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

Khabibullo Ibadullaev, Bekhzod Abulov

February 17, 2025

University of Bonn
Research Module in Econometrics and Statistics
Supervised by: Prof. Dr. Joachim Freyberger

Winter Semester 2024/2025

Contents

1 Introduction **1**

2 Theory **2**

2.1 Principal Component Regression (PCR) 2

2.1.1 Variance Reduction and Bias in PCR 3

2.1.2 Reconstructing β Coefficients in PCR 3

2.1.3 Unsupervised Nature of PCR 4

2.1.4 Conclusion 4

2.2 Partial Least Squares (PLS) 5

2.2.1 Overview 5

2.2.2 Mathematical Formulation 5

2.2.3 Bias-Variance Tradeoff in PLS 6

2.2.4 NIPALS Algorithm 6

2.2.5 SIMPLS Algorithm 7

2.2.6 Conclusion 7

2.3 Selection of the Optimal Number of Components in PCR and PLS 8

2.4 Elbow Method for PCR and PLS 8

2.4.1 Elbow Method for Principal Component Regression (PCR) 9

2.4.2 Elbow Method for Partial Least Squares (PLS) 9

2.4.3 Interpretation and Practical Considerations 10

2.4.4 Needle Algorithm for PCR and PLS 10

2.4.5 Interpretation and Practical Considerations 11

2.4.6 Cross-Validation for PCR and PLS 11

2.4.7 Practical Considerations 13

2.4.8 Comparison of PCR and PLS using Cross-Validation 13

2.4.9 Final Recommendation 13

2.4.10 Comparison of Methods 13

2.5 Final Thoughts on Component Selection 13

3 Empirical Methodology **14**

3.1 Application of PCR 14

3.2 Application of PLS 14

4 Simulation Study **14**

4.1 Motivation 14

4.2 Data Description 15

4.3 Simulation Results 16

5 Empirical Data Description **16**

5.1 Preprocessing 17

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:

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

Applying the eigen-decomposition:

$$X^T X = Q \Lambda Q^T$$

we obtain:

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

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 $\hat{\beta}_{\mathrm{OLS}}$.

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. 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):

$$X = U\Sigma V^T$$

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:

$$Z = XV_k$$

where 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.

3. Perform OLS regression on the transformed dataset:

$$Y = Z\gamma + \epsilon$$

where γ represents the regression coefficients associated with the principal components.

2.1.1 Variance Reduction and Bias in PCR

A key feature of PCR is its ability to reduce variance in estimated coefficients by eliminating components associated with small eigenvalues of $X^T X$. PCR estimates the coefficients as:

$$\hat{\beta}_{\text{PCR}} = V_k \Lambda_k^{-1} V_k^T X^T Y$$

where Λ_k contains only the top k eigenvalues of $X^T X$. 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 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.

2.1.2 Reconstructing β Coefficients in PCR

The standard OLS solution is given by:

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

which is sensitive to multicollinearity. To compare PCR coefficients with OLS, we rewrite the original regression coefficients as:

$$\beta = V \Lambda^{-1} V^T X^T Y$$

Since PCR retains only k components, we approximate β as:

$$\beta_{\text{PCR}} = V_k \Lambda_k^{-1} V_k^T X^T Y$$

which confirms that PCR coefficients are derived from OLS but exclude components with small variance contributions.

Convert the estimated coefficients back to the original predictor space:

$$\hat{\beta}_{\text{OLS}} = V_k \hat{\gamma}$$

ensuring that the final regression model remains interpretable in terms of the original predictor variables.

2.1.3 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:

$$X^T X = V \Lambda V^T$$

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.

