CAUCHY – Jurnal Matematika Murni dan Aplikasi
Volume 11 (2) (2026), Pages 1562-1579
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Robust PCA Using MCD and MM Estimators in MARS: A Simulation Study

Uswatun Hasanah*, Solimun, and Atiek Iriany

Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Brawijaya, Malang, Indonesia

Abstract

Multivariate Adaptive Regression Splines (MARS) models nonlinear relationships through adaptive basis functions but remain sensitive to outliers in the predictor variables. Existing robust extensions of MARS primarily address response outliers, while the few studies integrating Robust Principal Component Analysis (RPCA) with MARS use RPCA only for dimension reduction without comparing robust estimators. This study evaluates RPCA as a robust predictor transformation and systematically compares two robust covariance estimators—the Minimum Covariance Determinant (MCD) and the MM-estimator–within the RPCA-MARS framework. A full factorial simulation with 100 replications per condition covered 45 conditions: five sample sizes ($n = 50, 100, 200, 500, 1000$), three outlier proportions (5%, 10%, 25%), and three MARS interaction levels $(1, 2, 3)$ with eight predictor variables. Outliers were extreme values in a specified proportion of predictor observations. Performance was measured by Root Mean Square Error (RMSE). For analysis, the 45 conditions were collapsed into 15 scenarios by selecting the interaction level with the minimum RMSE for each sample size and outlier proportion. The MM estimator outperformed the MCD estimator in 8 of 15 scenarios, achieving lower RMSE under moderate-to-high outlier contamination ($10\% - 25\%$) with moderate sample sizes (n = 100–500). MCD performed better in the remaining 7 scenarios: under low contamination (5%) at $n \leq 200$ and $n \geq 1000$, and across all contamination levels at $n = 1000$. MCD showed higher variability at small samples with moderate-to-high contamination, while MM produced tighter confidence intervals and lower standard deviations. Within the RPCA-MARS framework, MM is recommended for moderately sized, highly contaminated data, while MCD is preferable under low contamination or in large-scale settings.

Keywords: Minimum Covariance Determinant; MM-estimator; Multivariate Adaptive; Outliers; Regression Splines; Robust Principal Component Analysis; Simulation

1. Introduction

Nonparametric regression methods, such as Multivariate Adaptive Regression Splines (MARS), are widely used for modelling complex nonlinear relationships and high-order interactions, yet their sensitivity to outliers in the predictor space remains an under-addressed limitation. To address these complex structures, the primary goal of these nonparametric methods shifts toward estimating a smoothing function that provides a more refined representation of the data, rather

*Corresponding author. E-mail: uswatunhasanahus10@gmail.com

Submitted: March 12, 2026 Reviewed: April 22, 2026 Accepted: June 18, 2026
DOI: https://doi.org/10.18860/cauchy.v11i2.41547

than focusing on estimating rigid regression coefficients [1]. This smoothing function helps capture the underlying trend between the dependent variable and the independent variables without assuming a specific functional form.

Multivariate Adaptive Regression Splines (MARS), introduced by Friedman (1991), works by splitting the predictor space into regions at knot points and builds piecewise linear spline functions. It uses a two-step procedure, that is, a forward step that adds basis functions and a backward step that prunes them [2]. Friedman demonstrated that MARS works especially well in moderate dimensional settings (typically $3 \leq p \leq 20$) with moderate sample sizes ($50 \leq n \leq 1000$), where it provides efficient automatic variable selection and interaction detection without excessive computation [2].

From a robustness perspective, MARS is almost immune to outliers in the response variable Y, so removing outliers in Y is generally unnecessary when using MARS [3]. By contrast, MARS remains vulnerable to outliers in the predictor variable (X). These predictor outliers can severely distort knot placement and basis function selection during MARS's forward stepwise procedure. Because MARS builds piecewise linear basis functions adaptively, an extreme predictor variable (X) value may force a knot at an unrepresentative location and propagate bias across the fitted surface.

