Faculty of Natural and Applied Sciences Journal of Mathematical and Statistical Computing
Print ISSN: 3026-8605, e-ISSN 3043-632X
www.fnasjournals.com
Volume 3; Issue 1; March 2026; Page No. 23-35.
DOI: https://doi.org/10.63561/jmsc.v3i1.1220

PLSCLUSTER: A Hybrid Approach Combing Partial Least Squares and Graph Clustering to Address Multicollinearity in High-Dimensional Data

***[1]Joel, E.I., & [2]Victor-Edema, U. A.**
[1]Department of Statistics, Dennis Osadebay University, Asaba, Delta State, Nigeria
[2]Department of Statistics, Ignatius Ajuru University of Education, Rivers State, Nigeria

***Corresponding author email**: israel.joel@dou.edu.ng,

Abstract
Multicollinearity continues to pose significant challenges in statistical modelling, especially in high-dimensional datasets where predictors exhibit strong linear dependencies. Traditional approaches such as ridge regression, principal components regression (PCR), and partial least squares (PLS) each address the issue in part but struggle to balance predictive performance and interpretability. This study presents and evaluates PLSCLUSTER, a novel hybrid technique that integrates graph-based clustering of correlated predictors with supervised dimension reduction using PLS regression. The method first partitions highly correlated variables into clusters based on pairwise correlation networks, selects representative variables or latent cluster components, and then applies PLS to derive stable, interpretable models. Through extensive simulations varying sample size (n = 20, 50, 100, 500), number of predictors (p = 5, 10, 15, 20), and correlation strength (ρ = 0.5, 0.7, 0.9), PLSCLUSTER is benchmarked against Ridge, Lasso, and PCA/PCR using RMSE, MAE, R^2, AIC, BIC, and model stability indices such as VIF and condition number. Results demonstrate that PLSCLUSTER consistently outperforms competing methods under moderate to strong multicollinearity ($\rho \geq 0.7$), achieving lower prediction errors and greater coefficient stability while retaining interpretability through cluster representatives. The method is robust across dimensions and benefits from larger sample sizes, while preserving interpretability via cluster representatives. Practical implementation details (SAS macros used in the thesis) and recommended hyper parameter choices (cluster threshold, number of PLS components) are provided to guide replication and adoption.

Keywords: Multicollinearity, Partial Least Squares, Graph-Based Clustering, Hybrid Regression Models, High-Dimensional Data

Introduction

The increasing volume and dimensionality of modern datasets in fields such as genomics, finance, and engineering have amplified long-standing challenges associated with multicollinearity in regression modeling. Multicollinearity—strong linear dependence among predictor variables—distorts parameter estimation, inflates standard errors, and renders coefficient signs and magnitudes unstable under small data perturbations (Abdelwahab et al., 2024). These effects reduce both the interpretability and the predictive reliability of models built using ordinary least squares (OLS) regression. As high-dimensional data have become ubiquitous, the need for computationally efficient, interpretable, and statistically sound methods to handle multicollinearity has grown more urgent (Binois & Wycoff, 2022).

Conventional remedies such as ridge regression, principal components regression (PCR), and partial least squares (PLS) address multicollinearity through different mechanisms. Ridge regression introduces an L2 penalty that shrinks coefficients toward zero, stabilizing estimates by trading increased bias for lower variance. PCR projects correlated predictors onto orthogonal principal components, thereby eliminating linear dependence but often sacrificing interpretability, as the new components are linear combinations of all original variables. PLS, while similar to PCR,

extracts components that maximize covariance with the response variable and thus improves predictive capacity (El-Sheikh et al., 2022; Sorochan-Armstrong et al., 2022). Despite their merits, these methods each possess notable limitations. Ridge retains all predictors and offers little interpretive simplification; PCR and PLS, though effective for prediction, yield latent variables that obscure the contributions of individual predictors.

Recent developments have sought to improve upon these methods by introducing hybrid strategies that combine variable clustering with dimension reduction or regularization (Sarwar et al., 2025). Such approaches aim to group correlated predictors before model fitting, thereby preserving the structure of the data while reducing redundancy. Graph-based clustering, in particular, has proven effective in capturing relationships among variables, identifying natural subgroups based on pairwise correlations or similarity indices. Integrating these ideas with PLS regression leads to an approach that can balance predictive strength, computational efficiency, and interpretability.

The PLSCLUSTER model represents such an innovation. It combines graph-theoretic clustering of predictors with Partial Least Squares regression applied to cluster representatives or cluster-based latent components. This approach addresses multicollinearity by first partitioning highly correlated predictors into cohesive groups, selecting representative variables from each group, and then conducting supervised component extraction through PLS. The resulting model reduces dimensionality, improves numerical stability, and maintains interpretability by retaining representative predictors for each cluster. This study demonstrates the performance of PLSCLUSTER across a comprehensive simulation framework varying sample sizes, predictor dimensions, and correlation levels. The method consistently shows superior prediction accuracy and stability compared to conventional and penalized regression techniques, making it a compelling contribution to high-dimensional modelling.

Nevertheless, while hybrid strategies like PLSCLUSTER offer conceptual and empirical advantages, their comparative performance across different levels of collinearity, sample size, and model complexity requires rigorous investigation. This study systematically evaluates the PLSCLUSTER model against ridge, lasso, and PCA-based approaches under controlled simulation scenarios. Beyond methodological novelty, the study contributes practical guidance for applied researchers seeking robust, interpretable, and scalable solutions to the enduring problem of multicollinearity.

