

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452

NAAS Rating (2026): 4.49

Maths 2026; 11(9): 01-08

© 2026 Stats & Maths

<https://www.mathsjournal.com>

Received: 01-07-2026

Accepted: 03-08-2026

Chioma T Eze-Golden

Department of Statistics,
Faculty of Physical Sciences,
Chukwuemeka Odumegwu
Ojukwu University, Anambra
State, Nigeria

Cecilia N Okoli

Department of Statistics,
Faculty of Physical Sciences,
Chukwuemeka Odumegwu
Ojukwu University, Anambra
State, Nigeria

SA-MJLR: A sparse multivariate joint latent regression statistical framework for ultra-high dimensional prediction: Monte Carlo evaluation under $(p, q) > n$ settings

Chioma T Eze-Golden and Cecilia N Okoli

DOI: <https://www.doi.org/10.22271/math.2026.v11.i9a.2681>

Abstract

High-dimensional multivariate method problems often involve complex data structures in which predictor-response relationships contain both sparse direct associations and hidden latent dependencies, but existing methods typically address either sparsity or latent structure separately, limiting their ability to simultaneously achieve accurate prediction and reliable coefficient recovery in ultra-high-dimensional settings. Hence, this study developed a Sparse Adaptive Multivariate Joint Latent Regression (SA-MJLR) framework, which is a hybrid statistical approach that integrates sparse direct-effect estimation with joint latent structure learning for high-dimensional multivariate regression. The proposed framework decomposed the regression coefficient matrix into sparse and latent components, enabling simultaneous variable selection and recovery of shared multivariate structures, and the performance of SA-MJLR was evaluated via a Monte Carlo simulation study involving five progressively challenging high-dimensional scenarios, with predictor and response dimensions greater than the sample size. The developed method was then compared with Ridge Regression, Elastic Net, Reduced Rank Regression (RRR), and Sparse Partial Least Squares (SPLS) using predictive accuracy measures [Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2)] as well as coefficient recovery measures [Frobenius Error and correlation]. The simulation results demonstrated that SA-MJLR achieved the best overall performance, obtaining the highest aggregate performance score of 6.9888, compared with 1.1420 for Elastic Net, -0.5756 for SPLS, -3.3432 for RRR, and -4.2120 for Ridge Regression, and across all simulation scenarios, SA-MJLR achieved the lowest average RMSE (0.9701), lowest MAE (0.7687), highest average R^2 (0.3331), lowest mean Frobenius Error (0.0294), and highest coefficient recovery correlation (0.7138). Again, under the most challenging ultra-high-dimensional scenario, SA-MJLR maintained superior prediction performance with an RMSE of 2.3839 and R^2 of 0.2275, demonstrating its robustness under weak signal conditions and increasing dimensionality. Thus, these findings indicated that SA-MJLR provides an effective and flexible framework for high-dimensional multivariate regression problems where sparse relationships coexist with latent structures, and the developed method offers potential applications in genomics, environmental modelling, biomedical research, climate science, and other fields involving complex multivariate data.

Keywords: High-dimensional regression, sparse modelling, latent variable models, multivariate regression, coefficient recovery, dimensionality reduction, Monte Carlo simulation, SA-MJLR

1. Introduction

The rapid expansion of high-throughput technologies has transformed many scientific disciplines by enabling the simultaneous measurement of thousands of variables from relatively small numbers of observations, and its applications in biomedical research, genomics, neuroscience, metabolomics, environmental sciences, and human phenotyping increasingly involve datasets where the number of measured characteristics is significantly larger than the available sample size. Datasets of this category create a challenging statistical environment commonly described as ultra-high dimensional modelling, in which traditional regression assumptions become inadequate because the information contained in the

Corresponding Author:

Chioma T Eze-Golden

Department of Statistics,
Faculty of Physical Sciences,
Chukwuemeka Odumegwu
Ojukwu University, Anambra
State, Nigeria

observations is not sufficient for direct estimation of the large number of unknown parameters (Li *et al.*, 2024; Fan *et al.*, 2020)^[6, 13].

Expressing a general multivariate regression framework can be:

$$Z = W\beta + \varepsilon \quad (1)$$

Where:

$$W \in \mathbb{R}^{n \times p} \text{ and } Z \in \mathbb{R}^{n \times q}$$

Represent the multivariate predictor and response matrices respectively,

$$\beta \in \mathbb{R}^{p \times q}$$

Represents the coefficient matrix being unknown, and the random errors is denoted by ε .

