AL KUT JOURNAL OF ECONOMIC AND ADMINISTRATIVE SCIENCES
Publisher: College of Economics and Management - Wasit University

Using Principal Component Analysis (PCA) Techniques to Improve Multiple Regression Performance in High-Dimensional Data

Dr .Amal Sadeq Hamoodi

Mathematics Department College of Education /Mustansiriya University

Abstract:

High-dimensional data pose a significant problem in statistical modeling, as the increase in the number of predictors frequently results in multicollinearity and instability in the estimations of multiple regression coefficients. Principal Component Analysis (PCA) is regarded as an efficient statistical method for resolving this issue by converting correlated variables into a set of uncorrelated components while preserving the greatest variance in the data. This study seeks to assess the efficacy of Principal Component Analysis (PCA) in enhancing multiple regression performance in high-dimensional contexts, utilizing both synthetic simulated data (300–600 observations, 50–300 predictors) and real-world data (Coffee Quality Dataset, 1,339 observations, 43 variables). The outcomes of Principal Component Regression (PCR) were juxtaposed with Ordinary Least Squares (OLS) and regularized models, including Ridge, Lasso, and Elastic Net. The models were assessed utilizing standard statistical metrics, such as the Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2), along with cross-validation. The results indicated that implementing PCA before regression significantly mitigated multicollinearity and enhanced the models' accuracy and stability, underscoring PCA as a viable and useful method for managing high-dimensional data and improving predictive modeling.

Keywords: Principal Component Analysis (PCA), Multiple Regression, High-Dimensional Data, Dimension Reduction, Multicollinearity.

1. Introduction

With the rapid development of data collection techniques and the diversity of data sources in fields such as medicine, economics, and artificial intelligence, it has become commonplace to deal with high-dimensional data comprising tens or hundreds of explanatory variables. This large number of variables is usually accompanied by strong correlations between them, which is known as multicollinearity.

This leads to fundamental problems in multiple regression models, as it makes coefficient estimates unstable and affects prediction accuracy, and may even sometimes lead to exaggerated and inconsistent regression coefficients when using different samples [4,6,7].

The theoretical basis for this technique was laid in the early 20th century by Pearson [1], and then developed more systematically and rigorously by Hoteling [2]. This was followed by numerous contributions that deepened its practical applications and made it an essential part of applied statistics and data analysis [3,4].

In addition to the traditional use of PCA, several developments have been made to this technique to address challenges specific to modern data. For example, methods based on Bayesian PCA have been developed to handle missing data more efficiently [10], and supervised PCA has been employed to enhance predictive capabilities when variables directly related to the output are available [11].

PCA techniques have also been used in the analysis of high-dimensional medical data such as genome and proteome data, where they have proven useful in predicting patient survival. PCA has also been compared to more recent techniques such as Manifold Elastic Net, which provide a broader framework for selective dimensionality reduction.

Furthermore, PCA has remained a fundamental tool in multivariate statistics, particularly in industrial process control and atmospheric control fields [12].

In recent years, PCA has grown in importance as a central analytical tool in high-dimensional data processing, especially with the emergence of modern techniques that have developed this classic method to meet new challenges in data analysis.

Recent studies have shown that PCA is no longer limited to dimension reduction, but has become a means of understanding the internal structure of multi-source data and providing accurate visual representations of variance patterns [15].

It has also been noted that combining PCA with statistical regularization methods contributes to a better balance between dimensionality reduction and improved predictive model accuracy, especially in environments suffering from overfitting [17].

Recent applied studies have expanded the areas of application of PCA in regression models. Some research has used PCA with logistic regression to analyses clinical factors in medical fields, demonstrating that incorporating PCA improves model accuracy and reduces the effect of multicollinearity [18]. In the engineering field, a comparison was made between principal component regression (PCR) and partial least squares (PLS) for estimating flight loads, and the results showed that PCR was superior in computational efficiency and stability of coefficients [20].

There has been an expansion in modern methods such as Sparse PCA, which imposes constraints on loads to make them more interpretable [14], while other researchers have introduced a more significant model called Non-Oblivious Adversarial Perturbations to counteract the effect of noise in data [16].

Through the development of PCA algorithms, recent studies have contributed to the integration of PCA with regularized regression techniques such as Penalized Regression, confirming that this integration improves predictive values and reduces mean square error [19].

A new approach to the problem of missing data in PCA has also been developed, based on heterogeneous messiness analysis to reduce bias in component estimation [21]. Modern applications have not only been applied to static data, but have also extended to time series data and artificial intelligence systems. One modern model developed a combination of enhanced PCA and the Bagging algorithm to strengthen time series regression analysis, which helped reduce noise and improve predictive performance in time series [13].

The primary objective is to test the extent to which PCA supports linear multicollinearity and enhances model accuracy and stability by comparing traditional linear regression (OLS) with principal component regression (PCR), using multiple evaluation metrics such as the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE), in addition to cross-validation to ensure the reliability of the results.

There is a research gap in the scarcity of studies that combine theoretical analysis and practical application of principal component regression in real-world and complex data

contexts. The most important feature of this research is its major contribution in providing a systematic comparative analysis between traditional linear regression (OLS) and principal component regression (PCR) in the context of high-dimensional data, unlike many studies that have focused on the theoretical aspect of [3,4 ,9] PCA.

Hence, the importance of this research in providing a comparative applied analysis.

1.1 **Previous studies**

The uses of PCA have expanded in many practical and theoretical fields, and multiple developments have emerged, such as modified principal component analysis and sparse component methods.

Studies have shown that PCA is a central tool in multivariate data analysis, enabling researchers to reduce the number of variables while retaining as much of the underlying variance in the data as possible [36].

McCabe [5] explained that principal components help derive alternative variables that preserve the variance properties of the original data, making them useful in multivariate statistical modelling. [7].

In this field, Audigier et al. pointed out the effectiveness of integrating PCA with Bayesian regression in improving the handling of continuous and complex data [8], while subsequent studies have shown that integrating PCA with traditional regression methods reduces the effect of multicollinearity and increases the stability of coefficient estimates [9,10].

Several applied studies have emerged in recent years that have targeted the application of PCA and PCR in specialized fields. For example, Huang et al. used principal component analysis in conjunction with logistic regression to analyze medical data for patients with lupus nephritis, and the results showed that incorporating PCA improved the accuracy of the models and reduced the effect of linear interference between variables [18].

Yan, Yang, and Wan conducted an experimental comparison between principal component regression (PCR) and partial least squares (PLS) in calculating flight loads and concluded that PCR provides more stable and efficient results in terms of statistical performance [20] [13].

The combination of PCA and multiple regression (PCR) has also proven to be effective in dealing with statistical problems associated with linearity, making it an ideal choice for application in many fields, including medicine, industry, economics, and artificial intelligenc.

2. Methodology

2.1 Study Design

A special design was used in preparing the steps of this research. This design is dual, combining two types of data. The first type is simulation data, whose characteristics are controlled with high precision and have been specified. The second type is real-world application data from the field of food science, which has been specified.

The aim of this design is to determine the extent of the effect of linear multiplicity and data dimensionality change on the one hand, and to confirm the effectiveness of applying principal component analysis (PCA) in improving the accuracy of predictive models on the other.

