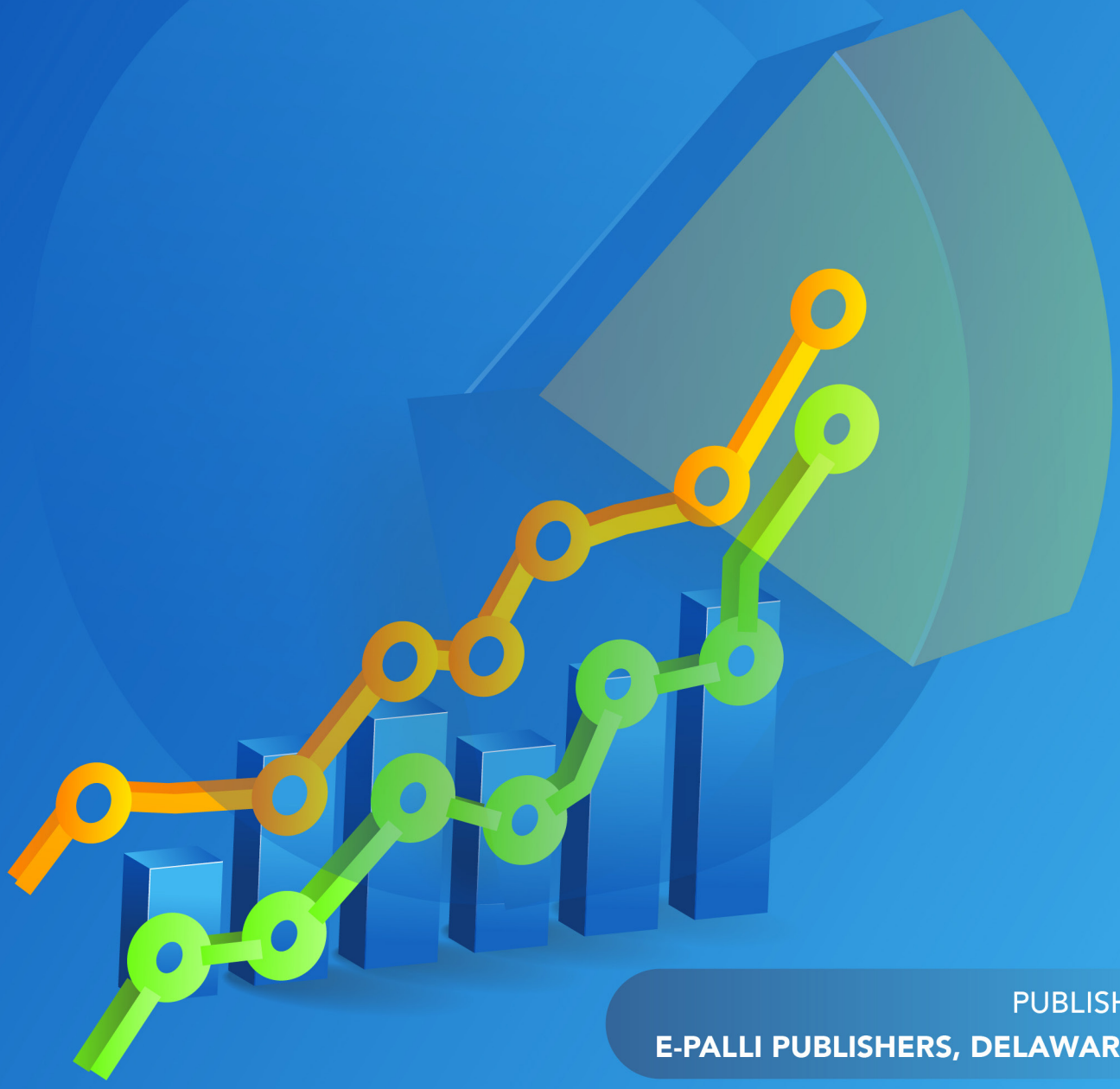
American Journal of Applied Statistics and Economics (AJASE)

ISSN: 2992-927X (ONLINE)

VOLUME 4 ISSUE 1 (2025)



PUBLISHED BY
E-PALLI PUBLISHERS, DELAWARE, USA

American Journal of Applied Statistics and Economics (AJASE)

Volume 4 Issue 1, Year 2025
ISSN: 2992-927X (Online)
DOI: https://doi.org/10.54536/ajase.v4i1.5938
https://journals.e-palli.com/home/index.php/ajase

Maximizing Predictive Regression and Dimensionality Reduction Techniques: Evidence from Monte Carlo's Simulation Study

Oluwafemi Clement Onifade[1*], Samuel Olayemi Olanrewaju[1], Emmanuel Segun Oguntade[1]

Article Information

Received: August 16, 2025

Accepted: September 19, 2025

Published: October 18, 2025

Keywords

Elastic Net, High-Dimensional Data, Lasso, Multicollinearity, Regularization, SCAD, SPCR

ABSTRACT

This study proposes a novel two-step sparse learning framework that combines Sparse Principal Component Regression (SPCR) with regularization methods, Lasso, Elastic Net, Ridge, and Smoothly Clipped Absolute Deviation (SCAD), to improve prediction and interpretability in high-dimensional settings. Simulation experiments were conducted under varying sample sizes, dimensionality levels, sparsity conditions, and predictor correlations to evaluate the performance of the hybrid estimators in comparison to traditional penalization approaches. Results show that SPCR-Lasso and SPCR-Enet consistently deliver superior accuracy and stability in high-dimensional, multicollinear contexts, with SPCR-Enet performing particularly well in extreme dimensionality. SPCR-SCAD demonstrated advantages in sparse, low-correlation scenarios, while Ridge regression contributed modest improvements. These findings underscore that estimator performance is strongly data-dependent and highlight the value of SPCR hybridization for mitigating multicollinearity while enhancing interpretability. The study offers practical guidance for applied researchers in fields such as genomics, finance, and climate science, and contributes methodologically by demonstrating the robustness of SPCR-based regularization in handling complex high-dimensional data structures.

INTRODUCTION

High-dimensional modelling has emerged as a critical and transformative area of research with profound implications across diverse domains, including data science, machine learning, and statistics. The prominence of high-dimensional data can be attributed to the prevalence of large-scale datasets and complex systems in various applications. In high-dimensional modelling, the term "high dimensional" refers to situations where the number of explanatory variables, denoted as p, exceeds the number of observations, n (i.e., $p > n$). This phenomenon has gained attraction due to the rapid advancements in technology, which enable the collection of a vast number of variables to better understand complex phenomena of interest. The applicability of high-dimensional modelling spans multiple fields, including computational chemistry, Chemometrics with spectral data, genomics, fMRI data analysis, large-scale healthcare analytics, text/image analysis, astronomy, and many others.

The versatility of high-dimensional modelling techniques has also been demonstrated in the field of drug discovery and development. For example, Priya *et al.* (2022) focus on the application of machine-learning approaches in chemo-informatics for drug discovery. Machine learning techniques, specifically QSAR (Quantitative Structure-Activity Relationship), have effectively modelled various physicochemical properties of drugs, including toxicity, absorption, and drug-drug interactions. These approaches, being a subset of artificial intelligence, show great potential in drug discovery by handling non-linear datasets and big data with increasing complexity. However, the curse of dimensionality, a well-known challenge in high-dimensional data, poses significant obstacles to accurate predictions and efficient parameter estimation. The exponential growth of data volume with increasing variables leads to sparse data points, which can hinder the effectiveness of traditional methods. Multicollinearity, a common issue in high-dimensional datasets, further complicates parameter estimation and can result in inflated confidence intervals.

To address these challenges, sophisticated techniques are required that can effectively handle the complexities posed by high-dimensional data. Dimensionality reduction and variable selection methods have emerged as attractive strategies to tackle high-dimensional studies. Over the last two decades, regularization approaches such as lasso, elastic net, ridge regression, and Smoothly Clipped Absolute Deviation (SCAD) have become the methods of choice for analyzing high-dimensional data. These regularization methods have been extensively applied in various disciplines, including statistics (Nwosu *et al.* 2024), chemo-informatics (Song *et al.,* 2024), epidemiology (Cleophas *et al.* 2024), and bioinformatics (Kitano *et al.,* 2024), and many others.

In recent years, Sparse Principal Component Regression (SPCR), and Sparse Partial Least Squares (SPLS), has garnered attention as a potential solution to improve predictive model accuracy. By identifying a small subset of the original predictor variables that capture most of the variance in the data, SPCR facilitates highly interpretable models with enhanced predictive accuracy. SPCR has demonstrated promising results in various fields,

[1] Department of Statistics, Faculty of Science, University of Abuja, Abuja, Nigeria
[*] Corresponding author's e-mail: onifade.oluwafemi@yahoo.com

including medical research, finance, and environmental sciences. SPCR has also proven successful in QSAR modelling by identifying the most relevant molecular descriptors that greatly influence biological activity or molecule properties. For example, Zhang *et al.* (2024) demonstrated the effectiveness of SPCR in identifying the most important features for predicting the antitumor activity of molecules, leading to more reliable QSAR models. While previous studies have predominantly focused on combining principal component regression (PCR) with regularization techniques in low-dimensional settings, where the number of predictors is less than the observations, there is a clear need for sparse PCR. Sparse PCR can be more advantageous in situations with a large number of predictor variables, as it identifies a smaller subset of the original predictors that are most crucial in predicting the response variable.

