

PCR and PLS: Methods for Dimensionality Reduction in High-Dimensional Data

Al Afia Ramzi

Guillemain Alexandre

Nowicki Yann

Contents

Introduction	2
1 Context: The Problem of High Dimensionality	2
2 Description of Principal Component Regression (PCR)	2
2.1 Step 1: Dimensionality Reduction with PCA (Unsupervised)	2
2.2 Step 2: Regression (Supervised)	3
2.3 The Fundamental Assumption... and its Major Risk	3
2.4 Practical Implementation: Standardization and Cross-Validation	3
2.5 Distinction from Selection Methods (Lasso)	4
3 Description of Partial Least Squares (PLS)	4
3.1 Constructing the First Factor Z_1 (The Supervised Step)	4
3.2 The Solution to the PCR Problem	4
3.3 The Iterative Process and Choosing M	5
3.4 Theoretical Justification and Properties	5
4 Summary: PCR vs. PLS Comparison	5

Introduction

The purpose of this document is to introduce Principal Component Regression (PCR) and Partial Least Squares (PLS) regression models. We aim to understand their implementation, objectives, and fundamental differences, without seeking a "miracle solution" that blindly answers all questions.

1 Context: The Problem of High Dimensionality

In modern data analysis, it is common to encounter situations of high dimensionality where the number of variables (predictors, p) is very large, often significantly exceeding the number of observations (n).

This scenario, denoted as $p \gg n$, poses two major problems for classical statistical modeling:

- **High Multicollinearity:** When many variables are present for few observations, it is almost guaranteed that groups of variables will be highly correlated. It becomes statistically unstable, or even impossible, to isolate the individual effect of each variable on the response Y .
- **Overfitting:** With more predictors than observations, a linear model can "memorize" the training data and achieve zero error. However, this model learns the *noise* rather than the *signal*. Consequently, its predictive performance collapses when exposed to new data.

To bypass these obstacles, we must avoid using the p variables directly. A robust solution is dimensionality reduction.

Rather than using complex "black box" models (like deep learning), which are difficult to interpret, this project focuses on latent variable models. Principal Component Regression (PCR) and Partial Least Squares (PLS) are two fundamental methods in this family, designed to extract a small number of relevant and interpretable components from high-dimensional data.

2 Description of Principal Component Regression (PCR)

Principal Component Regression (PCR) is a two-step statistical method. It does not seek to select variables, but rather to transform them into a smaller, uncorrelated set of components.

2.1 Step 1: Dimensionality Reduction with PCA (Unsupervised)

The "engine" of PCR is Principal Component Analysis (PCA), a fundamental dimensionality reduction technique initially developed by Karl Pearson (1901) and formalized by Harold Hotelling (1930).

PCA is an unsupervised method. Its objective is to summarize a large set of predictors X (say $p = 50$ variables) into a new, smaller set of M variables (say $M = 5$), called principal

components $(Z_1, \dots, Z_M)$.

- Each component Z is a linear combination (a "blend") of all original p variables in X .
- These components are constructed to capture the maximum variance of the data X :
 - Z_1 is the axis that captures the maximum possible variance.
 - Z_2 captures the maximum *remaining* variance while being orthogonal (uncorrelated) to Z_1 .
 - ...and so on.

2.2 Step 2: Regression (Supervised)

Once these M principal components (the Z 's) are created, PCR performs a classical linear regression. We "forget" the original p variables X and predict Y using only these M new components as predictors:

$$Y = \beta_0 + \beta_1 Z_1 + \dots + \beta_M Z_M + \epsilon$$

2.3 The Fundamental Assumption... and its Major Risk

PCR relies on a strong fundamental assumption:

The assumption of PCR is that the directions explaining the most variance in the predictors (X) are also the most relevant directions for predicting the response (Y).

This is the major risk of PCR. Step 1 (PCA) is "blind": it constructs the components Z by looking *only* at X , without ever taking into account the variable Y we want to predict.

There is no guarantee that the first component Z_1 (the one with the most variance) is the best for predicting Y . It is entirely possible that the most important "signal" for Y lies in an X variable with very low variance, which PCA might treat as "noise" and discard.

2.4 Practical Implementation: Standardization and Cross-Validation

For PCR to function correctly, two points are essential:

1. **Standardize the data:** Since PCA is based on variance, it is imperative to center and scale (standardize) all X variables beforehand. Without this, an "Expense" variable (ranging from 1 to 90,000) would completely crush the variance of an "Age" variable (ranging from 20 to 80), distorting the analysis.
2. **Choose M via Cross-Validation:** How do we know whether to keep $M = 3$ or $M = 10$ components? We use Cross-Validation. This method tests the predictive performance (MSE error) of the model for every possible M . We choose the M that yields the lowest prediction error, finding the best bias-variance balance.

2.5 Distinction from Selection Methods (Lasso)

Finally, it is important to note that PCR is **not a variable selection method**, unlike Lasso.

- **Lasso** selects a subset of X variables and sets the coefficients of others to zero (feature selection).
- **PCR** transforms the variables. It keeps all original p variables X but "*blends*" them (via linear combinations) into a smaller number of components Z (dimensionality reduction).

3 Description of Partial Least Squares (PLS)

Given the risk of PCR (its "blind" fundamental assumption), Partial Least Squares (PLS) Regression offers a supervised alternative.

Developed in the 1980s (notably by S. Wold) for chemometrics, PLS was specifically designed for prediction in $p \gg n$ contexts.

Its objective is not to maximize the *variance* of X (like PCR), but to maximize the covariance between the predictors X and the response Y . It accomplishes a "double mission": finding components that summarize X while being maximally linked to Y .