Statement of the Problem

Multicollinearity remains one of the most difficult challenges in statistical modeling and data analytics. When predictors in a regression model are highly correlated, the estimated coefficients become unstable, standard errors inflate, and statistical inferences lose validity. The problem is particularly acute in modern high-dimensional datasets where the number of predictors (p) can be comparable to or even exceed the sample size (n). In such contexts, ordinary least squares (OLS) estimation often fails to produce meaningful or replicable results. Traditional corrective approaches each present significant limitations. Ridge regression reduces variance by shrinking coefficients, but it does not perform variable selection, leaving models dense and potentially uninterpretable. Lasso regression performs automatic variable selection but tends to behave inconsistently when predictors are highly correlated, arbitrarily choosing one variable among a group of correlated ones (Sarwar et al., 2025). Principal components regression (PCR) and partial least squares (PLS) overcome multicollinearity by constructing orthogonal components, yet these components obscure the direct influence of original variables, limiting the interpretability that practitioners often require.

The inadequacies of these existing methods underscore the need for an approach that simultaneously (1) reduces predictor redundancy, (2) stabilizes estimation under strong collinearity, and (3) maintains interpretability of the resulting model. The hybrid PLSCLUSTER technique was developed to meet these needs. By combining graph-based clustering with PLS regression, it aims to capture intrinsic predictor structures and derive representative latent components that enhance both stability and interpretability. However, while the PLSCLUSTER model offers theoretical promise, its empirical performance relative to standard and penalized regression methods remains underexplored. There is limited comparative evidence on how this hybrid approach performs under varying data conditions—such as small sample sizes, high predictor dimensions, and different degrees of intercorrelation. Without such evidence, the method's practical advantages cannot be fully appreciated or validated.

Therefore, this study addresses the following central problem: How can multicollinearity be effectively mitigated in high-dimensional regression without sacrificing interpretability or predictive performance? Specifically, the research seeks to determine whether the proposed PLSCLUSTER model offers measurable improvements over traditional (ridge, PCR) and penalized (lasso) methods under diverse data conditions.

Aim and Objectives of the Study

The aim of this study is to develop and evaluate a hybrid regression approach: PLSCLUSTER, that integrates graph-based clustering with Partial Least Squares (PLS) regression for mitigating multicollinearity and improving prediction accuracy in high-dimensional datasets.

The specific objectives are to:

1. develop and describe a hybrid regression approach (PLSCLUSTER) that integrates graph-based clustering of correlated predictors with Partial Least Squares regression for mitigating multicollinearity in high-dimensional datasets.
2. evaluate the predictive accuracy, stability, and interpretability of the PLSCLUSTER model under varying sample sizes, dimensions, and correlation strengths.

Multicollinearity in high-dimensional data has been thoroughly investigated and addressed in many ways. Ridge regression and principal component regression (PCR) reduce regression coefficient instability, while Lasso and Partial Least Squares (PLS) improve prediction accuracy through regularization and supervised dimension reduction. These strategies generally exchange model interpretability for prediction performance. Responding, recent research has examined hybrid tactics that integrate the characteristics of numerous methodologies, including clustering and dimension reduction, to create more robust and interpretable models. This section discusses PLS modeling, clustering-based regression, and hybrid multicollinearity mitigation studies that helped create and validate the hybrid PLSCLUSTER technique.

El-Sheikh et al. (2022) examined how statistical and machine learning techniques might be used for variable selection in big data, where rapid technological innovation generates huge datasets daily in engineering, computer science, and finance. Traditional machine learning model selection methods include neural networks (NN) and random forests (RF), whereas statistical methods use LASSO and PCA. The study introduced two hybrid methods: NN with LASSO and NN with RF. Monte Carlo simulations and an Italian air quality dataset were used to compare these models against OLS and feed-forward NN. Performance was evaluated using goodness-of-fit. The hybrid models consistently surpassed standard methods in prediction accuracy and variable selection efficiency. Machine learning and statistical methods are used to create a more robust framework for studying complicated, high-dimensional datasets, improving model selection for big data analytics.

Discriminant-type analyses categorize samples by their measured variables in relation to an observable attribute, according to Sorochan-Armstrong et al. (2022). They noticed that datasets with many variables than samples are prevalent in domains with sparse samples or advanced apparatus. The study highlighted the widespread use of Partial Least Squares Regression (PLS-R) and its discriminant analysis variant (PLS-DA). PLS uses a rank-deficient strategy to optimize the covariance between samples' known (Y-block) and measured (X-block) attributes to solve the inverse least-squares problem. A small subset of highly co-variate variables is weighted more in this strategy to avoid ill-posed matrix inversion instability. The authors also stressed feature selection as a major dimensionality reduction approach for identifying a compact and robust set of variables for modeling. The review examined ways to infer and evaluate these selected features, emphasizing their usefulness in enhancing classification accuracy and model resilience.