To address this limitation, exploring the integration of robust estimation methods into the MARS framework is highly relevant. Classical PCA relies on decomposing the empirical covariance matrix by maximizing variance; however, both classical variance and covariance estimators are highly sensitive to anomalous observations [4]. Consequently, in the presence of predictor outliers commonly encountered in real-world data, classical PCA can produce distorted principal components that are heavily attracted toward outlying points. Using these unreliable components as inputs can subsequently weaken or mislead the performance of the MARS model.

As an alternative, Robust Principal Component Analysis (RPCA) replaces classical location and scatter estimators with robust counterparts, yielding principal components that are resistant to outliers. The initial idea behind RPCA is to combine Projection Pursuit with robust covariance estimation [4]. Directly applying robust regression within the MARS fitting stage primarily addresses contamination in the response variable (Y). However, it does not alter the candidate knot pool, which is drawn directly from the observed predictor values. Consequently, leverage points in X can still enter this pool and severely distort the adaptive partitioning mechanism. To overcome this structural vulnerability, using RPCA as a preprocessing step offers a more principled separation of the outlier-robust transformation from the nonlinear modeling.

Using RPCA as a preprocessing step provides a principled separation of the outlier-robust transformation from the nonlinear modeling; however, its advantage over alternative robust preprocessing strategies requires empirical validation. Here, RPCA is not used mainly to reduce dimensionality. Instead, it serves as a robust orthogonal transformation tool. RPCA aims to mitigate the effect of outliers in the predictor space by building more robust principal components. These components capture most of the variance in the data while reducing the influence of multivariate leverage points.

This approach is designed to provide MARS with a more resilient transformed coordinate system, aiming to stabilize its step-by-step knot selection and basis function estimation against extreme values. This two-phase approach—first, a robust transformation of the predictor space, and second, adaptive nonlinear modeling—supports this study to use RPCA as a preprocessing step, rather than using another robust estimator during the model fitting.

Two robust estimators are employed in this study to estimate location and covariance with the RPCA framework, that is, the Minimum Covariance Determinant (MCD) and the MM-estimator [5]. These two estimators are selected for their complementary properties. The MCD estimator delivers an affine equivariant covariance estimate and highly robust estimators for multivariate location and scatter. It is defined as the mean and covariance matrix of the h subset of observations with the smallest covariance determinant, achieving a breakdown value up

to 50% [6]. The MM-estimator, introduced by Yohai (1987), is defined through a three-stage procedure and possesses a high breakdown point of 50% and high efficiency of 95% under a normal model [7]. Outliers receive reduced weights through the re-descending ψ-function, so they have limited influence on the robust covariance matrix. Comparing these two estimators within the RPCA-MARS framework, therefore, allows us to assess the trade-off between breakdown robustness and statistical efficiency across sample sizes and varying contamination scenarios.

The integration of Robust Principal Component Analysis (RPCA) with Multivariate Adaptive Regression Splines (MARS) is a potentially promising approach for providing outlier-resistant predictor transformation prior to nonlinear modeling, though its advantage over standard MARS requires further empirical validation through baseline comparisons.

A previous study explored the integration of Robust Principal Component Analysis–Multivariate Adaptive Regression Splines (RPCA-MARS) to predict the total acid number (TAN) and total base number (TBN) in crude oil samples [8]. The study demonstrated that the combined RPCA-MARS approach outperformed Principal Component Regression (PCR) and Partial Least Squares Regression (PLS-R) based on the GCV-R2 and MSE. In their methodology, they utilized an RPCA method developed from the projection pursuit (PP) framework proposed by Croux and Ruiz-Gazen [9], which was primarily driven by the need for dimensionality reduction due to the large number of variables in crude oil data. However, it treated RPCA as a monolithic entity, failing to specify or compare the underlying robust covariance estimators.

