Results in Control and Optimization 21 (2025) 100609

Contents lists available at ScienceDirect

Results in Control and Optimization

journal homepage: www.elsevier.com/locate/rico

# Robust correlation feature selection based support vector machine approach for high dimensional datasets

Ishaq Abdullahi Baba [a],*, Mohammed Bappah Mohammed [b],
Kamal Bakari Jillahi [a], Aliyu Umar [a], Hasan Talib Hendi [c]

[a] *Department of Mathematical Sciences, Taraba State University Jalingo, Taraba State, Nigeria*
[b] *Mathematics and Statistics Department, Federal University of Kashere, 234, Gombe, Nigeria*
[c] *Institute for Mathematical Research, Universiti Putra Malaysia, 43400, Serdang, Selangor, Malaysia*

## ARTICLE INFO

*Keywords:*
Correlation
Outliers
High dimensional data
Feature selection
Support vector machine

## ABSTRACT

Correlation-based feature selection methods are popular tools used to select the most important variables to include the true model in the analysis of sparse and high-dimensional models. In application, the presence of anomalous observations in both predictors and responses can seriously jeopardize the prediction accuracy of the model, which in turn leads to misleading interpretations and conclusions if not correctly addressed. Furthermore, the cause of dimensionality is another serious difficulty facing many existing feature selection algorithms. To achieve more reliable feature selection and prediction accuracy, a weighted sure independence screening-based support vector machine for high-dimensional datasets is proposed. The key contribution of our proposed method is that it minimizes the influence of outliers in differentiating between significant and insignificant features and improves predictability and interpretability. Our method consists of three basic steps. In the first step, a weights-based modified reweighted fast, consistent, and high break-down point is computed. The second step utilizes the estimates of weights from the first step to select the most important variables for the model. The third step employs the support vector machine algorithm to calculate prediction values. To demonstrate the effectiveness of the developed procedure, we used both simulation and real-life data examples. Our results show that the proposed methods performs better with a clear margin compared to other procedures.

## 1. Introduction

The massive accumulation of high dimensional datasets as a result of the rapid development of computers, network technology, high-tech, and internet development has increased tremendously during the last two decades. The dimensions contain information about several variables, but many of these variables are insignificant or harmful to the prediction algorithm, which degrades its performance and increases its complexity. When building an efficient and accurate prediction model for machine learning with huge dimensions, a feature selection algorithm that selects a feature subset from the original features is the most effective screening procedure [1–3]. The quality of the selected variable subset is a key element in determining the accuracy of the prediction model. As a result, researchers in bioinformatics [4,5], text classification [6], glass classification [7], epidemiology [8] and related fields are more and more recognizing the importance of feature selection algorithms in building reliable prediction models.

* Corresponding author.
*E-mail address:* ishaqbaba@yahoo.com (I.A. Baba).

https://doi.org/10.1016/j.rico.2025.100609
Received 31 December 2023; Received in revised form 21 April 2025; Accepted 31 August 2025

Available online 11 September 2025

High-dimensional datasets are ubiquitous in various fields, such as genomics, proteomics, and image processing, where thousands or even millions of features need to be analyzed [9]. However, analyzing such datasets poses several challenges. The curse of dimensionality leads to sparsity, making it difficult to discern meaningful patterns from noise. Moreover, multicollinearity among features can distort model predictions, while outliers can significantly impact the performance of feature selection and machine learning algorithms [5,10]. Feature selection algorithms play a critical role in high-dimensional statistical regression modeling, which commonly appears in many fields of biomedical and engineering applications [11,12]. This procedure is used to calculate and select the most significant features for the proposed model. The imbalance in high dimensional space can lead to several issues, including overfitting, increased computational complexity, and difficulties in model interpretability [13]. By selecting a subset of relevant features, we can enhance model performance, reduce computational costs, and improve the interpretability of results. Despite the critical importance of feature selection, existing methods have significant limitations [14–16]. Classical correlation-based feature selection methods are particularly vulnerable to the effects of outliers and noise [17,18]. [19] studied application of variable selection for high dimensional survival data via Conditional partial likelihood screening algorithm. In the presence of high collinearity between variables, this method may select insignificant variables. [20] suggested a greedy adaptive quantile regression algorithm to conduct variable selection for high dimensional censored data. [10] proposed a two steps method called point biserial sure independence screening PB-SIS to filter unimportant variables. The method is fast and efficient, but can only be applied to the binary response dataset and perform poorly under highly correlated predictors.

While most of the existing sure independence screening algorithm identify significant variables that influence the output variables in high dimensional settings they suffer from the lack of robustness, sensitivity to high collinearity and unusual marginal screening limitations. This vulnerability undermines the performance of feature selection algorithms particularly in high-dimensional settings where the risk of overfitting is already high. Inspired by these limitations and the work of [21–24] we incorporated robustness into correlation-based feature selection and SVM to diminish the effect of outliers and increase the reliability of feature selection methods to recover the true model with maximum efficiency in high-dimensional data analysis. Robust correlation algorithms ensure that only significant features are selected, while support vector machines ensure that an efficient and accurate prediction model is constructed. This paper, proposes a framework to simultaneously perform feature selection and prediction using weighted Pearson correlation (Wpc) and support vector machines with optimal guarantees.

### 1.1. Motivation

The importance of correlation analysis in science, engineering, economics, and health sciences cannot be overemphasized. This statistic is widely used to create many vital statistics, such as penalized regression based feature selection, correlation based feature selection, and outlier diagnostic methods. The correlation coefficient measures the relationship between two or more variables. The coefficient value is measured in the range of +1 to −1. A correlation value of +1 indicates a direct positive relationship between variables, and −1 indicates an inverse relationship. In some cases, a small number of outliers can significantly reduce the efficiency of the correlation estimator. However, it is now evident that the Pearson correlation coefficient is greatly affected by outliers. An analytical study of the robust and traditional correlation coefficients shows that the Spearman and Kendall [25], a non-parametric based on maximum deviation [26], and Hampel medians of absolute deviations [27] correlation estimators are also sensitive to the presence of outliers, even though they possess a higher breakdown point than the Pearson correlation estimator. The weighted Pearson correlation, named LMS correlation, has proved to attain a 50% breakdown point compared to its competitors [21]. [28] noted that the LMS based correlation estimators tend to produce extreme correlation values since they use 0 or 1 as weights. As a result, a new robust correlation based on the weighted average correlation was proposed as an alternative by [28]. However, it is of interest to note that the above mentioned correlation method cannot be applied in high dimensional space. For high dimensional datasets where the number of features exceeds the sample, the curse of dimensionality, sparsity, and multicollinearity are the top challenges that hamper the application of feature selection techniques. [29,30] Recent studies have concentrated on proposing and applying numerous feature selection procedures combined with machine learning for variable selection and prediction. [31] provides a detailed survey of feature selection procedures by classifying them into filter, wrapper, and combined procedures. The authors emphasize the application of high-dimensional datasets, such as text mining and bioinformatics. [9] combined the genetic algorithm and mutual information procedure to perform feature selection for the classification of biomedical datasets. [32] introduces a new hybrid feature selection technique based on deep learning context. The procedure integrates a feature selection layer into the natural network to achieve model learning and the selection of significant features. [33] integrates the Least Absolute Shrinkage and Selection Operator (LASSO) with random forests for a high-dimensional dataset. [34] provides a review of recent developments in feature selection methods with applications in machine learning. [35] a novel feature selection method based on the Mahalanobis distance and it down weights observations with large distances. Their approach and remarkable results are challenged by the problems of computational complexity and scalability. Although all the feature selection methods discussed above work well in feature selection for high-dimensional data, none of these methods addresses the simultaneous problem of response and predictor outliers, which is well known to affect the estimate of correlation and, in turn, diminishes the performance of feature selection and prediction machine learning algorithms. To address the challenge, we unify the aforementioned ideas into a joint framework and develop a robust and efficient correlation algorithm and correlation feature selection-based support vector machine in high-dimensional settings.

### 1.2. Contribution of the study

Correlation based feature selection methods, such as the Kendall correlation based feature selection approach proposed by [36] are widely used in the literature [37,38]. This method is well known to be susceptible to outliers and noise in a dataset. However, their integration with SVM in the context of high-dimensional datasets has not been thoroughly explored. Addressing this gap is critical for advancing the performance and robustness of correlation based feature selection approaches in high dimensional data. In this paper, a novel weighted correlation feature selection based support vector machine approach is proposed for high dimensional data. This paper attempts to propose solutions to address the challenges facing existing correlation based feature selection methods, thereby proposing the following contributions:

- To construct a weighted Pearson correlation algorithm that is less sensitive to the presence of the outlying points
- To develop a robust correlation feature selection method that is less affected by outliers.
- To integrate the proposed robust correlation feature selection with support vector machines for prediction tasks in high-dimensional datasets
- To evaluate the performance and robustness of the proposed approach using simulation and benchmark datasets.