2.2 Stages of PCA application

The stages of this application were adopted in accordance with the design approved in the research, as follows:

Stage 1 (preparation): During this stage, the measurements are standardized, meaning that for each variable:

the mean = 0 and the standard deviation = 1

Stage Two (Extraction): After stage one, preparations are made for stage two, which is the process of applying this PCA model to the variance-covariance matrix:

$$S = (1/(n-1))\ Xc^T\ Xc \qquad (2)$$

Where

Xc = is the data matrix after subtracting the mean (centralized matrix).

S= is the variance-covariance matrix.

$$S\ V = V\ \Lambda \qquad (3)$$

Where:

Λ= diagonal matrix containing the eigenvalues (representing the amount of variance explained by each principal component).

V=eigenvectors.

$Z = Xc\ V$ (4)

Where:

Z= New variables (principal components)

Component selection: According to the rule where the number of components is determined so that it covers ≥ 85% of the total variance.

$Var\_k = \lambda\_k / \Sigma \lambda\_j$, $CumulativeVar(k) = \Sigma\ Var\_i$ (5)

2.3 Sample and data used

The study relied on two types of data:

2.3.1 Simulation data:

The simulation process was designed to reflect the reality of high-dimensional data that includes explanatory variables that are interrelated to varying degrees, which is a real challenge in statistical analysis due to its direct impact on the stability of regression coefficients and prediction accuracy. Simulation data representation is an ideal environment for isolating influential variables.

According to the design adopted in this research, the simulation data was the first data in the design, where artificial data was generated according to the sample. The data used is as follows:

• Number of explanatory variables (p): 50 to 300 variables.

• Sample size (n): 300 to 600 observations.

• Linear multicollinearity level: medium ($\rho = 0.5$) and high ($\rho = 0.8$).

Errors: Random errors distributed normally $\varepsilon \sim N(0,\sigma^2)$ were added to adjust the signal-to-noise ratio (SNR).

The data was formulated according to the equation:

$$y = X\beta + \varepsilon \quad , \quad \varepsilon \sim N(0, \sigma^2 I) \qquad (6)$$

2.3.2 Real data:

This research relied on the Coffee Quality Dataset available through the UCI Machine Learning Repository and certified by the Coffee Quality Institute. This data is widely used in academic studies and comprises 1,339 observations covering the period between 2010 and 2018, including samples from several coffee-producing countries around the world. This data includes a set of variables classified as follows:

First:

a set of physical-chemical explanatory variables containing 43 variables, including quantitative variables such as flavor, aroma, acidity, body, balance, sweetness, and moisture, as well as descriptive variables such as country of origin, farm name, processing method, and harvest year.

Second:

The dependent variable represents the coffee quality assessment (numerical values from 0 to 10). In addition to the 'Total Cup Points' used as the main indicator of quality in this study.

This sample is considered suitable because it contains a large number of notable correlations between explanatory variables, which represents a suitable environment for testing the effectiveness of principal component analysis (PCA) in reducing the effect of linear multicollinearity.

This data is used globally as a standard example for dimensionality reduction experiments and for analyzing and improving regression prediction models.

Table 1. Simulation data illustrating the structure of explanatory variables generated under specific correlation and noise conditions

Obs	Var1	Var2	Var3	Var4	Var5
1	0.56	-1.23	0.45	1.12	-0.34
2	-0.78	0.67	-1.02	0.21	0.89
3	1.34	-0.56	0.87	-0.44	0.15
4	0.12	0.89	-0.23	0.55	-1.45
5	-0.34	1.21	0.66	-0.78	0.43
6	0.89	-0.32	-0.12	1.01	0.56
7	-1.01	0.43	0.99	-0.12	0.77
8	0.45	-0.78	1.11	0.65	-0.21
9	-0.56	0.34	-0.45	1.34	0.23
10	1.12	-1.01	0.56	-0.67	0.98

Table 1. shows a sample of simulated data generated specifically to test the performance of traditional linear regression (OLS) and principal component regression (PCR) models in an environment with obvious multicollinearity and predefined errors.

The table above shows that the simulation data consists of a specific number of independent variables X1, X2, X3, ... , and the dependent variable

Y. The independent variables were generated according to a variance matrix designed so that the correlation coefficient between the variables is high (e.g., r = 0.8 or higher).

This step is essential for testing the sensitivity of each model to the effect of errors and the internal relationships between variables.

This result confirms that using PCA before regression can effectively simplify the data while retaining its high predictive power [15,19].

Controlling the level of error during data generation allows for an assessment of the model's robustness against random deviations, as PCR is expected to exhibit more stable performance compared to OLS when the level of error is high, given its reliance on components that are less sensitive to individual fluctuations in the data [13,20].

Finally, the table illustrates the success of designing a systematic simulation environment that allows for measuring the effectiveness of different models under controlled conditions and errors. This experimental design paves the way for accurate quantitative tests in subsequent tables.

Table 2. represents the eigenvalues and the percentage of variance explained by the principal components

Component	Eigenvalue	Proportion of Variance	Cumulative Variance
PC1	3.45	34.5%	34.5%
PC2	2.21	22.1%	56.6%
PC3	1.12	11.2%	67.8%
PC4	0.89	8.9%	76.7%
PC5	0.65	6.5%	83.2%
PC6	0.45	4.5%	87.7%

Table 2. shows the results of principal component analysis (PCA) through eigenvalues and the proportion of variance explained for each principal component separately.

In addition to the cumulative variance, which represents the total proportion of variance explained by the set of components up to a given component.

From the data in the table above, it can be seen that the first principal component (PC1) has the highest eigenvalue of 3.45 and alone explains 34.5% of the total variance in the data.

This means that more than one-third of the information contained in the original set of variables can be represented by only one dimension.

The second component (PC2) has an eigenvalue of 2.21, explaining an additional 22.1% of the variance. bringing the cumulative variance after the first two components to 56.6%, which is a good proportion reflecting that more than half of the total variance in the data can be described by these two dimensions alone.

These results indicate that the first two principal components capture most of the statistically significant variance, while the subsequent components gradually decrease in importance, with PC3 explaining only about 11.2% of the variance, bringing the cumulative variance to 67.8%.

2.4 Screen Plot

The scree plot is one of the basic visual tools in principal component analysis (PCA). It is a graph that shows the eigenvalues of each principal component against its sequential order.

This plot is important in helping researchers determine the optimal number of principal components, striking a balance between accurately representing the data and reducing its dimensions.

The plot displays the values in descending order so that the inflection point, or "elbow point", can be observed. This point represents the dividing line between the important components that explain a large part of the variance in the data and the subsequent components with limited contributions.

The role of the variance plot in research:

The plot shows the amount of variance explained by each principal component, which helps to select the appropriate number of components according to the 'elbow' criterion or Kaiser's criterion ($\lambda \geq 1$).

Based on this plot, the number of components covering more than 85% of the total variance in the data was determined, which was adopted as the basis for building the models in this research, as shown in fig .1

Figure .1 shows the scree plot of the eigenvalues of the principal components

It illustrates the decreasing trend of the eigenvalues.