This study addresses this gap by systematically comparing the MCD and MM estimators during the RPCA transformation stage. Clarifying this comparison is vital to improving the RPCA-MARS framework because MCD (hard-trimming) and MM (smooth downweighting) generate structurally distinct orthogonal scores, directly impacting how MARS adaptively positions its knots. By benchmarking these estimators across varying outlier proportions and sample sizes, this study transforms the framework into a data-adaptive pipeline, providing empirical guidance on when to prioritize breakdown robustness (MCD) versus statistical efficiency (MM).

To achieve this, this study aims to develop the integrated RPCA-MCD and RPCA-MM methods within the MARS framework and to compare their performance using RMSE as a metric through simulation studies with varying levels of outliers and sample sizes.

2. Methods

This study uses a quantitative research approach with a simulation study. Its goal is to develop and compare the RPCA-MCD-MARS and RPCA-MM-MARS methods. The proposed method combines the ability of Robust Principal Component Analysis (RPCA) to generate principal components that are robust to outliers with MCD and MM estimators, which are used as input variables in Multivariate Adaptive Regression Splines. This aims to adaptively model nonlinear relationships and complex interactions among predictor variables.

2.1. Design of the Simulation Study

This study comprises 100 simulation replications. In each replication, data are generated according to the specified data-generating process, outliers are introduced based on the scenarios outlined in Table 1, and models are evaluated using RMSE, along with its corresponding standard deviation (SD), and 95% confidence intervals (CI). The reported results represent the averages for all 100 replications. The simulation study was designed to compare the performance of RPCA-MCD-MARS and RPCA-MM-MARS across varying levels of outlier contamination, sample size, and MARS interaction order.

2.1.1. Data Generation Process

The data used consists of simulation datasets generated from a uniform distribution. The uniform distribution was chosen for its flexibility in generating data without assuming normality, its

alignment with the nonparametric context of MARS, and its heterogeneous-scale representation. This study includes one response variable and eight predictor variables, all of which are created with nonlinear relationships. Five predictor variables were generated independently from uniform distributions with different ranges, namely

$$X_j \sim U(a_j, b_j), \quad j = 1, 2, \ldots, 5$$

with parameter values $a_1 = 10, b_1 = 100; a_2 = 5, b_2 = 50; a_3 = 20, b_3 = 80; a_4 = 15, b_4 = 70$; and $a_5 = 8, b_5 = 60$.

Three additional predictor variables with moderate multicollinearity were created using the parameter $\lambda = 0.55$

$$X_{i6} = \lambda X_{i1} + \lambda X_{i2} + (1 - 2\lambda)U_{i6} + \varepsilon_{i6}$$
$$X_{i7} = \lambda X_{i3} + \lambda X_{i4} + (1 - 2\lambda)U_{i7} + \varepsilon_{i7}$$
$$X_{i8} = \lambda X_{i2} + \lambda X_{i5} + (1 - 2\lambda)U_{i8} + \varepsilon_{i8}$$

with $U_{i6} \sim U(10, 80)$, $U_{i7} \sim U(15, 75)$, $U_{i8} \sim U(12, 65)$, and ε_{i6}, ε_{i7}, $\varepsilon_{i8} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, 3^2)$. The response variable was generated using a nonlinear model:

$$Y = \sum_{j=1}^{8} \beta_j X_j + 0.4X_1^2 - 0.3X_2^2 + 0.6X_1X_3 - 0.4X_4X_5 + \epsilon,$$

where $\boldsymbol{\beta} = (2.5, -1.8, 3.2, 0.8, -2.1, 1.5, -0.9, 1.2)^T$, and $\epsilon \sim N(0, 25)$ with $\sigma_\epsilon = 5$

2.1.2. Outlier Contamination Mechanism

Outliers were introduced only in the predictor variables, not in the response variable. This design choice is based on two considerations: (1) MARS is documented as robust to contamination in Y, making response-variable outliers poor for method evaluation [2, 3]. (2) Our evaluation centers on comparing the relative predictive performance of the MCD and MM estimators when utilized as RPCA preprocessing steps for MARS under varying levels of predictor contamination. For each simulated data set, the number of contaminated observations is determined as:

$$n_{\text{out}} = \lfloor n \cdot r \rfloor$$

where n is the number of data sets, and r is the proportion of outliers.

