Distribution

The study assumes predictor variables follow a normal distribution, whereas real-world data often exhibit skewness, heavy tails, or heteroscedasticity. Since PCR and PLS rely on linear transformations, their performance may degrade under non-Gaussian conditions. Future work could explore robust versions incorporating transformations (e.g., Box-Cox) or heteroscedasticity-aware adaptations.

Limited Exploration of Nonlinear Relationships

PCR and PLS assume linearity between predictors and response variables, limiting their applicability in complex settings like genomics or finance. Nonlinear alternatives, such as Kernel PCA, Kernel PLS, and Deep PLS, as well as Functional PCR and Functional PLS could improve predictive power. Future research should evaluate these methods in high-dimensional, nonlinear contexts.

6 Appendix

6.1 Simulation Results: Classical Dataset (No Multicollinearity
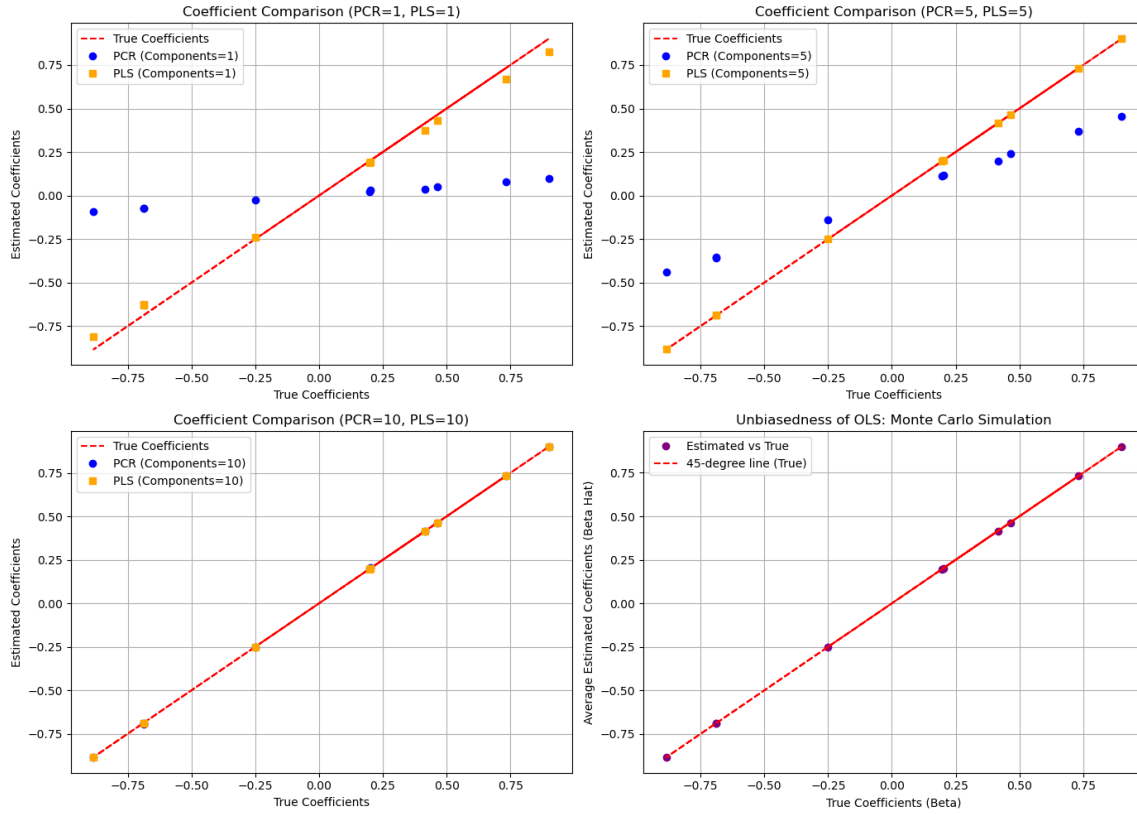
Figure 5: True vs Estimated coefficients: OLS, PCR and PLS
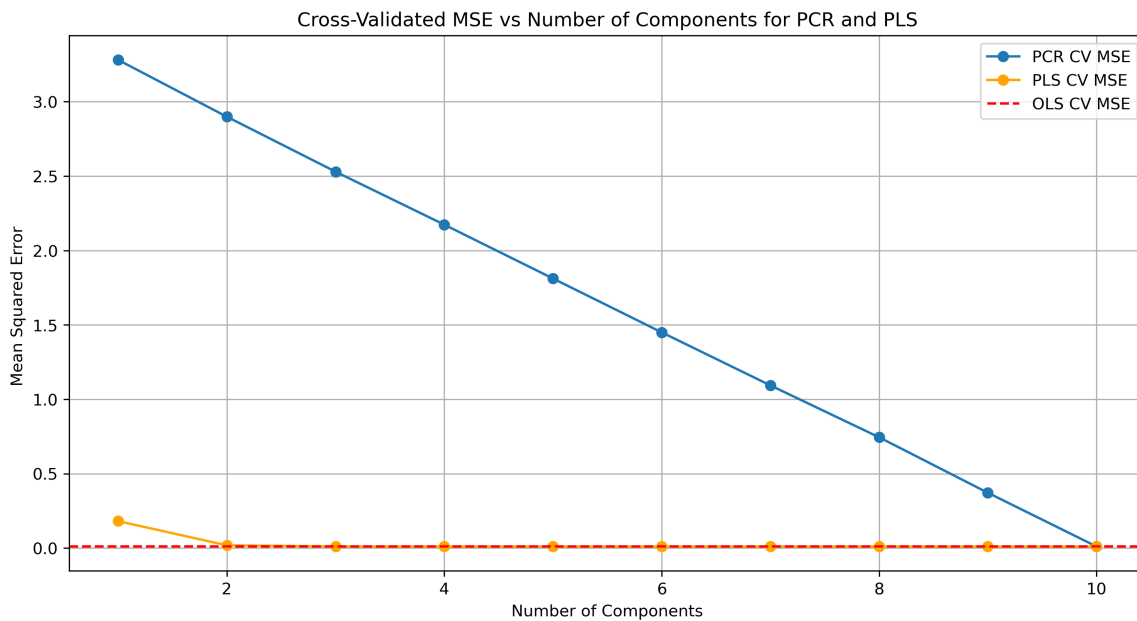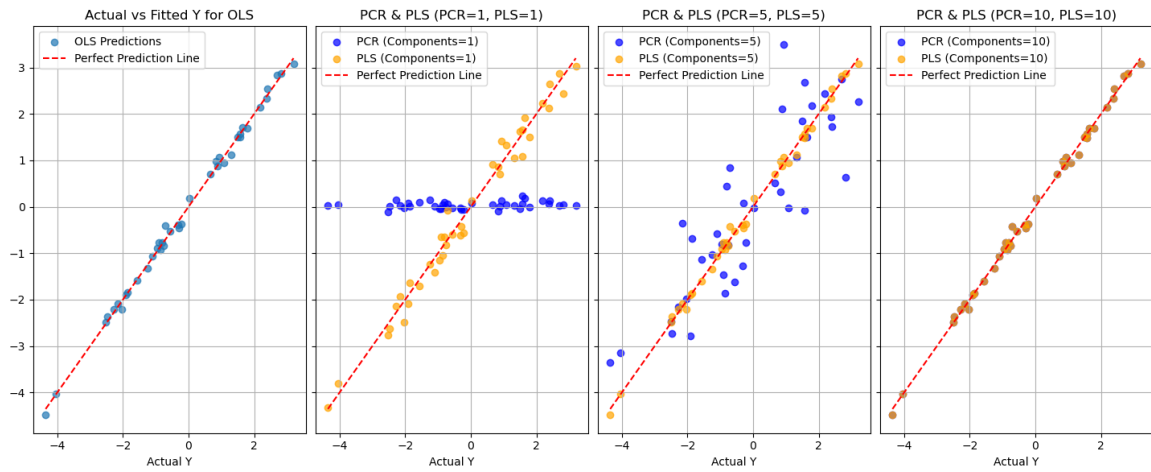

