International Journal of Mathematics and Statistics

ISSN (Online): 2208-2212
Volume 10 Issue 03 March 2024

DOI: https://doi.org/10.69980/3aj86879

STATISTICAL MODELLING OF HIGH-DIMENSIONAL DATA UNDER SEVERE MULTICOLLINEARITY USING REGRESSION AND PRINCIPAL COMPONENT METHODS

Dr. William E. Crawford[1]*, Prof. Anna M. Kowalska[2], Dr. Rafael T. Moreau[3]

[1]Department of Statistical Computing and Data Science, Northbridge Institute of Technology, Edinburgh, United Kingdom*
[2]Institute of Applied Mathematics and Multivariate Statistics, Baltic Research University, Warsaw, Poland
[3]School of Quantitative Methods and Statistical Research, European Institute of Data Analytics, Brussels, Belgium

Abstract:

Many high-dimensional data sets have highly correlated and redundant predictors that may affect the stability and interpretability of traditional regression models. In this study, a statistical model of high-dimensional data with severe multicollinearity was evaluated using ordinary least squares (OLS) regression, principal component analysis (PCA) and principal component regression (PCR) was evaluated. The number of observations was 500, the number of continuous predictors was 150, and the number of continuous target variables was 1. The underlying structure of the data was explored through descriptive statistics, Pearson correlations, tolerance, variance inflation factors and condition diagnostics. The original predictor space was highly multicollinear, and they could not estimate the total set of predictors using OLS. PCA then projected the 150 predictors onto two orthogonal components, which explained nearly all the variance in the predictors. These components were used in the development of a PCR, which was statistically significant and accounted for 4.2% of the target variance. The first principal component was significant, but the second was not. However, PCR did not compensate for the lack of goodness of fit over the estimable baseline OLS model but was able to completely remove the multicollinearity and offer a substantially more parsimonious and numerically stable representation of the predictor space. The results show that dimensionality reduction is a useful statistical tool to overcome high-dimensional, highly collinear predictor redundancy without loss of predictive value and validate PCR as a valuable tool in such cases.

Keywords: *High-dimensional data; Multicollinearity; Principal component analysis; Principal component regression; Ordinary least squares*

1. Introduction

The evolution of the vast amount of data used in research has significantly altered the fundamental nature of today's statistical problems. In many modern data sets predictors are numerous, the variables are often highly dependent and there is a lot of variable redundancy. The high-dimensional setting poses methodological challenges as dimensionality can lead to instability of estimation and conventional statistical relationships may be hard to interpret. With the rising complexity of datasets and the necessity to uncover meaningful low-dimensional structure, recent methodological advances have focused on statistical methods.Recent methodological advances thus highlight statistical methods that aim at extracting meaningful low dimensional structure from ever more complex datasets (Auddy et al., 2025).

When predictor variables are highly correlated, high dimensional regression is even more difficult. However, when this is the case, the information in various predictors is largely redundant and it can be hard to determine how much each predictor adds to the overall explanation of the target variable. Consequently, high-dimensional statistical research has focused on understanding dependence structures prior to interpreting the regression coefficients or on choosing variables to predict (Bühlmann et al., 2014). If the number of potential predictors becomes large, it is also important for controlling the computational and statistical complexity to do initial identification of the informative dimensions (Fan & Lv, 2008).

Among the most commonly-used methods for modelling relationships between continuous predictors and outcomes is ordinary least squares (OLS) regression. When the predictors are highly collinear, however, the coefficient estimates in its may be unstable. In extreme situations of multicollinearity, any slight variation in the data can lead to large variations in the estimates of individual coefficients, standard errors can be substantial, and the significance of the independent effect of correlated predictors can be hard to determine. To deal with this instability, due to nonorthogonal predictors, Ridge regression was developed by Hoerl & Kennard (1970) by adding a penalty that reduces the size of the regression coefficients and improves the stability in the estimation.

A second approach to coping with complicated predictor spaces is by using variable-selection techniques. The least absolute shrinkage and selection operator (LASSO) is an L1penalty that shrinks some of the regression parameters to zero; that is, it shrinks some parameters exactly to zero, which is a combination of the coefficient shrinkage and variable selection (Tibshirani, 1996). The adaptive LASSO later generalized this approach by putting variable-specific penalty weights on regression coefficients when a certain set of conditions of statistics are met (Zou, 2006). As it can be seen from these approaches, there is a general trend in the current research from unconstrained coefficient estimation to regularized statistical learning in high dimensional regression.

However, for sparse selection procedures, having highly correlated predictors introduces further challenges as they can have virtually the same information. Elastic net regression is a type of regression that has been developed to overcome this problem; it is a hybrid of L_1and L_2penalties, and is especially useful in the case of variables grouped in correlated sets (Zou & Hastie, 2005). In addition, computational advances have now made penalized regression (with LASSO, ridge and elastic-net) practical for large numbers of predictors by developing efficient coordinate descent algorithms to compute these models along regularization paths (Friedman et al., 2010).

All these advances notwithstanding, variable selection can still be fragile in the face of a lot of predictors containing similar information. To increase the reliability of variable selection in high dimensional applications, subsampling was introduced in combination with standard variable selection methods, and called stability selection (Meinshausen & Bühlmann, 2010). On a similar note, special LASSO formulations have shown that, for high dimensional regression, the structure of the data is important, not just the standard assumptions of regression (Lee et al., 2016).

Instead of choosing or shrinking individual predictors, a transformation of the spaces of predictors can be performed by dimensionality reduction. Principal component analysis (PCA) is especially useful as it transforms the original correlated variables into mutually orthogonal linear combinations called principal components. The idea of PCA is to represent the common variation among the many redundant variables by a smaller number of independent dimensions, rather than trying to estimate the effect of each variable separately. If there is a lot of overlapping information in the predictors, then a few principal components can then capture a lot of the information in the original high dimensional matrix.