Estimation is reliable when the number of observations is sufficiently larger than the number of variables in conventional regression settings, and modern scientific datasets increasingly exhibit the simultaneous high-dimensional condition; however, $p > n$ and $q > n$, that is, where the predictor and response dimensions are greater than the number of observations. This situation introduces severe statistical difficulties since the coefficient matrix contains pq unknown parameters; and when both predictor and response dimensions equal 1000, for instance, the model contains one million potential predictor-response relationships, while the available observations may be limited to only a few tens or hundreds.

Developments in high-dimensional regression recently have shown that these challenges cannot be addressed adequately through conventional multivariate regression since covariance matrices become unstable, parameter estimates become non-identifiable, and prediction performance deteriorates due to excessive model complexity (Ham *et al.*, 2024; Li *et al.*, 2024)^[9, 13]. These problems are particularly important in biomedical and biological applications, where thousands of molecular, physiological, or imaging measurements are collected together with multiple correlated outcomes. A major characteristic of such datasets is the presence of strong dependence structures, where predictor variables often represent groups of related biological mechanisms, while response variables frequently reflect interconnected physiological processes. Hence, modelling these systems requires approaches capable of simultaneously exploiting predictor correlation, response correlation, dimensional reduction, and variable selection.

It is clear that, several statistical strategies have been proposed to address high-dimensional regression problems, and one major direction is regularized estimation, in which penalty functions are introduced to stabilize parameter estimation, and the recent advances in sparse multivariate regression have focused on estimating large coefficient matrices while accounting for dependence among multiple response variables, and these approaches demonstrate that incorporating response correlation can improve estimation efficiency compared with treating responses independently (Ham *et al.*, 2024)^[9]. Regularized coefficient estimation, however, remains challenging when both dimensions increase simultaneously. In the setting where: $p > n$ and $q > n$, the coefficient matrix becomes extremely large, and direct estimation requires strong assumptions regarding sparsity,

covariance structure, or low-rank representation. Recent studies on high-dimensional output regression have emphasized that scalable methods must reduce the complexity of the output space while maintaining predictive information (Li *et al.*, 2024)^[13].

Sparse regression provides an attractive solution because many high-dimensional systems are believed to contain only a limited number of relevant predictors, and recent sparse multivariate regression frameworks have incorporated L_1 -based penalties to identify informative variables while controlling model complexity. Sparse regression methods that directly estimate regression coefficients may struggle, however, when predictors are strongly correlated because biologically meaningful variables often occur in groups rather than independently (Ham *et al.*, 2024)^[9]. Another important challenge is the modelling of multiple correlated responses, and when the response variables are associated, ignoring their dependence structure can lead to inefficient estimation. Studies recently have therefore investigated methods that jointly estimate regression coefficients and response covariance structures to improve prediction accuracy in high-dimensional multivariate regression (Ham *et al.*, 2024)^[9]; nevertheless, covariance-based approaches become computationally demanding when the number of responses is extremely large.

An alternative approach is to use the latent variable model, which assumes that high dimensional observations are generated from a lower dimensional latent space, methods that are appealing in particular are the ones which many complex scientific observations have hidden structures that cannot be directly observed. Factor-based regression techniques have been gaining more attention in recent years, as they are able to simultaneously reduce dimensionality and predict, and the recent advances in factor-augmented regression extend the applicability of dimensional reduction to prediction with both common variation and sparse effects present in the response structure (Fan *et al.*, 2023)^[7]. The reduced-rank regression has also become an important framework for multivariate prediction because it assumes that the coefficient matrix can be represented using a smaller number of latent dimensions. The sparse reduced-rank regression approaches recently have demonstrated improved predictive performance and interpretability by combining low-rank structure with variable selection, particularly in biological applications involving thousands of molecular predictors and multiple outcomes (Kobak *et al.*, 2021)^[12]. Even though latent techniques provide important advantages, existing techniques often focus primarily on either reducing predictor dimensionality or modelling response dependence, and they do not fully address the more difficult situation where both predictor and response spaces are simultaneously ultra-high dimensional.

The machine learning approaches have become increasingly important for high-dimensional prediction because of their ability to model nonlinear relationships and complex interactions, and among these methods, Extreme Gradient Boosting (XGBoost) has achieved considerable success due to its computational efficiency, regularization mechanisms, and ability to capture nonlinear patterns in large datasets (Chen & Guestrin, 2016)^[3]. Applications of machine learning recently in biomedical and scientific prediction demonstrate that ensemble tree methods can provide strong predictive performance when traditional statistical assumptions are violated, and therefore machine learning models generally focus on prediction rather than uncovering interpretable latent

relationships among correlated predictors and responses. As for multivariate problems involving hundreds or thousands of responses, predictive accuracy alone may not be sufficient, but researchers often require knowledge about which variables contribute to prediction and how groups of predictors relate to groups of outcomes. Hence, there remains a need for statistical approaches that combine predictive capability with interpretable latent structures.