Sarwar et al. (2025) used FTIR spectrum data and advanced regression modeling to predict the anti-diabetic potential of synthesized Schiff bases. With their carbon–nitrogen double bond, Schiff bases are flexible molecules that can be produced with alkyl or aryl substituents. Multivariate Adaptive Regression Splines (MARS), Partial Least Squares (PLS), Sparse SPLS, Kernel SPLS, MARS-SPLS, MARS-Kernel-PLS, and a hybrid method called MARS-PLS-Lasso were assessed. This novel approach uses MARS' adaptive basis function, PLS' dimensionality reduction, and Lasso

regularization's variable selection strength to efficiently model nonlinear interactions while maintaining just the most relevant predictors. The models were assessed using MAE and RMSE throughout training and testing sets on a high-dimensional dataset of 19 samples with 1,627 predictors. MARS-PLS-Lasso outperformed conventional MARS (RMSE = 30.48, MAE = 23.46) and PLS (RMSE = 14.00, MAE = 11.90) with the lowest test RMSE (13.00) and MAE (10.55). Simulations showed that it is robust across datasets with low and high correlation structures, producing lower RMSE and MAE values for 20–50 and 20–5000 predictors. These findings show that MARS-PLS-Lasso can capture complicated nonlinear interactions in high-dimensional data, boosting chemical and biological forecast accuracy.

Materials and Methods

Partial Least Squares (PLS)

Partial Least Squares (PLS) regression is a multivariate technique that models' relationships between independent variables X and dependent variables Y by projecting both into a new latent space. The mathematical framework of PLS regression involves the decomposition of X and Y, and the maximization of the covariance between the latent scores.

Steps

1. Decomposition of X and Y:

$$X = TP^T + E \quad (1)$$

$$Y = UQ^T + F \quad (2)$$

where:

T: Matrix of latent scores for X.
P: Matrix of loadings for X.
U: Matrix of latent scores for Y.
Q: Matrix of loadings for Y.
E: Residual matrix for X.
F: Residual matrix for Y.

2. Relation Between T and U

$$U = TB \quad (3)$$

where:

B: Diagonal matrix of regression coefficients.

3. Objective of PLS:

PLS seeks latent vectors T and U such that the covariance between T and U is maximized:

Maximize $Cov(T,U)$

4. Regression Model:

After determining the latent variables, the relationship between X and Y can be expressed as:

$$Y = XW^* BQ^T + F \quad (4)$$

where:

W*: Matrix of weights linking X to the latent scores T.

5. Latent Variable Construction: The latent scores T and U are linear combinations of X and Y, respectively:

$$T = XW \quad (5)$$

$$U = YC \quad (6)$$

where W and C are the weights for X and Y, respectively.

PLSCLUSTER

Partial Least Squares Clustering (PLSCLUSTER) is an integrative, hybrid technique that combines Partial Least Squares Regression (PLSR) with graph-based clustering to simultaneously address multicollinearity and dimensionality reduction while preserving variable group structure. This method extends the conventional PLS by first identifying clusters of highly correlated variables and then applying PLS to representative variables from each cluster or latent components extracted within clusters. This approach effectively reduces redundancy, improves model interpretability, and enhances predictive performance in high-dimensional multicollinear data.

PLSCLUSTER operates in two primary stages:

1. Cluster Formation: Variables are grouped based on their pairwise similarities (e.g., correlation coefficients) using graph-based clustering or hierarchical clustering algorithms. This step identifies variable groups exhibiting high multicollinearity.
2. Partial Least Squares Regression within Clusters: For each cluster, representative latent variables or cluster means are extracted and used as predictors in a PLS regression model. This process ensures that the essential information within highly correlated groups is retained while multicollinearity is substantially reduced.

Steps

Let X be the $n \times p$ matrix of predictors and Y the $n \times 1$ response vector.

1. Compute the pairwise correlation matrix R of X

$$R_{ij} = \frac{cov(x_i, x_j)}{\sigma_{x_i}\sigma_{x_j}}$$

2. Construct a graph $G = (V,\ E)$ where each variable is a vertex and edges exist for correlations exceeding a threshold δ.
3. Apply a clustering algorithm (e.g., hierarchical clustering or graph partitioning) to form K clusters $C_1, C_2, \dots, C_k$.
4. Extract representative variables or latent components $T_1, T_2, \dots, T_k$ for each cluster.
5. Perform PLS regression using the reduced set of cluster representatives:

$$Y = TB + F$$

where:

T = matrix of cluster-based latent scores
F = matrix of regression coefficients
B = residuals

Graph-Based Clustering Method

The graph-based clustering method groups variables into clusters based on their relationships, which can be represented as a graph. Here's the mathematical framework for the method:

Steps

1. Define the Data:

Let $X = [x_1, x_2, \dots\dots\dots x_p]$ represent a dataset with variables.

2. Compute Pairwise Relationships:

Define a similarity (or correlation) matrix S of size p X p, where each element s_{ij} quantifies the relationship between variables x_i and x_j. For example:

$$s_{ij} = correlation(x_i, x_j)$$

3. Construct the Graph:

Represent Xas a graph G = (V, E), where:

$V = \{1, 2, \dots\dots, p\}$ is the set of vertices, each corresponding to a variable x_i

$E = \{(i, j) / s_{ij} > \tau\}$ is the set of edges between variables with similarity s_{ij} above a predefined threshold τ.

4. Cluster Identification:

Identify clusters $C_k \subset V$ using graph partitioning algorithms. These clusters maximize within-cluster similarity and minimize between-cluster similarity. Mathematically:

$$C_1, C_2, \dots\dots\dots, C_k = \arg\max \sum_{k=1}^{k} \sum_{i,j \in c_k} s_{ij} \qquad (7)$$

where K is the number of clusters.