To the best of our knowledge, this is the first study that addresses the problem of predictor and response outliers in the correlation-based feature selection problem. To prove the efficiency of the proposed approach, a first comparison of the proposed weighted correlation with three existing correlation coefficient estimators is presented. Moreover, a comparison of the proposed weighted correlation feature selection based support vector machine approach is also presented. The rest of this paper is organized as follows: Section 2 presents the concept of the correlation coefficient estimator. Sections 3 and 4 explain the proposed weighted correlation and the weighted correlation feature selection procedure, while Section 5 presents the simulation experiment for both univariate and high-dimensional datasets and their results. Section 6 explores the application of feature selection-based support vector machines in high-dimensional space using glass data and Bardet–Biedl syndrome gene expression datasets. Section 7 gives a summary of the paper, including the limitations of the proposed technique and potential future research directions.

## 2. Measure of correlation coefficient

In practice, correlation coefficients are calculated from a small sample size when the number of observations is small. The vagueness about the true value of the coefficient can be high, particularly when the estimated correlation is small. Hence, it is paramount to measure the imprecision of the correlation coefficient to evaluate its significance and conduct sensitivity analysis. The correlation coefficient is applied for feature selection algorithms to study microarray gene expression and chemometric datasets, which frequently come with increasingly high-dimensional features and a small sample size. Outlier detection and robustness problems are challenging even when the number of predictor variables $p$ is of moderate size. A traditional robust estimator cannot be computed for $p > n$. Therefore, achieving a robust and efficient correlation estimator in high-dimensional data is a good contribution to the body of knowledge.

Given a dataset $\left(x_1, y_1\right), \left(x_n, y_n\right)$ which comes from a bivariate normal distribution with parameters $\left(\mu_x, \mu_y, \sigma_x, \sigma_y\right)$ where $\mu_y$ and $\sigma_y$ are the mean and variance values of $y$, $\mu_x$ and $\sigma_x$ are the mean and variance of $x$, and $\rho$ is the correlation coefficient between $x$ and $y$. The commonly used sample correlation method for estimating $\rho$ is the classical Pearson correlation (pc) coefficient estimator defined by [21,39,40]:

$$r_{\mathbf{pc}} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\left[\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2\right]^{\frac{1}{2}}} \tag{1}$$

It has been observed that classical correlation coefficients fail to produce reliable estimates because they rely on classical sample mean estimates, which are well known to be sensitive to outliers. To address this problem, Spearman rank correlation (sc) and Kendall correlation coefficient estimators are recommended. Therefore, the sample versions of the Spearman and Kendall correlation (kc) coefficients are defined as [41,42]:

$$r_{\mathbf{sc}} = 1 - \frac{6\sum_{i=1}^{n}\left(P_i - Q_i\right)^2}{n\left(n^2 - 1\right)} \tag{2}$$

$$r_{\mathbf{kc}} = \frac{\sum_{i\neq j}^{n}\sum_{j=1}^{n}\left(sgn\left(x_i - x_j\right) sgn\left(y_i - y_j\right)\right)}{n\left(n-1\right)} \tag{3}$$

The Kendall correlation, which is generally more robust than the Spearman correlation, can effectively handle heavy-tailed errors but not outliers in both dependent and independent variables [43–45]. As an alternative, a robust correlation coefficient estimator based on the least median squares (LMS) was introduced by [21]. This method has been shown to be more robust and efficient compared to the Spearman and Kendall correlation estimators. A comparison between LMS correlation and weighted average correlation, based on the leave-one-out test by [28] demonstrates that LMS-based correlation is robust and more efficient, even with 50% contamination. Penalized regression techniques are commonly used as an alternative to the correlation feature selection procedures [18,46,47] proposed a weighted least absolute deviation-based linear programming optimization method, emphasizing the importance of selecting proper weights to handle outlying points. [48,49] also discuss the significance of weight selection in this context. [50] investigated the effect of outlying observations on both predictors and response variables using a penalized weighted least absolute deviation estimator. Although this method shrinks less significant features toward zero, it is highly biased and computationally intensive. However, all the aforementioned weighted-based estimators, including robust correlation-based LMS estimators, are not applicable to high-dimensional settings ($p > n$). Therefore, introducing a robust correlation coefficient for high-dimensional datasets would be a significant contribution to the fields of statistics and other disciplines.

## 3. Proposed robust weighted correlation coefficient

The main objective behind robust estimation of univariate or multivariate statistics such as mean and covariance, distance measure, or correlation coefficient is to shrink the influence of outlier data points either by removing or downweighting them altogether using a specific rule [51,52]. It is well known that the presence of a single outlier can influence the estimate of the Pearson correlation coefficient badly. Moreover, direct computation of the robust correlation coefficient formula given in Equation (2.4) of [21] is quite difficult when the dimension of the dataset is high since it uses the least median squares (LMS) estimator. The restrictions of LMS in high-dimensional settings stem from its computational complexity, the need for a sufficiently large sample size relative to the number of variables, and the challenges of dealing with outliers in a high-dimensional space. These limitations suggest the need for an alternative, robust estimation method for high-dimensional data analysis.

Following [21], we formulate an equivalent weighted Pearson correlation (Wpc) algorithm that is simple to solve computationally in high-dimensional settings. For observation having d-variate points $\left(x_1, x_2, \ldots x_n\right)$ with nonnegative weights $\left(w_1, w_2, \ldots w_n\right)$, we defined their weighted average as:

$$\bar{x}_w = \frac{w_1 x_1 + w_2 x_2 + \ldots + w_n x_n}{w_1 + w_2 + \ldots + w_n} \tag{4}$$

The use of weights can vary depending on the scenario. In this paper, we focus on applying weights to reduce the influence of outlying observations when computing correlation coefficients [53,54]. Similar to Eq. (4), we define the weighted covariance matrix as follows:

$$\frac{\sum_{i=1}^{n} w_i \left(x_i - \bar{x}_w\right)\left(x_i - \bar{x}_w\right)\prime}{\sum_{i=1}^{n} w_i} \tag{5}$$

And the weighted Pearson correlation coefficient between $x$ and $y$ can be expressed as:

$$r_{\mathbf{Wpc}} = \frac{\sum_{i=1}^{n} w_i \left(x_i - \bar{x}_w\right)\left(y_i - \bar{y}_w\right)}{\left[\sum_{i=1}^{n} w_i \left(x_i - \bar{x}_w\right)^2 \sum_{i=1}^{n} w_i \left(y_i - \bar{y}_w\right)^2\right]^{\frac{1}{2}}} \tag{6}$$

The weighted Pearson correlation coefficient can be viewed as a Pearson correlation with appropriately transformed data. Calculating the breakdown of the weighted correlation coefficient with known weights is similar to determining the breakdown of the classical correlation in the presence of dataset contamination. Several weighting functions have been employed in the literature to minimize the effect of outliers. Examples include the bi-squared weight function [55], weights based on the Mahalanobis distance [56,57] and weights derived from a clean set [58,59]. However, before applying the standard Pearson correlation algorithm to compute the weighted Pearson correlation coefficient, we need to determine the weighting values. Following [60], we summarize the steps to calculate weights in high-dimensional settings as follows.

Define a matrix with dimensions $p \gg n$

(a) Obtain the location and scatter matrix from the modified reweighted fast consistent and high breakdown point (MRFCH) estimator $\hat{\mu}_{MRFCH}$ and $\hat{\Sigma}_{MRFCH}$

(b) Use the location and scatter matrix estimates from step (a) to calculate the robust Mahalanobis distance (RMD) as follows:

$$RMD_i = \frac{\sum \left(X_{ij} - \hat{\mu}_{j,MRFCH}\right)}{diag\left(\hat{\Sigma}_{MRFCH}\right)} \tag{7}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, p$

(c) Calculate the cut-off value according as:

$$cut = med\left(RMD_i\right) + 3\left(MAD\left(RMD_i\right)\right) \tag{8}$$

(d) Compute the weights as:

$$w_i = min\left[1, \frac{cut}{RMD_i}\right] \tag{9}$$

It is important to note that with the weights derived from the aforementioned algorithm, the weighted Pearson correlation formula in Eq. (6) is expected to be highly resistant to outliers and consistent as the MRFCH estimator. Our weighting function effectively reduces the impact of outlying observations by adjusting weights rather than assigning binary weights of zero or one to contaminated and non-contaminated observations. The performance of the proposed weighted Pearson correlation coefficient estimator will be discussed in the following section.