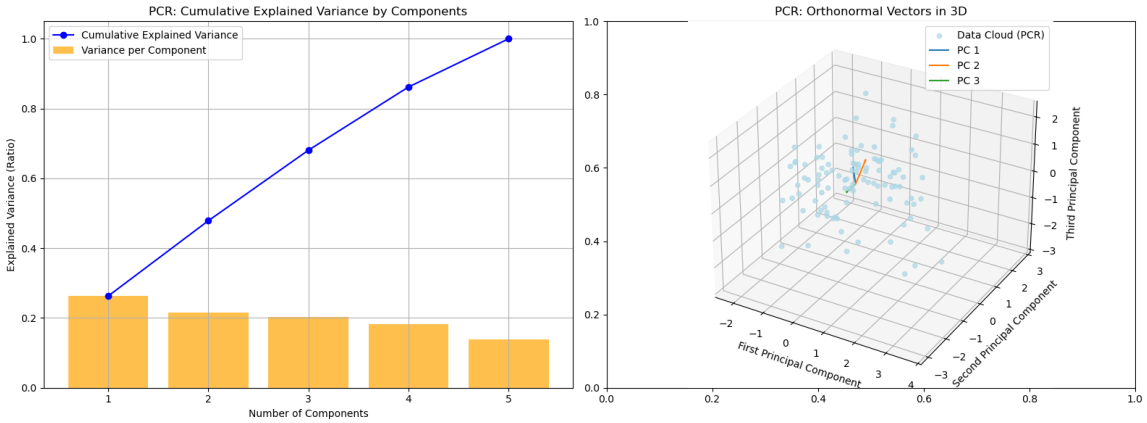

Figure 1: Principal Component Regression (PCR) Analysis Results

2.1.4 Conclusion

Principal Component Regression provides a robust approach for handling multicollinearity and high-dimensional data. By selecting principal components that capture the most variance, PCR ensures stable coefficient estimates with reduced variance. 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)

2.2.1 Overview

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.

PLS constructs latent variables (LVs) that maximize the covariance between the predictor (X) and response (Y) matrices. 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.

2.2.2 Mathematical Formulation

In the standard multiple linear regression framework, the relationship between the response Y and predictor matrix X is given by:

$$Y = X\beta + \epsilon$$

where:

- Y is the $n \times 1$ response vector,
- X is the $n \times p$ predictor matrix,
- β is the $p \times 1$ coefficient vector to be estimated,
- ϵ is the $n \times 1$ error term, assumed to be normally distributed with mean zero and variance $\sigma^2 I$.

Unlike OLS, which directly estimates β as:

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

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

$$T = XW, \quad U = YQ$$

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.

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

$$Y = TB + \epsilon$$

where B represents the regression coefficients for the latent variables. The estimated regression coefficients for the original predictors can be obtained as:

$$\hat{\beta}_{\text{PLS}} = W(P^T W)^{-1} B$$

2.2.3 Bias-Variance Tradeoff in PLS

PLS naturally introduces regularization by selecting a reduced set of components. The tradeoff between bias and variance is evident in:

$$\hat{\beta}_{\text{PLS}} = W(P^T W)^{-1} B$$

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

2.2.4 NIPALS Algorithm

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:
$$w = \frac{X^T u}{\|X^T u\|}$$
3. Compute the latent score vector:
$$t = Xw$$
4. Compute the loading vector:
$$p = \frac{X^T t}{t^T t}$$
5. Compute the regression coefficient:
$$b = \frac{u^T t}{t^T t}$$
6. Deflate X and Y:
$$X = X - tp^T, \quad Y = Y - btq^T$$
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, 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.

2.2.5 SIMPLS Algorithm

An alternative to NIPALS is the SIMPLS algorithm, which extracts PLS components without iterative deflation, making it computationally more efficient.

1. Compute the covariance matrix:
$$S = X^T Y$$
2. Extract the first singular vector r_1 from SVD of S.
3. Compute the score vector:
$$t_1 = X r_1$$
4. Compute the loading:
$$p_1 = X^T t_1 / (t_1^T t_1)$$
5. Compute the regression coefficient:
$$b_1 = (t_1^T Y)/(t_1^T t_1)$$
6. Deflate the covariance matrix:
$$S = S - r_1 (r_1^T S)$$
7. Repeat for additional components.

SIMPLS is preferred when computational efficiency is a concern, especially for very large datasets.

2.2.6 Conclusion

Partial Least Squares Regression offers a robust alternative to standard regression techniques by integrating dimensionality reduction with predictive modeling. By selecting components that maximize

covariance with Y, PLS ensures better predictive performance than PCR, which focuses solely on variance preservation in X. Its ability to address multicollinearity and balance the bias-variance tradeoff makes it an ideal method for high-dimensional regression problems.