Figure 6: CV MSE by components: OLS, PCR and PLS
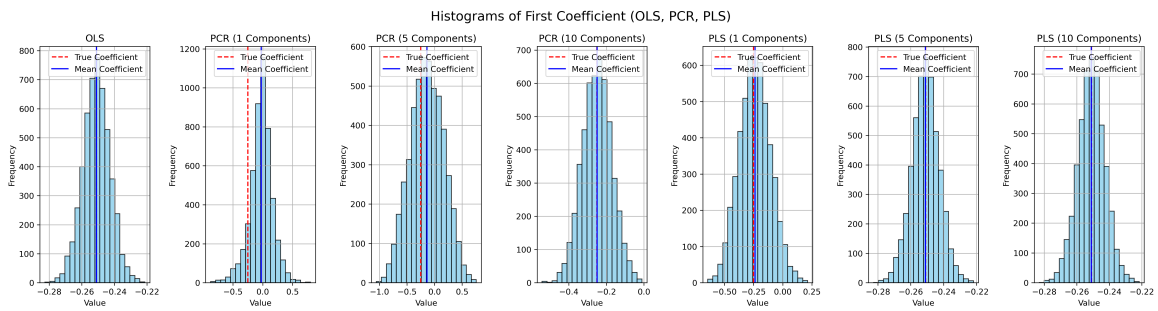
Figure 7: Model fit: OLS, PCR and PLS

Figure 8: Coefficient distribution: OLS, PCR and PLS

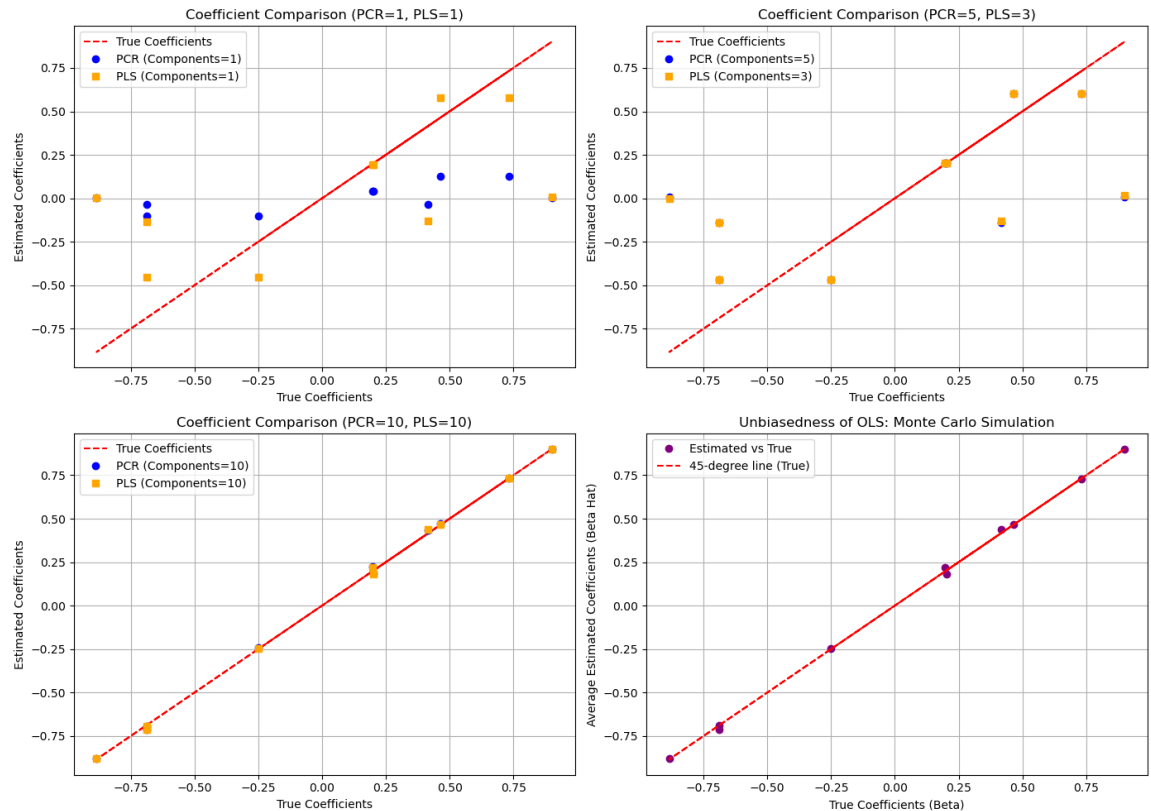## 6.2 Simulation Results: Moderate Multicollinearity Dataset