In this thesis, we aim to address the challenges posed by high-dimensional data through the application of regularization techniques and Sparse Principal Component Regression (SPCR). By developing a novel two-step sparse learning approach that integrates Sparse Principal Component Regression (SPCR) with regularization techniques (Ridge regression, Lasso, Elastic Net, and Smoothly Clipped Absolute Deviation). This combined approach seeks to enhance predictive accuracy and interpretability in high-dimensional datasets, particularly in scenarios involving multicollinearity and sparsity. The specific objectives include to:

i. Develop efficient framework that combine SPCR with regularization methods.

ii. Assess the performance of the combined approach using traditional modeling techniques like Lasso and Ridge through predictive accuracy measures, i.e. mean square error.

iii. Design a simulation study to demonstrate the robustness of the proposed approach across multiple high-dimensional datasets, varying in sample size, multicollinearity, and sparsity levels.

LITERATURE REVIEW

Empirical work on high-dimensional prediction has converged on two broadly successful strategies. The first reduces dimensionality via latent factors or components (e.g., Principal Component Regression — PCR), which mitigates multicollinearity and variance inflation (Jolliffe, 2002; Hastie *et al.*, 2009). The second directly penalizes regression coefficients to induce shrinkage and (sometimes) sparsity (Ridge, Lasso, Elastic Net, SCAD), which controls overfitting and performs variable selection when appropriate (Hoerl & Kennard, 1970; Tibshirani, 1996; Zou & Hastie, 2005; Fan & Li, 2001). Empirical comparisons show neither approach dominates across all data regimes: PCR is robust under extreme multicollinearity but produces components that are not tailored to prediction of the response, while penalized regressions are powerful for sparse signals but can struggle when predictors are highly correlated (Hastie *et al.,* 2009).

To bridge the gap between unsupervised dimension reduction and predictive goals, researchers developed Sparse Principal Component Analysis (SPCA) and Sparse Principal Component Regression (SPCR). SPCA (Zou *et al.*, 2006; Witten *et al.*, 2009) imposes sparsity on loadings so principal components involve only a subset of predictors, improving interpretability without discarding the variance-reduction benefit of PCA. Empirical studies in genomics, chemometrics, and neuroimaging have found SPCA yields components that are easier to interpret and often more useful as inputs for supervised tasks than dense PCA components.

SPCR, either formulated as a one-stage joint optimization of component extraction and regression loss or as a carefully tuned two-stage procedure, goes further by explicitly constructing components that optimize predictive performance (Kawano, 2018; Zou *et al.,* 2006). Empirical comparisons show SPCR often outperforms classical PCR when the directions of maximal predictor variance differ from the directions most predictive of the outcome (i.e., when supervised signal does not align with principal variance directions). Applications in biological data and other high-dimensional domains report improved prediction and sparser, more actionable component loadings (Zou *et al.,* 2006; Kawano, 2018).

There is substantial empirical evidence that different regularizers perform differently depending on correlation structure and sparsity. Ridge excels when many predictors carry signal but are highly correlated; it reduces variance without producing sparse solutions, often improving out-of-sample prediction in dense-signal, collinear settings (Hoerl & Kennard, 1970). Also, Lasso provides both shrinkage and variable selection and works well when the true model is sparse and predictors are not excessively collinear; empirical studies show it can fail to reliably select the "correct" group in the presence of strong predictor correlation (Tibshirani, 1996). Elastic Net empirically combines strengths of Ridge and Lasso, grouping correlated predictors while performing variable selection; simulation and applied work show Elastic Net often outperforms Lasso under grouped-correlated designs (Zou & Hastie, 2005). SCAD and other nonconvex penalties (Fan & Li, 2001) demonstrate favorable oracle properties in theory and often reduced bias empirically compared to Lasso, but they require careful tuning and are more sensitive to initialization and optimization choices. Empirical simulation studies repeatedly demonstrate there is no uniformly best penalty: performance depends on (i) sparsity level, (ii) inter-predictor correlation, (iii) signal strength, and (iv) sample size. This motivates this study that systematically maps performance across different scenarios rather than relying on single-case comparisons.

MATERIALS AND METHODS

Development of Novel Two-Step Sparse Learning Techniques

To address the challenges of high-dimensional data analysis, this study integrates Sparse Principal Component

Regression (SPCR) with regularization techniques such as Ridge, Lasso, Elastic Net, and SCAD. The proposed two-step sparse learning approach combines the strengths of dimensionality reduction (SPCR) with the variable selection and regularization capabilities of these methods, ensuring both interpretability and predictive accuracy.

Step 1: Dimensionality Reduction Using Sparse Principal Component Regression (SPCR) Workflow

Compute sparse principal components T=XW by solving:
minimize$\|Y-XW\alpha\|_2^2+\lambda_1 \|W\|_1+\lambda_2 \|W\|_F^2$
Where W is the matrix of sparse component weights, α is the regression coefficient vector, and λ_1 and λ_2 control sparsity and shrinkage.

i. Select the top k components based on the proportion of variance explained and their relevance to Y.

ii. Output the reduced dataset T, a sparse representation of X.

Step 2: Regularized Regression on Reduced Components

After dimensionality reduction, apply regularized regression techniques (Ridge, Lasso, Elastic Net, and SCAD) to the reduced dataset T to build predictive models while managing overfitting and multicollinearity. This means that the resulting equations involve combining the dimensionality reduction framework with the respective penalty functions of the chosen regularization technique.

Ridge Regression with SPCR

Objective Function:
minimize$\|Y-T\beta\|_2^2+\lambda\|\beta\|_2^2$
Where,
T=XW: Sparse components derived using SPCR,
λ: Regularization parameter controlling the degree of shrinkage.
β: Regression coefficients
$\|\beta\|_2^2$:L2-norm penalty that shrinks all coefficients toward zero but does not enforce sparsity,

Lasso Regression with SPCR

Objective Function:
minimize$\|Y-T\beta\|_2^2+\lambda\|\beta\|_1$,
Where:
T=XW: Sparse components from SPCR,
$\|\beta\|_1$: L1-norm penalty that enforces sparsity by shrinking some coefficients to exactly zero,
Lasso regularization enhances variable selection by retaining only the most relevant components or predictors.

Elastic Net with SPCR

Objective Function:
minimize
$\|Y-T\beta\|_2^2+\lambda_1\|\beta\|_1+\lambda_2\|\beta\|_2^2$
Where,
T=XW: Sparse components from SPCR,
β: Regression coefficients
$\|\beta\|_1$: Enforces sparsity (Lasso component),
$\|\beta\|_2^2$: Mitigates multicollinearity and provides stability (Ridge component),
λ_1: Controls sparsity,
λ_2: Controls shrinkage.
Elastic Net is particularly effective when predictors are highly correlated, as it selects groups of correlated components.

SCAD with SPCR

Objective Function:
minimize$\|Y-T\beta\|_2^2+\sum_{j=1}^{p} p_\lambda(|\beta_j|)$,
where, T=XW: Sparse components from SPCR,
β: Regression coefficients
$p_\lambda(|\beta_j|)$ is the SCAD penalty function,
λ controls the penalty's strength.

Advantages of Combining Regularization with SPCR

i. SPCR reduces dimensionality while Ridge or Lasso enhances the predictive power by handling multicollinearity or enforcing sparsity.

ii. SCAD further refines the predictor selection process, reducing bias for large coefficients while retaining sparse predictors.

Simulation Study

The simulation study aims to evaluate and compare the performance of sparse learning methods (Lasso, Ridge, Elastic Net, SCAD, and SPCR) under controlled and varied conditions. This study focuses on predictive accuracy, interpretability, and computational efficiency, providing insights into the strengths and weaknesses of these methods in high-dimensional settings. The following sections detail the design, dataset characteristics, evaluation metrics, comparative testing procedures, and the approach for data analysis and interpretation

Design of the Simulation Study

The simulation study replicates real-world challenges by systematically varying key parameters: sample size, predictor dimensionality, levels of multicollinearity, noise, and sparsity. These variations ensure a comprehensive evaluation of the methods' performance across diverse conditions, reflecting practical scenarios in high-dimensional data analysis.

Sample Sizes

Four small sample scenarios are considered: n=30, n=50, n=70 and n=100. The small sample sizes (n=30 and n=50) represent the most challenging setting where predictors (p) far exceed observations (p>n), crucial for assessing the methods' ability to avoid overfitting. While higher small sample sizes (i.e. n=70 and n=100) explore scalability and performance in more balanced or low-dimensional settings.

Predictor Dimensionality

Predictor dimensionality (p) varies from low (p=20) to high (p=200). Low-dimensional scenarios allow methods to demonstrate baseline predictive capabilities

without dimensionality-related challenges. Moderate and high-dimensional settings introduce significant computational and statistical challenges, such as sparsity and multicollinearity.

Multicollinearity

Multicollinearity is varied at four levels:

Low (ρ~0.1): Predictors are weakly correlated, minimizing interference among variables.

Moderate (ρ~0.5): Predictors form correlated blocks, testing methods like Elastic Net and SPRC designed to handle such scenarios.