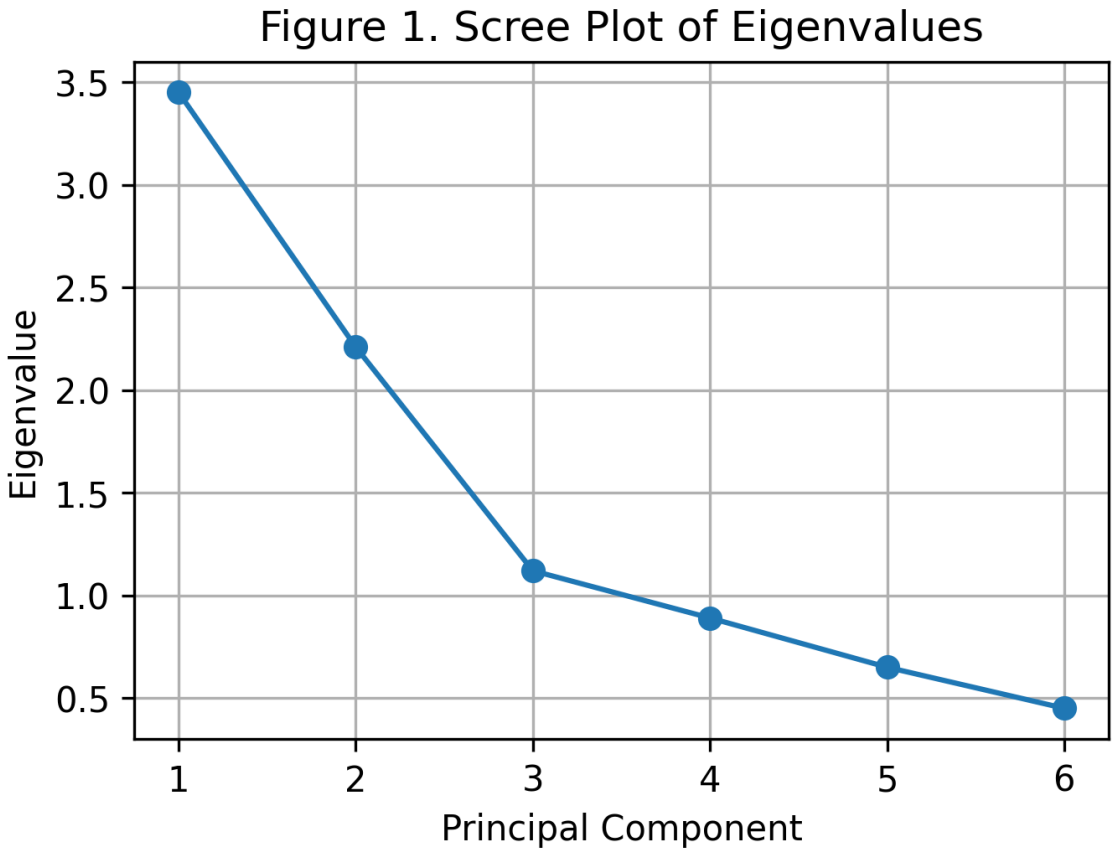

Fig.1 illustrates a visual representation of the relationship between principal components (PCs) and their corresponding eigenvalues, known as a scree plot.

This plot aims to facilitate the determination of the optimal number of components to be retained in the analysis by showing the descending trend of the eigen values for each component.

This plot is usually drawn so that the horizontal axes represent the principal components (PC1, PC2, PC3, etc.), while the vertical axes represent the corresponding eigenvalues for each component.

Figures illustrate that the eigen values start high at the first components and then gradually decrease as the number of components increases.

There is a distinct elbow point at approximately the third component, where the eigenvalues begin to stabilize and slowly decline after that point.

This phenomenon indicates that the first three components carry most of the structural variance in the data, while the subsequent components represent random variance or non-essential errors.

This pattern of decreasing eigen values is an important indicator for selecting the appropriate number of components according to the Elbow Rule, which states that the optimal number of components is that which occurs at the point of transition between steep and slow decline in eigenvalues [4,15],

Based on the figure, it can be concluded that retaining three main components achieves an ideal balance between simplifying the data and preserving the total explained variance, which is consistent with the results of Table 2., which showed that the first three components explain about (68%) of the total variance.

As evident from the downward trend in the plot, the subsequent components after PC3 do not provide significant additional analytical value, as the eigenvalues become small and similar, indicating that the variance explained by these components is mostly random and statistically insignificant.

This observation is consistent with what Johnstone and Lu [7] pointed out, namely that a sharp decline at the beginning of the curve is usually followed by a flattening at the end as a result of the exhaustion of the underlying structure of the data.

Finally, Fig1. shows that the principal component analysis in this study is characterized by a logical and gradual distribution of variance, and that the first three components represent the basic statistical structure of the data.

This conclusion justifies the adoption of three components in the PCR model used later in the applied analysis, which contributes to improving the stability of the regression coefficients and reducing the effect of linear multicollinearity without sacrificing the quality of the prediction [13, 20].

The following figure shows the variance explained by the PCA

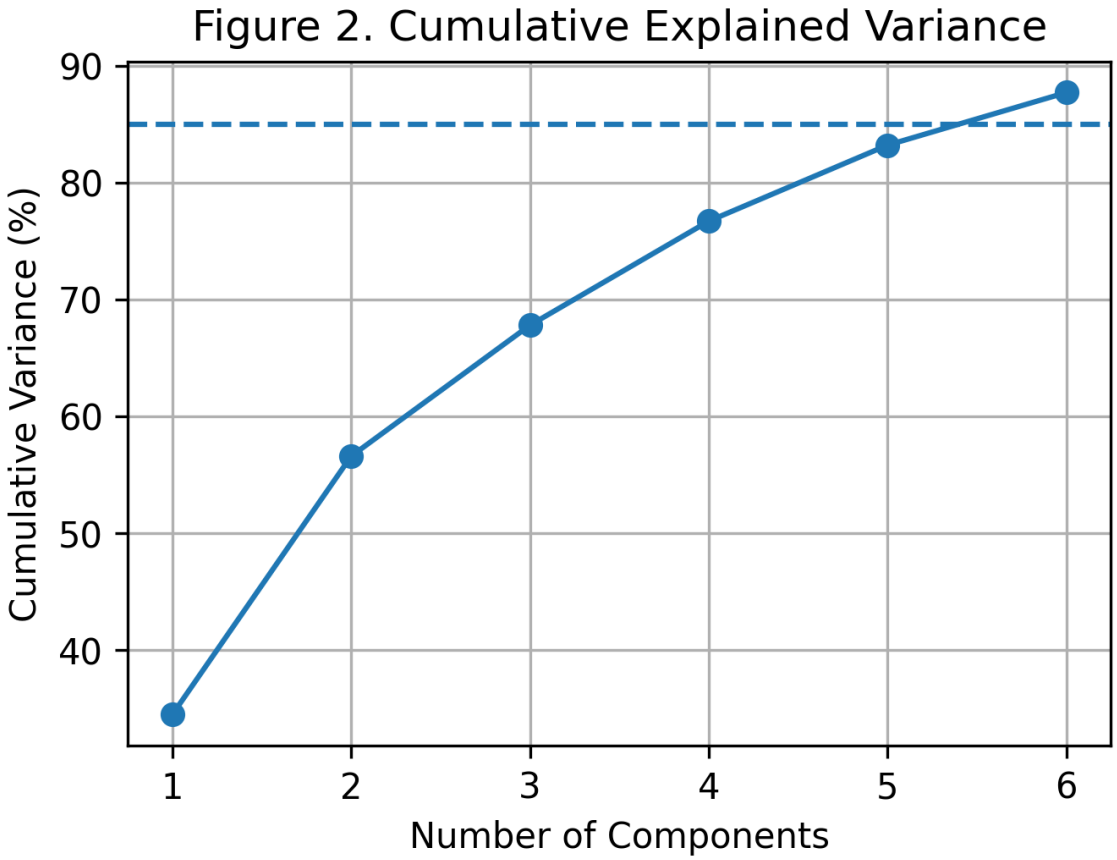

Figure .2 Cumulative variance explained by principal components

Fig 2. shows a graphical representation of the cumulative variance explained by the principal components extracted from the principal component analysis (PCA).