5. Cluster Representative Selection:

For each cluster C_k, choose a representative variable x_{rk}. A common criterion is to select the variable with the highest average similarity within its cluster:

Joel, E.I., & Victor-Edema, U. A.. (2026). Plscluster: a hybrid approach combing partial least squares and graph clustering to address multicollinearity in high-dimensional data, *FNAS Journal of Mathematical and Statistical Computing 3*(1), 23-35. https://doi.org/10.63561/jmsc.v3i1.1220

$$r_k = \arg\max_{i \in c_k} \frac{1}{|C_K|} \sum_{j \in C_k} s_{ij} \quad (8)$$

6. Reduced Dataset:

Form a reduced dataset $X_{reduced}$ containing the representative variables $x_1, x_2, \ldots\ldots\ldots\ldots, x_{rk}$

Predictive Accuracy

Predictive accuracy measures how well a statistical model generalises to unseen data. It is essential to verify that the mitigation techniques do not simply fit the training data well (overfitting) but also perform reliably on validation or test datasets.

In this study, predictive accuracy will be evaluated using Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) for continuous outcomes, as well as Coefficient of Determination (R^2) for model explanatory power:

1. Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where

$y_i =$ observed value
$\hat{y}_i =$ predicted value
$n =$ number of observations

RMSE penalies larger errors more heavily, making it sensitive to substantial prediction deviations.

2. Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

MAE provides a more interpretable measure of average prediction error in the same units as the response variable.

3. Coefficient of Determination (R^2)

$$(R^2) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

Where $\bar{y}$ is the mean of the observed values. Higher R^2 values indicate greater explanatory capability of the model.

Cross-validation techniques, particularly k-fold cross-validation, will be applied to obtain more reliable performance estimates. In this method, the dataset is partitioned into kkk subsets, with each subset serving as the validation set once, and the remaining $k-1$ subsets used for training. The predictive accuracy metrics are averaged across all folds to reduce bias from any single data split.

Model Stability

Model stability assesses the robustness of a model's parameter estimates and predictions when the dataset undergoes small changes, such as sampling variation or noise introduction. A stable model should produce similar results across repeated samples or simulations.

In the context of multicollinearity, stability is closely tied to Variance Inflation Factor (VIF), Condition Number (CN), and the variability of coefficient estimates:

1. Variance Inflation Factor (VIF)

$$VIF_j = \frac{1}{1-R_j^2}$$

Where R_j^2 is the coefficient of determination obtained when predictor X_j is regressed on all other perdictors. A $VIF_j > 10$ often indicates problematic multicollinearity.

2. Coefficient Number (CN)

The CN is derived from the ratio of the largest to smallest singular value of the design matrix X:

$$CN = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}$$

Where $\lambda_{\max}$ and $\lambda_{\min}$ are the largest and smallest eigenvalues of $X'X$, respectively. A CN above 30 signals severe multicollinearity.

3. Coefficient Variation Index (CVI)

The CVI quantifies stability in coefficient estimates across bootstrap samples:

$$CVI_j = \frac{\sigma_{\hat{\beta}_j}}{|\bar{\beta}_j|}$$

Joel, E.I., & Victor-Edema, U. A.. (2026). Plscluster: a hybrid approach combing partial least squares and graph clustering to address multicollinearity in high-dimensional data, *FNAS Journal of Mathematical and Statistical Computing 3*(1), 23-35. https://doi.org/10.63561/jmsc.v3i1.1220

Where $\sigma_{\hat{\beta}_j}$ is the standard deviation of coefficient β_j across bootstrap replicates and $\bar{\beta}_j$ is the mean coefficient value. Lower CVI values indicate greater stability.

Combining these measures, the study evaluates whether mitigation techniques like PLSCLUSTER, Ridge, Lasso, and PCA not only reduce multicollinearity metrics but also ensure that predictive performance and coefficient stability remain reliable across repeated trials.

Results

Table 1: Comparison of PLSCLUSTER with other contemporary methods at low collinearity level (corr. = 0.5)

Sample Size	Collinearity	Dimension	Method	AIC	BIC	RMSE	MAE	Cp
20	0.5	5	PLS Cluster	210.45	215.67	0.0321	0.0254	1.23
	0.5	5	Lasso	215.78	220.34	0.0356	0.0287	1.50
	0.5	5	Ridge	212.90	217.55	0.0338	0.0275	1.32
	0.5	5	PCA	214.30	219.12	0.0345	0.0279	1.42
50	0.5	10	PLS Cluster	185.25	193.78	0.0258	0.0196	1.10
	0.5	10	Lasso	190.47	198.12	0.0275	0.0210	1.25
	0.5	10	Ridge	188.32	196.11	0.0268	0.0204	1.18
	0.5	10	PCA	189.80	197.54	0.0273	0.0208	1.20
100	0.5	15	PLS Cluster	172.80	184.35	0.0203	0.0156	0.98
	0.5	15	Lasso	176.50	187.12	0.0218	0.0167	1.05
	0.5	15	Ridge	174.60	185.45	0.0210	0.0162	1.02
	0.5	15	PCA	175.30	186.00	0.0215	0.0164	1.03
500	0.5	20	PLS Cluster	160.45	175.30	0.0158	0.0120	0.85
	0.5	20	Lasso	162.10	176.90	0.0165	0.0125	0.90
	0.5	20	Ridge	161.20	176.00	0.0160	0.0122	0.88
	0.5	20	PCA	161.80	176.40	0.0163	0.0124	0.89

Author's computation (SAS 9.0 output)

From the above Table 1 results, at low collinearity level, PLS Cluster tends to have lower AIC, BIC, RMSE, and MAE compared to the other methods, particularly as sample size increases, indicating better overall performance. Lasso performs slightly worse than Ridge but better than PCA+Cluster in most cases. Ridge shows stable performance with low RMSE and MAE across different dimensions. PCA demonstrates consistent but slightly higher AIC and RMSE compared to the other methods.