High (ρ~0.9): Many predictors are extremely interrelated, challenging methods like Lasso, which may arbitrarily select variables from correlated groups.

Sparsity

Predictor sparsity is varied to test variable selection capabilities:

• Sparse (10% non-zero coefficients): Only a small fraction of predictors are relevant, providing a benchmark for variable selection.

• Dense (30% non-zero coefficients): Many predictors have small, non-zero effects, testing the methods' capacity to identify subtle contributions.

Data Generation Process

The data is generated as: $Y = X\beta+\epsilon$

where:

$X\sim N(0,\Sigma)$, with $\Sigma_{ij}=\rho$ for $i\neq j$), controlling the level of multicollinearity

β is a sparse vector with randomly assigned non-zero coefficients drawn from N(0,1),

$\epsilon\sim N(0,\sigma^2)$ is Gaussian noise adjusted to achieve desired R^2.

Evaluation Metrics

Performance will be assessed using the following metrics

Mean Squared Error (MSE)

MSE measures the average magnitude of the errors in the predicted values. It is often preferred over MSE as it is in the same units as the response variable, making it easier to interpret. The formula for MSE is:

$MSE=1/n \sum_{i=1}^{n}(y_i-\hat{y}_i)^2$

The estimation model with lowest MSE would be considered the best. Afterwards, the R-squared of the returned best model would be assessed.

R-squared (R^2)

Indicates the proportion of variance explained by the predictors, with higher values representing better model fit:

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

RESULTS AND DISCUSSION

Simulated Data Presentation and Preliminary Assessment

This section presents and discusses the simulated response variable and explanatory variables at different scenarios of high dimensionalities as shown in Table 1, Table 2, Table 3 and Table 4. The simulated response variable and explanatory variables at each scenarios were generated using $Y = X\beta+\epsilon$;

where:

$X\sim N(0,\Sigma)$, with $\Sigma_{ij}=\rho$ for $i\neq j$), controlling the level of multicollinearity

β is a sparse vector with randomly assigned non-zero coefficients drawn from N(0,1),

$\epsilon\sim N(0,\sigma^2)$is Gaussian noise adjusted to achieve desired R^2.

Table 1: Summary Statistics of the Simulated Response Variables and Last 7-Explanatory Variables at Sample Size 30

No of Predictors	**50**							
Variables	**Y**	**X44**	**X45**	**X46**	**X47**	**X48**	**X49**	**X50**
Mean	1.324	0.0594	-0.010	0.0731	-0.074	-0.022	0.228	0.147
Median	1.966	0.115	-0.060	0.117	-0.104	0.142	0.472	0.090
Min.	-7.318	-1.916	-2.043	-1.994	-2.250	-2.129	-2.08	-1.615
Max.	11.453	2.401	2.479	2.309	2.065	1.410	2.158	2.236
No of Predictors	**70**							
Variables	**Y**	**X64**	**X65**	**X66**	**X67**	**X68**	**X69**	**X70**
Mean	1.386	-0.098	-0.025	0.039	0.039	0.038	0.082	0.009
Median	2.610	-0.092	0.039	0.090	0.080	0.048	-0.124	-0.055
Min.	-12.311	-1.823	-1.564	-1.552	-1.609	-1.823	-1.678	-2.005
Max.	14.267	1.833	1.956	2.013	1.941	1.833	2.217	1.948
No of Predictors	**200**							
Variables	**Y**	**X194**	**X195**	**X196**	**X197**	**X198**	**X199**	**X200**
Mean	-0.497	0.036	0.047	-0.006	0.088	0.153	0.194	0.147
Median	0.344	-0.202	-0.095	-0.175	-0.152	0.063	-0.048	-0.115
Min.	-15.905	-2.442	-2.609	-2.907	-2.132	-2.079	-1.888	-2.019
Max.	13.879	2.539	2.532	2.787	2.764	2.834	2.566	2.488

Source: Researchers' Compilations from R-Output

Table 2: Summary Statistics of the Simulated Response Variables and Last 7 Independent Variables at Sample Size 50

No of Predictors	**70**							
Variables	**Y**	**X44**	**X45**	**X46**	**X47**	**X48**	**X49**	**X50**
Mean	-0.048	0.062	-0.101	-0.125	-0.099	0.067	-0.106	-0.014
Median	0.237	0.065	-0.064	-0.145	0.114	-0.217	0.057	-0.167
Min.	-8.288	-2.070	-2.035	-2.378	-3.486	-1.716	-3.099	-1.494
Max.	7.115	2.594	2.245	2.372	2.079	2.811	1.618	1.872
No of Predictors	**100**							
Variables	**Y**	**X94**	**X95**	**X96**	**X97**	**X98**	**X99**	**X100**
Mean	-1.232	0.163	-0.106	-0.013	0.159	0.234	0.206	0.054
Median	-0.968	0.161	-0.023	-0.024	0.066	0.128	0.188	-0.086
Min.	-20.891	-1.975	-3.081	-1.798	-3.041	-2.014	-2.255	-1.978
Max.	25.595	2.494	2.305	1.374	2.568	3.348	2.241	2.529
No of Predictors	**200**							
Variables	**Y**	**X194**	**X195**	**X196**	**X197**	**X198**	**X199**	**X200**
Mean	-0.375	0.069	0.096	0.071	0.080	0.065	0.063	0.071
Median	-0.808	0.147	0.235	0.376	0.313	0.349	0.349	0.368
Min.	-10.372	-2.383	-2.492	-2.409	-2.333	-2.398	-2.475	-2.467
Max.	20.619	1.669	1.574	1.607	1.627	1.641	1.577	1.572

Source: Researchers' Compilations from R-Output

Table 3: Summary Statistics of the Simulated Response Variables and Last 7 Independent Variables at Sample Size 70

No of Predictors	**100**							
Variables	**Y**	**X94**	**X95**	**X96**	**X97**	**X98**	**X99**	**X100**
Mean	-0.472	0.027	0.006	0.085	0.024	0.026	-0.075	0.034
Median	-1.048	-0.001	0.117	0.107	0.008	-0.079	-0.031	-0.051
Min.	-10.022	-2.215	-2.177	-2.139	-1.665	-2.279	-3406	-2.084
Max.	10.800	2.041	1.928	1.884	2.376	2.274	2.203	2.419
No of Predictors	**150**							
Variables	**Y**	**X144**	**X145**	**X146**	**X147**	**X148**	**X149**	**X150**
Mean	0.537	-0.114	-0.148	-0.157	-0.159	-0.178	-0.148	-0.147
Median	1.510	-0.245	-0.202	-0.195	-0.243	-0.225	-0.202	-0.215
Min.	-15.961	-2.112	-1.982	-1.688	-1.980	-1.995	-1.960	-1.995
Max.	19.764	2.894	3.079	3.063	2.869	3.028	3.328	3.134
No of Predictors	**200**							
Variables	**Y**	**X194**	**X195**	**X196**	**X197**	**X198**	**X199**	**X200**
Mean	-0.444	0.022	0.003	0.001	0.013	-0.043	-0.026	-0.0001
Median	-0.695	-0.124	-0.110	-0.194	-0.109	-0.077	-0.077	-0.101
Min.	-7.515	-1.702	-1.643	-1.661	-1.711	-1.659	-1.575	-1.492
Max.	7.247	2.599	2.542	2.495	2.553	2.492	2.144	2.257

Source: Researchers' Compilations from R-Output

Table 4: Summary Statistics of the Simulated Response Variables and Last 7 Independent Variables at Sample

No of Predictors	**120**							
Variables	**Y**	**X114**	**X115**	**X116**	**X117**	**X118**	**X119**	**X120**
Mean	-0.086	0.132	0.085	-0.052	0.044	-0.126	-0.065	-0.008
Median	1.026	0.158	-0.005	-0.001	0.051	-0.006	0.016	0.089
Min.	-24.882	-2.411	-2.646	-3.345	-2.594	-2.825	-3.218	-2.975
Max.	23.759	1.936	2.944	3.344	2.349	2.028	2.685	3.341
No of Predictors	**150**							
Variables	**Y**	**X144**	**X145**	**X146**	**X147**	**X148**	**X149**	**X150**
Mean	0.164	0.047	-0.042	-0.053	0.089	-0.086	-0.202	-0.045
Median	0.588	0.020	0.049	-0.178	0.109	-0.144	-0.263	-0.018
Min.	-10.883	-2.446	-2.409	-2.554	-2.487	-2.239	-2.630	-2.769
Max.	12.130	3.384	2.371	2.677	2.796	2.784	2.064	2.635
No of Predictors	**200**							
Variables	**Y**	**X194**	**X195**	**X196**	**X197**	**X198**	**X199**	**X200**
Mean	0.449	-0.146	-0.047	-0.079	-0.077	-0.076	-0.070	0.070
Median	0.333	-0.125	0.064	-0.089	-0.036	-0.184	0.053	0.082
Min.	-11.441	-2.284	-2.749	-3.593	-2.269	-1.879	-3.431	-2.265
Max.	12.423	2.781	3.419	3.087	1.808	2.134	3.174	3.102

Source: Researchers' Compilations from R-Output

Explicitly, Table 1 presents the summary statistics of simulated response-variables and last 7-explanatory variables at different scenarios of p>(n=30), as p was varied across 50, 70 and 200. The n=30 is our first small sample settings representing the most challenging setting where predictors (p) far exceed observations. According to the table at (p=50)>(n=30), the simulated response-variable has mean of 1.32 ranges between -17.32 and 11.45. Also, the table show that at (p=70)>(n=30) the simulated response-variable has mean of 1.39 ranges between -12.31 and 14.27. Likewise, Table 1 reveals that at (p=200)>(n=30) the simulated response-variable has mean of -0.497 ranges between -15.91 and 13.88.