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.

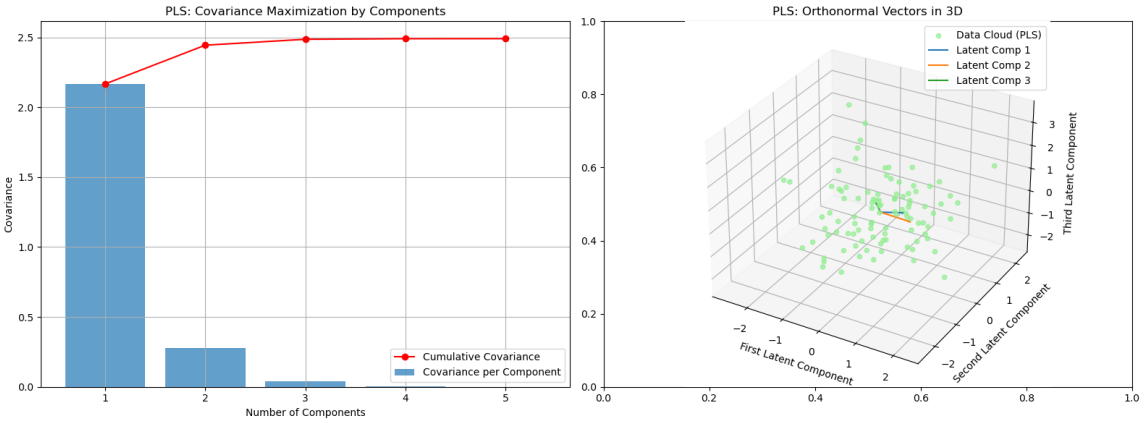

Figure 2: Partial Least Squares (PLS) Analysis Results

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. 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. Various methods are available to guide this selection:

2.4 Elbow Method for PCR and PLS

The Elbow Method is a widely used heuristic 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. The optimal number of components is chosen at the "elbow point", where the marginal gain in variance or covariance explained diminishes significantly.

2.4.1 Elbow Method for Principal Component Regression (PCR)

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:

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

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:

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

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

2.4.2 Elbow Method for Partial Least Squares (PLS)

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:

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

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:

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

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

2.4.3 Interpretation and Practical Considerations

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.

In summary: - For PCR, components are selected based on explained variance, ensuring that the most important directions in X are retained. - For PLS, components are selected based on explained covariance, ensuring that the most predictive directions for Y are retained.

A combination of these methods, along with domain knowledge, should guide the final selection of k to balance model complexity and predictive performance.

2.4.4 Needle Algorithm for PCR and PLS

The Needle Algorithm is an alternative method for determining the optimal number of components in Principal Component Regression (PCR) and Partial Least Squares (PLS). Unlike the Elbow Method, which relies on a visual inspection of a variance or covariance plot, the Needle Algorithm 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.

Needle Algorithm for Principal Component Regression (PCR) In Principal Component Regression (PCR), the goal is to select the number of components k that retain most of the variance in X while avoiding unnecessary complexity. 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:

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

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

The **Needle Criterion** is defined as:

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

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

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

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

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

Needle Algorithm for Partial Least Squares (PLS) For Partial Least Squares (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:

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

where t_i are the latent variables extracted by PLS.

The Needle Criterion for PLS is:

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

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

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

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

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

2.4.5 Interpretation and Practical Considerations

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." However:

- **For PCR**, it helps determine the number of principal components that explain sufficient variance in X.
- **For PLS**, it selects components based on their predictive power by maximizing the covariance between X and Y.
- The threshold ϵ should be chosen carefully to balance between model complexity and predictive accuracy.

A combination of methods, including the Elbow Method, Needle Algorithm, and Cross-Validation, is recommended to ensure an optimal choice of k.

2.4.6 Cross-Validation for PCR and PLS

Cross-Validation (CV) is a data-driven approach to selecting the optimal number of components in Principal Component Regression (PCR) and Partial Least Squares (PLS). Unlike heuristic methods such as 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.

Cross-Validation for Principal Component Regression (PCR) 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:

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

   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):

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

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.