## 4. Proposed feature selection based weighted correlation method

In this section, we develop an independent feature selection algorithm built upon the Pearson correlation and weights based on robust Mahalanobis distances. Let $x = \left(x_1, \ldots, x_p\right)^T$ be the covariate vector, and $y = \left(y_1, \ldots, y_n\right)^T$ be the response vector. In high-dimensional space, where the number of covariates exceeds the sample size $n$, it is natural to assume that only a few covariates are significant to the response variable y. Following [37,61], we defined the active and inactive sets of the covariate by

$$\mathcal{D} = \{j : F(y|x) \text{ functionally depends on } X_j\}, \tag{10}$$

and

$$\mathcal{I} = \{j : F(y|x) \text{ does not functionally depend on } X_j\}. \tag{11}$$

Our primary interest is to select active set of covariates $\mathcal{D}$ using weighted feature selection procedures for high dimensional data analysis.Following [37,61], we define the active and inactive sets of covariates as:

$$\omega_j = corr_w\left(X_j, y\right). \tag{12}$$

where positive weights $w > 0$ are used to construct robust estimates. The weighted correlation $\omega_j$ measures the linear relationship between $X_j$ and $y$, and $\omega_j = 0$ indicates independence.

We then select significant features as

$$\hat{D}_\gamma = \{j : |\omega_j| \text{ is among the top } [\gamma n] \text{ largest of all predictors}\},$$

where $[\gamma n]$ is the integer part of $\gamma n$. This reduces the dimensionality from $\{1, \dots, p\}$ to a submodel $\hat{D}_\gamma$ of size $d = [\gamma n] < n$. Hence, the weighted Pearson correlation (Wpc) approach generalizes the traditional Pearson correlation-based sure independence screening, while offering improved robustness and flexibility. Fig. 1 shows the flowchart for computing weighted sure independence screening and weighted sure independence screening based support vector machine algorithms.

### 4.1. Statistical properties

In this section, we discuss some important theoretical properties under appropriate conditions. The weighted Pearson correlation is a generalization of the traditional Pearson correlation [62] when $w_i = \frac{1}{n}$ for all i, implying that the proposed Wpc is an extension of the classical Pearson correlation coefficient that incorporates a chosen weight vector to each observation in the dataset [21,47]. The $w_i$ in Eqs. (6) are applied to minimize the effect of outlying observations in calculating means and variances of $X$ and $y$ as well as correlation estimates. In addition, under non-constant variance, the Wpc assigned small weights to noise data points, marking it robust against outliers and guaranteeing better estimates of the true population parameters as the amount of data collected increases. For consistency properties, the $w_i$ are positive values such that $\sum_{i=1}^{n} w_i \rightarrow 1$ as $n \rightarrow n$ and that the observations are independent and identically distributed generated from the bivariate distribution function. Empirically, we prove that the correlation coefficient estimated using WPC converges to the true values as the sample size n and the number of covariate variables grow to infinity. Figs. 3(a) and 3(b) clearly illustrate the strong consistency of the Wpc estimator supported by the decrease in bias with increasing sample size $n$. Figs. 3(c) to 3(d) further highlight the robustness of the weighted correlation method by showcasing its performance in terms of mean squared error. The proposed Wpc consistently achieves low bias indicating that it provide more accurate point estimate compare to nonparametric Kendall and Spearman correlation that show moderate bias. The Pearson produces high bias highlighting it sensitivity to the presence of contamination. Similarly The Wpc has lowest MSE with Pearson showing MSE growing rapidly. In overall the proposed method show robustness and consistency in both bias and MSE even under contamination and non-constant variance see [22,24]. Under these conditions, we show that the Wpc is unbiased and consistent.

### 4.2. Robustness and consistency of the weighted correlation

(i) For robustness: we defined $(y_o, X_o)$ as an outlier with large magnitude, and the traditional Pearson correlation given Eq. (1) is very sensitive. Then, if $y_o$ and $X_o$ are outliers, $\bar{y}$ and $\bar{X}$ will be heavily affected, leading to bias. In Wpc applying $w_i$ $(w_i \rightarrow 0)$ minimizes the influence of $(y_o, X_o)$, thereby diminishing bias. Hence the Wpc is robust since

$$\lim_{w_i \to 0} r_{Wpc} = r$$

(ii) For consistency: a correlation estimator $r_n$ is consistent if

$$\lim_{n \to \infty} r_n = r_\rho$$

where $r_\rho$ is the true population correlation. For the Wpc estimator

$$E[r_{Wcp}] = r_\rho + O(O^{-\frac{1}{2}}) \tag{13}$$

under standard regularity conditions the weights sum to 1 and the data is iid, then law of large number ensures that

$$\lim_{n \to \infty} r_{Wpc} = r_\rho$$

Hence the Wpc is unbiased and consistent

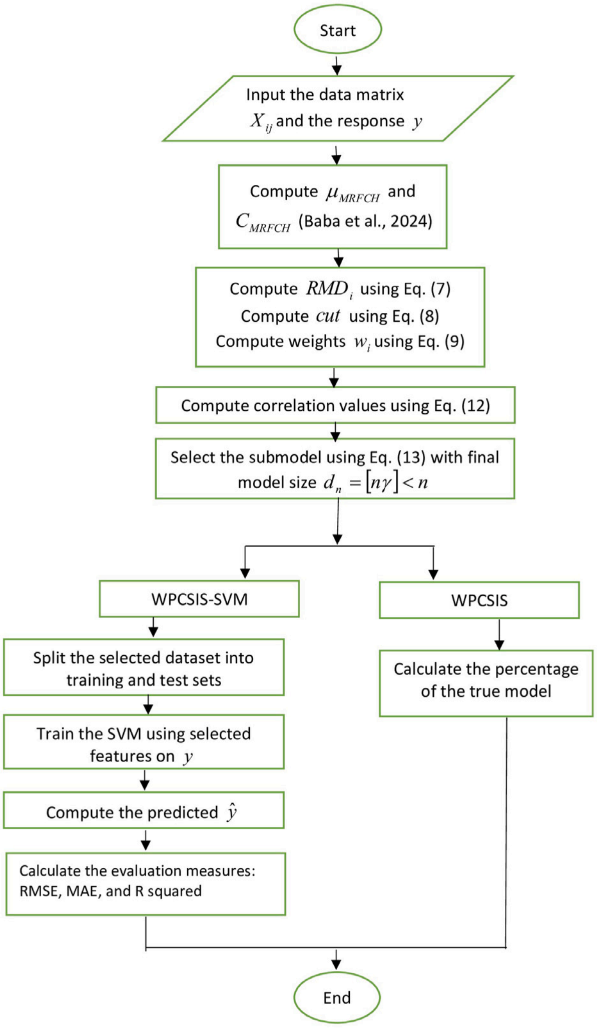

Start
Input the data matrix $X_{ij}$ and the response $y$
Compute $\mu_{MRFCH}$ and $C_{MRFCH}$ (Baba et al., 2024)
Compute $RMD_i$ using Eq. (7)
Compute $cut$ using Eq. (8)
Compute weights $w_i$ using Eq. (9)
Compute correlation values using Eq. (12)
Select the submodel using Eq. (13) with final model size $d_n = [n\gamma] < n$
WPCSIS-SVM
WPCSIS
Split the selected dataset into training and test sets
Train the SVM using selected features on $y$
Compute the predicted $\hat{y}$
Calculate the evaluation measures: RMSE, MAE, and R squared
Calculate the percentage of the true model
End

**Fig. 1.** Flowchart for computing WPCSIS and WPCSIS-SVM estimators.

### 4.3. Screening properties

We now turn to sure screening and ranking consistency of the weighted sure independence screening (WPC-SIS) procedure as $n \to \infty$ and $p \to \infty$ with $p \gg n$. WPC-SIS extends the original SIS by incorporating observation weights, thereby improving feature selection in the presence of outliers and heteroscedasticity. Under conditions analogous to those in Fan and Lv [61] and robust extensions such as Zhong [23], WPC-SIS achieves (i) the *sure screening* property (the active set is retained with probability tending to one) and (ii) *ranking consistency* (active predictors are ranked ahead of inactive ones with high probability). Simulation studies presented in Section 5.3 support these theoretical guarantees and demonstrate the practical utility of WPC-SIS in real-world settings.