2. Material and Methods

2.1 Problem Formulation

This study considers an ultra-high-dimensional multivariate regression problem in which the predictor and response dimensions are both greater than the number of observations. Let:

$$W \in \mathbb{R}^{n \times p} \text{ and } Z \in \mathbb{R}^{n \times q}$$

Represent the multivariate predictor and response matrices respectively, where

$$p > n, \quad q > n$$

Then, the general multivariate regression association is represented as stated in Equation (1). In ultra-high dimensional settings, the direct estimation of β becomes infeasible since the number of unknown parameters is:

$$pq$$

Which may greatly be much greater than the available observations, and additionally, if $p > n$, the predictor covariance matrix:

$$W^T W$$

is singular, preventing the use of ordinary least squares, and similarly, if $q > n$, the response covariance structure becomes rank deficient, limiting the effectiveness of conventional multivariate regression approaches.

To overcome these challenges, this study develops a Sparse Adaptive Multivariate Joint Latent Regression (SA-MJLR) framework that assumes that the coefficient matrix has two complementary structures, which are: a sparse direct-effect component and a low-rank latent component. Thus, the proposed coefficient structure is mathematically defined as follows:

$$B = B_s + Q_w A Q_z^T \quad (2)$$

Where B_s denotes sparse direct predictor-response associations; Q_w denotes predictor latent loading vectors; Q_z denotes response latent loading vectors; and A denotes the latent regression association. Thus, the response model is therefore:

$$Z = W B_s + W Q_w A Q_z^T + E \quad (3)$$

This decomposition enables SA-MJLR to simultaneously identify individual predictor-response associations and shared latent dependencies.

2.1 Mechanism for Data Generation

The predictor and response matrices were generated according to the proposed sparse-latent structure for simulation experiments, and the predictor matrix is generated as:

$$W = T Q_w^T + E_w \quad (4)$$

Where

- $T \in \mathbb{R}^{n \times r}$ denotes latent scores;
- $Q_w \in \mathbb{R}^{p \times r}$ denotes sparse predictor loadings;
- E_w denotes the predictor noise.

On the other hand, the response matrix is generated as follows:

$$Z = T A Q_z^T + \gamma W B_s + E_z \quad (5)$$

Where;

- $A \in \mathbb{R}^{r \times r}$ denotes latent regression coefficients;
- $Q_z \in \mathbb{R}^{q \times r}$ denotes sparse response loadings;
- $B_s \in \mathbb{R}^{p \times q}$ denotes sparse direct coefficients;
- γ controls the magnitude of direct sparse effects;
- E_z denotes the response noise.

Thus, the true coefficient matrix employed for recovery evaluation is:

$$B_{true} = Q_w A Q_z^T + \gamma B_s \quad (6)$$

2.3 Sparse Direct-Effect Estimation

The first component of SA-MJLR estimates sparse direct predictor-response associations, and for each response variable $j = 1, 2, \dots, q$, a LASSO regression model is fitted as:

$$\hat{B}_{s,j} = \arg \min_{\beta_j} \left\{ \|Z_j - W \beta_j\|_2^2 + \lambda_B \|\beta_j\|_1 \right\} \quad (7)$$

Where λ_B controls sparsity, $\|\beta_j\|_1$ denotes the $L1$ penalty, and the resulting sparse coefficient matrix is:

$$\hat{B}_s = \begin{bmatrix} \hat{\beta}_{s,1} & \hat{\beta}_{s,2} & \cdots & \hat{\beta}_{s,q} \end{bmatrix} \quad (8)$$

After the sparse direct component estimation, its contribution is removed, that is:

$$Z_r = Z - W \hat{B}_s \quad (9)$$

Where Z_r denotes the remaining response variation that is explained by latent structures.