Cross-Validation for Partial Least Squares (PLS) Unlike PCR, where components are selected based only on variance in X, 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:

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

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

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

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

2.4.7 Practical Considerations

While cross-validation is the most rigorous method for selecting k, it has some practical considerations:

- **Computational Cost**: Cross-validation is computationally expensive since it requires training multiple models for different values of k across multiple folds.
- **Bias-Variance Tradeoff**: Smaller k values lead to higher bias, while larger k values lead to higher variance. CV helps find the optimal balance.
- **Choice of K**: Typically, $K = 5$ or $K = 10$ is used. A smaller K reduces computation time, but larger K provides a more reliable error estimate.

2.4.8 Comparison of PCR and PLS using Cross-Validation

- **PCR** often requires more components than PLS to achieve the same predictive performance because it does not consider Y when selecting components.
- **PLS** typically achieves a lower optimal k^* due to its supervised nature, which aligns components with Y.
- Both methods benefit from CV in avoiding overfitting by selecting an appropriate number of components.

2.4.9 Final Recommendation

Since CV provides a direct measure of predictive performance, it is considered the gold standard for selecting the number of components in both PCR and PLS, therefore CV is applied in our simulations and empirical estimation.

2.4.10 Comparison of Methods

- The **Elbow Method** is computationally simple but may be subjective.
- The **Needle Algorithm** provides a more formal criterion but may still require expert judgment.
- **Cross-Validation** is the most robust, as it directly evaluates predictive performance, but it is computationally expensive.

2.5 Final Thoughts on Component Selection

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.

3 Empirical Methodology

To assess the predictive performance of Principal Component Regression (PCR) and Partial Least Squares (PLS), we implement the following steps:

3.1 Application of PCR

1. Compute the **Principal Component Analysis (PCA)** on the predictor matrix X.
2. Select the first k principal components that explain at least 95% of the variance.
3. Fit an **OLS regression** model using the selected components as predictors.
4. Evaluate the model using **Root Mean Squared Error (RMSE)** and **Mean Squared Error (MSE)**.

3.2 Application of PLS

1. Use the **NIPALS algorithm** to compute PLS components.
2. Select the optimal number of latent components using cross-validation.
3. Fit a **PLS regression** model on the transformed predictor matrix.
4. Evaluate the model performance using **RMSE** and **MSE**.

4 Simulation Study

4.1 Motivation

The purpose of this simulation study is to assess the performance of **Principal Component Regression (PCR)** and **Partial Least Squares (PLS)** under different data-generating scenarios. Specifically, we aim to analyze the predictive accuracy and robustness of both methods in the presence of **multicollinearity** and **high-dimensional predictor spaces**. The study follows a Monte Carlo simulation approach, allowing for a systematic evaluation of bias-variance tradeoffs and mean squared prediction error (MSPE).

The simulation study is structured to mimic real-world high-dimensional regression problems with three distinct setups:

1. **Classical Data (No Multicollinearity):** Predictors are independently generated from a normal distribution.
2. **Moderate Multicollinearity:** Predictors consist of base features and additional highly correlated features.
3. **Severe Multicollinearity (Low Observations):** Predictors exhibit strong correlation among features, and only a subset of them contribute to generating the response variable.

4.2 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.

Impact of Reduced Observations and Multicollinearity 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.

Each scenario is replicated 1000 times for statistical reliability.

4.3 Simulation Results

5 Empirical Data Description

For the empirical analysis, we utilize a dataset that exhibits high-dimensional characteristics, commonly found in fields such as chemometrics and finance. Specifically, we analyze a dataset containing spectral data, where the number of predictors (p) is significantly larger than the number of observations (n).

The dataset consists of:

- **Response Variable (Y)**: A continuous outcome variable, representing a target measurement (e.g., concentration of a chemical compound or financial index value).

- **Predictors (X)**: A set of highly correlated explanatory variables, typically derived from spectral wavelengths or macroeconomic indicators.

- **Sample Size (n)**: 100 observations.

- **Number of Predictors (p)**: 200 variables.

5.1 Preprocessing

Before applying regression models, we preprocess the data by:

- Standardizing all predictors to have zero mean and unit variance.
- Splitting the dataset into **training** (80%) and **testing** (20%) subsets.
- Applying cross-validation for model selection.