## 5. Simulation study

### 5.1. Simulation design for bivariate scenario

To perform a simulation for a bivariate normal distribution, we specify the mean, variance, and correlation parameters. We then use the *mvrnorm* function from the *MASS* package in R to generate the dataset. Our example begins by generating observations from a multivariate normal distribution based on the following procedure:

$$X_{n\times p} \sim MN\left(\begin{bmatrix}0\\1\end{bmatrix}, \begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}\right), \tag{14}$$

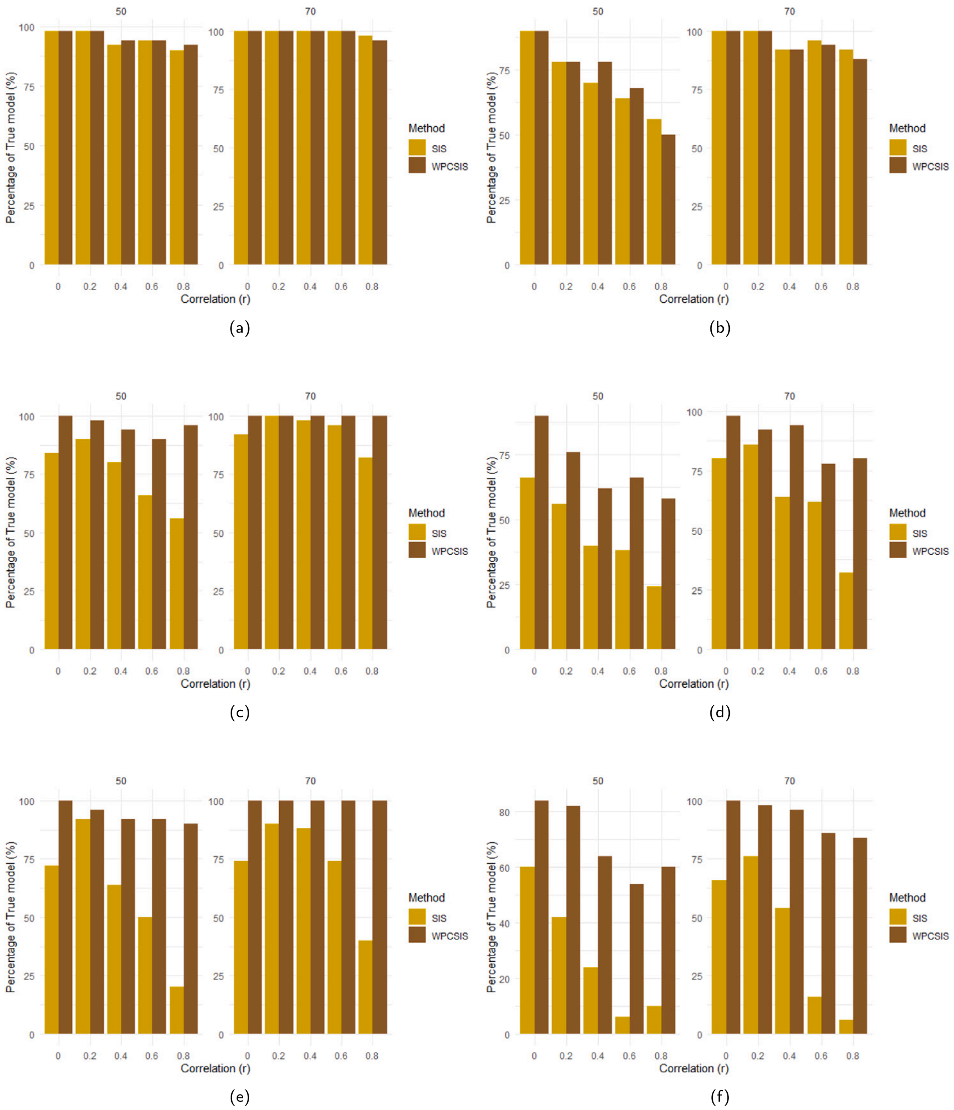


**Fig. 2.** Accuracy of SIS and WPCSIS in including the true model: (a) and (b) for $p = 100$ and $p = 1000$ at 0 contamination, (c) and (d) for $p = 100$ and $p = 1000$ at 0.02 contamination and (e) and (f) for $p = 100$ and $p = 1000$ at 0.05 contamination.

For contaminated set, the following model is applied:

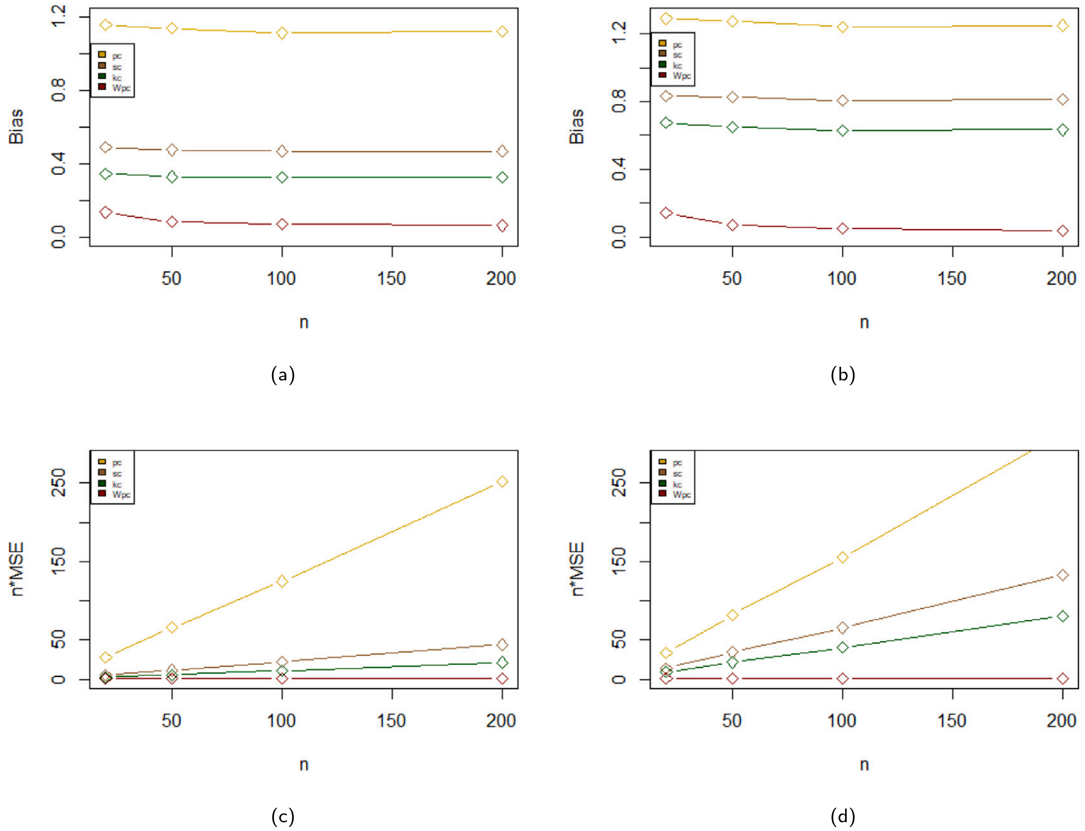
$$X_{n\times p} \sim MN\left(\begin{bmatrix}5\\-5\end{bmatrix}, 5\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}\right), \tag{15}$$

where $p$ represents the dimensions of the data matrix and $n$ is the sample size with values specified at 20, 50, 100, and 200. Let the percentage of the outliers be denoted as $\alpha$ with values 0.1 and 0.2. For a given percentage of outliers $\alpha$, let $m = [\alpha \times n]$ denote the number of bad observations and symbolize the integer part. For each sample, we computed the correlation coefficient $\hat{\rho}_k$, $(k = 1, \ldots, 100)$ and employed the absolute bias and mean squared error (MSE) criteria to evaluate the performance of the selected correlation methods according to the next formulas:

$$MSE = \frac{1}{M}\sum_{j=1}^{M}\left(\hat{\rho}_k - \rho\right)^2 \tag{16}$$

$$Bias = |\hat{\rho}_k - \rho| \tag{17}$$

It is important to emphasize that a correlation coefficient with low MSE and bias is considered a more accurate estimator. Fig. 3 shows the biases of the correlation methods considered. It can be observed that when the contamination is between 0.1 and 0.2, the



(a)

(b)

(c)

(d)

**Fig. 3.** (a) and (b): Plot of Bias vs sample size ($n$) and (c) and (d): n*MSE vs sample size ($n$) for bivariate data ($p = 2$).

proposed weighted Pearson correlation estimator (Wpc) provides superior estimates, followed by the Kendall correlation estimator (kc). In contrast, the Pearson correlation (pc) is the worst performing method in sample sizes of 20, 50, 100, and 200, respectively. Similar results were obtained in Figs. 3(c) and 3(d), indicating that our proposed Wpc method yields fine results by producing low MSE values compared to the other methods considered in this study, regardless of the sample size and contamination level. Furthermore, for the other methods, their performance deteriorates as the sample size increases, particularly when the contamination is 0.2 and the sample size is 20, 50, 100, and 200, respectively. Based on these findings, we can conclude that the proposed weighted Pearson correlation (Wpc) demonstrates high efficiency and robustness across different combinations of characteristic parameters (sample size, contamination, and dimensions).