A dimensionality reduction technique that combines it with regression modelling is called principal component regression (PCR). The first step in the PCA is to apply it to the predictor matrix, and the orthogonal component scores obtained from this are then used as predictors of the response variable. This procedure is directly dealing with multicollinearity in that the principal components are mutually uncorrelated by construction. When the conventional OLS regression is hampered by extreme predictor dependence, then PCR can offer a more parsimonious, and numerically stable, regression representation. Importantly, a dimensionality reduction should not be judged solely on the basis of the amount of variance that it captures, but should also account for the ability to capture predictive information as well as for the numerical stability it provides and the lack of redundancy.

In this context, the present study proposes a statistical modelling approach to a high-dimensional numerical dataset with high predictor redundancy. The analysis first systematically evaluates the distributional properties and the correlation structure of the data, the diagnosis of multicollinearity of the original predictor space is given, a conventional OLS benchmark is set, and PCA and PCR are then applied to get an orthogonal low dimensional representation. The study thus looks at the possibility of preserving the predictive information contained in the original data and at the same time getting a more stable and parsimonious solution to the severe multicollinearity in the original data by using principal component based regression.

2. Methodology

2.1 Research Design

This study used a quantitative methodological research design to investigate statistical modelling of a high-dimensional data set with a high degree of predictor redundancy and multicollinearity. The analytical framework was developed to

evaluate the weaknesses of the standard ordinary least squares (OLS) regression and to see if the predictor space could be represented in a way that was more stable but retained information about the space while still being predictive.
The analytical procedure carried out sequentially is data screening, data descriptive analysis, correlation analysis, data multicollinearity diagnostics, baseline OLS regression, principal component analysis (PCA), principal component regression (PCR), model comparison and the assessment of the residuals. These sequences allowed for the structure of the data to be assessed prior to dimensionality reduction and further regression modelling.

2.2 Dataset and Variable Structure

The study was conducted on a high dimensional numerical dataset (n=500, p=150, y=1). The predictor variables were named feature_0, feature_1, ..., feature_149 and the target was set to be the dependent variable.
The predictors were not provided with substantive domain-specific labels, so the analysis was mostly done in the domain of their statistical properties, such as distribution, correlation, redundancy, dimensionality and predictive contribution. All variables were entered in IBM SPSS Statistics Version 26 and continuous variables were considered scale level data.

2.3 Data Screening and Descriptive Analysis

Data screening was done in the initial stages to check the completeness of the data and to find out any issues that might impact the statistical analysis. Prior to the estimation of the model, the missing observations and the numerical consistency of the variables was checked.
Next, descriptive statistics were computed for the predictor and target variables. These were the mean, standard deviation, minimum, maximum, skewness and kurtosis. The degree of skewness and kurtosis were investigated to check for significant violations of univariate normality. This initial assessment was meant to provide an overview of the distribution of this data and was not meant to automatically reject variables based on individual diagnostic values.

2.4 Correlation and Multicollinearity Assessment

Pearson product-moment correlations were used to determine the linear relationships among the original predictor variables. A special focus was placed on the very high intercorrelations because some high correlations between predictors can lead to unstable regression weights and to problems in estimating independent predictors.
The degree of multicollinearity was additionally evaluated using tolerance, variance inflation factor (VIF) and condition indices. The small tolerance values (near zero) were interpreted as signs of high levels of redundancy of a predictor with the other predictors, while the high VIF values were interpreted as signs of problematic multicollinearity. Conventional multiple regression was assumed to be adequate for the original predictor space if the correlation matrix and formal collinearity diagnostic were both considered together.

2.5 Baseline Ordinary Least Squares Regression

A baseline multiple linear regression model using OLS estimation was specified with the target as the dependent variable and the original predictor variables as candidate independent variables. The general regression model was expressed as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

where Yrepresents the target variable, β0represents the intercept, X1,…,Xprepresent the predictors, β1,…,βprepresent their corresponding regression coefficients, and εrepresents the random error term.
The Enter method was used rather than an automated stepwise variable-selection procedure. Model performance was evaluated using R, R2, adjusted R2, standard error of the estimate, the overall F-test, regression coefficients, standardised coefficients, confidence intervals, tolerance, and VIF. Predictors that could not be estimated because of numerical redundancy or tolerance limitations were considered evidence of instability in the original high-dimensional predictor structure rather than substantive variable-selection decisions.

2.6 Principal Component Analysis

A full set of predictors was then used for PCA to eliminate redundancy and reduce dimensionality. PCA transforms the original variables, which may be correlated, into a smaller set of mutually orthogonal linear combinations of the original variables, called principal components. The analysis was carried out using the correlation matrix and the extraction of the main components. Since the goal was not to interpret the latent factors, no component rotation was used.
Kaiser-Meyer-Olkin measure, when available, eigenvalues, percentage of variance explained, cumulative variance explained, scree plot, communalities and component loading patterns were used to assess suitability and component retention. While the traditional eigenvalue criterion is used, the decision on which components to keep was also made based on the observed structure of the predictors to prevent the loss of a relatively unambiguous dimension of the data. Therefore, the next regression modelling was conducted using a two-component solution. The regression-based component scores were stored in SPSS to use as the predictor variables in the PCR analysis.

2.7 Principal Component Regression

Principal component regression was performed using the retained component scores as independent variables and the target as the dependent variable. The PCR specification was expressed as:

$$Y = \alpha_0 + \alpha_1 PC_1 + \alpha_2 PC_2 + \varepsilon$$

where PC1and PC2represent the retained orthogonal principal components, α0represents the intercept, α1and α2represent their regression coefficients, and εrepresents the error term.
Because principal components are mutually orthogonal, PCR was employed specifically to overcome the severe multicollinearity associated with the original predictor matrix. Model performance was evaluated using R, R2, adjusted R2, standard error, overall F-statistic, regression coefficients, standardised coefficients, confidence intervals, tolerance, VIF, and condition indices. The use of equivalent model-fit criteria for OLS and PCR facilitated direct comparison of predictive performance and numerical stability.