Similarly, Table 2 presents the summary statistics of simulated response-variables and last 7-explanatory variables at different scenarios of p>(n=50), as p was varied across 70, 100 and 200. The n=50 is our second small sample settings also representing the most challenging setting where predictors (p) far exceed observations. According to the table at (p=70)>(n=50), the simulated response-variable has mean of -0.048 ranges between -8.29 and 7.12. Also, the table show that at (p=100)>(n=50) the simulated response-variable has mean of -1.232 ranges between -20.89 and 25.59. In addition, Table 2 reveals that at (p=200)>(n=50) the simulated response-variable has mean of -0.375 ranges between -10.37 and 20.62.

Furthermore, Table 3 presents the summary statistics of simulated response-variables and last 7-explanatory variables at different scenarios of p>(n=70), as p was varied across 100, 150 and 200. The n=70 is our first higher small sample size considered to explore scalability and performance in more balanced or low-dimensional settings. According to the table at (p=100)>(n=70), the simulated response-variable has mean of -0.472 ranges between -10.022 and 10.800. Also, the table show that at (p=150)>(n=70) the simulated response-variable has mean of 0.537 ranges between -15.961 and 19.764. Table 3 further reveals that at (p=200)>(n=70) the simulated response-variable has mean of -0.444 ranges between -7.515 and 7.247.

Moreover, Table 4 presents the summary statistics of simulated response-variables and last 7-explanatory variables at different scenarios of p>(n=100), as p was varied across 120, 150 and 200. The n=100 is our second higher small sample size considered to explore scalability and performance in more balanced or low-dimensional settings. According to the table at (p=120)>(n=100), the simulated response-variable has mean of -0.086 ranges between -24.882 and 23.759. Also, the table show that at (p=150)>(n=100) the simulated response-variable has mean of 0.164 ranges between -10.883 and 12.130. Table 4 further reveals that at (p=200)>(n=100) the simulated response-variable has mean of 0.449 ranges between -11.441 and 12.423.

Based on the foregoing it is quite evident that simulated dataset obviously exhibits high-dimensionality problem (i.e. p>n), thus necessitate advanced methods of regression estimation other than the OLS.

Performance Assessment of Ridge, Lasso, Elastic Net, SCAD and the Novel Two-Step Sparse Learning Methods under High Dimensionality and Multicollinearity

This section presents and discusses the performances of the celebrated ridge, lasso, elastic net, SCAD and our

four novel two-steps sparse regression models towards providing a robust regression model for the simulated response-variables under the high dimensionality scenarios (as presented in the previous section) and multicollinearity problems.

Table 5 presents the assessment results (i.e. MSEs) of each Ridge, Lasso, Elastic-Net, SCAD and the novel two-step sparse learning regression models under problem of high dimensionality and multicollinearity at small sample sizes (i.e. 30 and 50). Explicitly, for sample size 30 at 10% sparsity Table 5 reveals lowest MSEs of 0.001, 0.0167 and 0.00006 for Lasso estimator when p=50 (i.e. low high-dimensional) at low correlation (r=0.1), moderate correlation (r=0.5) and when p=200 at moderate correlation (r=0.5) levels respectively. The table further depicts lowest MSEs for the novel SPCR-Lasso estimator when p=50; r=0.9 (mse = 0.0143) i.e. low high-dimension with high correlation, p=70; r=0.1 (mse=0.0071),p=70;r=0.5 (mse=0.0025) & p=70; r=0.9 (mse=0.0671) i.e. moderate high-dimensional with any

Table 5: Summary Statistics of the Simulated Response Variables and Last 7 Independent Variables at Sample

n	**Sparsity**	**10**								
	p	**50**			**70**			**200**		
	r	**0.1**	**0.5**	**0.9**	**0.1**	**0.5**	**0.9**	**0.1**	**0.5**	**0.9**
30	Ridge	22.2261	14.7271	5.5648	4.7699	7.8126	4.7783	2.8892	8.78595	4.3978
	Lasso	0.0010	0.0167	0.1172	0.0638	0.0211	0.1235	0.000019	0.000063	0.00516
	Enet	0.0041	0.0269	0.1988	0.0334	0.0192	0.2428	0.000082	0.000372	0.01729
	SCAD	2.5382	0.7175	1.2063	0.1656	0.3210	0.9581	0.1454	0.2581	1.34518
	SPCR-Ridge	1.7273	1.5237	0.5399	0.6035	0.2873	0.5755	0.0539	0.4525	0.17657
	SPCR-Lasso	0.3215	0.1458	0.0143	0.0071	0.0025	0.0671	0.0000004	0.007986	0.003849
	SPCR-Enet	1.0575	0.2012	0.0351	0.0672	0.0342	0.0997	0.000696	0.07211	0.011083
	SPCR-SCAD	0.2566	0.0348	0.0577	0.1156	0.0055	0.1057	0.0001347	0.04886	0.044825
	Sparsity	**30**								
	Ridge	38.9309	34.5543	5.0801	51.7066	26.5211	11.1444	15.8539	12.43887	4.99299
	Lasso	0.0078	0.7775	0.7791	0.0516	0.0462	0.0665	0.04935	0.000873	0.00397
	Enet	0.0871	1.2908	0.4971	0.1924	0.0218	0.1709	0.17285	0.000262	0.01215
	SCAD	3.9459	0.9614	1.2167	1.6763	0.9853	4.9219	0.21469	0.81469	3.17956
	SPCR-Ridge	38.7535	28.7759	17.0134	34.2091	22.8729	175.7768	15.5972	33.1529	15.2773
	SPCR-Lasso	0.9465	1.8579	0.3021	0.0269	0.0173	0.0511	0.04379	0.06934	0.17663
	SPCR-Enet	2.1343	5.9747	0.9162	7.4662	0.0632	0.2739	0.11784	0.30761	0.43285
	SPCR-SCAD	0.9149	1.6942	1.1840	1.4363	0.0753	0.4019	0.08979	0.07658	0.26952
	Sparsity	**10**								
	p	**50**			**70**			**200**		
50	Ridge	14.27529	13.14787	3.99444	7.59169	11.5585	2.51970	10.93393	17.68289	3.68972
	Lasso	0.024283	0.01111	0.26847	0.001597	0.00047	0.12309	0.000056	0.00016	0.00761
	Enet	0.035625	0.03055	0.32753	0.002292	0.00199	0.10761	0.005829	0.00089	0.03019
	SCAD	0.137990	0.37474	0.71736	0.135452	0.28281	0.34494	1.51859	0.20688	0.81663
	SPCR-Ridge	1.59547	1.50312	0.69733	0.961469	0.41246	0.3449	1.16773	0.14019	0.19149
	SPCR-Lasso	0.071019	0.20712	0.07296	0.031501	0.00401	0.14790	0.0000021	0.02395	0.00096
	SPCR-Enet	01.80131	0.450867	0.08955	0.065121	0.03633	0.20077	0.009640	0.06199	0.00076
	SPCR-SCAD	0.002342	0.20008	0.27585	0.008672	0.00831	0.05008	0.001545	0.00298	0.00113
	Sparsity	**30**								
	Ridge	28.72535	24.78267	10.92322	46.11221	39.1727	9.20752	25.78999	27.94949	12.2576
	Lasso	0.001752	0.04827	0.45687	0.00135	0.00376	0.12701	0.03429	0.05793	0.04912
	Enet	0.00692	0.12742	0.76926	0.00513	0.01369	0.28181	0.35494	0.20659	0.27805
	SCAD	0.39620	1.59569	4.17981	2.22021	4.83648	1.49247	0.49354	0.36052	23.3093
	SPCR-Ridge	1.790068	2.43327	1.00552	1.90385	0.99053	0.91456	19.86799	2.06924	0.70623
	SPCR-Lasso	0.15863	0.31728	0.25098	0.04437	0.04119	0.01670	0.30157	0.09142	0.00785
	SPCR-Enet	0.08939	0.54537	0.30180	0.26648	0.05336	0.01793	1.04003	0.23314	0.00523
	SPCR-SCAD	0.05241	0.79069	0.23878	0.07853	1.32367	0.01216	0.42382	0.08032	0.00686

Source: Researchers' Compilations from R-Outputs

correlation levels, and p=200; r=0.1 (mse=0.0000004) & p=200; r=0.9 (mse=0.003849) i.e. high-dimensional with low and high correlation levels. In the same vein, Figure 1 presents performance ranks of each estimator under varied levels of high dimensionality and multicollinearity for sample size 30 with 10% sparse. The figure similarly, ranks Lasso estimator best when p=50 & r=0.1, p=50 & r=0.5, and p=200 & r=0.5 while SPCR-Lasso returned best rank estimator when p=50 & r=0.9, p=70 & r=0.1, p=70 & r=0.1, p=70 & r=0.5, p=70 & r=0.9, p=200 & r=0.1, and p=200 & r=0.9.