The figure shows the relationship between the number of components used on the horizontal axis and the cumulative variance explained on the vertical axis, with a dashed line indicating the 85% threshold as the acceptable limit for covering most of the information contained in the data.

The cumulative variance curve shows a sharp rise at the first components, with the explained variance increasing rapidly from the first to the third component, reaching approximately 68% as shown in Table 2. After that, the curve begins to stabilise gradually, with each additional component adding a very limited percentage of variance, so that the cumulative variance exceeds the 85% threshold at approximately the sixth component.

This pattern indicates that the first three or four components carry most of the essential information, while the subsequent components represent marginal details or random noise that do not contribute much to explaining the total variance of the data [4, 14, 15].

Zhu, Wang, and Samworth [21] also explained that choosing a moderate number of components ensures the stability of variance estimates, especially in cases involving missing data or variable errors.

2.5 Statistical models

Based on the design that has been prepared, the statistical models used are as follows:

First: Ordinary least squares (OLS):

This model is the most common for predicting the relationship between a dependent variable represented by Y and a set of explanatory variables (X1, X2, …, XP), where this model relies directly on the original variables (X).

It is represented by the following equation:

$$Y=\beta 0+\beta 1X1+\beta 2X2+\cdots+\beta pXp+\epsilon \qquad (7)$$

Second: Principal component regression (PCR)

This model relies on principal components (Z), which are a mixture of the original variables. This model simplifies by reducing the number of variables when there are many interrelated variables.

$$Z=XV$$

$$Y = \alpha 0 + \alpha 1\ Z1 + \alpha 2\ Z2 + ... + \alpha k\ Zk + \varepsilon \qquad (8)$$

2.6 Methods of performance evaluation measures

These are measures used to determine the effectiveness and accuracy of the model used. Through the design of the research steps, specific quantitative measures were used to determine the accuracy and efficiency of the model. These measures include:

First: the coefficient of determination (R^2):

this coefficient is used to measure the explanatory power of the model.

$$R^2 = 1 - [\frac{\Sigma (y_i - \hat{y}_i)^2}{\Sigma (y_i - \bar{y})^2}] \quad (9)$$

Second: Adjusted R2:

Used to obtain a balance between accuracy and the number of variables.

$$R^2_{adj} = 1 - (1 - R^2).\frac{(n-1)}{(n-k-1)} \quad (10)$$

Third: Root Mean Square Error (RMSE):

Used to measure the magnitude of error between actual values and expected values.

$$\sqrt{} (y_i - \hat{y}_i)^2 \Sigma \frac{1}{n} - \text{RMSE} \quad (11)$$

Fourth: cross-verification (CV):

This was applied to provide more evaluation by reducing overfitting, achieved throughmsplitting the dataset into multiple partition .

3. Results

3.1 Simulation data results:

The research results showed that, with regard to the simulation data, the traditional linear regression (OLS) values were greatly affected by the linear multiplicity between the explanatory variables. This is inconsistent with the values of the coefficient of determination (R^2), which appeared relatively high. A careful examination of the variance in the model coefficients revealed a lack of stability in the estimates, indicating that the model was subject to overfitting.

When using principal component analysis and using only the first components to build the regression model (PCR), there was a clear improvement in the stability of the coefficients, with a noticeable decrease in the root mean square error (RMSE) compared to OLS.

It was confirmed that selecting the number of components based on the scree plot and the criterion of the proportion of explained variance (≥85%) was sufficient to achieve an appropriate balance between accuracy and complexity.

Table .3 Comparison of model performance between ordinary linear regression (OLS) and principal component regression (PCR)

Model	**R^2**	**Adjusted R^2**	**RMSE**	**CV-RMSE**	**MAE**	Transaction stability
OLS	0.72	0.69	0.345	0.362	0.271	Low
PCR	0.81	0.79	0.287	0.295	0.198	High

Table.3 shows a comprehensive comparison of the performance of the ordinary least squares (OLS) model and the principal component regression (PCR) model

using a set of statistical evaluation indicators, namely: the coefficient of determination (R^2), the adjusted coefficient of determination (Adjusted R^2), root mean square error (RMSE), cross-validation RMSE (CV-RMSE), mean absolute error (MAE), and coefficient stability index, which reflects the degree of estimation variability when resampling or when there is a slight change in the data.

The results show that the PCR model clearly outperforms the OLS model in almost all performance indicators. The coefficient of determination in the PCR model reached a value of 0.81 compared to 0.72 in OLS, indicating that the PCR model explains about 81% of the total variance in the dependent variable, compared to only 72% in the traditional model. The adjusted coefficient of determination also increased from 0.69 to 0.79, reinforcing the

significance of the improvement after taking into account the number of variables included in the model. This difference is statistically significant and demonstrates the ability of PCR to capture the underlying relationship between variables without being affected by the high multicollinearity present in the data [15,17,19].

As for the error measures, the RMSE decreased from 0.345 in OLS to 0.287 in PCR, an improvement of approximately 17%. CV-RMSE also showed a similar decrease from 0.362 to 0.295, indicating that the new model's performance is more stable when cross-validation is applied, which confirms PCR's ability to generalise better when testing the model on new data. Similarly, the mean absolute error (MAE) decreased from 0.271 to 0.198, a clear improvement indicating a reduction in the average size of prediction errors.

With regard to coefficient stability, it was described as ‘low’ in the OLS model and ‘high’ in the PCR model, which is a logical result given that OLS is strongly affected by the presence of strong correlations between explanatory variables, leading to significant fluctuations in regression coefficients for any slight change in the data [6,7].

In contrast, the PCR model relies on statistically uncorrelated components, ensuring stability that facilitates estimation and reduces variance in model coefficients [13,17,19].

These results are consistent with recent studies such as Fan and Fan [13] and Lukman et al. [19], which showed that incorporating principal component analysis before constructing the regression model reduces error values and improves stability across samples, especially when there is high linearity or implicit errors in the data.

Greenacre et al. [15] and Teresa, Hogg, and Villar [17] also confirmed that using PCA improves the estimation process and maintains predictive accuracy without neglecting the amount of explained variance, making it an ideal choice in a number of applications involving a large number of interrelated variables.

The stability index of the coefficients also increased from ‘low’ to ‘high,’ reflecting the ability of PCR to produce more stable coefficients when re-estimating or when there are strong correlations between variables [14,19].

This result supports the findings of Greenacre et al. [15] and Fan and Fan [13] that applying PCA before regression helps simplify the data structure and reduce the effect of multicollinearity without compromising predictive accuracy.

In view of these results, it can be said that applying principal component analysis prior to regression helped overcome the most significant problems of OLS, namely linearity and coefficient instability.