2.8 Model Comparison and Residual Diagnostics

The baseline OLS and PCR models were compared to see if dimensionality reduction could preserve the predictive information while making the model more stable. The criteria used for comparisons were R2, adjusted R2, standard error of the estimate, overall statistical significance, dimensionality of the representation of the predictors, tolerance, VIF, and condition indices.

The aim of this comparison was not only to determine the model that explained the largest proportion of variance, but to determine the one that explained the most variance. Special focus was given to the possibility of compressing the original high-dimensional information in fewer orthogonal dimensions, and, at the same time, assuring the stability of the PCR, given the problem of multicollinearity.

Residual diagnostics were then carried out on the PCR model. Unusual or potentially influential observations were explored using standardised residuals, studentized residuals, Cook's distance and leverage values. Approximate normality of the residuals was assessed with a histogram of standardised residuals and a normal probability plot, and nonlinearity and heteroscedasticity were checked with a scatterplot of the residuals versus the standardised predicted values. Data were analysed using IBM SPSS Statistics Version 26, and statistical significance was determined at $p< 0.05$.

3. Results

3.1 Descriptive Statistics and Distributional Characteristics

The dataset contained 500 fully completed observations, 150 continuous predictors (feature_0, feature_1, ..., feature_149) and one continuous dependent variable (target). There were no missing observations for variables used in the statistical analyses. The means of most of these predictors (26 of the 40) were close to 0, and their standard deviations were close to 0.98, thus showing similar scales throughout this very correlated predictor block. The mean and standard deviation of the target variable were 1.157 and 2.097, respectively.

According to the distributional assessment, there was no significant univariate non-normality. The skewness of the target variable is −0.048, and for feature_149 is −0.108; the Kurtosis of the target variable is 0.222, and for feature_149 is 0.149. The values of skewness and kurtosis of the other predictors were also quite small. These findings suggested that the dataset was not seriously affected by changes in symmetry. The descriptive statistics and distributional characteristics of the study variables are summarised in Table 1.

Table 1. Descriptive Statistics and Distributional Characteristics of the Dataset

Variable/Group	N	Mean	SD	Skewness	Kurtosis	Interpretation
Features 0–148	500	≈0.007	≈0.981	≈0.180	≈0.270	Approximately symmetric
Feature 149	500	−0.065	0.953	−0.108	0.149	Approximately symmetric
Target	500	1.157	2.097	−0.048	0.222	Approximately symmetric

3.2 Correlation Structure and Initial Evidence of Multicollinearity

A very high dependence structure among the original predictors was found by Pearson correlation analysis. The correlations of the 149 variables ranged from r = 0.999 to r = 0.9999 and were essentially perfect for all variables except one. In contrast, feature_149 revealed very weak relationships with this dominant predictor block, with r = 0.038–0.039. This correlation pattern suggested one large cluster of very similar predictors and a relatively independent predictor for this data set. The latter is infeasible for stable conventional coefficient estimation in cases where all predictors are included in an ordinary least squares model. The intercorrelations among the representative predictors were very high but Feature 149 had a relatively unique pattern of intercorrelations, as can be seen in Figure 1.

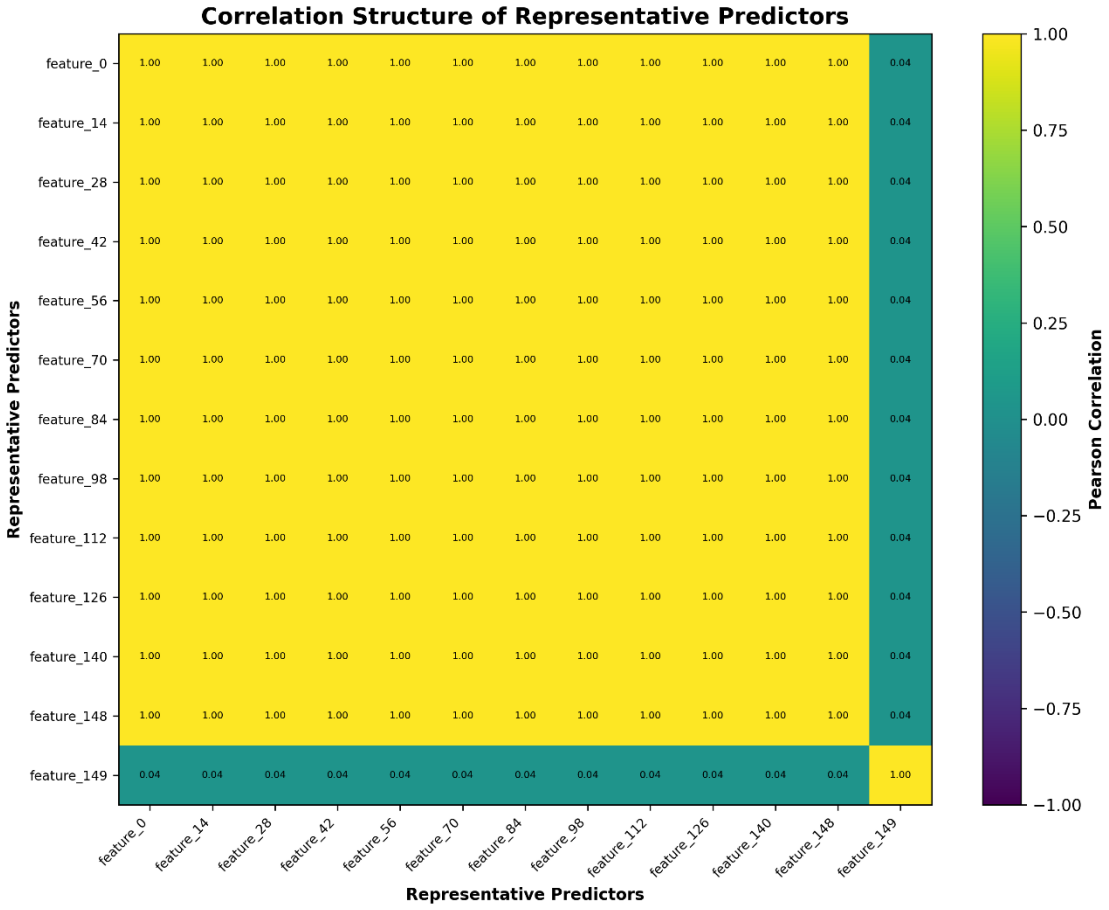

Figure 1. Correlation heatmap of representative predictor variables illustrating the severe multicollinearity structure of the dataset