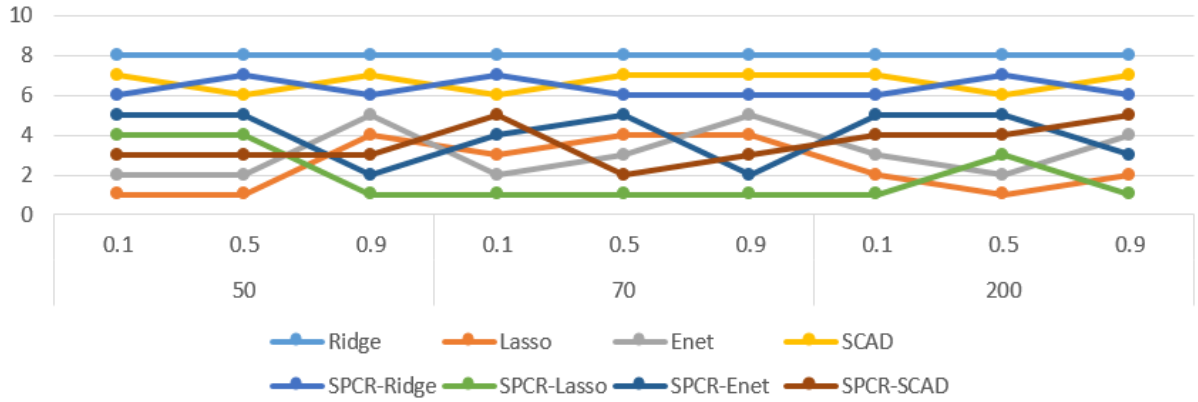


Figure 1: Performance Ranks of Each Estimator under Varied Levels of High Dimensionality and Multicollinearity for Sample Size 30 with 10% Sparse

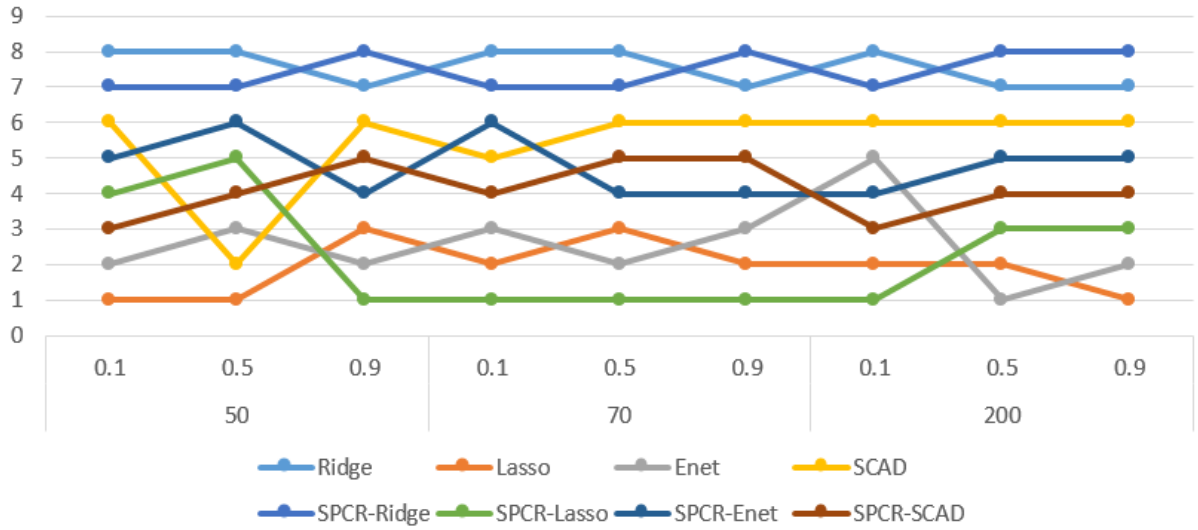


Figure 2: Performance Ranks of Each Estimator under Varied Levels of High Dimensionality and Multicollinearity for Sample Size 30 with 30% Sparse

Similarly, for sample size 30 at 30% sparsity Table 5 reveals lowest MSEs of 0.0078, 0.7775, 0.00087 and 0.00397 for Lasso estimator when p=50 (i.e. low high-dimensional) at low correlation (r=0.1), moderate correlation (r=0.5) and high correlation (r=0.9) levels respectively. While when p=200 at moderate correlation (r=0.5), the table returned Elastic Net (Enet) estimator with lowest MSE of 0.00026. The table also depicts lowest MSEs for the novel SPCR-Lasso estimator when p=50; r=0.9 (mse = 0.3021) i.e. low high-dimension with high correlation, p=70; r=0.1 (mse=0.0269),p=70;r=0.5 (mse=0.0173) & p=70; r=0.9 (mse=0.0511) i.e. moderate high-dimensional with any correlation levels, and p=200; r=0.1 (mse=0.04379) i.e. high-dimensional with low correlation level. In the same vein, Figure 2 presents performance ranks of each estimator under varied levels of high dimensionality and multicollinearity for sample size 30 with 30% sparse. The figure similarly, ranks Lasso estimator best when p=50 & r=0.1, p=50 & r=0.5, and p=200 & r=0.9. It also ranks Elastic Net estimator best when p=200 & r=0.5 while SPCR-Lasso returned best rank estimator when p=50 & r=0.9, p=70 & r=0.1, p=70 & r=0.1, p=70 & r=0.5, p=70 & r=0.9, and p=200 & r=0.1.

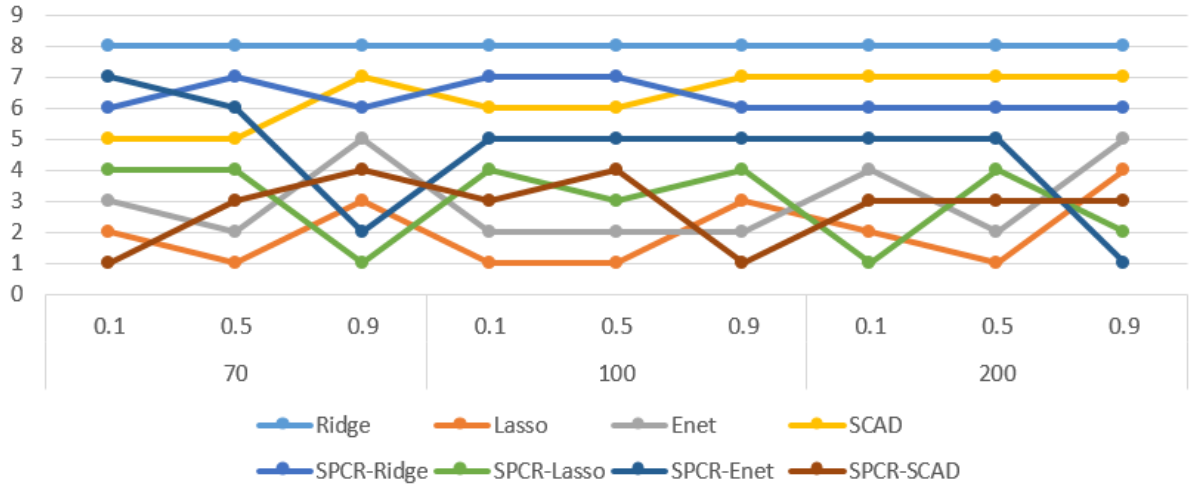


Figure 3: Performance Ranks of Each Estimator under Varied Levels of High Dimensionality and Multicollinearity for Sample Size 50 with 10% Sparse

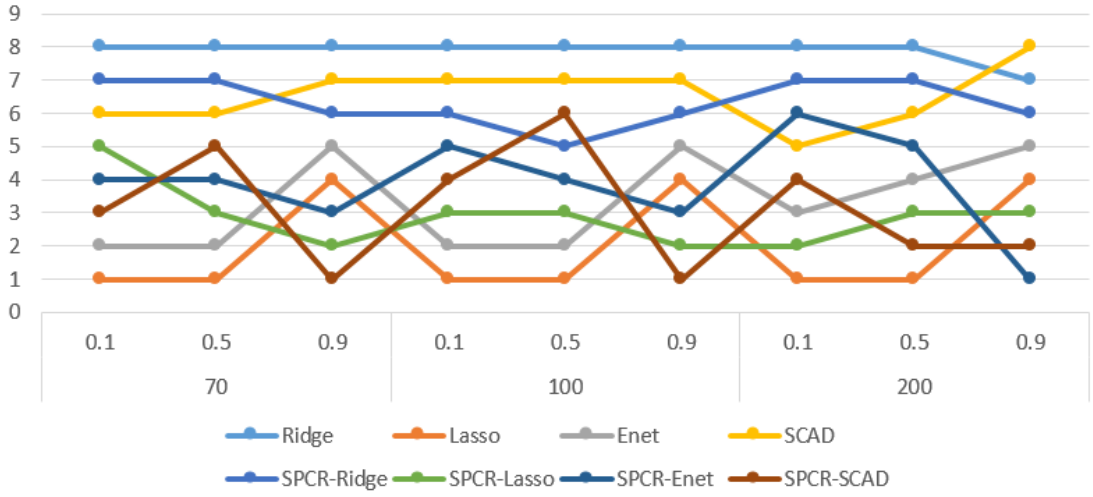


Figure 4: Performance Ranks of Each Estimator under Varied Levels of High Dimensionality and Multicollinearity for Sample Size 50 with 30% Sparse