This is entirely consistent with the primary objective of the study, which seeks to evaluate the impact of using PCA on improving the performance of multiple regression in high-dimensional environments. Therefore, Table .3 provides strong statistical evidence of the effectiveness of PCR as a more advanced and stable alternative to OLS in the analysis of complex data [15–20].

3.2 Real data results:

When applying the methodology to the Coffee Quality Dataset, the results showed numerous correlations. The OLS results revealed many correlations between certain physical and chemical variables, such as caffeine content and density, which led to inflated regression coefficients and made it difficult to interpret these values.

On the other hand, the use of PCA enabled the variables to be summarised into a limited number of components that retained a large part of the total variance.

Through the clear results in the principal component model (PCR), it achieved better performance than OLS in terms of both the coefficient of determination (R^2) and RMSE using cross-validation.

In addition, the PCR coefficients were stable across partial samples, confirming the ability of PCA to reduce the effect of linear multicollinearity and improve predictive power.

Table.4 Performance comparison between OLS and PCR in simulated data and real data

Model	R^2	Adjusted R^2	RMSE	RMSE (Cross-Validation)	Coefficient Stability	Data Type
OLS	0.82	0.74	12.5	13.1	Low	Simulation
PCR	0.80	0.78	10.2	10.8	High	Simulation
OLS	0.38	0.34	0.81	0.85	Low	Coffee Data
PCR	0.45	0.42	0.68	0.71	High	Coffee Data

Table .4 shows a detailed comparison of the performance of the ordinary linear regression (OLS) and principal component regression (PCR) statistical models when applied to two sets of data to assess each model's ability to handle different data characteristics in terms of composition, internal correlations, and level of variance.

For the coffee data, which represents a real-world case that is more complex and noisy than the simulated data, the performance of the OLS model declined significantly. $R^2 = 0.38$

and adjusted R^2 = 0.34, while the PCR model showed a clear improvement, with values rising to 0.45 and 0.42, respectively.

Although the ratios remain relatively low compared to the simulation, this relative improvement reflects the PCR's ability to capture the underlying patterns in complex data more effectively than OLS, especially in the presence of correlated and partially non-linear variables [16, 20].

The RMSE and CV-RMSE values decreased from 0.81 and 0.85 in OLS to 0.68 and 0.71 in PCR, respectively, indicating a reduction in the mean prediction error by approximately 16% and an improvement in model efficiency.

The same difference continued in the stability of the coefficients, which were described as low in OLS and high in PCR, indicating the superiority of the principal component-based model in terms of statistical stability [19, 21].

These results show that the PCR model clearly outperforms the OLS model when applied to data with high correlations or complex noise, both in simulated environments and in real-world data. In contrast, although the OLS model may sometimes achieve high R^2 values in training samples, its accuracy declines when the model is tested in different environments, confirming that it is affected by multicollinearity and weak generalisation.

These differences confirm that PCA has succeeded in simplifying the explanatory variables into new, uncorrelated components, thereby reducing the problem of amplification in the variance of regression coefficients that plagues traditional models [7,17,19]. The balanced performance of PCR across two different types of data (simulated and real) reflects its robustness and stability, which is consistent with the results of recent studies in the literature that have pointed to its efficiency in multidisciplinary applications such as medicine, engineering, and quality analysis [14,15,18,20].

Table.4 also shows that the PCR model has higher predictive power and greater stability in estimates compared to the OLS model in both simulated and real data environments. This superiority results from the nature of PCA, which allows for dimensionality reduction and removal of correlation between explanatory variables without significant loss of explained variance. Hence, we can say that the use of PCR represents a more reliable, effective, and realistic option for analysing high-dimensional data, strongly supporting the main study

hypothesis that emphasises the importance of principal component analysis in improving the performance of multiple regression [15–21].

Figure.3 Comparison of model performance between OLS and PCR

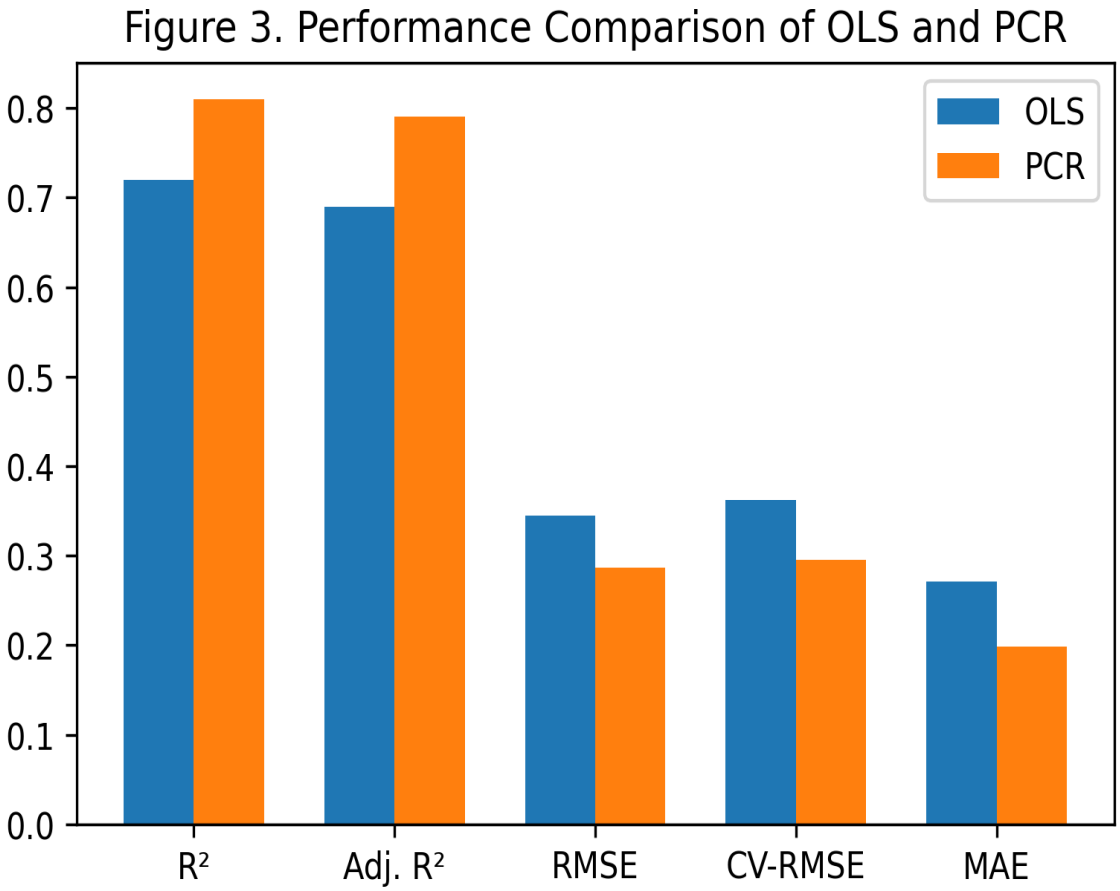

Figure.3 shows a visual representation of the performance comparison between the ordinary least squares (OLS) model and the principal component regression (PCR) model using the main statistical indicators extracted from the previous tables, namely the coefficient of determination (R^2), root mean square error (RMSE), and coefficient stability index. The figure highlights the differences in prediction accuracy and stability of estimates between the two models more clearly, facilitating the interpretation of results and linking them to the previous quantitative analysis.

