What is Principal Components Regression?

Authored by

stats writer

December 18, 2025

Principal Components Regression (PCR) is a powerful technique utilized in statistical modeling and machine learning, specifically designed to address challenges arising from high-dimensional datasets. At its core, PCR integrates two fundamental statistical methodologies: Principal Component Analysis (PCA) for feature extraction and linear regression for predictive modeling. This method involves transforming the original set of highly correlated predictor variables into a smaller set of uncorrelated variables, known as the **principal components**. These components capture the maximum possible variance of the original data.

The primary objective of PCR is to perform effective dimension reduction before fitting a regression model. By focusing on the components that explain the most variation in the features, PCR simplifies the modeling task. This approach proves particularly valuable when dealing with scenarios where the number of predictors (p) is large, or even exceeds the number of observations (n). Furthermore, PCR is an excellent tool for mitigating issues caused by multicollinearity--a condition where predictor variables are highly correlated with one another, often leading to unstable and unreliable regression coefficient estimates.

Understanding the Multicollinearity Challenge

One of the most persistent and detrimental issues that data scientists encounter when constructing robust predictive models is high levels of multicollinearity. This statistical phenomenon occurs when two or more independent (predictor) variables in a dataset exhibit a strong linear relationship. While multicollinearity does not necessarily affect the predictive power of the model as a whole, it drastically inflates the variance of the coefficient estimates, making it difficult to interpret the individual impact of each predictor.

When coefficient estimates are unstable due to high correlation among features, the model becomes highly sensitive to minor fluctuations in the training data. This sensitivity significantly increases the risk of overfitting. An overfit model performs exceptionally well on the data it was trained on but fails spectacularly when introduced to a new, previously unseen dataset. The model essentially memorizes the noise and specific idiosyncrasies of the training set rather than learning the underlying, generalized relationship between the predictors and the response variable.

To combat this, various strategies have been developed, broadly categorized into methods involving selection, regularization, and dimension reduction. While selection and regularization directly manipulate or constrain the coefficient values of the original predictors, PCR offers an entirely distinct approach by fundamentally changing the input variables themselves before the regression modeling even begins.

Addressing High Dimensionality: Subset Selection vs. Regularization

When faced with excessive predictors or high correlation, two traditional methods are often employed before turning to techniques like PCR. The first category is **subset selection**, which aims to identify and retain only the most impactful predictors while discarding the irrelevant or redundant ones. This results in a simpler, more interpretable model.

Common techniques within subset selection include:

Best subset selection
Stepwise selection

These methods seek to improve model generalization by eliminating unnecessary variance introduced by irrelevant features. However, if the correlation among the remaining features remains high, the stability of the coefficient estimates may still be compromised.

The second major category involves regularization, often referred to as shrinkage methods. These techniques constrain or penalize the magnitude of the model coefficients during the fitting process, effectively biasing the estimates towards zero. This reduction in coefficient magnitude serves to decrease model variance, leading to better out-of-sample performance.

Popular regularization methods include:

Ridge Regression
Lasso Regression

While highly effective, regularization methods still operate using the original, potentially correlated predictors. Principal Components Regression, in contrast, bypasses the issue of feature correlation entirely by creating entirely new, uncorrelated features upon which the regression is built.

The Detailed Mechanism of Principal Components Regression

Understanding the flow of PCR requires detailing the three primary stages, from feature transformation to model fitting. Suppose we have a dataset with p original predictors: $X_1, X_2, dots, X_p$. The goal is to predict a response variable Y.

The process begins with the calculation of the principal components $Z_1, dots, Z_M$, where $M < p$. Each component Z_m is derived as a linear combination of the original variables, weighted by coefficients (loadings) Phi_{jm}. This approach is fundamentally known as **dimension reduction**.

The construction of these components works as follows:

Suppose a given dataset contains p predictors: $X_1, X_2, dots , X_p$.

Calculate $Z_1, dots , Z_M$ to be the M linear combinations of the original p predictors.