### *5.2. Simulation design for high dimensional scenario*

To further establish the efficiency and robustness of the proposed weighted Pearson correlation (Wpc) estimators in high-dimensional cases ($p > n$), we followed a simulation scheme similar to that of Croux and Dehon (2010) and Boudt et al. (2012): First, we generated clean observations from the multivariate normal distribution $N_p \sim (0, R)$, where $R$ represents the correlation matrix with all diagonal elements equal to 1 and all off-diagonal elements equal to 0.8. Secondly, for contaminated data points, we generated a fraction of 0%, 10%, and 20% of contaminated sample data points from the multivariate distribution $N_p \sim ((5, -5, 5, -5, \dots), 5 * R)$. The simulation was repeated 100 times with $p = 200$ and $p = 1000$ dimensions and sample sizes of $n = 20, 50$, and 100. Table 1 shows the MSE for $p = 200, n = 20, 50$, and 100, and $p = 1000, n = 20, 50$, and 100, at different contamination levels (0%, 10%, and 20%).

Fig. 4 displays the MSE values at different correlation values ($r = 0.2, 0.4, 0.6, 0.8$) for $p = 100, 1000$, and $n = 20$ at 10% and 20% contamination, demonstrate the relationship between MSE and correlation ($r$) values. In the case of contamination, as the correlation value ($r$) increases, the MSE also increases for Pearson, Spearman, and Kendall correlation estimators. However, for our weighted Pearson correlation estimator, the MSE value decreases with an increase in the correlation value, indicating that our proposed Wpc is superior to the compared approaches in terms of MSE value. The results in Table 1 indicate that all the methods perform well. When $n = 20$ at 0 contamination, the performance of pc dramatically degrades, with sc and kc showing moderate performance, while the Wpc remains stable even at 0.1 and 0.2 contamination, indicating strong robustness under smaller sample settings. Similar performance was observed for $n = 50$ and 100 when $p = 200$ and 1000. Across all settings, Wpc maintains low MSE and robustness but the traditional approaches become very sensitive in the presence of contamination.

**Table 1**
MSE for different sample for high dimensional scenarios.

| $n$ | con | $p = 200$ | | | | $p = 1000$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | pc | sc | kc | Wpc | pc | sc | kc | Wpc |
| 20 | 0 | 0.62 | 0.71 | 0.96 | 0.84 | 0.72 | 0.82 | 1.05 | 0.87 |
| | 0.1 | 10.39 | 2.17 | 1.94 | 0.94 | 10.47 | 2.21 | 1.98 | 0.84 |
| | 0.2 | 13.18 | 5.55 | 3.73 | 0.93 | 13.11 | 5.51 | 3.72 | 0.91 |
| 50 | 0 | 0.55 | 0.63 | 1.63 | 0.98 | 0.57 | 0.69 | 1.79 | 1.09 |
| | 0.1 | 25.29 | 4.62 | 4.24 | 1 | 25.68 | 4.89 | 4.56 | 1.15 |
| | 0.2 | 30.48 | 13.01 | 8.56 | 1.08 | 31.33 | 13.22 | 8.75 | 1.1 |
| 100 | 0 | 0.57 | 0.67 | 3.09 | 1.24 | 0.57 | 0.68 | 3.1 | 1.22 |
| | 0.1 | 48.75 | 8.66 | 8.21 | 1.08 | 49.1 | 8.74 | 8.21 | 1.09 |
| | 0.2 | 61.08 | 25.52 | 16.53 | 1.04 | 60.35 | 25.67 | 16.76 | 1.25 |

At all levels of correlation, the Wpc outperforms the traditional methods achieving smaller MSE. The pc is the worst performing, followed by sc and kc methods. This shows that our proposed weighted Pearson correlation (Wpc) estimator is not only suitable for low-dimensional datasets but can also be applied in high-dimensional settings, even in the presence of contamination.

### 5.3. Simulation design for feature selection algorithm

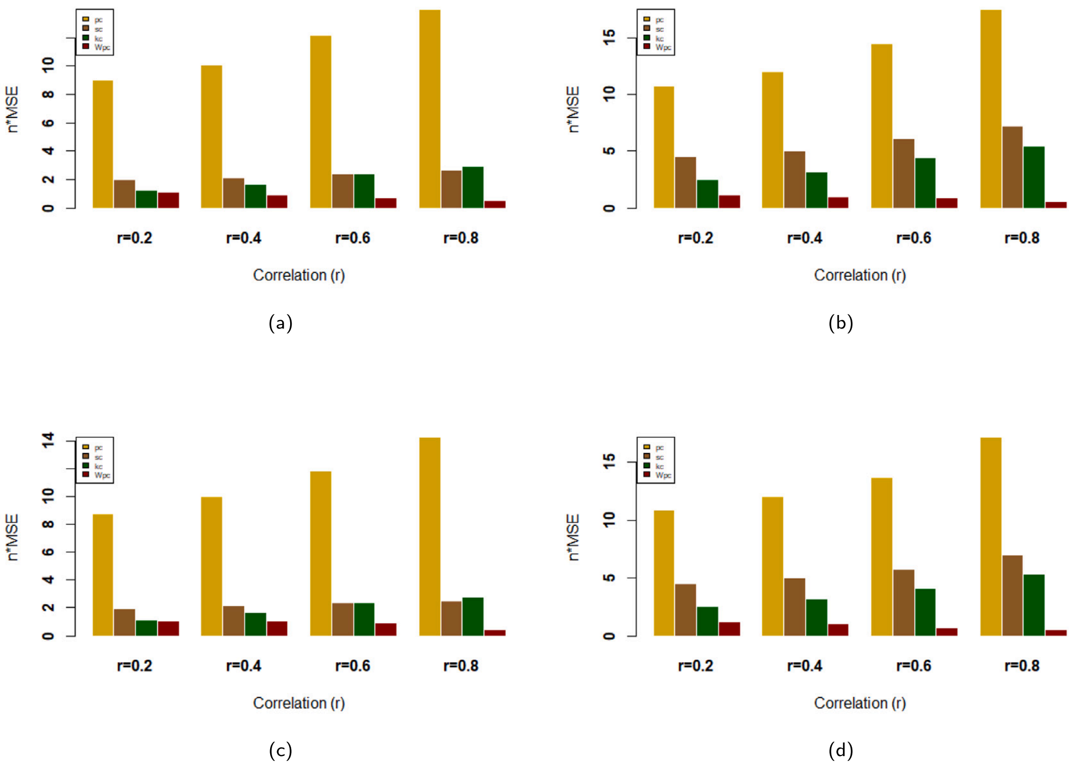
To demonstrate the benefits of the weighted correlation based feature selection algorithm over the unweighted algorithm, we modified the simulation example I in [61] to a more challenging setting to include contamination. Our goal is to evaluate the sure screening property of the proposed method, defined as the proportion of simulations where all the true features are successfully nominated. The details of the simulation procedure are as follows: Following [61] a response variable is generated according to the linear regression model:

$$y = 5x_1 + 5x_2 + 5x_3 + e \tag{18}$$

where $x_1, \dots, x_p$ are the $p$ independent variables and $e$ is the error term drawn from normal distribution with mean 0 and standard deviation of 1. For this simulation, a sample of $(x_1, \dots, x_p)$ having n samples generated from a multivariate normal distribution $N(0, \Sigma)$ with covariance structure $\Sigma = (\sigma ij)_{p\times p}$ where $\sigma_{ij} = 1$ for $i = 1, \dots, p$ and $\sigma_{ij} = \rho$ for $i \neq j$. We set $p = 100, 1000, n = 50, 70, \rho = 0, 0.2, 0.4, 0.6, 0.8$ respectively. To quantify the resistance of the proposed method against contaminated observations, we introduced contamination in the first $m = [\epsilon n]$ observations $(X_i, y_i)$, where $\epsilon = 0, 0.02, 0.05$ is the contamination rate. The first $n - m$ data is generated from the scaled distribution $t_3 \times \gamma$, with $\gamma = 8$ [63]. The scaled $t$ distribution allows us to generate heavy-tailed observations in the predictor and response variables simultaneously. Finally, we applied the SIS and the proposed WPCSSIS to the final dataset and tried to select all the significant variables using $d_n = n/log(n)$ to select the top features based on the absolute correlation values. For each scenario we repeated the simulation 50 times. Figures 2(a) to 2(e) displayed the percentage of SIS and WPCSIS that include the true model at increasing levels of correlation between true predictors with sample size $n = 50$ and $n = 70$. Different level of correlation was used to assess the effectiveness of the proposed method in its ability to recover the true model under the simultaneous problem of collinearity and outliers. Evidence from Fig. 2(a) to Fig. 2(e) show that the performance of the SIS drop sharply as correlation value increases at 0.02 and 0.05 contaminations. In contrast, the proposed WPCSIS has a near perfect recovery of (95–100)% when $p = 100$ and (75–85) % when $p = 1000$ even under high correlation between predictor variables. For $p = 100$, $n = 50$ and 70, at 0 contamination both SIS and WPCSIS perform very closely, achieving (95–100) % across all correlation values. For $p = 1000$, $n = 50$ and 70 at 0 contamination the WPCSIS show high robustness compared to SIS as the correlation increases, especially at $\rho = 0.6$ and 0.8. Overall in ultra high dimensional setting for example, where $p = 1000$ the WPCSIS show advantages over SIS especially under increasing contamination, correlation level and smaller sample sizes. Even for clean dataset the proposed weighted method provides stability when correlation is high and $p \gg n$ (see Fig. 4).

## 6. Application: feature selection based support vector machine approach

In recent years, feature selection algorithms have been widely adapted in many fields of science and engineering, such as bioinformatics [64,65], text classification [6], and glass classification [66]. A literature search shows successful application of the combined feature selection and machine learning algorithms in model prediction [67,68]. [69] suggested a correlation-based feature selection using the clustering technique. [70] used feature selection and a machine learning algorithm to estimate forest height. [71] applied a feature selection correlation-based algorithm with application in machine learning procedures to nuclear power plant data to perform feature selection and classification. [72] proposed machine learning-based feature selection methods to detect and predict chronic kidney diseases. Machine learning and feature selection algorithms to predict the performance of wastewater treatment plants were studied by [73,74]. Similarly, a binary classification feature selection using supervised and unsupervised methods was applied to select the most significant variables [75]. A rank correlation-based feature selection algorithm is only robust against heavy-tailed errors but not resistant to outliers in both responses and the predictor variables [76–78]. In this paper, we combined sure independence screening (SIS) and support vector machine algorithms for feature selection to investigate the predictive and interpretability of our proposed algorithm.



(a)

(b)

(c)

(d)

**Fig. 4.** Fig. 2. Barplot of n*MSE vs correlation (r) values for pc, sc, kc and wpc estimator (a) with $p = 100$, $n = 20$. And contamination at 0.2 (b) with $p = 100$, $n = 20$, and contamination at 0.1, (c) with $p = 1000$, $n = 20$ and contamination at 0.1 and (d) with $p = 1000$, $n = 20$ and contamination at 0.2 all obtained using 50 iterations.

To demonstrate the robustness of our method against outliers, we analyzed two real-life datasets: the glass data and Bardet–Biedl syndrome gene expression data. The glass dataset is available in the cellWise package in R software and contains 1920 features of wavelengths collected from 180 samples of archaeological glass [79]. Previous researchers, including [80,81], indicated that the glass dataset contains 38 outlying observations. Following [81], we removed columns with median absolute deviation (MAD) values equal to zero, then applied the remaining 1905 features to the weighted sure independence screening-based support vector machine algorithm to investigate its prediction performance in solving feature selection problems.

For the Bardet–Biedl syndrome gene expression data, available as the trim32 dataset in the R package Abbess, the data comprise 120 samples from a microarray experiment of mammalian eye tissue and 501 variables, with the first being the expression level of the TRIM32 gene and the remaining 500 being gene probes. To the best of our knowledge, no outlier procedure has been used to diagnose this dataset. Using the Mahalanobis outlier detection-based MRFCH, we identified 16 outliers out of 120 observations. Fig. 5 displays the boxplots of the original datasets: (a) glass data and (b) Bardet–Biedl syndrome gene expression data. The boxplots indicate a dense region and spread among the observations in the datasets. Similarly, we applied weighted sure independence screening and support vector machines to demonstrate the predictive performance of our newly proposed algorithm.

To achieve our objective, we used the screening and svm functions from the R software to obtain a weighted Pearson correlation sure independence screening based support vector machine (WPCSIS-SVM) algorithm for robust feature selection and prediction in high-dimensional settings.

Fig. 6 shows the bar plots for the sure independence screening-based support vector machine (SIS) and the weighted sure independence screening (WPCSIS)-based support vector machine. These plots are based on a model size of $d = n/log(n)d$, with the datasets split into 65/35 training and test sets for the glass dataset and 75/25 for the Bardet–Biedl syndrome gene expression data. The vertical axis includes the values of $R^2$, RMSE, and MAE, which are computed based on the formulas used by [82]. Notably, our proposed algorithm, which utilizes weights from the MRFCH algorithm, performs well on both datasets, especially regarding the $R^2$, values. The proposed algorithm achieved higher $R^2$, values, indicating a better model fit compared to the SIS based SVM method.

## 7. Conclusion

It is now evident that the cause of dimensionality, combined with the presence of outliers in both predictors and responses, can have a severe effect on modeling high-dimensional regression datasets. Several studies have been conducted to address the

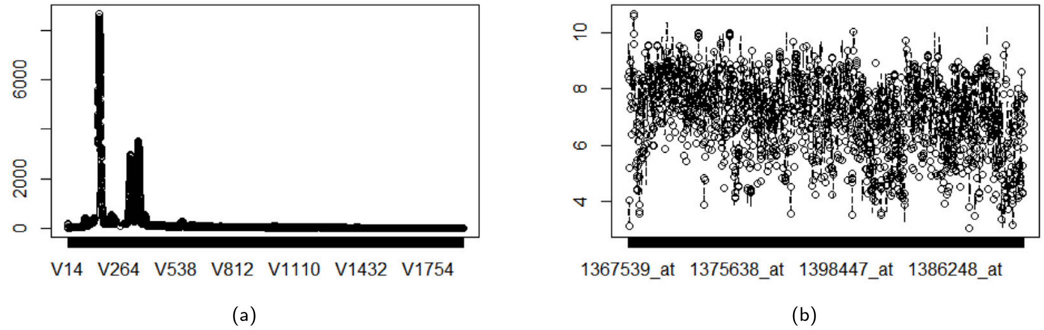


**Fig. 5.** Boxplot matrix for diagnostic of original (a) glass and (b) Bardet–Biedl syndrome gene expression datasets.

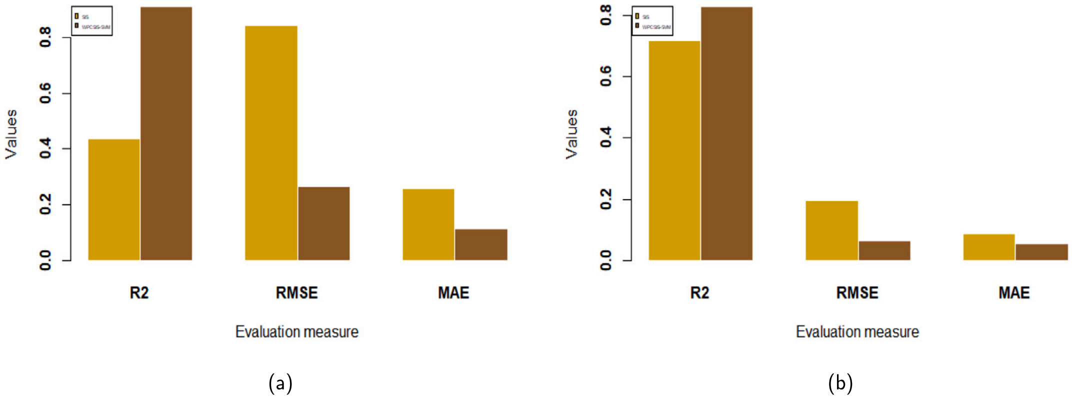


**Fig. 6.** Barplot of the R2, RMSE and MAE for SIS and WPCSIS-based SVM prediction algorithm (a) using glass dataset and (b) using Bardet–Biedl syndrome gene expression data.