2.4 Sparse Joint Latent Loading Estimation

The latent component of SA-MJLR is being estimated from the residual response matrix Z_r , whereas the initial predictor-response association matrix and singular value decomposition applied are respectively given in Equations (10) and (11) as:

$$C = W^T Z_r \quad (10)$$

$$C = U \Sigma V^T \quad (11)$$

The initial loading matrices are obtained using Equations (12) and (13):

$$Q_w = U_r \quad (12)$$

$$Q_z = V_r \quad (13)$$

Where r denotes the selected latent rank, and to introduce sparsity, SA-MJLR applies alternating soft-thresholding updates, whereby the soft-thresholding operator is defined as:

$$S_{\lambda}(x) = \text{sign}(x) \max(|x| - \lambda, 0) \quad (14)$$

The predictor loading update and the response loading update are respectively represented in Equations (15) and (16) as:

$$Q_w^{(t+1)} = S_{\lambda_Q}(W^T Z_r Q_w^{(t)}) \quad (15)$$

$$Q_z^{(t+1)} = S_{\lambda_Q}(Z_r^T W Q_w^{(t+1)}) \quad (16)$$

After each update, the loading vectors are normalized as:

$$Q_j = \frac{Q_j}{\sqrt{Q_j^T Q_j}} \quad (17)$$

The alternating procedure will continue for a fixed number of iterations until a stable latent loading estimate is achieved.

2.5 Latent Regression Estimation

Latent predictor and response scores are computed, following latent loading estimation, as in Equations (18) and (19):

$$T = W Q_w \quad (18)$$

$$S = Z_r Q_z \quad (19)$$

Then, the latent regression association becomes:

$$S = TA + E \quad (20)$$

Using ridge regularization, the latent coefficient matrix is estimated as:

$$\hat{A} = (T^T T + \lambda_A I)^{-1} T^T S \quad (21)$$

Where λ_A controls numerical stability and I denotes the identity matrix.

2.6 Coefficient Reconstruction and Prediction

The latent coefficient component is reconstructed using:

$$\hat{B}_L = \hat{Q}_w \hat{A} \hat{Q}_z^T \quad (22)$$

The final SA-MJLR coefficient matrix is given as:

$$\hat{B} = \hat{B}_S + \hat{B}_L \quad (23)$$

or written equivalently as

$$\hat{B} = \hat{B}_S + \hat{Q}_w \hat{A} \hat{Q}_z^T \quad (24)$$

Predictions are obtained for new predictor observations, W_{new} as:

$$\hat{Z}_{new} = W_{new} \hat{B} \quad (25)$$

This formulation, however allows SA-MJLR in order to add both sparse direct effects and latent multivariate associations.

2.7 Benchmark Regression Methods

The proposed SA-MJLR technique is compared with four existing high-dimensional regression methods.

Ridge Regression

The Ridge estimates coefficients as stated in Equation (26):

$$\hat{B}_R = (W^T W + \lambda I)^{-1} W^T Z \quad (26)$$

in which λ controls shrinkage.

Elastic Net Regression

Elastic Net combines $L1$ and $L2$ regularization:

$$\hat{B}_{EN} = \arg \min_B \{\|Z - WB\|_2^2 + \lambda[\alpha \|B\|_1 + (1-\alpha) \|B\|_2^2]\} \quad (27)$$

Where $\alpha = 0.5$ is the implementation employed.

Reduced Rank Regression

The Reduced Rank Regression (RRR) estimates a low-rank approximation as:

$$B_{RRR} = U_r \Sigma_r V_r^T \quad (28)$$

Where r denotes the retained rank.

Sparse Partial Least Squares

The SPLS estimates latent predictor and response directions through singular value decomposition of $W^T Z$ and reconstructs:

$$B_{SPLS} = Q_w A Q_z^T \quad (29)$$

2.8 Computational Algorithm of SA-MJLR

The SA-MJLR algorithm provides a unified framework for modelling high-dimensional multivariate systems which contain both sparse direct associations and latent dependency structures, and is summarized as follows:

The input is W, Z whereas the output is $\hat{Z}$

- **Step one:** Obtain/generate the predictor and the response matrices
- **Step two:** Using LASSO, estimate sparse direct coefficients $\hat{B}_s$
- **Step three:** Expunge sparse direct contribution, that is: $Z_r = Z - W\hat{B}_s$
- **Step four:** Using SVD initialization and alternating soft thresholding, estimate joint sparse latent loadings, that is: $\hat{Q}_w, \hat{Q}_z$
- **Step five:** Calculate latent scores, that is: $T = W\hat{Q}_w$ and $S = Z_r\hat{Q}_z$
- **Step six:** Estimate the latent regression coefficient matrix: $\hat{A}$
- **Step seven:** Reconstruct coefficient matrix, that is: $\hat{B} = \hat{B}_s + \hat{Q}_w\hat{A}\hat{Q}_z^T$

Step eight: Then produce/generate predictions: $\hat{Z} = W\hat{B}$

2.9 Monte Carlo Simulation Design

2.9.1 Objective of Simulation Study: The Monte Carlo simulation study is designed to evaluate the finite-sample performance of the proposed Sparse Adaptive Multivariate Joint Latent Regression (SA-MJLR) method under controlled high-dimensional settings, and the objective is to examine whether SA-MJLR can effectively recover regression structures and provide accurate predictions whenever the predictor and response dimensions exceed the available sample size. Thus, the simulation framework is specifically designed to reflect situations where the true coefficient matrix contains two simultaneous components, which are:

1. A sparse direct-effect component representing a small number of important predictor-response relationships, and

2. A low-rank latent component representing hidden shared variation between predictor and response variables.

This structure reflects many modern high-dimensional applications in which observed variables are influenced by both individual effects and underlying latent processes.