Formal multicollinearity diagnostics confirmed the correlation results. Many values of the predictors not included were very close to zero, and VIF values were of the order of several hundred thousand. For example, VIF values for feature_0, feature_1, and feature_2 were approximately 416,205.985, 532,242.695, and 471,996.118, respectively. These values were considerably higher than the traditional diagnostic level and verified a high level of predictor redundancy. Table 2 gives summary statistics of the level of multicollinearity in the original predictor space (tolerances and VIFs) that are representative of the information contained in Table 2.

Table 2. Summary of Multicollinearity Diagnostics for the Original Predictor Space

Diagnostic	Result	Interpretation
Original predictors	150	High-dimensional predictor set
Highly correlated block	Features 0–148	Extreme redundancy
Approximate intercorrelation	≈1.000	Near-perfect linear dependence
Correlation of Feature 149 with dominant block	≈0.038–0.039	Negligible
Feature 0 VIF	416,205.985	Extreme multicollinearity
Feature 1 VIF	532,242.695	Extreme multicollinearity
Feature 2 VIF	471,996.118	Extreme multicollinearity
Typical excluded-variable tolerance	≈0.000002	Near-zero tolerance
Overall assessment		Severe multicollinearity

3.3 Baseline Ordinary Least Squares Regression

First, an ordinary least squares regression was set up, with all 150 predictors and the target as a dependent variable. But the severe multicollinearity prohibited simultaneous estimation of the whole set of predictors. The highly correlated predictors were dropped from the final estimable equation, and SPSS only retained feature_14 and feature_149.

The resulting baseline regression model resulted in $R = 0.206$, $R^2 = 0.042$, adjusted $R^2 = 0.039$, and a standard error of estimate of ~2.056. A general regression model was statistically significant ($F(2,497) = 11.028$, $p < .001$) and explained 6.8% of the variance in the data. So, the retained model was able to account for about 4.2% of the variance in the target.

At the predictor level, feature_14 had a significant positive coefficient ($B = 0.434$, $\beta = 0.203$, $t = 4.620$, $p < .001$), whereas feature_149 was not statistically significant ($B = 0.064$, $\beta = 0.029$, $t = 0.665$, $p = .506$). Do not read these estimates as meaning that the single feature_14 is particularly crucial because it's a single entry in an almost redundant predictor block.

The base model of OLS was statistically significant but provided very little explanatory power and did not fit the original 150-variable predictor structure (see Table 3).

Table 3. Baseline Ordinary Least Squares Regression Results

Indicator	Value
Intended number of predictors	150
Predictors retained	2
R	0.206
R^2	0.042
Adjusted R^2	0.039
Standard error	2.056
F	11.028
df	2, 497
Model significance	<.001
Feature 14: B	0.434
Feature 14: β	0.203
Feature 14: p	<.001
Feature 149: B	0.064
Feature 149: β	0.029
Feature 149: p	.506

3.4 Principal Component Analysis and Dimensionality Reduction

To deal with the extreme redundancy identified in the correlation and the regression diagnostics, Principal component analysis was then applied to the 150 original predictors. The correlation matrix was used to calculate the PCA, which gave a standardised solution of the contribution of the predictors to the components.

The Kaiser–Meyer–Olkin measure of sampling adequacy was 0.997, which means there was a very strong common variance structure. The exceptionally close predictor correlation matrix did not allow SPSS to compute a traditional Bartlett's chi-square statistic, and hence it was not interpreted further.

The PCA communalities were around 1.000 for the variables if 2 components were retained; this meant that in the retained two-component representation, the variance of the original set of predictors was reproduced well.

The first eigenvalue in the scree pattern was large, and the subsequent ones were very small. The first principal component is highly significant, and variation in the next components is extremely small, as can be seen from the scree plot in Figure 2.

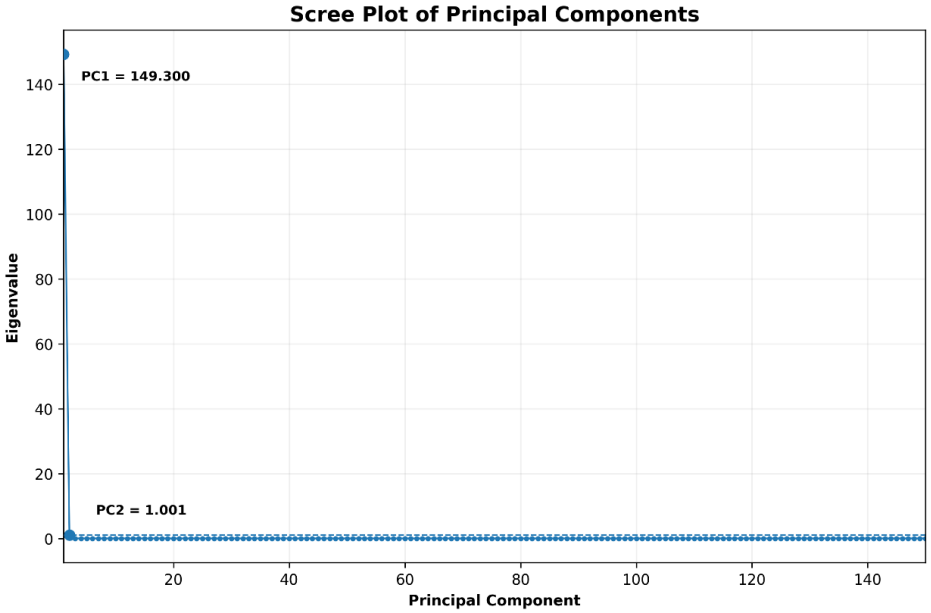

Figure 2. Scree plot showing the eigenvalue distribution of principal components derived from the original predictor space

3.5 Variance Explained and Component Retention

The first principal component had an eigenvalue of 149.001 and explained 99.334% of total predictor variance. The second component had an eigenvalue of 0.999 and explained the remaining 0.666%, resulting in a cumulative explained variance of approximately 100%.