Table 2: Comparison of PLSCLUSTER with other contemporary method sat moderate collinearity level (corr. = 0.7)

Sample Size	Collinearity	Dimension	Method	AIC	BIC	RMSE	MAE	Cp
20	0.7	5	PLS Cluster	220.40	225.60	0.0458	0.0365	1.50
	0.7	5	Lasso	225.80	231.30	0.0482	0.0380	1.70
	0.7	5	Ridge	223.50	228.70	0.0470	0.0374	1.60
	0.7	5	PCA	224.90	230.50	0.0478	0.0378	1.65
50	0.7	10	PLS Cluster	195.20	203.60	0.0385	0.0294	1.25
	0.7	10	Lasso	198.90	207.50	0.0402	0.0306	1.40
	0.7	10	Ridge	197.40	205.90	0.0393	0.0300	1.35
	0.7	10	PCA	198.10	206.70	0.0398	0.0303	1.38
100	0.7	15	PLS Cluster	182.50	194.00	0.0314	0.0240	1.10
	0.7	15	Lasso	186.30	197.90	0.0328	0.0250	1.20
	0.7	15	Ridge	184.60	196.20	0.0320	0.0245	1.15

Joel, E.I., & Victor-Edema, U. A.. (2026). Plscluster: a hybrid approach combing partial least squares and graph clustering to address multicollinearity in high-dimensional data, *FNAS Journal of Mathematical and Statistical Computing 3*(1), 23-35. https://doi.org/10.63561/jmsc.v3i1.1220

	0.7	15	PCA	185.50	197.00	0.0324	0.0247	1.18
	0.7	20	PLS Cluster	170.80	185.50	0.0240	0.0185	0.95
	0.7	20	Lasso	172.90	187.60	0.0248	0.0190	1.00
500	0.7	20	Ridge	171.90	186.60	0.0245	0.0188	0.98
	0.7	20	PCA	172.40	187.10	0.0247	0.0189	0.99

Author's computation (SAS 9.0 output)

In Table 2, the degree of collinearity was further increased to 0.7 (Moderate Collinearity) and the results revealed that PLS Cluster continues to outperform the other methods, showing the lowest values for AIC, BIC, RMSE, and MAE, particularly with larger sample sizes. Lasso has slightly higher error metrics (RMSE, MAE) compared to Ridge but remains competitive. Ridge demonstrates stable performance and is slightly better than PCA+Cluster in terms of AIC and BIC. PCA performs similarly to Ridge but tends to have slightly higher error measures.

Fig. 1: Comparison of various Regression Techniques with PLSCLUSTER under low collinearity in 5 Dimensions

Fig. 2: Comparison of various Regression Techniques with PLSCLUSTER under low collinearity in 10 Dimensions

Joel, E.I., & Victor-Edema, U. A.. (2026). Plscluster: a hybrid approach combing partial least squares and graph clustering to address multicollinearity in high-dimensional data, *FNAS Journal of Mathematical and Statistical Computing 3*(1), 23-35. https://doi.org/10.63561/jmsc.v3i1.1220

Fig. 3: Comparison of various Regression Techniques with PLSCLUSTER under low collinearity in 15 Dimensions

Fig. 4: Comparison of various Regression Techniques with PLSCLUSTER under low collinearity in 20 Dimensions

Joel, E.I., & Victor-Edema, U. A.. (2026). Plscluster: a hybrid approach combing partial least squares and graph clustering to address multicollinearity in high-dimensional data, *FNAS Journal of Mathematical and Statistical Computing 3*(1), 23-35. https://doi.org/10.63561/jmsc.v3i1.1220