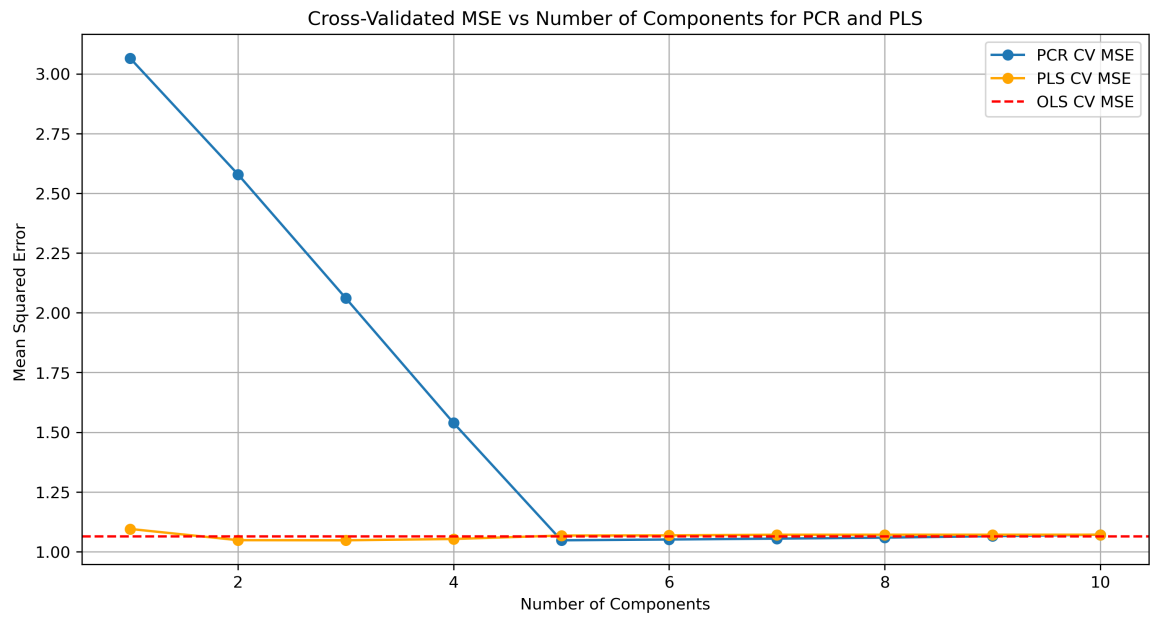Figure 9: True vs Estimated coefficients: OLS, PCR and PLS
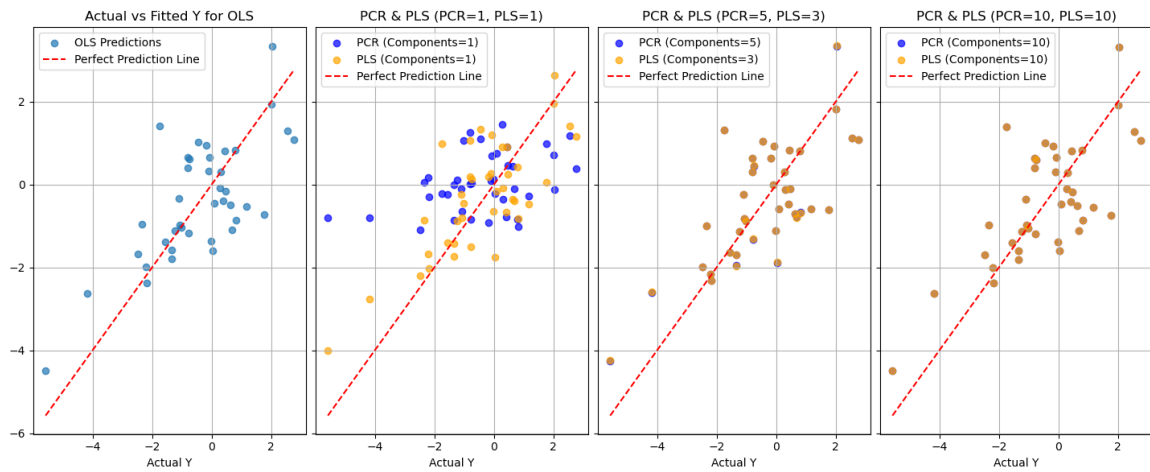
Figure 10: CV MSE by components: OLS, PCR and PLS
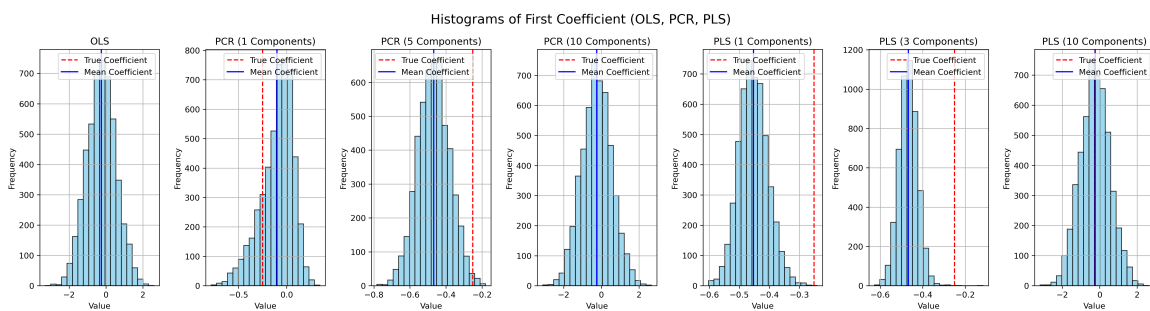