problem of predictor outliers or response outliers in feature selection-based correlation problems. In statistical analysis, eliminating outlying observations or predictors may be considered detrimental. Firstly, this paper proposes a weighted Pearson correlation estimator to achieve efficiency and robustness. In this context, we examine the effect of outliers at different contamination levels via a simulation study. We evaluated four different correlation methods using MSE and bias as performance measures. Our results show that the proposed Wcp scales through with good performance both in terms of MSE and bias. Secondly, a weighted sure independence screening (WPCSIS)-based support vector machine to systematically handle the problem of response and predictor outliers in a high-dimensional dataset is developed. We applied two real life datasets to evaluate the prediction accuracy of the classical SIS-based support vector machine and WPCSIS-based support vector machine methods. These results also indicate that the new WPCSIS outperforms the SIS method in terms of the $R^2$, RMSE, and MAE values. Figures 2(a) to 2(e) present the percentage of SIS and WPCSIS that include the true model. It can be seen that the WPCSIS consistently outperforms SIS across varying levels of correlation, contamination, sample sizes, and dimensions of the dataset. Overall, the proposed Wpc and WPCSIS algorithms proved to be most effective compared to their competitors for all scenarios considered. Although our proposed algorithm performs well, it cannot be used for a categorical response, making it impossible to use on a disease classification dataset. This is a good future research topic. Certainly, judicious selection of weights plays an important role in achieving a weighted pc estimator and a weighted SIS with a much higher breakdown point. Similar to the SIS, the WPCSIS may fail to select some important features that are marginally independent of the response variable. Hence, it will be of interest to propose an iterative WPCSIS to address such a problem. The weighted correlation algorithm can be used with principal component analysis to achieve more robust outlier detection and classification. Further research direction will be to extend the result of this paper to a cell-wise approach, as this may improve efficiency and increase accuracy in high dimensional datasets. Weighted canonical correlation based feature selection is another area one can explore. Integrating weighted correlation based feature selection with penalized prediction models such as LASSO, least absolute deviation LASSO estimators, elastic net and adaptive LASSO to deal with the problem of predictors and response outliers is an important area for future research.

List of Abbreviation

| | |
|---|---|
| **pc** | Pearson correlation |
| **sc** | Spearman correlation |

**kc** Kendall correlation

**Wpc** Weighted Pearson correlation

**WPCSIS** Weighted Pearson correlation sure independence screening

**PB-SIS** point biserial sure independence screening

**SIS** Sure independence screening

**MRFCH** Modified reweighted fast consistent and high breakdown point estimator

**RMD** Robust Mahalanobis distance

**SVM** Support vector machine

**WPCSIS-SVM** Weighted Pearson correlation sure independence based support vector machine

**MAE** Mean absolute error

**MSE** Mean squared error

**RMSE** Root mean squared error

## Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Ishaq Abdullahi Baba reports financial support was provided by Tertiary Education Trust Fund (TetFund), Federal government of Nigeria. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data will be made available on request.

## References

[1] Chandrashekar G, Sahin F. A survey on feature selection methods. Comput Electr Eng 2014;40(1):16–28.
[2] Miao J, Niu L. A survey on feature selection. Procedia Comput Sci 2016;91:919–26.
[3] Alelyani S, Tang J, Liu H. Feature selection for clustering: A review. Data Clust 2018;29–60.
[4] Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, Liu H. Feature selection: A data perspective. ACM Comput Surv 2017;50(6):1–45.
[5] Chan JYL, Leow SMH, Bea KT, Cheng WK, Phoong SW, Hong ZW, Chen YL. Mitigating the multicollinearity problem and its machine learning approach: a review. Mathematics 2022;10(8):1283.
[6] Mamdouh Farghaly H, Abd El-Hafeez T. A high-quality feature selection method based on frequent and correlated items for text classification. Soft Comput 2023;27(16):11259–74.
[7] Zhao X, Cao Y, Zhang T, Li F. An improve feature selection algorithm for defect detection of glass bottles. Appl Acoust 2021;174:107794.
[8] Bentout S, Chekroun A, Kuniya T. Parameter estimation and prediction for coronavirus disease outbreak 2019 (COVID-19) in Algeria. AIMS Public Heal 2020;7(2):306.
[9] Agushaka JO, Akinola O, Ezugwu AE, Oyelade ON. A novel binary greater cane rat algorithm for feature selection. Results Control Optim 2023;11:100225.
[10] Jana DK, Bhunia P, Adhikary SD, Mishra A. Analyzing of salient features and classification of wine type based on quality through various neural network and support vector machine classifiers. Results Control Optim 2023;11:100219.
[11] Peng Y, Wu Z, Jiang J. A novel feature selection approach for biomedical data classification. J Biomed Inform 2010;43(1):15–23.
[12] Islam A, Seth S, Bhadra T, Mallik S, Roy A, Li A, Sarkar M. Feature selection, clustering and IoMT on biomedical engineering for COVID-19 pandemic: A comprehensive review. J Data Sci Intell Syst 2023.
[13] Liu R, Gillies DF. Overfitting in linear feature extraction for classification of high-dimensional image data. Pattern Recognit 2016;53:73–86.
[14] Khalid S, Khalil T, Nasreen S. A survey of feature selection and feature extraction techniques in machine learning. In: 2014 science and information conference. IEEE; 2014, p. 372–8.
[15] Hira ZM, Gillies DF. A review of feature selection and feature extraction methods applied on microarray data. Adv Bioinform 2015;2015(1):198363.
[16] Pudjihartono N, Fadason T, Kempa-Liehr AW, O'Sullivan JM. A review of feature selection methods for machine learning-based disease risk prediction. Front Bioinform 2022;2:927312.
[17] Larabi-Marie-Sainte S. Outlier detection based feature selection exploiting bio-inspired optimization algorithms. Appl Sci 2021;11(15):6769.
[18] Baba IA, Midi H, June LW, Ibragimove G. Penalized LAD-SCAD estimator based on robust wrapped correlation screening method for high dimensional models. Pertanika J Sci Technol 2021;29(2).
[19] Qu L, Wang X, Sun L. Variable screening for varying coefficient models with ultrahigh-dimensional survival data. Comput Statist Data Anal 2022;172:107498.
[20] Rahaman Khan MH, Nishat MNS. Variable selection for censored data with greedy algorithm based adaptive quantile regression models. Comm Statist Simulation Comput 2025;1–18.