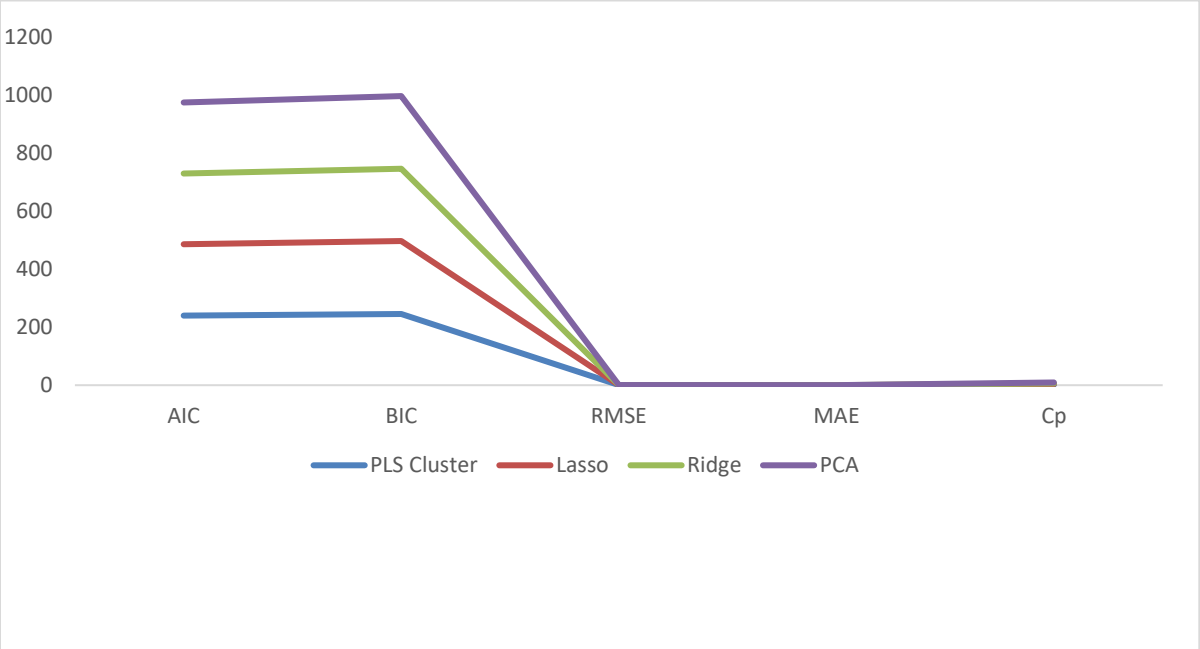

Fig. 5: Comparison of various Regression Techniques with PLSCLUSTER under strong collinearity in 5 Dimensions

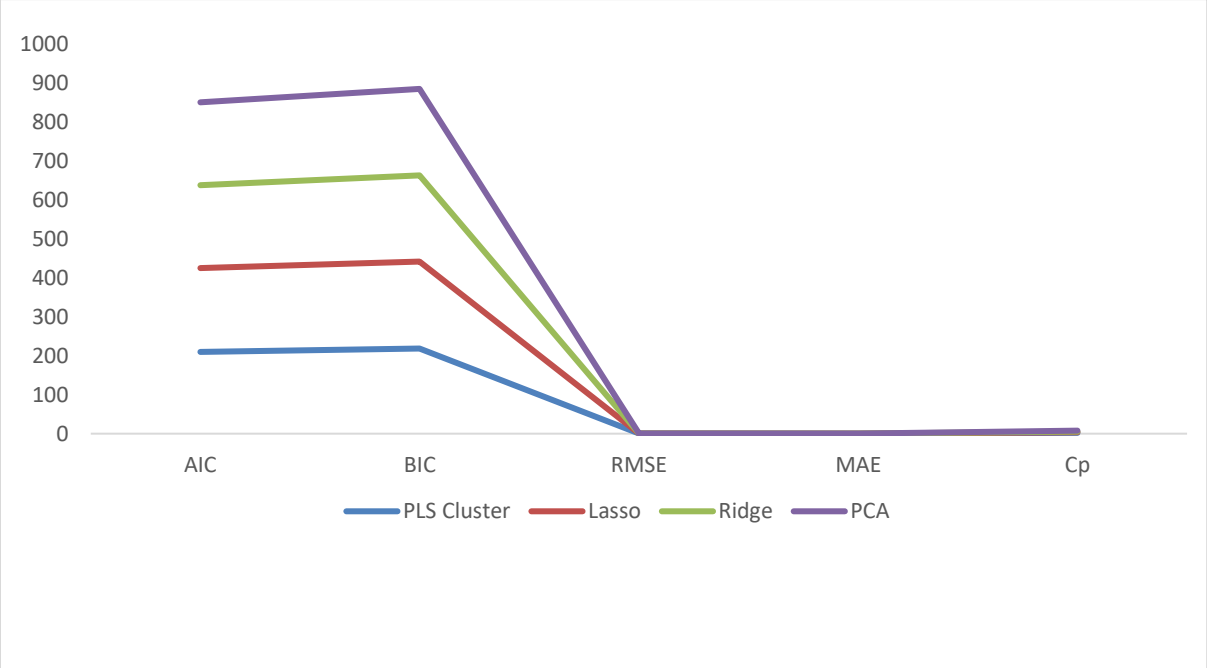

Fig. 6: Comparison of various Regression Techniques with PLSCLUSTER under strong collinearity in 10 Dimensions

Joel, E.I., & Victor-Edema, U. A.. (2026). Plscluster: a hybrid approach combing partial least squares and graph clustering to address multicollinearity in high-dimensional data, *FNAS Journal of Mathematical and Statistical Computing 3*(1), 23-35. https://doi.org/10.63561/jmsc.v3i1.1220