Figure 11: Model fit: OLS, PCR and PLS

Figure 12: Coefficient distribution: OLS, PCR and PLS

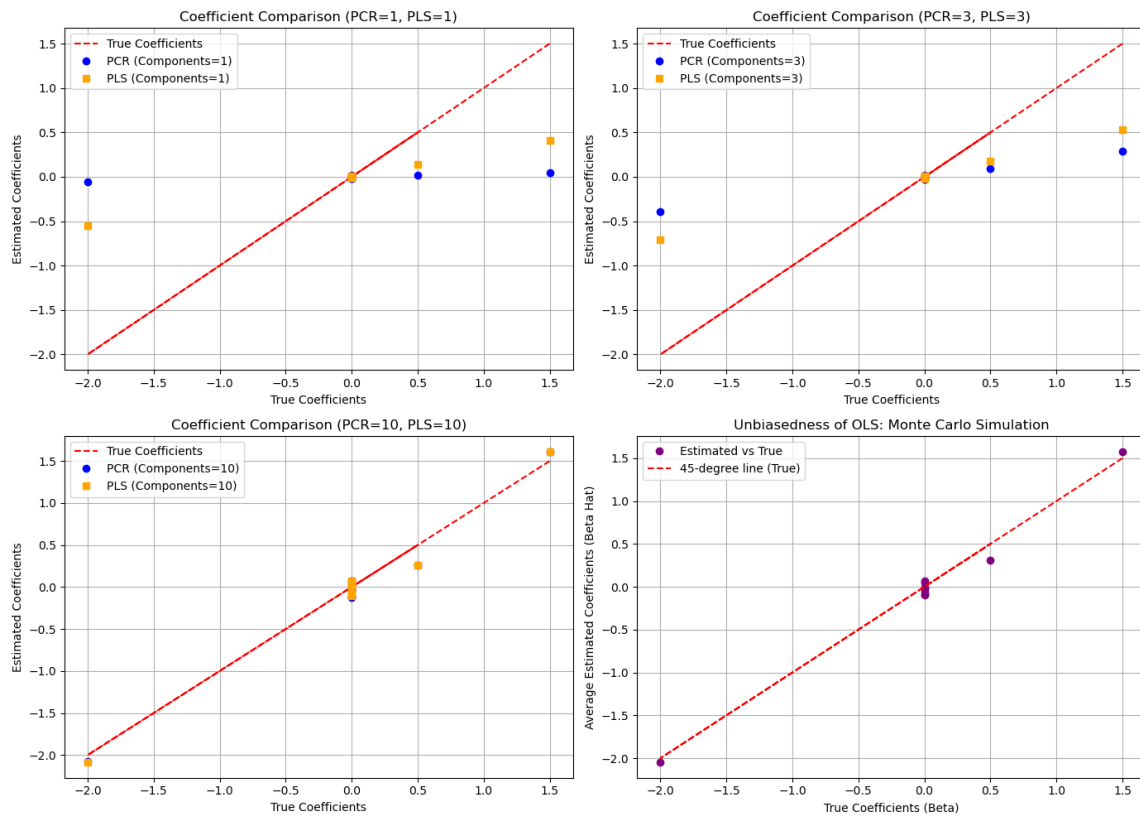## 6.3 Simulation Results: Severe Multicollinearity with Low Observations