2.9.2 Data-Generating Mechanism

In each simulation replicate; the predictor and the response matrices $W \in \mathbb{R}^{n \times p}$ and $Z \in \mathbb{R}^{n \times q}$ respectively were generated based on the model as follows:

$$Z = WB_s + WQ_wAQ_z^T + \varepsilon_z \quad (30)$$

Where W denotes the high-dimensional predictor matrix, B_s denotes a sparse direct-effect coefficient matrix, Q_w and Q_z denote, respectively, sparse latent loading matrices for predictor and response spaces, A denote the latent regression relationship, whereas ε denotes random response noise; and the predictor matrix is produced (generated) using:

$$W = TQ_w^T + \varepsilon_w \quad (31)$$

Where T represents latent scores and ε_w denotes predictor noise, and the true coefficient matrix used for evaluation is defined as:

$$B_{true} = Q_wAQ_z^T + \gamma B_s \quad (32)$$

Where γ controls the strength of the sparse direct contribution.

2.9.3 Simulation Parameters and Scenarios

In the experiment, five simulation scenarios were considered to evaluate model robustness under increasing dimensional complexity, and the scenarios were designed such that both predictor and response dimensions exceeded sample size, representing genuine high-dimensional problems. Thus, the simulation parameters are summarized in Table 1.

Table 1: Monte Carlo Simulation Settings

Scenario	Sample size (n)	Predictors (p)	Responses (q)	Rank	Sparse predictor loading	Sparse response loading	Sparse B	SNR	γ
SC1	50	100	100	5	0.05	0.05	0.01	5	0.5
SC2	80	300	500	5	0.03	0.03	0.005	3	1.0
SC3	120	600	800	5	0.02	0.02	0.003	2	1.0
SC4	150	1000	800	5	0.01	0.01	0.002	1	1.5
SC5	200	1500	1200	5	0.005	0.005	0.001	0.5	1.5

Table 1 shows the Monte Carlo simulation settings in which the scenarios progressively increased dimensionality while reducing signal-to-noise ratio and sparsity levels, and SC1 represents a relatively favourable high-dimensional condition, whereas SC5 represents an extreme setting with weak signals and highly sparse underlying relationships.

2.9.4 Competing Methods

The proposed SA-MJLR method is compared against four established multivariate regression approaches, known as Ridge Regression (representing dense shrinkage-based

regression), Elastic Net Regression (representing sparse regularized regression), Reduced Rank Regression (representing low-rank multivariate modeling), and Sparse Partial Least Squares (representing sparse latent projection modeling). These methods were selected because they represent major modelling strategies commonly used for high-dimensional regression problems.

2.9.5 Monte Carlo Procedure

For each scenario, the simulation experiment was repeated 10 times using independent randomly generated datasets, and a

fixed random seed was applied to ensure reproducibility. Then, for each replication, the following were executed, but the final reported results represent the average performance across all replications:

- Synthetic predictor and response matrices were generated.
- Each competing method was fitted using 80% of observations for training.
- The remaining 20% of observations were used for prediction evaluation.
- Estimated coefficient matrices were compared with the known true coefficient matrix.
- The prediction and recovery metrics were computed.

2.9.6 Evaluation Criteria

Two categories of evaluation metrics were used, which are prediction performance and Coefficient Recovery Performance. The prediction accuracy was assessed using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2), whereas coefficient recovery performance, which is the ability to recover the true coefficient structure, was evaluated using Frobenius Error (FE) and Correlation (ρ), and their computational formulas are stated in Equations (33) to (37).

$$RMSE = \sqrt{\frac{1}{nq} \sum_{i=1}^n \sum_{j=1}^q (Z_{ij} - \hat{Z}_{ij})^2} \quad (33)$$

$$MAE = \frac{1}{nq} \sum_{i=1}^n \sum_{j=1}^q |Z_{ij} - \hat{Z}_{ij}| \quad (34)$$

$$R^2 = 1 - \frac{\sum (Z - \hat{Z})^2}{\sum (Z - \bar{Z})^2} \quad (35)$$

$$FE = \sqrt{\frac{1}{pq} \|B_{true} - \hat{B}\|_F^2} \quad (36)$$

Here, a lower value indicates a more accurate coefficient recovery.