[21] Abdullah MB. On a robust correlation coefficient. J R Stat Soc Ser D: Stat 1990;39(4):455–60.
[22] Maturi TA, Abdelfattah EH. A new weighted rank correlation. J Math Stat 2008;4(4).
[23] Zhong W. Robust sure independence screening for ultrahigh dimensional non-normal data. Acta Math Sin Engl Ser 2014;30(11):1885–96.
[24] Yu H, Hutson AD. Inferential procedures based on the weighted pearson correlation coefficient test statistic. J Appl Stat 2024;51(3):481–96.
[25] Raymaekers J, Rousseeuw PJ. Fast robust correlation for high-dimensional data. Technometrics 2021;63(2):184–98.
[26] Gideon RA, Hollister RA. A rank correlation coefficient resistant to outliers. J Amer Statist Assoc 1987;82(398):656–66.
[27] Shevlyakov GL. On robust estimation of a correlation coefficient. J Math Sci 1997;83:434–8.
[28] Niven EB, Deutsch CV. Calculating a robust correlation coefficient and quantifying its uncertainty. Comput Geosci 2012;40:1–9.
[29] Kondo M, Bezemer CP, Kamei Y, Hassan AE, Mizuno O. The impact of feature reduction techniques on defect prediction models. Empir Softw Eng 2019;24:1925–63.
[30] Saheed YK. Effective dimensionality reduction model with machine learning classification for microarray gene expression data. In: Data science for genomics. Elsevier; 2023, p. 153–64.
[31] Xie J, Sage M, Zhao YF. Feature selection and feature learning in machine learning applications for gas turbines: A review. Eng Appl Artif Intell 2023;117:105591.
[32] Zhang D, Chen Y, Chen Y, Ye S, Cai W, Jiang J, Xu Y, Zheng G, Chen M. [Retracted] heart disease prediction based on the embedded feature selection method and deep neural network. J Heal Eng 2021;2021(1):6260022.
[33] Zhou L, Wang H. A combined feature screening approach of random forest and filterbased methods for ultra-high dimensional data. Curr Bioinform 2022;17(4):344–57.
[34] Htun HH, Biehl M, Petkov N. Survey of feature selection and extraction techniques for stock market prediction. Financ Innov 2023;9(1):26.
[35] Wahid A, Khan DM, Hussain I, Khan SA, Khan Z. Unsupervised feature selection with robust data reconstruction (UFS-RDR) and outlier detection. Expert Syst Appl 2022;201:117008.
[36] Li G, Peng H, Zhang J, Zhu L. Robust rank correlation based screening. 2012.
[37] Li R, Zhong W, Zhu L. Feature screening via distance correlation learning. J Amer Statist Assoc 2012;107(499):1129–39.
[38] Wang M, Song L, Tian G-l. SCAD-penalized least absolute deviation regression in high-dimensional models. Comm Statist Theory Methods 2015;44(12):2452–72.
[39] Pearson K. Notes on the history of correlation. Biometrika 1920;13(1):25–45.
[40] Zheng G, Modarres R. A robust estimate of the correlation coefficient for bivariate normal distribution using ranked set sampling. J Statist Plann Inference 2006;136(1):298–309.
[41] Ma R, Xu W, Wang Q, Chen W. Robustness analysis of three classical correlation coefficients under contaminated Gaussian model. Signal Process 2014;104:51–8.
[42] Sadeghi B. Chatterjee correlation coefficient: A robust alternative for classic correlation methods in geochemical studies-(including ''TripleCpy'' Python package). Ore Geol Rev 2022;146:104954.
[43] Devlin SJ, Gnanadesikan R, Kettenring JR. Robust estimation and outlier detection with correlation coefficients. Biometrika 1975;62(3):531–45.
[44] Shevlyakov GL, Vilchevski NO. Robustness in data analysis: criteria and methods. Walter de Gruyter; 2011.
[45] Yuen KV, Ortiz GA. Outlier detection and robust regression for correlated data. Comput Methods Appl Mech Engrg 2017;313:632–46.
[46] Wang H, Li G, Jiang G. Robust regression shrinkage and consistent variable selection through the LAD-Lasso. J Bus Econom Statist 2007;25(3):347–55.
[47] Giloni A, Simonoff JS, Sengupta B. Robust weighted LAD regression. Comput Statist Data Anal 2006;50(11):3124–40.
[48] Gao X, Feng Y. Penalized weighted least absolute deviation regression. Stat Interface 2017;11(1):79–89.
[49] Jiang Y, Wang Y, Zhang J, Xie B, Liao J, Liao W. Outlier detection and robust variable selection via the penalized weighted LAD-LASSO method. J Appl Stat 2021;48(2):234–46.
[50] Arslan O. Weighted LAD-LASSO method for robust parameter estimation and variable selection in regression. Comput Statist Data Anal 2012;56(6):1952–65.
[51] Hadi AS, Simonoff JS. Procedures for the identification of multiple outliers in linear models. J Amer Statist Assoc 1993;88(424):1264–72.
[52] Jesús Sánchez M, Pena D. The identification of multiple outliers in ARIMA models. Comm Statist Theory Methods 2003;32(6):1265–87.
[53] Docherty P, Gray R, Mansell E. Reducing the effect of outlying data on the identification of insulinaemic pharmacokinetic parameters with an adapted Gauss-Newton approach. IFAC Proc Vol 2014;47(3):5635–40.
[54] Jesiļevska S. Iterative method for reducing the impact of outlying data points: Ensuring data quality. Stat J IAOS 2016;32(2):257–63.
[55] Khan DM, Ali M, Ahmad Z, Manzoor S, Hussain S. A new efficient redescending M-estimator for robust fitting of linear regression models in the presence of outliers. Math Probl Eng 2021;2021(1):3090537.
[56] Hubert M, Rousseeuw PJ. Robust regression with both continuous and binary regressors. J Statist Plann Inference 1997;57(1):153–63.
[57] Yin L, Lv L, Wang D, Qu Y, Chen H, Deng W. Spectral clustering approach with K-nearest neighbor and weighted mahalanobis distance for data mining. Electronics 2023;12(15):3284.
[58] Billor N, Hadi AS, Velleman PF. BACON: blocked adaptive computationally efficient outlier nominators. Comput Statist Data Anal 2000;34(3):279–98.
[59] Wang X, Cao Z, Liu C, Wang M. Group selection via adjusted weighted least absolute deviation regression. J Comput Appl Math 2020;378:112924.
[60] Baba IA, Midi H, June LW, Ibragimov G. A modified reweighted fast consistent and high-breakdown estimator for high-dimensional datasets. Decis Anal J 2024;10:100424.
[61] Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. J R Stat Soc Ser B Stat Methodol 2008;70(5):849–911.
[62] Li Z, Yang Y, Li L, Wang D. A weighted pearson correlation coefficient based multi-fault comprehensive diagnosis for battery circuits. J Energy Storage 2023;60:106584.
[63] Su P, Tarr G, Muller S, Wang S. CR-Lasso: Robust cellwise regularized sparse regression. Comput Statist Data Anal 2024;197:107971.
[64] Chen Z, Pang M, Zhao Z, Li S, Miao R, Zhang Y, Feng X, Feng X, Zhang Y, Duan M, et al. Feature selection may improve deep neural networks for the bioinformatics problems. Bioinformatics 2020;36(5):1542–52.
[65] Li L, Algabri YA, Liu ZP. Identifying diagnostic biomarkers of breast cancer based on gene expression data and ensemble feature selection. Curr Bioinform 2023;18(3):232–46.
[66] Li Z, Long Z, Lei S, Zhang T, Liu X, Kuang D. Predicting the glass formation of metallic glasses using machine learning approaches. Comput Mater Sci 2021;197:110656.
[67] Chong J, Tjurin P, Niemelä M, Jämsä T, Farrahi V. Machine-learning models for activity class prediction: A comparative study of feature selection and classification algorithms. Gait Posture 2021;89:45–53.
[68] Sivaranjani S, Ananya S, Aravinth J, Karthika R. Diabetes prediction using machine learning algorithms with feature selection and dimensionality reduction. In: 2021 7th international conference on advanced computing and communication systems, vol. 1, IEEE; 2021, p. 141–6.
[69] Chormunge S, Jena S. Correlation based feature selection with clustering for high dimensional data. J Electr Syst Inf Technol 2018;5(3):542–9.
[70] Arjasakusuma S, Swahyu Kusuma S, Phinn S. Evaluating variable selection and machine learning algorithms for estimating forest heights by combining lidar and hyperspectral data. ISPRS Int J Geo-Inform 2020;9(9):507.
[71] He Y, Yu H, Yu R, Song J, Lian H, He J, Yuan J. A correlation-based feature selection algorithm for operating data of nuclear power plants. Sci Technol Nucl Install 2021;2021(1):9994340.

[72] Ebiaredoh-Mienye SA, Swart TG, Esenogho E, Mienye ID. A machine learning method with filter-based feature selection for improved prediction of chronic kidney disease. Bioengineering 2022;9(8):350.

[73] El-Rawy M, Abd-Ellah MK, Fathi H, Ahmed AKA. Forecasting effluent and performance of wastewater treatment plant using different machine learning techniques. J Water Process Eng 2021;44:102380.

[74] Ekinci E, Özbay B, Omurca Sİ, Sayın FE, Özbay İ. Application of machine learning algorithms and feature selection methods for better prediction of sludge production in a real advanced biological wastewater treatment plant. J Environ Manag 2023;348:119448.

[75] Hallajian B, Motameni H, Akbari E. Ensemble feature selection using distance-based supervised and unsupervised methods in binary classification. Expert Syst Appl 2022;200:116794.

[76] Delaigle A, Hall P. Effect of heavy tails on ultra high dimensional variable ranking methods. Statist Sinica 2012;909–32.

[77] Chen X, Chen X, Wang H. Robust feature screening for ultra-high dimensional right censored data via distance correlation. Comput Statist Data Anal 2018;119:118–38.

[78] Giordano F, Milito S, Parrella M. A nonparametric procedure for linear and nonlinear variable screening. J Nonparametr Stat 2022;34(4):859–94.

[79] Smucler E, Yohai VJ. Robust and sparse estimators for linear regression models. Comput Statist Data Anal 2017;111:116–30.

[80] Hubert M, Rousseeuw PJ, Vanden Branden K. ROBPCA: a new approach to robust principal component analysis. Technometrics 2005;47(1):64–79.

[81] Filzmoser P, Maronna R, Werner M. Outlier identification in high dimensions. Comput Statist Data Anal 2008;52(3):1694–711.

[82] Liu J, Wang Y, Fu C, Guo J, Yu Q. A robust regression based on weighted LSSVM and penalized trimmed squares. Chaos Solitons Fractals 2016;89:328–34.