Figure 13: True vs Estimated coefficients: OLS, PCR and PLS
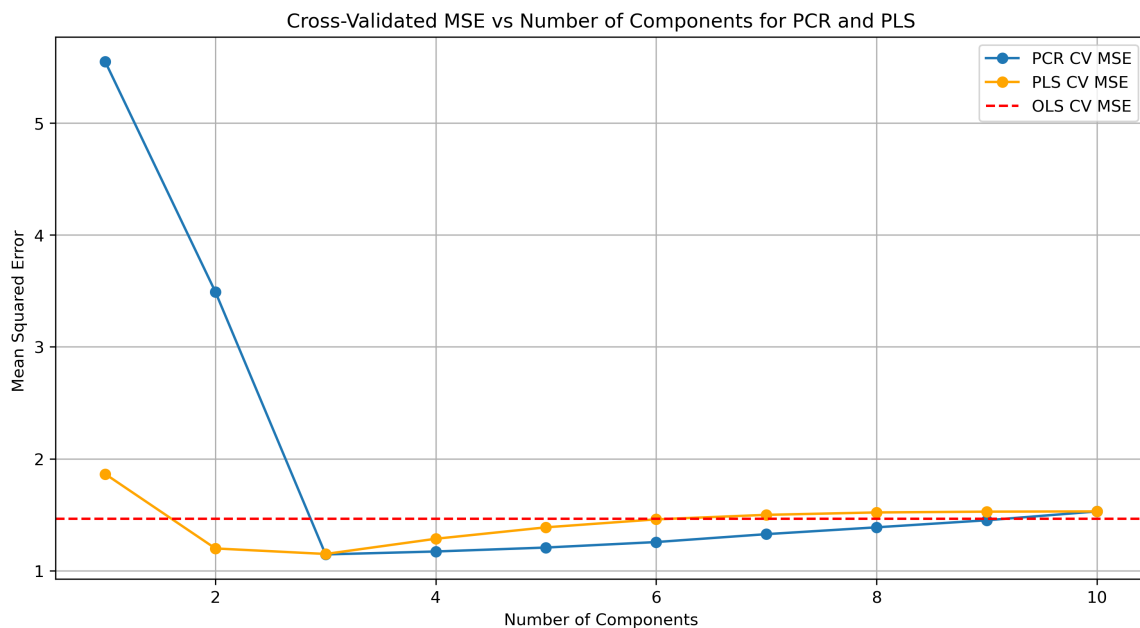

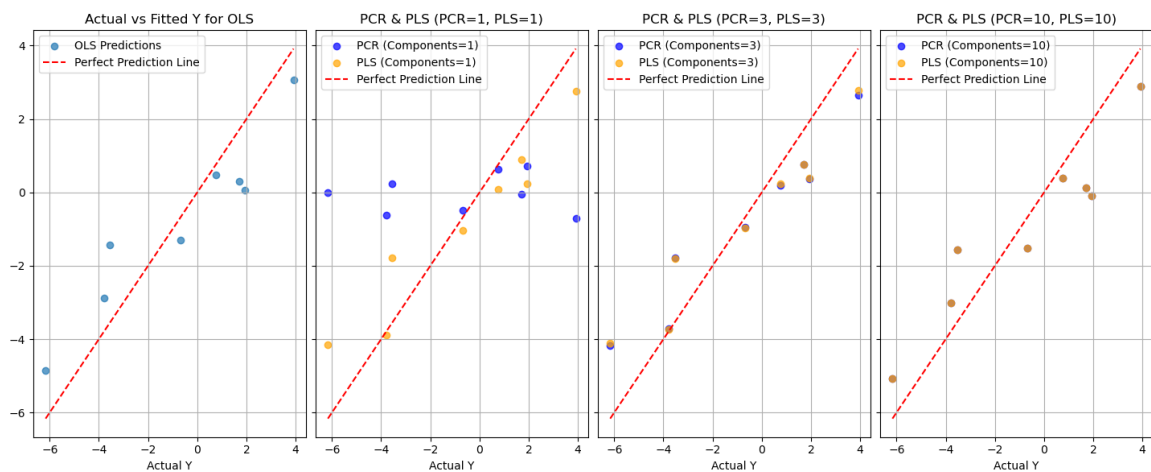Figure 14: CV MSE by components: OLS, PCR and PLS
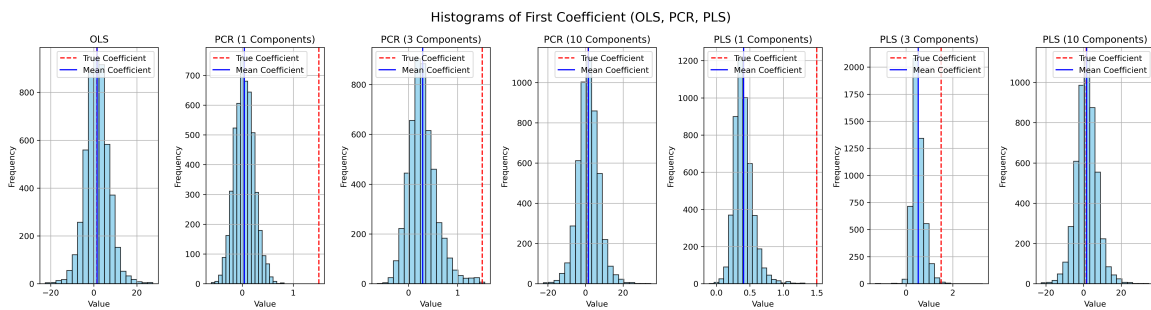
Figure 15: Model fit: OLS, PCR and PLS