$$\rho = \text{cor}(\text{vec}(B_{true}), \text{vec}(\hat{B})) \quad (37)$$

Here, a higher value indicates a better structural recovery

2.10 Monte Carlo Simulation RStudio Codes for the Five Methods: The Monte Carlo Simulation for the proposed method, as well as the four competing methods, is committed to a GitHub repository accessible from the link <https://github.com/judend/Monte-Carlo-Simulation-RStudio-Codes-for-SA-MJLR-and-Four-Competing-Methods/blob/main/SAVE%20EZE-GOLDEN.docx>

3. Results

Table 2: Predictive Performance of Competing Techniques via Simulation Scenarios

Scenario	Method	RMSE Mean	RMSE SD	MAE Mean	MAE SD	R ² Mean	R ² SD
SC1	Elastic Net	0.3215	0.0173	0.2420	0.0065	0.6320	0.0950
SC1	RRR	0.3283	0.0170	0.2476	0.0121	0.6168	0.1031
SC1	Ridge	0.3005	0.0124	0.2335	0.0069	0.6796	0.0781
SC1	SA-MJLR	0.2955	0.0221	0.2213	0.0111	0.6901	0.0775
SC1	SPLS	0.2495	0.0136	0.1925	0.0066	0.7793	0.0532
SC2	Elastic Net	0.4431	0.0076	0.3494	0.0058	0.0475	0.0449
SC2	RRR	0.4293	0.0108	0.3327	0.0076	0.1061	0.0481
SC2	Ridge	0.4380	0.0089	0.3446	0.0066	0.0691	0.0470
SC2	SA-MJLR	0.3927	0.0067	0.3094	0.0047	0.2519	0.0351
SC2	SPLS	0.4123	0.0088	0.3219	0.0062	0.1754	0.0347
SC3	Elastic Net	0.6378	0.0040	0.5062	0.0026	-0.0177	0.0290
SC3	RRR	0.6261	0.0062	0.4893	0.0045	0.0195	0.0155
SC3	Ridge	0.6432	0.0050	0.5064	0.0037	-0.0349	0.0216
SC3	SA-MJLR	0.5843	0.0045	0.4634	0.0031	0.1457	0.0248
SC3	SPLS	0.6218	0.0067	0.4860	0.0049	0.0330	0.0141
SC4	Elastic Net	1.2589	0.0057	1.0007	0.0039	0.2782	0.0181
SC4	RRR	1.4804	0.0186	1.1338	0.0120	0.0022	0.0025
SC4	Ridge	1.4798	0.0174	1.1422	0.0100	0.0031	0.0045
SC4	SA-MJLR	1.1941	0.0043	0.9492	0.0030	0.3506	0.0165
SC4	SPLS	1.4795	0.0190	1.1326	0.0122	0.0034	0.0012
SC5	Elastic Net	2.4396	0.0095	1.9438	0.0067	0.1909	0.0122
SC5	RRR	2.7132	0.0203	2.0849	0.0142	-0.0006	0.0011
SC5	Ridge	2.7356	0.0192	2.1193	0.0150	-0.0172	0.0028
SC5	SA-MJLR	2.3839	0.0101	1.9001	0.0083	0.2275	0.0116
SC5	SPLS	2.7109	0.0194	2.0825	0.0137	0.0011	0.0007

The results in Table 2 show the predictive behaviour of all competing models under increasing dimensional complexity. For instance, in SC1, which represents the least challenging scenario, SPLS achieved the strongest predictive performance, producing the lowest RMSE (0.2495), MAE (0.1925), and highest R^2 (0.7793), which indicates that when the underlying structure is predominantly low-rank, sparse latent projection methods can effectively capture the predictor-response relationship. As dimensionality increased from SC2 to SC5, however, the performance advantage shifted toward SA-MJLR, and across SC2, SC3, SC4, and

SC5, SA-MJLR consistently achieved the lowest prediction errors and highest explanatory power, particularly under SC5, where the problem represents an extreme high-dimensional environment; SA-MJLR maintained the best predictive performance with an RMSE of 2.3839 compared with 2.7109 for SPLS and 2.7356 for Ridge. Thus, the consistently higher R^2 values obtained by SA-MJLR demonstrate its ability to preserve predictive information when conventional models experience deterioration due to increased dimensionality and reduced signal-to-noise ratio.