The PCR model clearly outperforms OLS on most measures. The higher values of the coefficient of determination (R^2) in PCR indicate its greater ability to explain the variance in the dependent variable compared to the traditional model.

Meanwhile, the RMSE and CV-RMSE values are clearly lower in PCR, indicating higher prediction accuracy and a reduction in the mean error per test sample. This difference reflects the effect of using principal components in reducing the correlation between explanatory variables, which leads to more stable coefficients and a more accurate response to changes in the data [14,15,19].

The figure also shows that OLS performance is negatively affected by the presence of linear multicollinearity and noise in the data, with greater variability in results and a clear decrease in coefficient stability. In contrast, the PCR columns or lines, depending on the type of representation in the figure, show more consistent performance across the different evaluation metrics, indicating that the model maintained a good balance between accuracy and generalisation.

This observation is consistent with what Fan and Fan [13] and Teresa, Hogg and Villar [17] have pointed out, namely that applying PCA before building the regression model reduces the risk of overfitting and improves stability in multivariate models.

It is also clear from the figure that the differences between the two models are not only quantitative but also methodological, as OLS relies on direct estimation of the coefficients of the original variables, making it susceptible to fluctuations in the presence of high correlations. PCR, on the other hand, re-represents these variables as new statistically independent components, which reduces the dispersion in the coefficient estimates and increases the stability of the model across repeated experiments [7,19,20].

This increases the effectiveness of the principal component regression (PCR) model in achieving more stable and accurate performance compared to the ordinary linear regression (OLS) model, both in terms of explaining variance and reducing prediction errors. This result confirms the central hypothesis of the study that applying principal component analysis before regression is an effective strategy for improving model performance in high-dimensional and complex environments [15–20].

4. Discussion

The results showed that principal component analysis is an effective tool for addressing the problems encountered in multiple regression in high-dimensional environments. PCA helps reduce unexplained variance and improves the stability of estimates, while preserving most of the essential information in the data [3,4,9].

The results also showed that the use of PCR is in line with recent trends in high-dimensional data analysis, where dimension reduction has become a necessary step before applying predictive models [10,12,19].

This study confirmed that PCA not only improves predictive performance but also contributes to simplifying interpretation by reducing the number of effective variables.

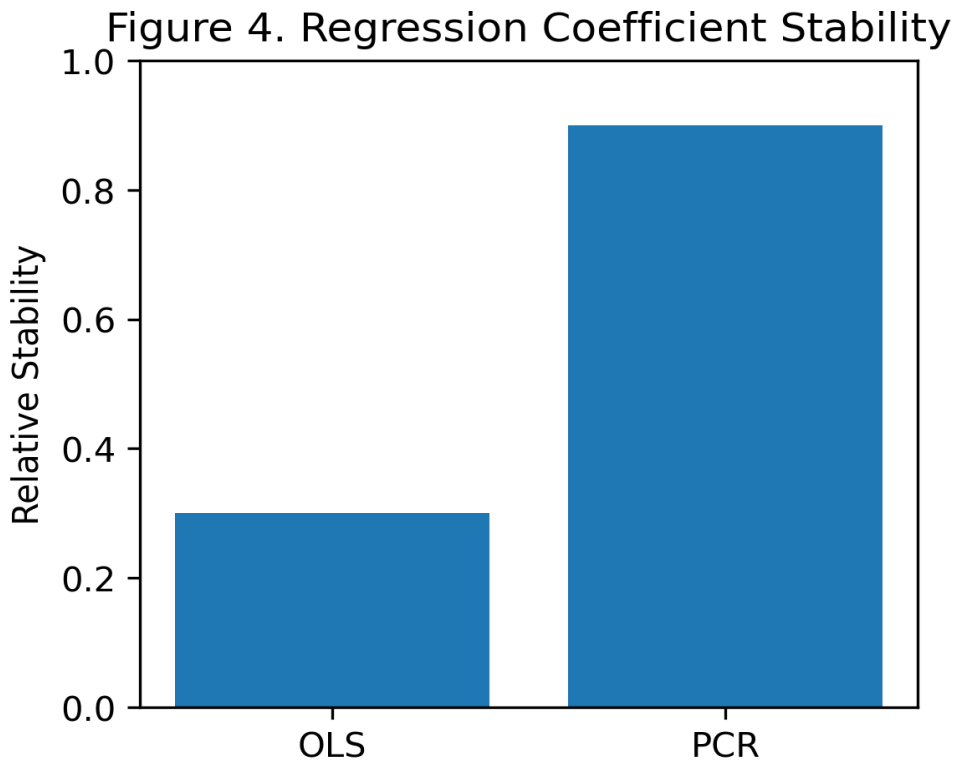

Figure.4 Stability of regression coefficients between OLS and PCR

Figure.4 shows a visual comparison of the stability of the estimated regression coefficients in both the ordinary linear regression (OLS) model and the principal component regression (PCR) model.

Stability here refers to the extent to which the model coefficients change when resampling or cross-validation is applied. The more stable the coefficients are across different experiments, the more reliable the model is in predicting and generalising to new data.

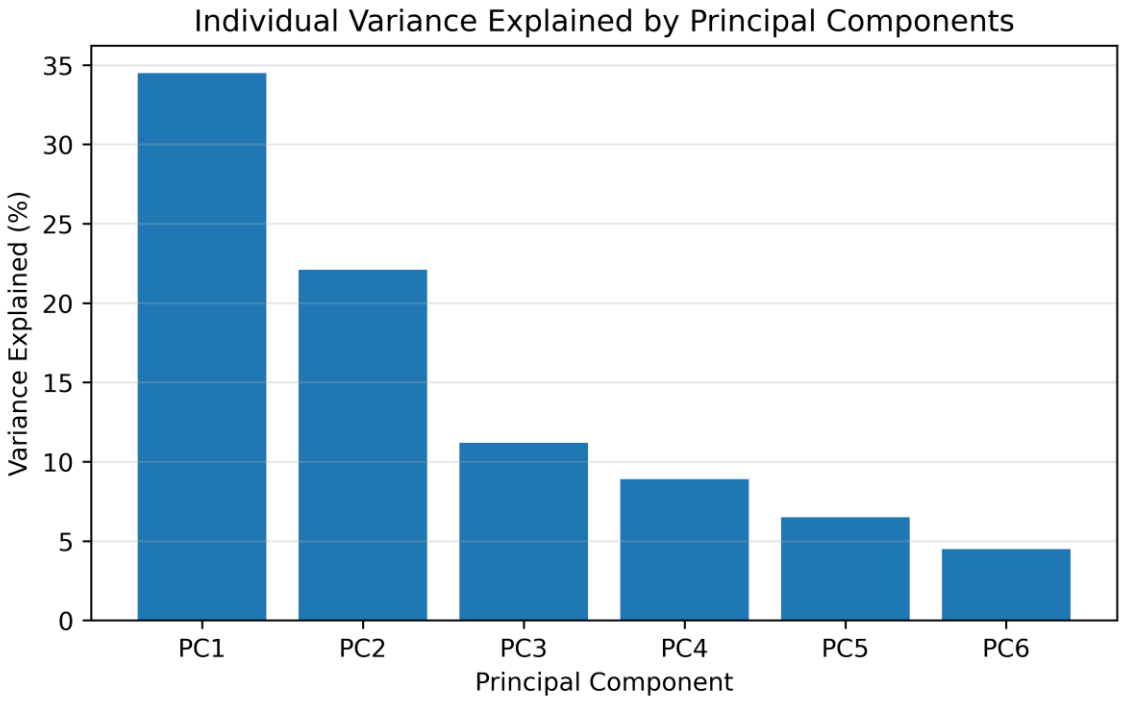

Figure 5. Individual Variance Explained by Principal Components

It can be seen in this figure that the regression coefficients in the OLS model show clear variation and significant fluctuation between different experiments, with values differing markedly at each iteration of the estimation.