Figure 16: Coefficient distribution: OLS, PCR and PLS

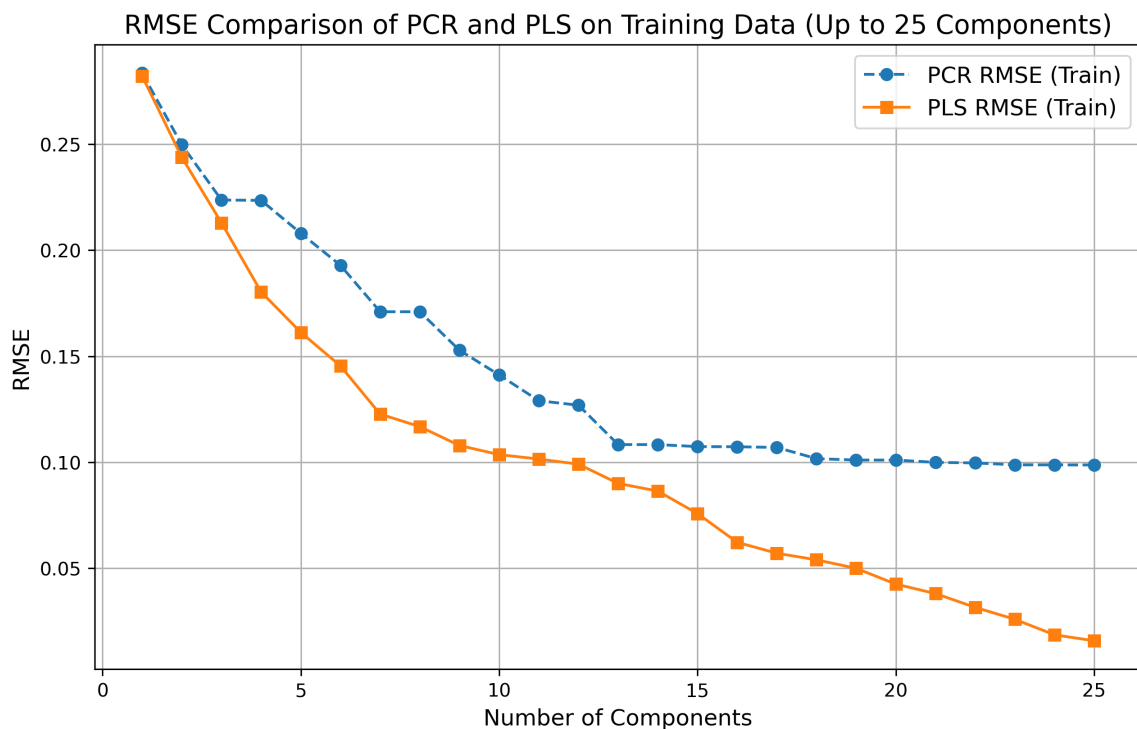## 6.4 Analysis of NIR Data: Corn Moisture Prediction Using PLS and PCR
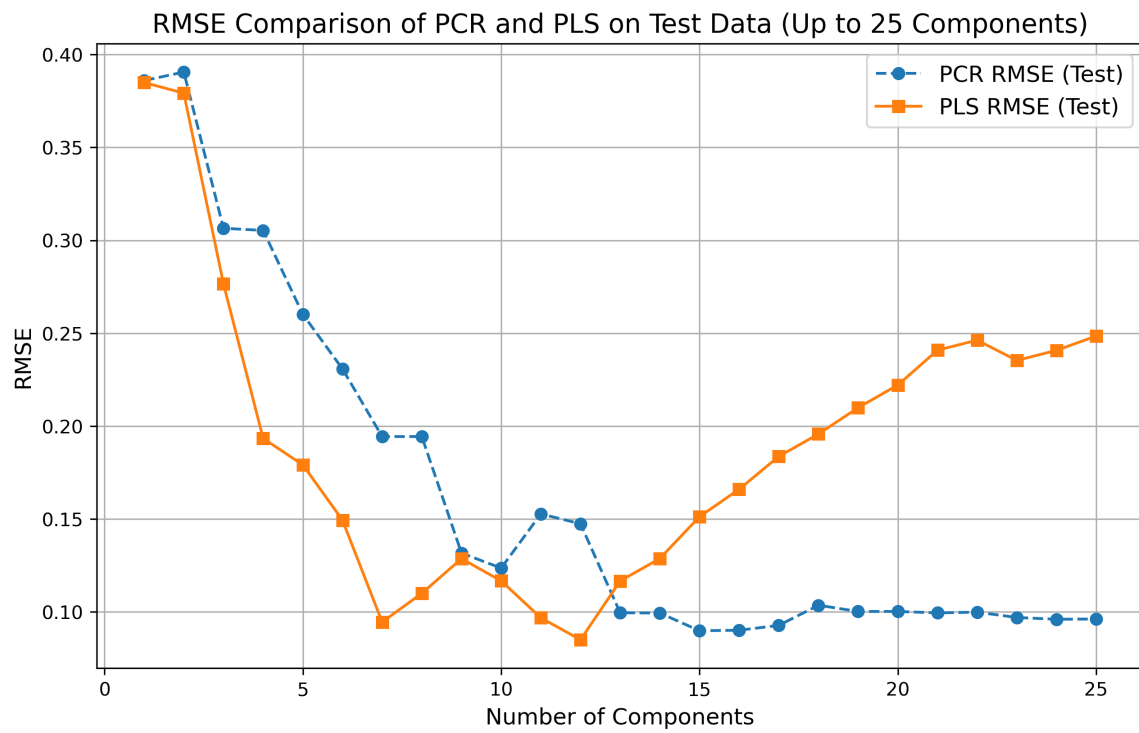