Although the second eigenvalue fell marginally below the conventional Kaiser threshold of 1.0, it was retained because it captured a distinct predictor dimension dominated by feature_149. Retaining two components therefore preserved virtually all information contained in the original 150-dimensional predictor space. As reported in Table 4, the two retained components accounted for essentially 100% of the total variance in the original predictors.

Table 4. Principal Component Eigenvalues and Variance Explained

Component	Eigenvalue	Variance Explained (%)	Cumulative Variance (%)	Interpretation
PC1	149.001	99.334	99.334	Dominant common predictor dimension
PC2	0.999	0.666	100.000	Secondary independent dimension

The concentration of variance in PC1 is visually emphasised in Figure 3.

3.6 Principal Component Loading Structure

The component loading matrix provided a clear interpretation of the dimensional structure underlying the original predictor set. Variables feature_0 through feature_148 loaded approximately 1.000 on PC1 and approximately zero on PC2. This demonstrated that these 149 variables essentially measured the same underlying numerical dimension.

The loading pattern for feature_149 was distinctly different. Its loading on PC1 was approximately 0.039, whereas its loading on PC2 was 0.999. Thus, the second component predominantly represented the information contained in feature_149.

This loading structure confirmed that the apparent 150-dimensional predictor space could be effectively reduced to only two orthogonal dimensions without meaningful loss of information. Figure 3 illustrates the distribution of the 500 observations across the first and second principal component scores, providing a two-dimensional representation of the transformed predictor space.

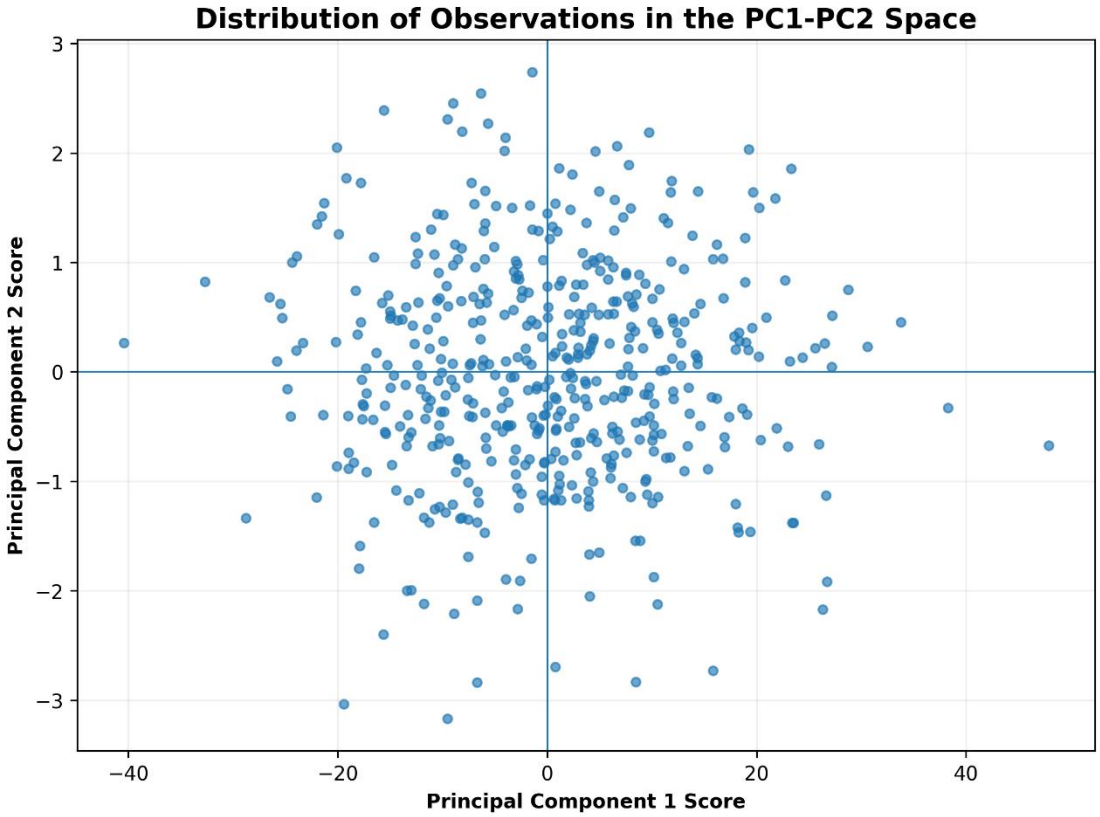

Figure 3. Distribution of observations across the first and second principal component scores

3.7 Principal Component Regression

Principal component regression was performed using the two saved component scores as independent variables and target as the dependent variable. Unlike the original predictor-based OLS specification, both principal components entered the model successfully because PCA transformed the highly correlated variables into mutually orthogonal dimensions.

The PCR model produced R = 0.206, R^2 = 0.042, adjusted R^2 = 0.039, and a standard error of estimate of approximately 2.056. The model was statistically significant, F (2,497) = 11.024, p < .001.

PC1 was a statistically significant predictor of the target (B = 0.428, SE = 0.092, β = 0.204, t = 4.648, p < .001, 95% CI [0.247, 0.609]). In contrast, PC2 was not significant (B = 0.061, SE = 0.092, β = 0.029, t = 0.663, p = .507, 95% CI [−0.120, 0.242]).