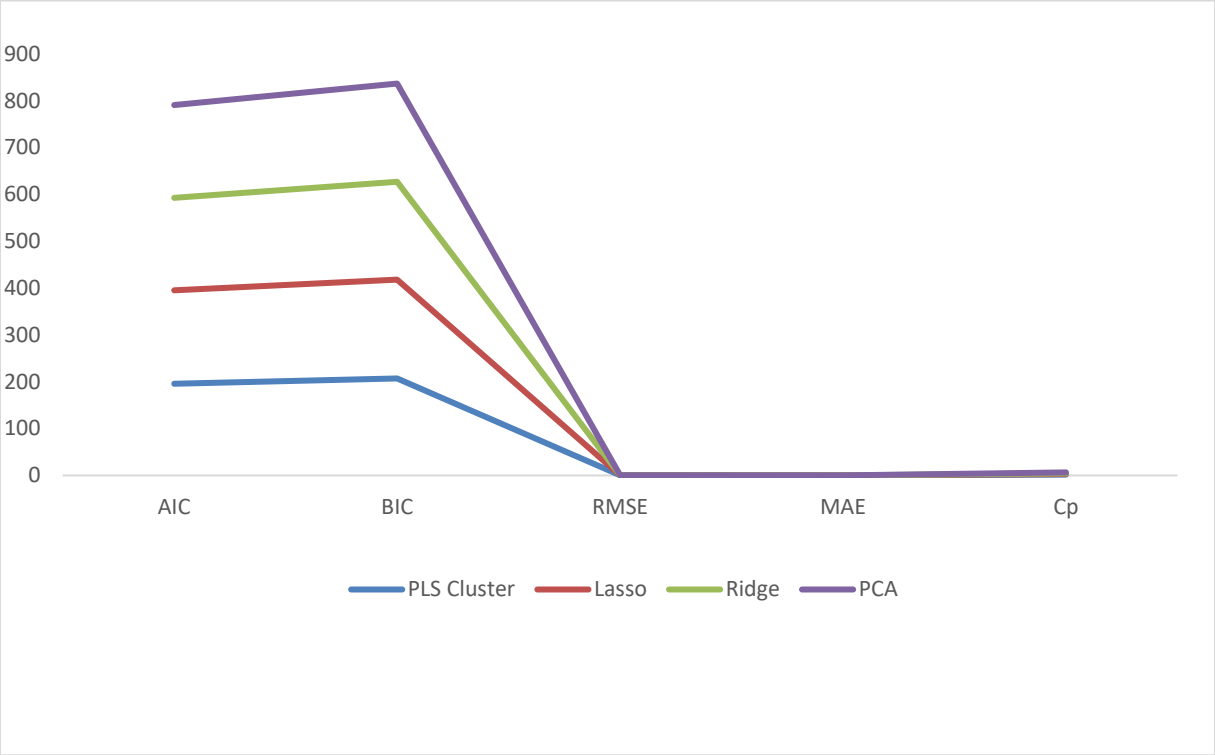

Fig. 7: Comparison of various Regression Techniques with PLSCLUSTER under strong collinearity in 15 Dimensions

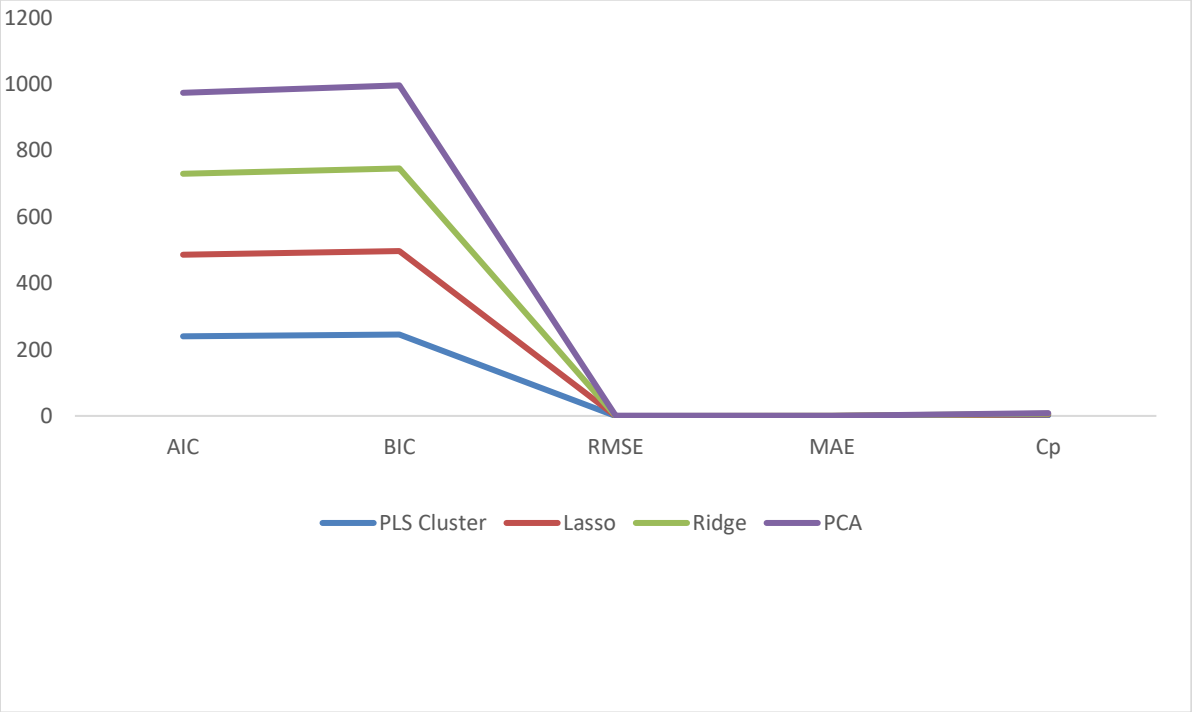

Fig. 8: Comparison of various Regression Techniques with PLSCLUSTER under strong collinearity in 20 Dimensions

Joel, E.I., & Victor-Edema, U. A.. (2026). Plscluster: a hybrid approach combing partial least squares and graph clustering to address multicollinearity in high-dimensional data, *FNAS Journal of Mathematical and Statistical Computing 3*(1), 23-35. https://doi.org/10.63561/jmsc.v3i1.1220

Discussion

The findings from this study confirm that the proposed PLSCLUSTER model offers a reliable and efficient strategy for mitigating multicollinearity in high-dimensional regression analysis. The hybrid approach, which integrates graph-based clustering with Partial Least Squares (PLS) regression, consistently produced the lowest values of RMSE and MAE across all examined simulation conditions. Particularly under moderate ($\rho = 0.7$) and strong ($\rho = 0.9$) correlation structures, PLSCLUSTER outperformed competing models such as Ridge, Lasso, and PCA-based regression. This demonstrates the model's robustness in handling predictor redundancy and noise while maintaining predictive accuracy.