Figure 17: RMSE of PCR and PLS on training data

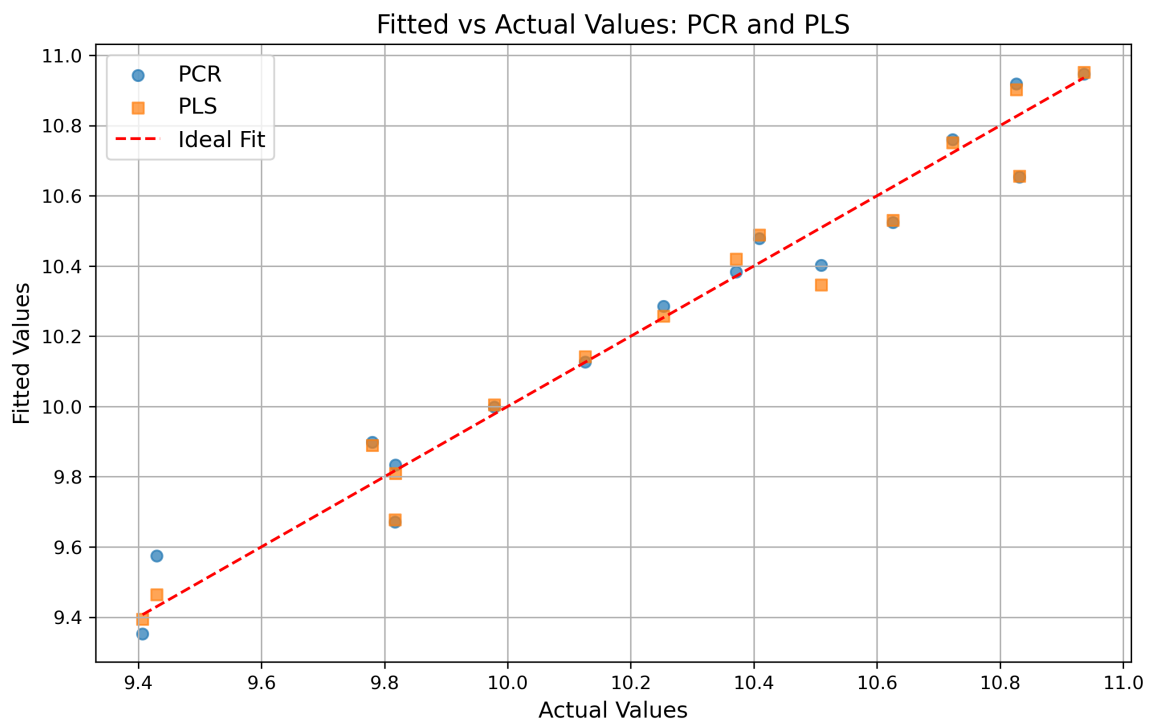Figure 18: RMSE of PCR and PLS on test data

Figure 19: Actual and prediction models performance

6.5 Quantitative Results

Methods	Optimal components	Avg CV MSE of prediction	Dataset: Multicollinearity level
OLS	NA	0.0106	
PCR	10	0.0106	No multicollinearity
PLS	10	0.0106	
OLS	NA	1.0642	
PCR	5	1.0477	Moderate
PLS	3	1.0477	
OLS	NA	1.4639	
PCR	3	1.1452	Severe
PLS	3	1.1492	
OLS	NA	NA	
PCR	15	0.0898	NIR spectroscopy data
PLS	12	0.0849	

Table 1: Prediction accuracy and optimal components across different multicollinearity levels

Methods	Avg MSE of regression coefficients	Bias^2	Variance	Dataset: Multicollinearity level
OLS	0.0001	0.0	0.0001	
PCR	0.0063	0.0	0.0063	No multicollinearity
PLS	0.0001	0.0	0.0001	
OLS	0.6598	0.0002	0.6596	
PCR	0.2414	0.2334	0.0080	Moderate
PLS	0.2305	0.2285	0.0020	
OLS	12.7383	0.0072	12.7310	
PCR	0.5993	0.4221	0.1772	Severe
PLS	0.4118	0.2719	0.1400	

Table 2: MSE coefficient decomposition across different different multicollinearity levels

6.6 Definitions and proofs

Eckart-Young Theorem for Frobenius Norm

Definition: Let $A \in \mathbb{R}^{m \times n}$ be a real (possibly rectangular) matrix with $m \leq n$. Suppose that the singular value decomposition (SVD) of A is given by:

$$A = U \Sigma V^\top, \tag{47}$$

where:

- $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal matrices,

- $\Sigma \in \mathbb{R}^{m \times n}$ is a diagonal matrix containing the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_m \geq 0$.

The best rank-k approximation to A in the Frobenius norm, denoted by $\| \cdot \|_F$, is given by:

$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^\top, \tag{48}$$

where u_i and v_i denote the ith columns of U and V, respectively.

Proof: First, note that we have:

$$\|A - A_k\|_F^2 = \left\| \sum_{i=k+1}^{n} \sigma_i u_i v_i^\top \right\|_F^2. \tag{49}$$

Using the orthonormality of u_i and v_i, this simplifies to:

$$\sum_{i=k+1}^{n} \sigma_i^2. \tag{50}$$

Now, let $B_k = XY^\top$ be any rank-k matrix where X and Y have k columns. We need to show that:

$$\|A - A_k\|_F^2 = \sum_{i=k+1}^{n} \sigma_i^2 \leq \|A - B_k\|_F^2. \tag{51}$$

Using the triangle inequality with the spectral norm, if $A = A' + A''$, then:

$$\sigma_1(A) \leq \sigma_1(A') + \sigma_1(A''). \tag{52}$$

Suppose A_k' and A_k'' respectively denote the rank-k approximation to A' and A'' obtained via SVD. Then, for any $i, j \geq 1$:

$$\sigma_i(A') + \sigma_j(A'') \geq \sigma_{i+j-1}(A). \tag{53}$$

Since $\sigma_{k+1}(B_k) = 0$, when $A' = A - B_k$ and $A'' = B_k$, we conclude that for $i \geq 1, j = k + 1$:

$$\sigma_i(A - B_k) \geq \sigma_{k+i}(A). \tag{54}$$

Thus,

$$\|A - B_k\|_F^2 = \sum_{i=1}^{n} \sigma_i(A - B_k)^2 \geq \sum_{i=k+1}^{n} \sigma_i(A)^2 = \|A - A_k\|_F^2. \tag{55}$$

Therefore, A_k achieves the lowest possible Frobenius norm error, proving that it is the best rank-k approximation. $\qquad\square$

The Gram-Schmidt Process

The vector projection of a vector $\mathbf{v}$ on a nonzero vector $\mathbf{u}$ is defined as:

$$\operatorname{proj}_{\mathbf{u}}(\mathbf{v}) = \frac{\langle \mathbf{v}, \mathbf{u} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} \mathbf{u},$$

where $\langle \mathbf{v}, \mathbf{u} \rangle$ denotes the inner product of the vectors $\mathbf{u}$ and $\mathbf{v}$. This means that $\operatorname{proj}_{\mathbf{u}}(\mathbf{v})$ is the orthogonal projection of $\mathbf{v}$ onto the line spanned by $\mathbf{u}$. If $\mathbf{u}$ is the zero vector, then $\operatorname{proj}_{\mathbf{u}}(\mathbf{v})$ is defined as the zero vector.