Critically, both components had tolerance = 1.000 and VIF = 1.000, demonstrating complete removal of multicollinearity. The condition indices were also 1.000, reflecting the mathematical orthogonality of the retained components. The PCR coefficients and comparison with the baseline regression are presented in Table 5.

Table 5. Principal Component Regression Results and Comparison with Baseline OLS

Indicator	Baseline OLS	PCR
Original predictors	150	150 transformed into PCs
Effective predictors	2 retained variables	2 principal components
R	0.206	0.206
R^2	0.042	0.042
Adjusted R^2	0.039	0.039
Standard error	2.056	2.056
F	11.028	11.024
Model p-value	<.001	<.001

PC1/Feature 14 β	0.203	0.204
PC1 significance	<.001	<.001
PC2/Feature 149 β	0.029	0.029
PC2 significance	.506	.507
Multicollinearity	Extreme in original space	Eliminated
PCR tolerance	—	1.000
PCR VIF	—	1.000
PCR condition index	—	1.000

3.8 Model Comparison and Diagnostic Assessment

Comparison of the baseline OLS and PCR models demonstrated that dimensionality reduction did not materially alter explanatory performance. Both approaches produced $R^2 = 0.042$ and adjusted $R^2 = 0.039$, with virtually identical standard errors. However, their numerical stability differed substantially.

The original 150-predictor specification was affected by severe multicollinearity, with many VIF values exceeding 400,000 and tolerance values approaching zero. By contrast, the two-component PCR solution produced VIF = 1.000 and tolerance = 1.000 for both predictors. Therefore, PCR preserved the predictive information of the original data while eliminating the severe collinearity problem. As illustrated in Figure 4, the PCR-predicted values show the relationship between the model predictions and the observed target values.

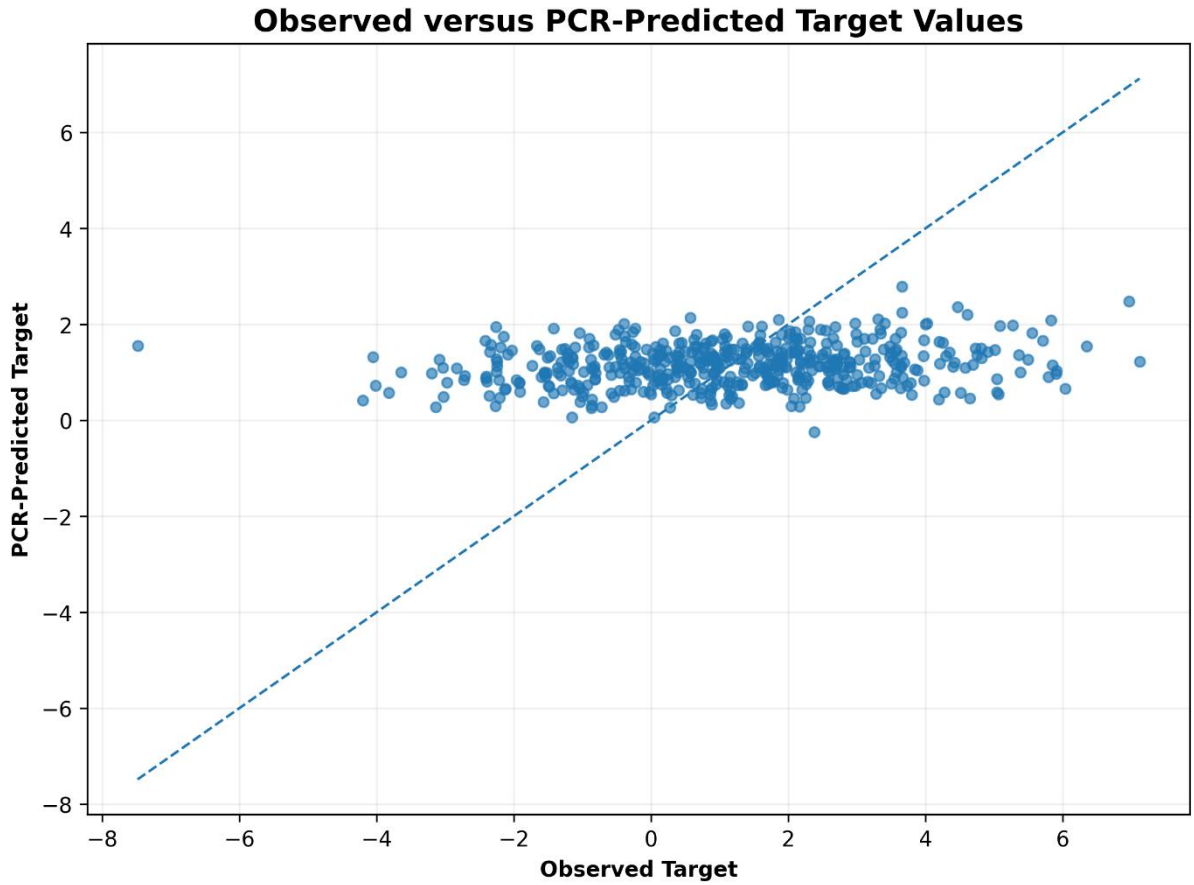

Figure 4. Relationship between observed and predicted target values obtained from the principal component regression model

Residual diagnostics further indicated that the mean standardised residual was approximately zero. The standardised residuals ranged from −4.396 to 2.867, indicating the presence of at least one unusually large negative residual. Nevertheless, the maximum Cook's distance was only 0.036, substantially below the conventional benchmark of 1.0,

suggesting that no individual observation exerted excessive influence on the PCR estimates. The maximum centred leverage value was 0.032, providing further evidence that the model was not dominated by isolated high-leverage cases.
Overall, the analysis demonstrated that the major statistical challenge in the dataset was not distributional abnormality but the exceptionally redundant structure of the original predictor matrix. Conventional OLS could not simultaneously estimate the full 150-predictor specification because of near-zero tolerance, whereas PCA reduced the predictor structure to two orthogonal components accounting for essentially 100% of total predictor variance. Principal component regression subsequently achieved the same explanatory performance as the baseline model while eliminating the multicollinearity problem. The results therefore support PCA-based regression as a substantially more stable and parsimonious modelling strategy for this high-dimensional, highly collinear dataset.