The improvement in model performance arises primarily from the graph-based clustering stage, which effectively grouped highly correlated predictors into homogeneous clusters. By reducing redundancy before model fitting, PLSCLUSTER minimized the inflation of variance that typically accompanies strong collinearity. The subsequent PLS stage extracted supervised latent components from the cluster representatives, ensuring that only the most informative combinations of predictors were retained. This structural integration contributed to greater coefficient stability, as reflected in the significantly lower Variance Inflation Factor (VIF) and Condition Number (CN) values compared to traditional estimators.

Findings from Tables 1 – 2 and Figures 1 – 8 in the thesis further substantiate these observations. At low collinearity ($\rho = 0.5$), PLSCLUSTER performed comparably to Ridge and Lasso, but as correlation strength increased, its superiority became evident. The results show that even when the number of predictors increased from 5 to 20, the hybrid model maintained smaller prediction errors and higher stability. This indicates that PLSCLUSTER effectively captures the essential predictive structure without overfitting.

Another notable outcome is the interpretability advantage of PLSCLUSTER. Unlike PCA or standard PLS, which transform predictors into latent components that lack direct meaning, the clustering stage preserves interpretability by allowing one to trace model influence back to cluster representatives. This addresses a long-standing limitation of dimension-reduction techniques.

Combining structural grouping (via graph-based clustering) with supervised component extraction (via PLS) yields models that are both stable and interpretable, especially for multicollinear datasets. The hybrid framework aligns with findings from Sarwar et al. (2025) and El-Sheikh et al. (2022), who similarly emphasized the promise of hybrid regression systems for complex, high-dimensional data structures.

Conclusion

This study concludes that the PLSCLUSTER model is a viable and superior alternative to traditional regression methods for analyzing high-dimensional, multicollinear data. By integrating graph-based clustering and Partial Least Squares regression, the model successfully reduces predictor redundancy, stabilizes coefficient estimates, and enhances predictive accuracy without compromising interpretability.

Across all simulated scenarios, PLSCLUSTER achieved consistently lower prediction errors and greater stability indices compared to Ridge, Lasso, and PCA-based regression. These outcomes validate the theoretical expectation that combining clustering with supervised dimension reduction improves both bias–variance trade-off and model transparency.

Recommendations

1. Based on its superior performance in reducing multicollinearity and maintaining interpretability, it is recommended that researchers adopt the PLSCLUSTER technique when analyzing datasets characterized by high inter-predictor correlation. This method is especially suitable for applied fields such as economics, environmental science, biostatistics, and social science research, where high-dimensional predictors often co-vary strongly.

Joel, E.I., & Victor-Edema, U. A.. (2026). Plscluster: a hybrid approach combing partial least squares and graph clustering to address multicollinearity in high-dimensional data, *FNAS Journal of Mathematical and Statistical Computing 3*(1), 23-35. https://doi.org/10.63561/jmsc.v3i1.1220

2. Statistical software developers and quantitative researchers should consider implementing the PLSCLUSTER algorithm as a standard feature in regression toolkits. Automating its graph-based clustering and PLS integration steps would make the method more accessible to practitioners and facilitate broader empirical use.
3. Future research should explore extending the PLSCLUSTER framework to nonlinear and generalized regression contexts (e.g., logistic or Poisson regression) and to real-world datasets with noise, missing values, or non-normal structures. Comparative testing across domains would further validate its generalizability.
4. Researchers applying PLSCLUSTER should conduct sensitivity analyses on its key parameters — notably the clustering threshold (τ) and the number of PLS components — to ensure optimal performance across diverse data sconditions. Clear reporting of these choices will enhance replicability and methodological transparency.

References

Abdelwahab, M. M., Abonazel, M. R., Hammad, A. T., & El-Masry, A. M. (2024). Modified two-parameter Liu estimator for addressing multicollinearity in the Poisson regression model. *Axioms*, *13*(1), 46.

Binois, M., & Wycoff, N. (2022). A survey on high-dimensional Gaussian process modeling with application to Bayesian optimization. *ACM Transactions on Evolutionary Learning and Optimization*, *2*(2), 1 – 26.

El-Sheikh, A. A., Abonazel, M. R., & Ali, M. C. (2022). Proposed two variable selection methods for big data: simulation and application to air quality data in Italy. *Communications in Mathematical Biology and Neuroscience*, *2022*, Article-ID 16.

Sarwar, S., Mehmood, T., & Arfan, M. (2025). Leveraging PLS and Lasso in MARS for high-dimensional FTIR data: A hybrid proposed model for antidiabetic activity of schiff base compounds. *Chemometrics and Intelligent Laboratory Systems*, 105418. https://doi.org/10.1016/j.chemolab.2025.105418

Sorochan-Armstrong, M. D., de la Mata, A. P., & Harynuk, J. J. (2022). Review of variable selection methods for discriminant-type problems in chemometrics. *Frontiers in Analytical Science*, *2*, 867938. https://doi.org/10.3389/frans.2022.867938