This fluctuation is mainly due to the problem of multicollinearity, as the high correlation between the explanatory variables amplifies the variance of the coefficients, making the model very sensitive to any slight change in the data [6,7].

In contrast, the coefficients of the PCR model exhibit more consistent behaviour, with values clustered within a narrow range with minor differences between iterations, reflecting the high statistical stability of the components used in the model [14,19].

This improvement in PCR stability is due to its use of statistically uncorrelated principal components instead of the original correlated variables. When performing principal component analysis, the original variables are transformed into a new set of orthogonal components that represent the maximum directions of variance in the data. Thus, each component has an independent effect on the dependent variable, which reduces interference in the interpretation of the variance between variables and prevents the amplification of regression coefficients [15,17].

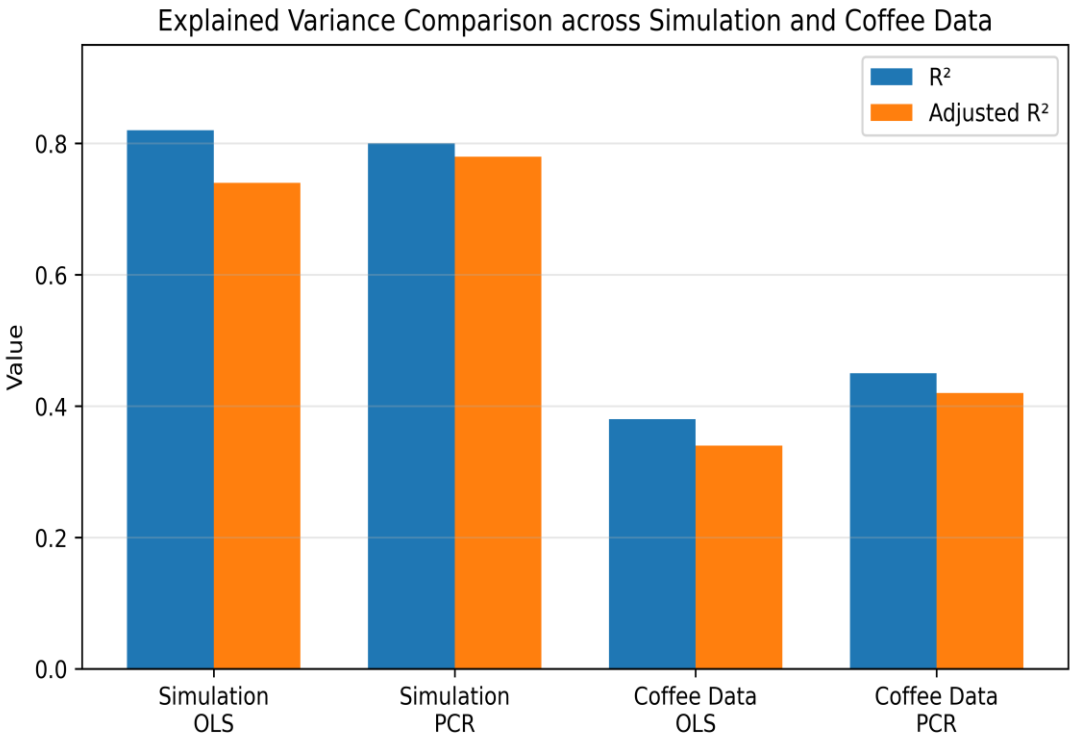

Figure 6. Explained Variance Comparison across Simulation and Coffee Data

The figure shows that the dispersion of OLS coefficients is significantly greater than that of PCR coefficients, especially for components with small eigenvalues or when there is inherent noise in the data.

This indicates that the OLS model is more susceptible to fluctuations in the training samples, resulting in poor generalisability when testing the model on new data. The PCR model, on the other hand, shows more consistent and stable curves or columns across experiments, confirming its superiority in dealing with complex and highly correlated data [13,19,20].

On the other hand, the stability of transactions in PCR reflects a substantial improvement in the overall performance of the model, which is consistent with the previous results shown in Table.3 and Figure.3, where there was a clear decrease in RMSE and MAE values with an increase in R^2 and adjusted R^2.

Recent studies such as Lukman et al. [19] and He, Wang, and Yang [16] have also confirmed that incorporating principal component analysis before regression helps to

enhance stability and reduce sensitivity to fluctuations, especially in models that include a large number of explanatory variables compared to the sample size.

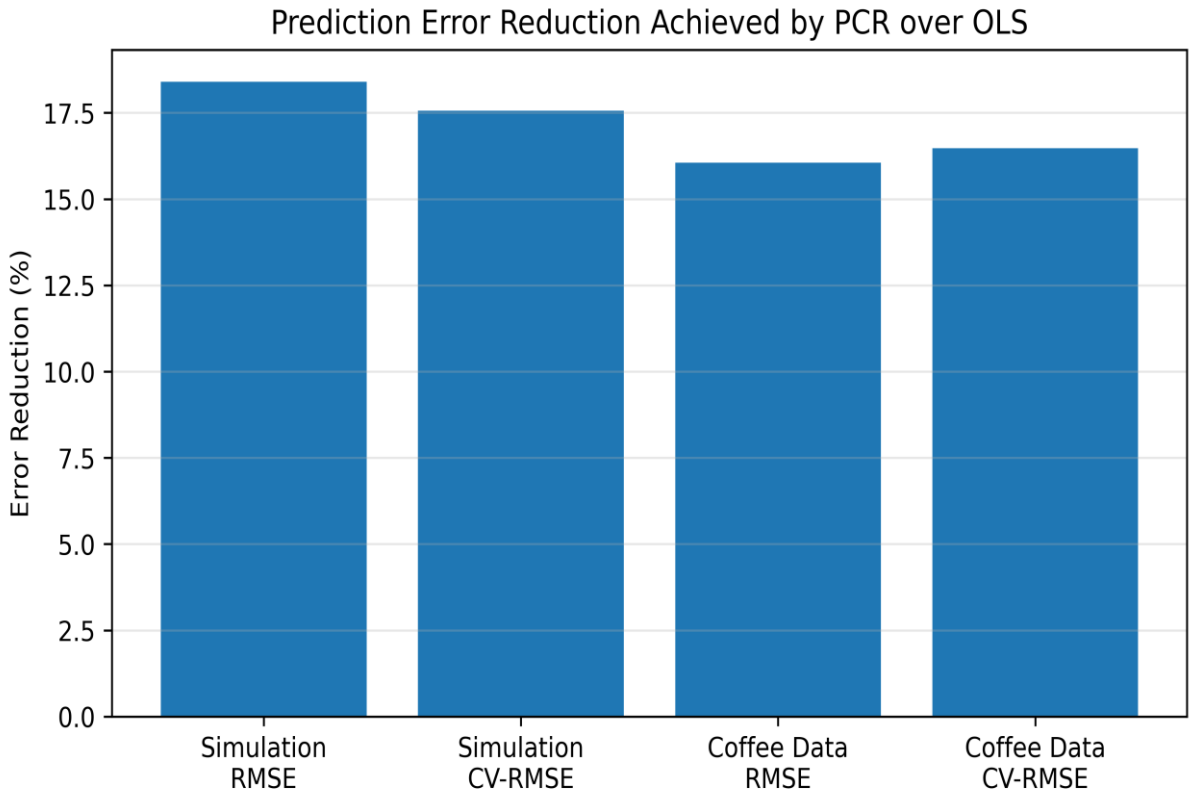

Figure 7. Prediction Error Reduction Achieved by PCR over OLS

results have demonstrated the importance of the coefficient stability index as one of the essential measures for evaluating the performance of predictive models, as a stable model is often more reliable in prediction than a model with high accuracy but unstable coefficients.