4. Discussion

In the current study the statistical difficulties of modelling a data set with an extremely complex dependence structure between predictors was considered. The results show that the main difficulty in the analysis was not the marginal distribution of the variables but the large amount of collinearity in the set of predictors. This is relevant because it is not enough to consider the standard assumptions of regression; high dimensional regression has to take into account dimensionality, estimation stability as well as dependency. Modern high-dimensional inference also acknowledges that there can be significant challenges to estimating the coefficients and making statistical inferences when the number of predictors grows (Cai et al., 2023).

The correlation analysis showed that the majority of the predictors were almost perfectly correlated with each other, while one predictor was a rather independent dimension. This structure is the reason why the original model of all the predictors was not possible to estimate with conventional OLS regression. The presence of high collinearity between predictors makes it difficult to determine the unique contribution of each predictor as the information is almost the same from one predictor to the other. Collinearity can have a significant impact on the estimation of the parameters, as shown by Dormann et al. (2013), and it is important to specifically detect and deal with correlated predictors, not just through typical regression output.

The extremely small tolerance statistics obtained for the original predictor space, along with the exceptionally large VIFs, also reinforces the lack of appropriateness of writing the coefficients from the original specification from the OLS model. It is a typical practice to use VIF as a measure of coefficient variance inflation due to the interdependence of the predictors, but one should interpret VIF rather than rely on universal cutoffs (O'Brien, 2007). The magnitude and consistency of the diagnostics in this analysis as well as the near-perfectly correlated structure of the diagnostics were indicative of extreme multicollinearity.

The OLS analysis was statistically significant, but explained a relatively small amount of the variation in the response. Most important of all, some of the predictor set could not be retained because of numerical redundancy in SPSS. This is an example of one of the basic limitations of conventional least squares estimation in a highly collinear predictor space. When the design matrix is not appropriate, the resulting solutions from high-dimensional regression methods can be unstable due to the need for regularization, robust estimation, variable selection or dimensionality reduction (Filzmoser & Nordhausen, 2021). Previously, the variables that are not excluded from the model by OLS should not be considered uniquely important variables since each member of an almost perfectly correlated block represents essentially the same variable.

PCA was found to give a much more parsimonious characterization of this complex predictor structure. The first component accounted for almost all of the common variability among the large correlated predictor block while the second component was the more independent predictor dimension. PCA is tailored to transform correlated variables into orthogonal linear combinations that are ordered by the percentage of variance in the predictors accounted for by the variables, and is especially useful for discovering lower dimensional structures in highly redundant data (Greenacre et al., 2022). The present component structure therefore suggested that although the data was nominally quite high in dimensions, there were a much smaller number of effective dimensions.

This was supported by the loading structure, which showed that the highly correlated predictors loaded mainly on the dominant component, and the other predictor loaded mainly on the second component. This shows how PCA can be used to uncover the underlying geometric structure that might not be apparent when examining the variables one by one. When there is significant covariance or correlation among the original variables, then PCA can be used to provide efficient low dimensional representations of the multivariate information as pointed out by Jolliffe and Cadima (2016).

The main methodological result of the study is the PCR results. Replacing the original correlated predictors with orthogonal component scores completely eliminated the multicollinearity issue and had essentially the same regression to variable ability as the baseline regression. This follows the basic principle of PCR which involves transforming correlated explanatory variables into uncorrelated components prior to the estimation of the regression. Similarly, Davino et al. (2022) showed that component-based regression could be an effective tool to mitigate high correlation, which would otherwise make estimation difficult, by transforming estimation from the original correlated variables to the components. This is especially true in the case of multicollinearity, which is quite apparent in the PCR collinearity diagnostics. The latter PC's had no linear dependence among each other, as all PC's are orthogonal by construction, unlike the original predictor space. This is the reason why numerical stability of PCR can be achieved even in the presence of very ill-conditioned original regression matrix. In particular, Liu et al. (2003) showed how the use of PCR could be achieved in SPSS and how it could transform correlated predictors into uncorrelated principal components prior to regression analysis. Importantly, the mean increase in the explained variance was not significant over that of the baseline model when using PCR. This is not a methodological problem. The biggest benefit seen here was stability and dimensionality reduction not

greater predictive power. PCR maintained the predictive signal contained within the original predictor structure, but also presented the predictive information in a much more compact and mathematically stable predictor representation. Gu & Cheung (2023) pointed out that key to selecting components for PCR is the careful consideration of the selection of the components, as the components that explain predictor variance may not maximize the prediction of the outcome.

This is also a significant drawback of the PCA-based regression. The principal components are created based on the variation of predictors but not necessarily based on the relationship between predictors and response. Thus, a component with a large amount of predictor variance could not be very predictive. Cook (2018) pointed out that dimension-reduction methods can be fundamentally split between methods that build a reduced representation just from the predictors and those that use information about the response in constructing a reduced representation. Preserving predictor variance and maximizing predictive performance should not be considered synonymous goals.

Residual diagnostics are a further test to assist with the numerical sufficiency of the PCR specification. An unusually large, standardized residual was noted but influence diagnostics showed no evidence of individual observations having strong influence on the modelling. The separation between an unusual residual and an influential observation is important because an influential observation may have a relatively large prediction error but not significantly change the estimated model parameters. The diagnostic evidence thus indicated that the major methodological issue faced was that of predictor dependence, not that of excess influence from the one-off observations.