Furthermore, considering small sample size of 50 at 10% sparsity Table 5 reveals the Lasso estimator with least MSEs of 0.0111, 0.0016, 0.0005 and 0.00016 when p=70 & r=0.5, p=100 & r=0.1, p=70 & r=0.5, and p=200 & r=0.5 respectively. Meanwhile the table depicts the novel; SPCR-Lasso estimator with least MSEs when p=70 & r=0.9 (mse=0.07296) and p=200 & r=0.1 (mse=0.0000032), SPCR-SCAD estimator with least MSEs when p=70 & r=0.1 (mse=0.002342) and p=100 & r=0.9 (mse=0.05008), and SPCR-Enet estimator with lowest MSE when p=200 & r=0.9 (mse=0.00076). Similarly, Figure 3 presents the performance ranks of each estimator under varied levels of high dimensionality and multicollinearity for sample size 50 with 10% sparse. According to the figure, the Lasso estimator was ranked best (i.e. 1st) on four occasions namely, p=70 & r=0.5, p=100 & r=0.1, p=70 & r=0.5, and p=200 & r=0.5. the novel SPCR-Lasso estimator was ranked best on two occasions namely, p=70 & r=0.9 and p=200 & r=0.1 . Also, the novel SPCR-SCAD was ranked best on two occasions namely p=70 & r=0.1 and p=100 & r=0.9. As well as our novel SPCR-Enet was ranked best when p=200 & r=0.9.

Considering small sample size of 50 at 30% sparsity Table 5 and Figure 4 reveal Lasso estimator returned with least MSE and 1st ranking on six occasions namely p=70 & r=0.1 (mse=0.001752), p=70 & r=0.5 (mse=0.04827), p=100 & r=0.1 (mse=0.00135), p=100 & r=0.5 (mse=0.00376), p=200 & r=0.1 (mse=0.03429) and p=200 & r=0.5 (mse=0.00376). Additionally, Table 5 and Figure 4 depict our novel SPCR-SCAD estimator returned with least MSE and 1st ranking on two occasions namely p=70 & r=0.9 (mse=0.23878) and p=100 & r=0.9 (mse=0.01216). Also, according to Table 5 and Figure 4 our novel SPCR-Enet returned with least MSE and ranked 1st when =200 & r=0.9 (mse=0.00523).

Moreover, Table 6 presents the assessment results (i.e. MSEs) of each Ridge, Lasso, Elastic-Net, SCAD and the novel two-step sparse learning regression models under problem of high dimensionality and multicollinearity at

higher small sample sizes (i.e. 70 and 100). According to Table 6 and Figure 5 when considering sample size 70 with 10% sparsity, Lasso estimator returned with least MSE and ranked 1st on three occasions namely p=100 & r=0.1 (mse=0.01239), p=150 & r=0.1 (mse=0.000253) and p=150 & r=0.5 (mse=0.00860). Also, our novel SPCR-Enet returned with the least MSE and ranked best (1st) on two occasions namely, p=150 & r=0.9 (mse=0.01112) and p=200 & r=0.9 (mse=0.01801). Similarly, our novel SPCR-SCAD returned with the least MSE and ranked best (1st) on two occasions namely, p=200 & r=0.1 (mse=0.0000029) and p=200 & r=0.5 (mse=0.00497). Table 6 and Figure 5 reveal our novel SPCR-Lasso with the least MSE and 1st ranking when p=100 & r=0.1 (mse=0.04099).

Table 6: MSE of Ridge, Lasso, Elastic-Net, SCAD and the Novel Two-Step Sparse Learning Regression Models under Problem of High Dimensionality and Multicollinearity at Higher Small Sample Size

n	**Sparsity**	**10**								
	p	**100**			**150**			**200**		
	r	**0.1**	**0.5**	**0.9**	**0.1**	**0.5**	**0.9**	**0.1**	**0.5**	**0.9**
70	Ridge	16.2612	12.23781	4.03865	11.26075	11.60848	3.39042	8.64446	9.76361	2.67325
	Lasso	0.01239	0.06057	0.18933	0.000253	0.00860	0.08186	0.00108	0.00857	0.11078
	Enet	0.04216	0.04236	0.30014	0.008587	0.02651	0.16821	0.00337	0.01337	0.10399
	SCAD	0.91239	0.42240	1.54271	0.061469	0.09442	0.70653	0.05592	0.04387	0.37114
	SPCR-Ridge	7.81776	2.56198	0.65102	20.5621	0.10744	0.48016	0.17769	0.77086	0.37926
	SPCR-Lasso	0.25674	0.26353	0.04099	0.39296	0.05496	0.01196	0.01360	0.01134	0.05829
	SPCR-Enet	0.52689	0.39883	0.05059	0.55724	0.12907	0.01112	0.02494	0.05916	0.01801
	SPCR-SCAD	0.19465	0.60541	0.14201	0.14136	0.05721	0.01919	0.0000029	0.00497	0.23454
	Sparsity	**30**								
	Ridge	38.13883	37.66447	12.62623	24.65487	36.8325	11.61202	43.27625	40.24778	14.48422
	Lasso	0.00662	0.00338	0.37802	0.000527	0.00093	0.06559	0.00039	0.00097	0.03899
	Enet	0.01231	0.01996	0.58596	0.01063	0.00397	0.15603	0.00169	0.01392	0.10182
	SCAD	0.54454	1.58249	7.48471	0.18944	0.75494	0.74484	0.38007	1.28358	1.70183
	SPCR-Ridge	1.05483	1.59691	1.40662	1.38173	2.20109	0.63355	20.95064	41.6646	0.58520
	SPCR-Lasso	0.01037	0.11502	0.10606	0.11955	0.08147	0.00544	0.37343	0.42448	0.00979
	SPCR-Enet	0.10300	0.23069	0.04797	0.17829	0.14697	0.01071	1.25418	0.87494	0.01865
	SPCR-SCAD	0.01718	0.06238	0.12671	0.15924	0.14211	0.06971	0.52191	0.50069	0.02307
	Sparsity	**10**								
	p	**120**			**150**			**200**		
100	Ridge	7.76100	12.6273	2.51623	15.30857	12.1868	3.17887	12.09962	12.77974	2.54699
	Lasso	0.03929	0.09684	0.50299	0.01276	0.00601	0.20818	0.00835	0.00496	0.08117
	Enet	0.04664	0.10807	0.51229	0.03871	0.02988	0.32757	0.01998	0.01209	0.13614
	SCAD	0.09555	0.26393	0.97131	0.28428	0.18236	0.61504	0.07291	0.19182	0.60169
	SPCR-Ridge	0.60385	2.70444	0.80414	1.10457	0.45551	0.67572	3.83138	0.55069	0.42212
	SPCR-Lasso	0.10859	0.32131	0.04616	0.04605	0.11171	0.10353	0.27173	0.03944	0.00243
	SPCR-Enet	0.20375	0.86351	0.15666	0.05838	0.08879	0.24179	0.74755	0.04253	0.14797
	SPCR-SCAD	0.11849	0.47327	0.37146	0.03035	0.23513	0.42612	0.29195	0.00853	0.00307
	Sparsity	**30**								
	Ridge	44.73326	43.83492	8.42825	25.9745	36.23507	9.66694	40.51713	26.12172	10.96855
	Lasso	0.00553	0.02191	0.25439	0.00293	0.00363	0.16858	0.00104	0.00111	0.12742
	Enet	0.01770	0.05789	0.38731	0.00866	0.01271	0.30531	0.00383	0.00347	0.26067
	SCAD	0.42485	1.01778	1.44601	0.76685	0.50686	2.20988	0.28645	0.66346	0.88463
	SPCR-Ridge	4.84146	1.15390	0.93489	1.80877	3.69042	0.79279	9.05412	3.75005	0.97242
	SPCR-Lasso	0.09466	0.10141	0.07998	0.09151	0.19978	0.05461	1.69562	0.75744	0.05299
	SPCR-Enet	0.63696	0.06535	0.16816	0.18488	0.50735	0.06652	4.53791	0.79444	0.06279
	SPCR-SCAD	0.50354	0.32955	0.15189	0.20898	0.28904	0.06717	0.00176	0.46458	0.06440