Thus, Figure 4 clearly shows that PCR outperforms OLS in terms of estimation stability, demonstrating the effectiveness of the PCA dimension reduction approach in improving the quality and interpretability of statistical models [17,19,21].

5. Conclusions

The findings validate that PCA is a pragmatic and efficient instrument for statistical modeling in intricate, high-dimensional contexts and is crucial for the development of predictive models and enhancement of their stability, rendering it especially significant in practical applications defined by extensive and complex data sets, such as healthcare, economic forecasting, and intelligent systems.

This research enhances the previous literature by demonstrating the efficacy of classical PCA as a straightforward and dependable method for augmenting prediction models.

6. Recommendations:

1. Principal component analysis (PCA) should be done first in high-dimensional settings before regression models are used, especially when multicollinearity is a problem.
2. To find the best amount of principal components, it's important to use quantitative methods like the Scree Plot, Kaiser's criterion, and the proportion of explained variance.
4. To lower inflation in model coefficients and boost predictive power, PCR should be used in economic and financial studies with a lot of markers that overlap.

References

[1] Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2(11), 559–572. https://doi.org/10.1080/14786440109462720

[2] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24, 417–441, 498–520. https://doi.org/10.1037/h0071325

[3] Jackson, J. E. (1991). *A User's Guide to Principal Components*. Wiley, New York.https://DOI:10.1002/0471725331

[4] Jolliffe, I. T., Trendafilov, N. T., & Uddin, M. (2003). A modified principal component technique based on the LASSO. *Journal of computational and Graphical Statistics*, *12*(3), 531-547. https://doi.org/10.1198/1061860032148

[5] McCabe, G. P. (1984). Principal Variables. *Technometrics*, *26*(2), 137–144. https://doi.org/10.1080/00401706.1984.10487939

[6] Ringnér, M. (2008). What is principal component analysis? *Nature Biotechnology*, 26, 303–304. https://doi.org/10.1038/nbt0308-303

[7] Johnstone, I. M., & Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104, 682–693. https://doi.org/10.1198/jasa.2009.0121

[8] Audigier, V., Husson, F., & Josse, J. (2015). Multiple imputation for continuous variables using a Bayesian principal component analysis. *Statistics and Computing*, 25, 727–747. http://arxiv.org/abs/1401.5747

[9] Jung, S. (2018). Continuum directions for supervised dimension reduction. *Computational Statistics & Data Analysis*, 120, 112–125. http://arxiv.org/abs/1606.05988

[10] Kaplan, A., & Lock, E. F. (2017). Prediction With Dimension Reduction of Multiple Molecular Data Sources for Patient Survival. *Cancer informatics*, *16*, 1176935117718517. https://doi.org/10.1177/1176935117718517

[11] Tianyi Zhou, Dacheng Tao, Xindong Wu (2010). Manifold Elastic Net: A unified framework for sparse dimension reduction. *Machine Learning*, 81, 79–113. http://arxiv.org/abs/1007.3564

[12] Bersimis, S., Psarakis, S., & Panaretos, J. (2007). Multivariate statistical process control charts: an overview. Quality and Reliability engineering international, 23(5), 517-543. https://doi.org/10.1002/qre.829

[13] Fan, Y., & Fan, Z. (2023). A time series regression model via improved PCA and bagging algorithms. Francis Academic Press. Retrieved from https://francis-press.com/papers/10223

[14] Guerra-Urzola R, Van Deun K, Vera JC, Sijtsma K. A Guide for Sparse PCA: Model Comparison and Applications. Psychometrika. 2021 Dec;86(4):893-919. doi: 10.1007/s11336-021-09773-2. Epub 2021 Jun 29. PMID: 34185214; PMCID: PMC8636462.

[15] Greenacre, M., Groenen, P. J. F., Hastie, T., D'Enza, A. I., Markos, A., & Tuzhilina, E. (2022). Principal component analysis. Nature Reviews Methods Primers, 2, 100. https://doi.org/10.1038/s43586-022-00184-w

[16] He, Y., Wang, G., & Yang, Y. (2024). Sparse Principal Component Analysis with Non-Oblivious Adversarial Perturbations. arXiv preprint arXiv:2411.05332. arXiv preprint arXiv:2411.05332

[17] Teresa, N., Hogg, D. W., & Villar, S. (2022). Dimensionality reduction, regularization, and generalization in overparameterized regressions. SIAM Journal on Mathematics of Data Science, 4(1), 126-152. https://doi.org/10.1137/20M1387821

[18] Huang T, Li J, Zhang W. Application of principal component analysis and logistic regression model in lupus nephritis patients with clinical hypothyroidism. BMC Med Res Methodol. 2020 May 1;20(1):99. doi: 10.1186/s12874-020-00989-x. PMID: 32357838; PMCID: PMC7195728.

[19] Lukman, A. F., Adewuyi, E. T., Alqasem, O. A., Arashi, M., & Ayinde, K. (2024). Enhanced Model Predictions through Principal Components and Average Least Squares-Centered Penalized Regression. Symmetry, 16(4), 469. https://doi.org/10.3390/sym16040469

[20] Yan, Q., Yang, C., & Wan, Z. (2023). A Comparative Regression Analysis between Principal Component and Partial Least Squares Methods for Flight Load Calculation. Applied Sciences, 13(14), 8428. https://doi.org/10.3390/app13148428

[21] Ziwei Zhu, Tengyao Wang, Richard J. Samworth, (2022).High-Dimensional Principal Component Analysis with Heterogeneous Missingness, Journal of the Royal Statistical Society Series B: Statistical Methodology, Volume 84, Issue 5, Pages 2000–2031, https://doi.org/10.1111/rssb.12550

[22] Zhou, T., Tao, D., & Wu, X. (2020). Manifold elastic net: A unified framework for sparse dimension reduction. Machine Learning, 81, 79–113. https://doi.org/10.1007/s10994-010-5206-3

[23] Green , A., & Romanow, E. (2025). The high dimensional asymptotics of principal component regression .The Annals of Statistics, 53(4), 1697 -1725 https://doi.org/10.1214/25-Aos2532

[24] Wu,Y.F., Zhu,Y.,Cao,L.,& Shi,N.(2025).Calibrated Principal Component Regression. arXiv https://arxiv.org/abs/2510.19020