Given k vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$, the Gram–Schmidt process defines the vectors $\mathbf{u}_1, \ldots, \mathbf{u}_k$ as follows:

$$\mathbf{u}_1 = \mathbf{v}_1, \quad \mathbf{e}_1 = \frac{\mathbf{u}_1}{\|\mathbf{u}_1\|}$$

$$\mathbf{u}_2 = \mathbf{v}_2 - \operatorname{proj}_{\mathbf{u}_1}(\mathbf{v}_2), \quad \mathbf{e}_2 = \frac{\mathbf{u}_2}{\|\mathbf{u}_2\|}$$

$$\mathbf{u}_3 = \mathbf{v}_3 - \operatorname{proj}_{\mathbf{u}_1}(\mathbf{v}_3) - \operatorname{proj}_{\mathbf{u}_2}(\mathbf{v}_3), \quad \mathbf{e}_3 = \frac{\mathbf{u}_3}{\|\mathbf{u}_3\|}$$

$$\mathbf{u}_4 = \mathbf{v}_4 - \operatorname{proj}_{\mathbf{u}_1}(\mathbf{v}_4) - \operatorname{proj}_{\mathbf{u}_2}(\mathbf{v}_4) - \operatorname{proj}_{\mathbf{u}_3}(\mathbf{v}_4), \quad \mathbf{e}_4 = \frac{\mathbf{u}_4}{\|\mathbf{u}_4\|}$$

$$\vdots$$

$$\mathbf{u}_k = \mathbf{v}_k - \sum_{j=1}^{k-1} \operatorname{proj}_{\mathbf{u}_j}(\mathbf{v}_k), \quad \mathbf{e}_k = \frac{\mathbf{u}_k}{\|\mathbf{u}_k\|}.$$

The sequence $\mathbf{u}_1, \ldots, \mathbf{u}_k$ forms an orthogonal system, and the normalized vectors $\mathbf{e}_1, \ldots, \mathbf{e}_k$ create an orthonormal set. The calculation of $\mathbf{u}_1, \ldots, \mathbf{u}_k$ is known as Gram–Schmidt orthogonalization, while computing $\mathbf{e}_1, \ldots, \mathbf{e}_k$ is referred to as Gram–Schmidt orthonormalization.

7 Bibliography

[1] Hastie, Trevor, Tibshirani, Robert, and Friedman, Jerome. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer. DOI: 10.1007/978-0-387-84858-7

[2] James, Gareth, Witten, Daniela, Hastie, Trevor, and Tibshirani, Robert. (2021). *An Introduction to Statistical Learning: with Applications in R.* Springer. DOI: 10.1007/978-1-0716-1418-1

[3] Brunton, Steven L., and Kutz, J. Nathan. (2019). *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control.* Cambridge University Press. DOI: 10.1017/9781108380690

[4] Johnstone, Iain M., and Lu, Arthur Yu. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486), 682–693. DOI: 10.1198/jasa.2009.0121

[5] Greenberg, Edward. (1975). Minimum variance properties of principal component regression. *Journal of the American Statistical Association*, 70(349), 194–197. DOI: 10.1080/01621459.1975.10480287

[6] Agarwal, Anish, Shah, Devavrat, Shen, Dennis, and Song, Dogyoon. (2021). On robustness of principal component regression. *Journal of the American Statistical Association*, 116(536), 1731–1745. DOI: 10.1080/01621459.2021.1928513

[7] Rosipal, Roman, and Krämer, Nicole. (2006). Overview and recent advances in partial least squares. In Saunders, Craig, Grobelnik, Marko, Gunn, Steve, and Shawe-Taylor, John (Eds.), *Subspace, Latent Structure and Feature Selection* (pp. 34–51). Springer, Berlin, Heidelberg. DOI: 10.1007/11752790_2

[8] Garthwaite, Paul H. (1994). An interpretation of partial least squares. *Journal of the American Statistical Association*, 89(425), 122–127. DOI: 10.1080/01621459.1994.10476452

[9] Göktaş, Atila, and Akkuş, Özge. (2020). Comparison of partial least squares with other prediction methods via generated data. *Journal of Statistical Computation and Simulation*, 90(16), 3009–3024. DOI: 10.1080/00949655.2020.1793342

[10] Esposito Vinzi, Vincenzo, and Russolillo, Giorgio. (2013). Partial least squares algorithms and methods. *WIREs Computational Statistics*, 5(1), 1–19. DOI: 10.1002/wics.1239

[11] de Jong, Sijmen. (1993). SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18(3), 251–263. DOI: 10.1016/0169-7439(93)85002-X

[12] Lin, You-Wu, Deng, Bai-Chuan, Xu, Qing-Song, Yun, Yong-Huan, and Liang, Yi-Zeng. (2016). The equivalence of partial least squares and principal component regression in the sufficient dimension reduction framework. *Chemometrics and Intelligent Laboratory Systems*, 150, 58–64. DOI: 10.1016/j.chemolab.2015.11.003