3.1 Constructing the First Factor Z_1 (The Supervised Step)

Like PCR, PLS creates components (called "latent factors") that are linear combinations of X :

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{p1}X_p$$

But to determine the weights (ϕ), PLS does not use the variance of X alone. More formally, the first weight vector w_1 (containing the ϕ_j) is directly proportional to the empirical covariance vector $\hat{\sigma} = X^TY/n$.

Conceptually, this means it "auditions" each predictor X_j :

- Variables X_j that have the strongest correlation (covariance) with Y will be assigned the highest weight (ϕ_{j1}) in the construction of Z_1 .
- An X_k variable uncorrelated with Y (even if it has high variance) will receive a low or zero weight.

Thus, Z_1 is, by construction, the direction (the "blend" of X) most aligned with the target Y .

3.2 The Solution to the PCR Problem

PLS directly solves the problem of PCR's fundamental assumption.

If a relevant "signal" for Y has low variance (the failure case for PCR), it does not matter. As long as this signal has a strong correlation with Y , PLS will detect it, assign it a high weight, and include it in its first factor Z_1 .

3.3 The Iterative Process and Choosing M

To find the subsequent components $(Z_2, Z_3, \dots)$, PLS uses an iterative approach:

1. It computes Z_1 (based on $X^T Y$).
2. It "cleans" the data X and Y of the information already explained by Z_1 . This is the deflation step (taking residuals).
3. It repeats the process on these residuals to find Z_2 , and so on.

As with PCR, the optimal number of components M (or K) to keep is a crucial hyperparameter. It is determined by Cross-Validation by choosing the M that minimizes prediction error (MSE).

3.4 Theoretical Justification and Properties

Although the principle is simple, theoretical analysis is complex because the PLS estimator is inherently *non-linear* (since it depends on Y for its construction).

Recent studies, such as Castelli et al. (2024), provide non-asymptotic analysis of PLS, which is essential to justify its performance in the $p \gg n$ setting.

These mathematical analyses confirm several key points:

- **Signal Condition:** PLS is only relevant if the signal-to-noise ratio is sufficient. The algorithm must detect a signal in the first component that exceeds a certain noise threshold.
- **Error Decomposition:** Under this condition, the prediction error is bounded and can be formally decomposed into a *bias* term (due to dimensionality reduction) and a *variance* term (due to noise ϵ).
- **Performance Interpretation:** Estimator performance depends on what the paper calls "inverse relative inertia". Simply put: prediction is better if the first PLS component (Z_1) captures a large part of the total inertia (variance) of the covariates.
- **Foundation for Sparsity:** This theoretical approach allows extending PLS to **Sparse PLS (sPLS)**, which integrates a LASSO-type penalty to perform simultaneous variable selection. It is proven that this method achieves performance guarantees similar to LASSO.

4 Summary: PCR vs. PLS Comparison

The following table summarizes the key distinctions between the two methods:

graphicx caption float

Feature	PCR (Principal Component Reg.)	PLS (Partial Squares)	Least
Nature	Unsupervised Step + Supervised Step	Fully Supervised	
Objective	Maximize $\text{Var}(X)$	Maximize $\text{Cov}(X, Y)$	
Computation	Simultaneous (Eigendecomposition)	Iterative (Deflation)	
Risk	May discard low-variance predictive signals	Slightly higher variance than PCR	
Best Use	When X -variance drives the process	When specific $X - Y$ correlations matter	

Table 1: Comparison of PCR and PLS Methods

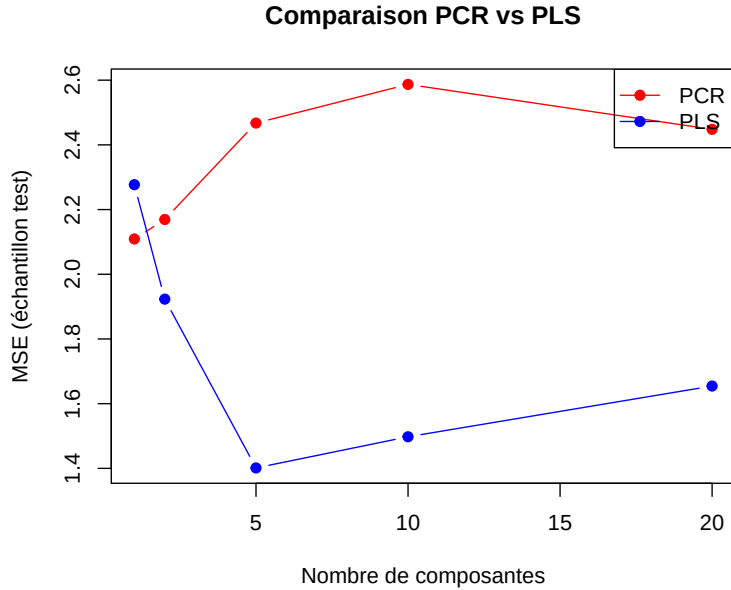


Figure 1. Test mean squared error (MSE) for Principal Component Regression (PCR) and Partial Least Squares (PLS) as a function of the number of components.

This figure provides a numerical illustration of the conceptual difference between PCR and PLS. In this simulated setting, the predictive signal lies in a low-variance direction of the covariates. As a consequence, PCR which relies on variance-based dimension reduction exhibits poor predictive performance for a small number of components, as relevant information may be discarded. In contrast, PLS explicitly accounts for the relationship between the predictors and the response, allowing it to recover the predictive signal already in the first components. In practical terms, this means that when the goal is to make reliable predictions, methods that take into account the relationship between the data and the target variable such as PLS can be more appropriate than approaches that focus solely on summarizing data variability, like PCR.