Source: Researchers' Compilations from R-Outputs

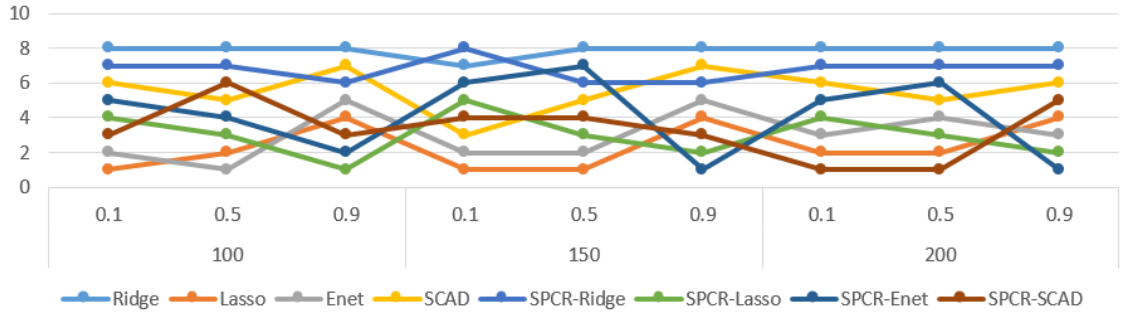


Figure 5: Performance Ranks of Each Estimator under Varied Levels of High Dimensionality and Multicollinearity for Sample Size 70 with 10% Sparse

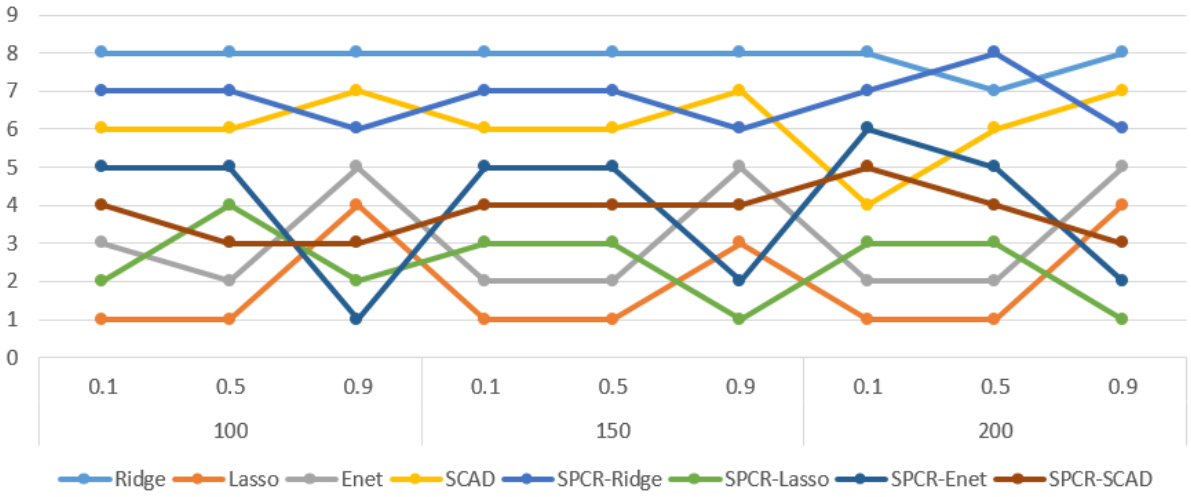


Figure 6: Performance Ranks of Each Estimator under Varied Levels of High Dimensionality and Multicollinearity for Sample Size 70 with 30% Sparse

In addition, considering sample size 70 with 30% sparsity, Table 6 and Figure 6 reveal Lasso estimator with lowest MSE and ranked 1st on six occasions namely p=100 & r=0.1 (MSE=0.00662), p=100 & r=0.5 (MSE=0.00338),p=150 & r=0.1 (MSE=0.000527), p=150 & r=0.5 (MSE=0.00093), p=200 & r=0.1 (MSE=0.00039), and p=200 & r=0.5 (MSE=0.00097). The table and figure further depict our novel SPCR-Lasso estimator with the lowest MSE and best ranking estimator when p=150 & r=0.9 (MSE=0.00544), and p=200 & r=0.9 (MSE=0.00979) as well as SPCR-Enet when p=100 & r=0.9 (MSE=0.04797).

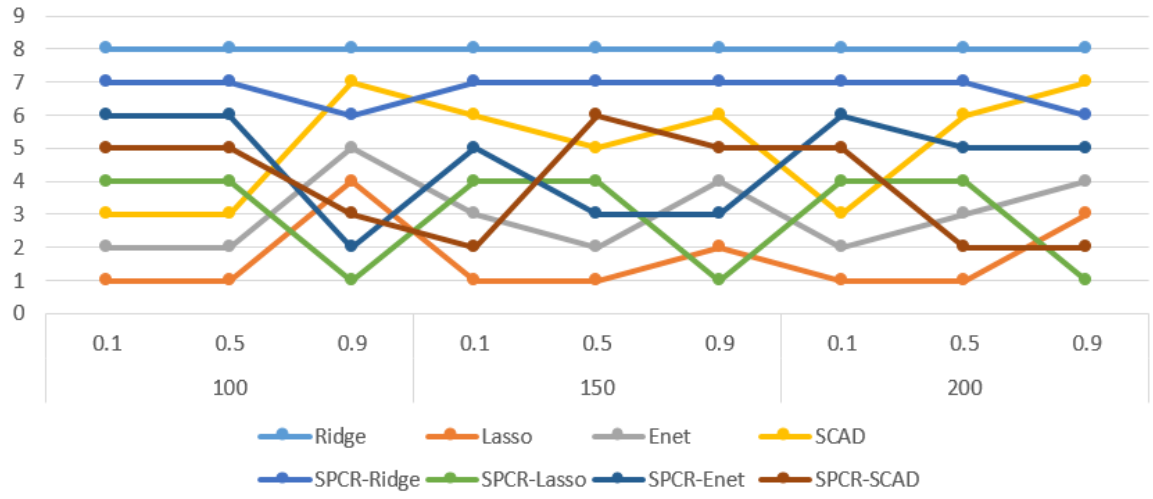


Figure 7: Performance Ranks of Each Estimator under Varied Levels of High Dimensionality and Multicollinearity for Sample Size 100 with 10% Sparse

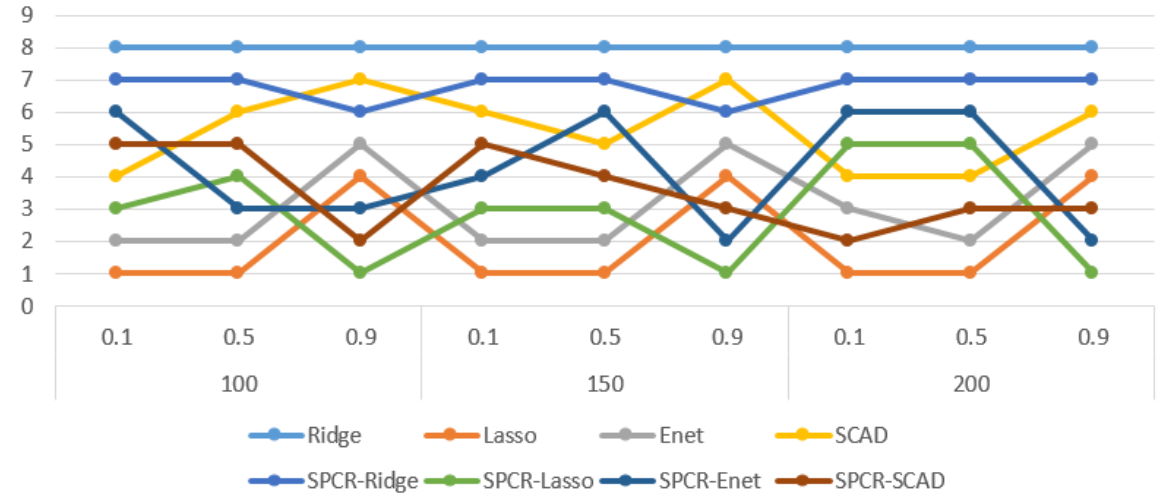


Figure 8: Performance Ranks of Each Estimator under Varied Levels of High Dimensionality and Multicollinearity for Sample Size 100 with 30% Sparse

Furthermore, considering sample size 100 with 10% sparsity, Table 6 and Figure 7 reveal Lasso estimator with lowest MSE and ranked 1st on six occasions namely p=120 & r=0.1 (MSE=0.00662), p=120 & r=0.5 (MSE=0.09684),p=150 & r=0.1 (MSE=0.01276), p=150 & r=0.5 (MSE=0.00601), p=200 & r=0.1 (MSE=0.00835), and p=200 & r=0.5 (MSE=0.00496). The table and figure establish our novel SPCR-Lasso with the least MSE and best ranking estimator when p=120 & r=0.9 (MSE=0.07998), p=150 & r=0.9 (MSE=0.10353) and p=200 & r=0.9 (MSE=0.00243).

Similarly, considering sample size 100 with 30% sparsity, Table 6 and Figure 8 reveal Lasso estimator with lowest MSE and ranked 1st on six occasions namely p=120 & r=0.1 (MSE=0.00553), p=120 & r=0.5 (MSE=0.02191),p=150 & r=0.1 (MSE=0.00293), p=150 & r=0.5 (MSE=0.00363), p=200 & r=0.1 (MSE=0.00104), and p=200 & r=0.5 (MSE=0.00111). The table and figure establish our novel SPCR-Lasso with the least MSE and best ranking estimator when p=120 & r=0.9 (MSE=0.04616), p=150 & r=0.9 (MSE=0.05461) and p=200 & r=0.9 (MSE=0.05299).