Table 3: Coefficient Recovery Performance via Simulation Scenarios

Scenario	Method	Frobenius Error Mean	Frobenius SD	Correlation Mean	Correlation SD
SC1	Elastic Net	0.1024	0.0060	0.3926	0.0389
SC1	RRR	0.0660	0.0034	0.5090	0.0559
SC1	Ridge	0.0882	0.0027	0.4640	0.0366
SC1	SA-MJLR	0.0565	0.0049	0.6475	0.0462
SC1	SPLS	0.0313	0.0020	0.8329	0.0250
SC2	Elastic Net	0.0464	0.0006	0.4612	0.0152
SC2	RRR	0.0382	0.0012	0.1667	0.0358
SC2	Ridge	0.0439	0.0009	0.2735	0.0123
SC2	SA-MJLR	0.0298	0.0008	0.6701	0.0190
SC2	SPLS	0.0359	0.0012	0.2685	0.0326
SC3	Elastic Net	0.0311	0.0002	0.5218	0.0086
SC3	RRR	0.0287	0.0005	0.1090	0.0186
SC3	Ridge	0.0314	0.0004	0.2304	0.0070
SC3	SA-MJLR	0.0229	0.0003	0.6757	0.0083
SC3	SPLS	0.0281	0.0005	0.1529	0.0179
SC4	Elastic Net	0.0241	0.0002	0.7469	0.0092
SC4	RRR	0.0344	0.0007	0.0815	0.0034
SC4	Ridge	0.0343	0.0006	0.2463	0.0025
SC4	SA-MJLR	0.0207	0.0002	0.8142	0.0077
SC4	SPLS	0.0343	0.0007	0.0825	0.0011
SC5	Elastic Net	0.0183	0.0001	0.7159	0.0054
SC5	RRR	0.0239	0.0003	0.0610	0.0026
SC5	Ridge	0.0243	0.0003	0.2119	0.0017
SC5	SA-MJLR	0.0171	0.0001	0.7617	0.0051
SC5	SPLS	0.0238	0.0003	0.0629	0.0018

The result in Table 3 shows the coefficient recovery analysis, which provides stronger evidence of the structural advantage of SA-MJLR; while some methods achieved competitive prediction in simple settings, accurate recovery of the underlying coefficient matrix requires successful identification of both sparse and latent components. From SC2 to SC5, SA-MJLR consistently achieved the lowest Frobenius Error and highest coefficient correlation, and this

indicates that the proposed method provides a more accurate reconstruction of the true regression structure. The improvement becomes particularly evident under SC4 and SC5, and in SC5, SA-MJLR achieved a Frobenius Error of 0.0171 and correlation of 0.7617, which outperforms Elastic Net, RRR, Ridge, and SPLS, confirming that separating sparse direct effects from latent shared variation improves parameter estimation in ultra-high-dimensional settings.

Table 4: Overall Ranking of Competing Methods

Method	Mean RMSE	Mean MAE	Mean R ²	Mean Frobenius Error	Mean Correlation	Performance Score
SA-MJLR	0.9701	0.7687	0.3331	0.0294	0.7138	6.9888
Elastic Net	1.0202	0.8084	0.2262	0.0445	0.5677	1.1420
SPLS	1.0948	0.8431	0.1985	0.0307	0.2799	-0.5756
RRR	1.1155	0.8577	0.1488	0.0382	0.1854	-3.3432
Ridge	1.1194	0.8692	0.1399	0.0444	0.2852	-4.2120

The results in Table 4 shows the aggregated ranking of the four existing methods against the proposed method (SA-MJLR), which confirms the overall superiority of SA-MJLR, and the proposed method achieved the highest performance score (6.9888), substantially exceeding all competing approaches. The SA-MJLR obtained the best overall predictive accuracy, structural recovery, and coefficient correlation, and the results demonstrate that combining sparse estimation with latent joint modelling provides a more complete representation of complex multivariate relationships than methods relying exclusively on shrinkage or low-rank assumptions. In overall, the simulation study provides strong empirical support that SA-MJLR is a robust framework for high-dimensional multivariate regression problems characterized by simultaneous sparsity and latent dependency structures.