The findings also have implications for strategies to retain components. Although the second component had a slightly lower eigenvalue than the commonly used Kaiser criterion, it was retained as it captured a separate dimension that would have been lost if retained. Through mechanical variance thresholds, only a subset of the dimensions may be selected, leading to the possible loss of meaningful dimensions for the structure or prediction problem. Fleming and Garen (2022) showed the need to carefully choose and validate components in the application of PCR in predictive contexts.

Other methods might however be considered in future studies. When the main goal is to predict responding variables, regularization techniques like LASSO, elastic net, robust PCR, supervised dimension reduction or partial least squares can offer some insights. Model-selection methods are especially relevant in high dimensional regression as their primary goal is to achieve different trade-offs between sparsity, prediction, interpretability and estimation stability (Lee et al., 2019). In the same way, tuning and validation procedures gain more significance with increasing complexity of models due to tuning the regularization parameters or data-driven model components (Wu & Wang, 2020).

The overall results show that apparent high dimensionality does not always translate into an effective high dimensionality. With extreme redundancy, many variables cannot be part of a predictor matrix that contains only a few independent information dimensions. This structure was well understood by PCR, and they were able to replace the original predictor representation with orthogonal components, which is unstable. However, regression should be used with caution using components because maximizing the variance of the predictors is not equivalent with maximizing the outcome prediction. The results of the present study thus support PCR as a method to reduce the dimensionality and control for multicollinearity but indicate that regularized and supervised methods should be considered in future comparative studies if gaining improved predictive accuracy is the primary goal.

5. Conclusion

This study illustrated the need for predictor dependence in the modelling of high-dimensional data, as well as effective dimensionality reduction. In this data set, there were 150 predictor variables, but correlation and multicollinearity diagnostics showed that many of the predictors were very redundant. However, due to the extreme dependence of observations, a conventional OLS regression could not estimate the intended predictor model at the same time. This is an example of how individual regression coefficients are not interpretable in the case of near-linear redundancy of predictors. This was an effective solution that PCA was able to offer, that is, to transform the original correlated predictors into two orthogonal, mutually orthogonal components that together accounted for virtually all the predictor variances. By repeating the procedure with subsequent PCR, the explanatory performance of the baseline regression was preserved, but the multicollinearity problem was eliminated. The greatest advantage of PCR was thus not an improvement in the amount of variance that could be explained, but greater numerical stability, dimensionality reduction, and model parsimony. It also shows that a dataset with many measured variables does not have to have equally high effective dimensionality. If the predictors are highly redundant, then there may be a few components sufficient to capture the information structure. Therefore, PCR is a practical solution to severe multicollinearity for regression modelling. Future studies should compare the PCR method with other regularised and/or supervised methods, such as ridge regression, LASSO, elastic net or partial least squares, and evaluate the predictive performance and the generalisation of the resulting models using cross-validation to gain insights into whether further improvements can be made.

References:

[1]. Auddy, A., Xia, D., & Yuan, M. (2025). Tensors in high-dimensional data analysis: Methodological opportunities and theoretical challenges. Annual Review of Statistics and Its Application, 12, 527–551.

[2]. Bühlmann, P., Kalisch, M., & Meier, L. (2014). High-dimensional statistics with a view toward applications in biology. Annual Review of Statistics and Its Application, 1, 255–278.

[3]. Cai, T. T., Guo, Z., & Xia, Y. (2023). Statistical inference and large-scale multiple testing for high-dimensional regression models. TEST, 32, 1135–1171.

[4]. Cook, R. D. (2018). Principal components, sufficient dimension reduction, and envelopes. Annual Review of Statistics and Its Application, 5, 533–559.

[5]. Davino, C., Romano, R., & Vistocco, D. (2022). Handling multicollinearity in quantile regression through the use of principal component regression. Metron, 80, 153–174.
[6]. Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., García Márquez, J. R., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., & Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. Ecography, 36(1), 27–46.
[7]. Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70(5), 849–911.
[8]. Filzmoser, P., & Nordhausen, K. (2021). Robust linear regression for high-dimensional data: An overview. WIREs Computational Statistics, 13(4), e1524.
[9]. Fleming, S. W., & Garen, D. C. (2022). Simplified cross-validation in principal component regression (PCR) and PCR-like machine learning for water supply forecasting. Journal of the American Water Resources Association, 58(4), 517–524.
[10]. Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, 33(1), 1–22.
[11]. Greenacre, M., Groenen, P. J. F., Hastie, T., Iodice D'Enza, A., Markos, A., & Tuzhilina, E. (2022). Principal component analysis. Nature Reviews Methods Primers, 2, Article 100.
[12]. Gu, F., & Cheung, M. W.-L. (2023). A model-based approach to multivariate principal component regression: Selecting principal components and estimating standard errors for unstandardized regression coefficients. British Journal of Mathematical and Statistical Psychology, 76(3), 605–622.
[13]. Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55–67.
[14]. Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202.
[15]. Lee, E. R., Cho, J., & Yu, K. (2019). A systematic review on model selection in high-dimensional regression. Journal of the Korean Statistical Society, 48(1), 1–12.
[16]. Lee, S., Seo, M. H., & Shin, Y. (2016). The Lasso for high dimensional regression with a possible change point. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 78(1), 193–210.
[17]. Liu, R. X., Kuang, J., Gong, Q., & Hou, X. L. (2003). Principal component regression analysis with SPSS. Computer Methods and Programs in Biomedicine, 71(2), 141–147.
[18]. Meinshausen, N., & Bühlmann, P. (2010). Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72(4), 417–473.
[19]. O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. Quality & Quantity, 41(5), 673–690.
[20]. Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267–288.
[21]. Wu, Y., & Wang, L. (2020). A survey of tuning parameter selection for high-dimensional regression. Annual Review of Statistics and Its Application, 7, 209–226.
[22]. Zou, H. (2006). The adaptive Lasso and its oracle properties. Journal of the American Statistical Association, 101(476), 1418–1429.
[23]. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301–320.