Findings Summary, Discussion of Findings, And Conclusion

Findings by Small Sample Sizes and Dimensionality

At extremely small sample sizes (n=30), SPCR-Lasso consistently outperformed all other estimators, especially when dimensionality was high (p=7 or p=200). This demonstrates the strength of SPCR-Lasso in small-sample, high-dimensional contexts, where traditional Lasso, Ridge, or Elastic Net tend to become unstable. At moderately small sample sizes (n=50), results showed variation across conditions: SPCR-SCAD excelled in contexts of low sparsity and low correlation. SPCR-Lasso and SPCR-Enet provided superior performance under higher correlation and dimensionality. As sample sizes increased further (n≥70), SPCR-Lasso and SPCR-Enet emerged as the most consistent and robust estimators across both moderate and high correlations. Notably, SPCR-Enet showed particular strength in very high-dimensional scenarios (p=200), reflecting its ability to balance shrinkage and group variable selection.

Findings by Multicollinearity and Sparsity

The findings also highlight clear interactions between predictor correlation and sparsity:

- Under low correlation (r=0.1), traditional Lasso sometimes matched or exceeded SPCR-based methods in low-dimensional settings, suggesting SPCR hybridization may not always be necessary in weakly collinear designs.
- Under moderate (r=0.5) or high correlation (r=0.9), SPCR-Lasso and SPCR-Enet decisively outperformed alternatives, confirming the necessity of the SPCR step for mitigating multicollinearity.
- With respect to sparsity, SPCR-SCAD performed best in highly sparse, low-correlation conditions, while SPCR-Lasso and SPCR-Enet proved more adaptable across both sparse and dense regimes.

Table 7: Overview of Best Estimators under Different Considered Small Sample Sizes, High-Dimensionality, Multicollinearity and Sparsity Levels

n	r	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9
	p	50			70			200		
30	10%	Lasso	Lasso	SPCR-Lasso	SPCR-Lasso	SPCR-Lasso	SPCR-Lasso	SPCR-Lasso	Lasso	SPCR-Lasso
	30%	Lasso	Lasso	SPCR-Lasso	SPCR-Lasso	SPCR-Lasso	SPCR-Lasso	SPCR-Lasso	Enet	Lasso

		70			100			200		
50	10%	SPCR-SCAD	Lasso	SPCR-Lasso	Lasso	Lasso	SPCR-SCAD	SPCR-Lasso	Lasso	SPCR-Enet
	30%	Lasso	Lasso	SPCR-SCAD	Lasso	Lasso	SPCR-SCAD	Lasso	Lasso	SPCR-Enet
		100			**150**			**200**		
70	10%	Lasso	Enet	SPCR-Lasso	Lasso	Lasso	SPCR-Enet	SPCR-SCAD	SPCR-SCAD	SPCR-Enet
	30%	Lasso	Lasso	SPCR-Enet	Lasso	Lasso	SPCR-Lasso	Lasso	Lasso	SPCR-Lasso
		120			**150**			**200**		
100	**10%**	**Lasso**	**Lasso**	**SPCR-Lasso**	**Lasso**	**Lasso**	**SPCR-Lasso**	**Lasso**	**Lasso**	**SPCR-Lasso**
	30%	Lasso	Lasso	SPCR-Lasso	Lasso	Lasso	SPCR-Lasso	Lasso	Lasso	SPCR-Lasso

Source: Researchers' Compilations

Discussion of Findings

Theoretical and Methodological Insights

The results validate the rationale for hybridizing SPCR with regularization penalties. SPCR effectively reduces dimensionality while preserving predictive features, and the addition of regularization stabilizes estimates in the presence of multicollinearity. Together, this hybrid approach delivers stronger predictive accuracy and interpretability than either dimension reduction or regularization alone.

The study also demonstrates that penalty choice must be data-dependent. Specifically:

- SPCR-Lasso and SPCR-Enet are best suited for high-dimensional, correlated designs.
- SPCR-SCAD retains value under extreme sparsity with low correlation.
- SPCR-Ridge, while stabilizing, offers limited benefits compared to its sparse counterparts.

These findings align with empirical evidence in high-dimensional statistics but extend prior work by systematically comparing multiple regularizers within an SPCR framework across diverse simulation conditions.

Practical Implications for Applied Research

For applied researchers working in genomics, finance, climate science, and social sciences, the study's findings provide clear practical guidance:

- Use SPCR-Lasso or SPCR-Enet when predictors are highly correlated or dimensionality is large.
- Employ SPCR-SCAD in cases of extreme sparsity with weak predictor correlation.
- Expect interpretability benefits from SPCR, as sparse principal components link outcomes to identifiable subsets of predictors rather than opaque linear combinations.

This guidance equips researchers with a decision-making framework to select the most effective hybrid estimator given the structural characteristics of their data.

Policy and Applied Modeling Implications

The findings also have implications for applied modeling in policy-relevant domains. Policymakers and analysts working with high-dimensional, multicollinear data (e.g., in economic forecasting, climate modeling, or epidemiological surveillance) can adopt SPCR-based methods to achieve more reliable predictions. By improving both accuracy and interpretability, these methods enhance the credibility of evidence-based policy decisions.

Implications for Future Research

The findings suggest several avenues for further inquiry:

i. Extending the hybrid SPCR framework to nonlinear models (e.g., kernel methods, deep learning).

ii. Applying SPCR-regularization pipelines to real-world datasets in genomics, finance, and environmental science to validate simulation results.

iii. Investigating stability selection and uncertainty quantification after SPCR to improve robustness of variable selection in practice.

iv. Exploring time-series extensions of SPCR hybridization for forecasting applications.

CONCLUSION

This section has discussed the findings of the simulation study and their implications for statistical methodology, applied practice, and policy. The results confirm that hybrid SPCR estimators substantially outperform traditional penalization methods in small-sample, high-dimensional, and multicollinear conditions. Among these, SPCR-Lasso and SPCR-Enet emerge as the most versatile and reliable, while SPCR-SCAD shows targeted advantages in sparse, low-correlation settings. Collectively, the findings highlight the significance and necessity of hybrid SPCR approaches as a methodological advancement for high-dimensional data analysis.

REFERENCES

Ali, H., Shahzad, M., Sarfraz, S., Sewell, K. B., Alqalyoobi, S., & Mohan, B. P. (2023). Application and impact of Lasso regression in gastroenterology: a systematic

review. *Indian Journal of Gastroenterology, 42*(6), 780-790.

Chatterjee, I., & Baumgärtner, L. (2024). Unveiling Functional Biomarkers in Schizophrenia: Insights from Region of Interest Analysis Using Machine Learning. *Journal of Integrative Neuroscience, 23*(9).

Chen, J., Yang, S., Wang, Z., & Mao, H. (2021). Efficient sparse representation for learning with high-dimensional data. *IEEE Transactions on Neural Networks and Learning Systems, 34*(8), 4208-4222.

Cleophas, T. J., & Zwinderman, A. H. (2024). *Application of Regularized Regressions to Identify Novel Predictors in Clinical Research*. Springer Nature.

Fan, J., & Li, R. (2001). *Variable selection via nonconcave penalized likelihood and its oracle properties*. JASA.

Gupta, V., Chen, Y., & Wan, M. (2024). Predictability of weakly turbulent systems from spatially sparse observations using data assimilation and machine learning. *arXiv preprint arXiv:2407*.10088.

Hoerl, A. E., & Kennard, R. W. (1970). *Ridge regression: Biased estimation for nonorthogonal problems*. Technometrics.

Jolliffe, I. T. (2002). *Principal component analysis*. Springer.

Kawano, S. (2018). *Sparse Principal Component Regression (one-stage SPCR literature)*.

Kitano, T., & Noma, H. (2024). *Ridge, lasso, and elastic-net estimations of the modified Poisson and least-squares regressions for binary outcome data*. arXiv preprint arXiv:2408.13474.

Manzhos, S., & Ihara, M. (2022). Advanced machine learning methods for learning from sparse data in high-dimensional spaces: A perspective on uses in the upstream of development of novel energy technologies. *Physchem, 2*(2), 72-95.

Meinshausen, N. (2007). *Relaxed Lasso and stability selection literature*.

Nwosu, A., Aimufua, G. I. O., Ajayi, B. A., & Olalere, M. (2024). The Impact of Regularization on Linear Regression Based Model. *Journal of Artificial Intelligence and Computer Science, 1*(1).

Priya, A. K., Gnanasekaran, L., Rajendran, S., Qin, J., & Vasseghian, Y. (2022). Occurrences and removal of pharmaceutical and personal care products from aquatic systems using advanced treatment-A review. *Environmental Research, 204*, 112298.

Song, J., Xu, L., & Wang, X. (2024, July). A Regularization Method for Enhancing the Robustness of Regression Networks. In *2024 43rd Chinese Control Conference (CCC)* (pp. 8524-8529). IEEE.

Tibshirani, R. (1996). *Regression shrinkage and selection via the Lasso*. JRSS-B.

Witten, D. M., Tibshirani, R., & Hastie, T. (2009). *A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis*. Biostatistics.

Zhang, X., Sun, Q., & Kong, D. (2024). Supervised Principal Component Regression for Functional Responses with High Dimensional Predictors. *Journal of Computational and Graphical Statistics, 33*(1), 242-249.

Zou, H., & Hastie, T. (2005). *Regularization and variable selection via the Elastic Net*. JRSS-B.

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*.