$Z_m = sum_{j=1}^{p} Phi_{jm}X_j$ for some constants $Phi_{1m}, Phi_{2m}, dots, Phi_{pm}$, where $m = 1, dots, M$.
Z_1 is the linear combination of the predictors that captures the most variance possible.
Z_2 is the next linear combination of the predictors that captures the most variance while being orthogonal (i.e., uncorrelated) to Z_1.
Z_3 is then the next linear combination of the predictors that captures the most variance while being orthogonal to Z_1 and Z_2.
And so on, ensuring all selected principal components are mutually orthogonal.

Use the method of least squares to fit a linear regression model using the first M principal components $Z_1, dots, Z_M$ as predictors.

The core benefit of **dimension reduction** comes from the fact that this method only has to estimate $M+1$ coefficients (one intercept and M slopes) instead of the original $p+1$ coefficients, where $M < p$. In many cases where multicollinearity is present in a dataset, principal components regression is able to produce a more stable and generalized model than conventional multiple linear regression.

Practical Steps for Implementing Principal Components Regression

In practice, the following systematic steps are used to perform principal components regression effectively:

Standardize the Predictors.

First, we typically standardize the data such that each predictor variable has a mean value of 0 and a standard deviation of 1. This crucial step prevents one predictor from being overly influential in the PCA decomposition, especially if predictors are measured in vastly different units (e.g., if X_1 is measured in inches and X_2 is measured in yards). Standardization ensures fair weighting based on underlying variability.

Calculate the Principal Components and Perform Linear Regression.

Next, we calculate the principal components Z and use the method of least squares to fit a linear regression model using the first M principal components $Z_1, dots, Z_M$ as predictors.

Decide How Many Principal Components to Keep (M).

Finally, we use resampling techniques, typically k-fold cross-validation, to find the optimal number

of principal components (M) to retain in the model. The "optimal" number of components is determined empirically; it is the M value that produces the lowest test mean-squared error (MSE), indicating the best performance when generalizing to unseen data. This step balances the complexity of the model against predictive accuracy.

Advantages of Principal Components Regression

Principal Components Regression (PCR) offers the following significant **pros**:

PCR tends to perform exceptionally well when the first few principal components are able to capture most of the variation in the predictors along with the relationship with the response variable.
PCR can perform well even when the predictor variables are highly correlated because it produces principal components that are orthogonal (i.e., uncorrelated) to each other, thus resolving the coefficient instability caused by multicollinearity.
PCR doesn't require manual variable selection; since each principal component is a linear combination of all original predictor variables, information from all features is implicitly utilized.
PCR can be used when there are more predictor variables (p) than observations (n), a high-dimensional scenario that makes standard multiple linear regression impossible to estimate.

Key Limitations of Principal Components Regression

However, PCR comes with one major methodological **con**:

PCR does not consider the response variable (Y) when deciding which principal components to retain or drop. Instead, the PCA step is entirely unsupervised, focused only on maximizing the magnitude of the variance among the predictor variables (X). It is theoretically possible that the principal components associated with the largest variances are not the components that best predict the response variable, leading to sub-optimal modeling if a low-variance component holds significant predictive power.

This limitation means that while PCR effectively tackles multicollinearity and high dimensionality, its component selection criteria are blind to the predictive target. For this reason, rigorous tuning using k-fold cross-validation is essential to mitigate the risk of inadvertently discarding important predictive information.

Conclusion: Finding the Optimal Model

In practice, the best modeling strategy involves fitting many different types of models (PCR, Ridge, Lasso, Multiple Linear Regression, etc.) and using robust validation techniques like k-fold cross-validation to identify the model that produces the lowest test MSE on new data.

In cases where multicollinearity is present in the original dataset (which is common in real-world data), Principal Components Regression tends to perform better than ordinary least squares regression because it dramatically reduces variance and prevents overfitting. However, it is always a sound strategy to fit several different models so that you can empirically identify the one that generalizes best to unseen data.

PCR in Practice: Tools and Resources

The following tutorials show how to perform principal components regression in R and Python using standard data science libraries:

Principal Components Regression in R (Step-by-Step)
Principal Components Regression in Python (Step-by-Step)