4 Conclusion

In this study, the Sparse Adaptive Multivariate Joint Latent Regression (SA-MJLR) framework was proposed specifically

for high-dimensional multivariate regression problems characterized by the simultaneous presence of sparse direct effects and latent dependency structures. But the motivation for developing SA-MJLR is based on the lacunas or limitations of existing approaches, which often focus exclusively on either sparse variable selection or latent structure extraction, while many real-world high-dimensional systems exhibit both characteristics simultaneously. An extensive Monte Carlo simulation study was conducted to examine the performance of SA-MJLR under progressively challenging high-dimensional scenarios, and the simulation framework considered situations where both predictor and response dimensions were more than the sample size, with increasing dimensionality, decreasing sparsity levels, and reduced signal-to-noise ratios. The comparison was based on the developed technique, as against the four existing techniques known as Ridge Regression, Elastic Net Regression, Reduced Rank Regression (RRR), and Sparse Partial Least Squares (SPLS), and the findings revealed that SA-MJLR provided superior overall performance, particularly

under complex high-dimensional conditions. Although SPLS showed competitive performance in the simplest simulation scenario, SA-MJLR consistently achieved better predictive accuracy and coefficient recovery as the dimensional complexity increased, but the proposed method obtained lower RMSE and MAE values, higher R^2 values, lower Frobenius reconstruction errors, and stronger correlations between estimated and true coefficient matrices across most challenging scenarios.

Hence, the improved performance of SA-MJLR could be attributed to its hybrid modelling strategy, which separates sparse direct relationships from shared latent structures; unlike conventional regularization methods that assume either dense or purely sparse relationships, or latent projection methods that primarily capture hidden variation, SA-MJLR simultaneously exploits both sources of information, and this enables more accurate estimation of complex predictor-response relationships while maintaining interpretability through sparse coefficient recovery. Thus, the overall ranking analysis further confirmed the robustness of SA-MJLR, with the proposed method achieving the highest aggregate performance score among all competing approaches, and these results provide evidence that SA-MJLR represents a promising framework for high-dimensional multivariate regression applications where traditional approaches may suffer from information loss, poor prediction, or inaccurate structural recovery.

References

1. Bunea F, She Y, Wegkamp MH. Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics*. 2011;39(2):1282-1309. DOI:10.1214/10-AOS859.
2. Chen J, Huang JZ. Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association*. 2012;107(500):1533-1545. DOI:10.1080/01621459.2012.716613.
3. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery; 2016. p. 785-794. DOI:10.1145/2939672.2939785.
4. Chun H, Keleş S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2010;72(1):3-25. DOI:10.1111/j.1467-9868.2009.00723.x.
5. Fan J, Han F, Liu H. Challenges of big data analysis. *National Science Review*. 2014;1(2):293-314. DOI:10.1093/nsr/nwt032.
6. Fan J, Li R, Zhang CH, Zou H. Statistical foundations of data science. *Annual Review of Statistics and Its Application*. 2020;7:1-29. DOI:10.1146/annurev-statistics-031219-041205.
7. Fan J, Li R, Liao H, Wang Y. Learning high-dimensional factor models with structured sparsity. *Journal of the American Statistical Association*. 2023;118(542):1002-1016.
8. Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning: data mining, inference, and prediction*. 2nd ed. New York: Springer; 2017.
9. Ham Y, Price BS, Rothman AJ. Sparse multivariate regression with structured covariance estimation for high-dimensional responses. *Journal of Computational and Graphical Statistics*. 2024;33(2):451-465.
10. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*. 1970;12(1):55-67. DOI:10.1080/00401706.1970.10488634.
11. Izenman AJ. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*. 1975;5(2):248-264. DOI:10.1016/0047-259X(75)90058-4.
12. Kobak D, Lomond R, Sanchez A. Sparse reduced-rank regression for high-dimensional multivariate data. *Statistics and Computing*. 2021;31:76.
13. Li Y, Li L, Wang H, Zhang X. High-dimensional multivariate regression with large-scale response variables: sparse estimation and prediction. *Journal of Multivariate Analysis*. 2024;204:105341.
14. Rothman AJ, Levina E, Zhu J. A new approach to Cholesky-based covariance regularization in high dimensions. *Biometrika*. 2010;97(3):539-550. DOI:10.1093/biomet/asq022.
15. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1996;58(1):267-288.
16. Wainberg M, Alipanahi B, Frey BJ. Are random forests truly the best classifiers? *Journal of Machine Learning Research*. 2016;17(1):3837-3841.
17. Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*. 2001;58(2):109-130. DOI:10.1016/S0169-7439(01)00155-1.
18. Yuan M, Zhang CH. On tensor completion via nuclear norm minimization. *Foundations of Computational Mathematics*. 2012;12:279-297.
19. Zhao P, Yu B. On model selection consistency of Lasso. *Journal of Machine Learning Research*. 2006;7:2541-2563.
20. Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2005;67(2):301-320. DOI:10.1111/j.1467-9868.2005.00503.x.