The contaminated observations were randomly selected, and for each selected observation, between two and five predictors were randomly chosen from the eight available predictors. The values of these variables were then modified by adding or subtracting large random numbers drawn from $\delta_{ij} \sim U(40, 80)$, with the direction of the perturbation randomly assigned. Thus, the contaminated predictor value becomes

$$x_{ij}^* = x_{ij} + s_{ij}\delta_{ij}$$

where:

- x_{ij}^*: the contaminated value of the j-th predictor for the i-th observation
- x_{ij}: the original value of the j-th predictor for the i-th observation
- $s_{ij} \in \{-1, 1\}$: a random sign indicator variable, where $P(s_{ij} = 1) = 0.5$ and $P(s_{ij} = -1) = 0.5$, ensuring that the direction of the contamination is assigned completely at random
- $\delta_{ij} \sim U(40, 80)$: the magnitude of the random perturbation drawn from a uniform distribution between 40 and 80.

This conditional formulation guarantees that the direction of the contamination for each chosen predictor is assigned with an equal chance without assuming a specific data direction. The response variable Y was not contaminated, allowing us to clearly evaluate and compare how the MCD and MM estimators extreme predictor values (high-leverage points) within the MARS framework.

2.1.3. Simulation Scenarios

In this simulation, the number of predictor variables was fixed at $p = 8$. This choice of p yields a sufficiently complex multivariate setting to evaluate the performance of the proposed RPCA-MARS framework under various data conditions. The simulation parameters were systematically varied on three main factors, resulting in 45 distinct experimental scenarios as shown in Table 1. All computations were performed in RStudio software. The simulation scenarios used are as follows:

Table 1: Simulation study scenarios

Factor	**Level**
Sample Size (n)	50, 100, 200, 500, 1000
Outlier Proportions (r)	5%, 10%, 25%
Maximum Interaction MARS	1, 2, 3

First-order interactions imply that the model uses only basis functions that contain one predictor variable. Second-order interactions mean that the model includes basis functions of the form of a product of two basis functions involving two different predictor variables. Meanwhile, third-order interactions mean that the model includes basis functions of the form a product of three basis functions involving three different predictor variables. Note that the baseline model within each replication is tuned via minimum GCV, while the final optimal interaction level for each aggregated scenario is determined via minimum mean RMSE across all replications.

2.2. Multivariate Adaptive Regression Splines

Multivariate Adaptive Regression Spline (MARS) is a nonparametric regression method that builds complex, nonlinear models relating response and predictor variables using a combination of piecewise linear basis functions [2]. In the proposed framework, the MARS algorithm is applied directly to the robust principal component scores obtained from the RPCA stage, enabling the model to handle both outliers and nonlinear relationships simultaneously. Therefore, the input variables for MARS are the transformed scores, denoted as $\mathbf{z}_i$, for each observation i $(i = 1, 2, \ldots, n)$, the robust principal component score vector is $z_i = (z_{i1}, z_{i2}, \ldots, z_{ip})^T = \left(PC_{i1}^{\text{rob}}, PC_{i2}^{\text{rob}}, \ldots, PC_{ip}^{\text{rob}}\right)^T$, and $z_{ij} = PC_{ij}^{\text{rob}}$ represents the j-th robust principal component score for the i-th observation. Generally, the MARS model is defined as follows [10]:

$$y_i = \alpha_0 + \sum_{m=1}^{M} \alpha_m BF_m^q(\mathbf{z}_i) + \varepsilon_i, \quad i = 1, 2, ..., n$$

with

$$BF_m^q(\mathbf{z}) = \prod_{k=1}^{K_m} \left[s_{k,m}(z_{v(k,m)} - t_{k,m}) \right]_+, \quad k = 1, 2, ..., K_m \tag{1}$$

where:

- $\mathbf{z}_i = (z_{i1}, z_{i2}, ..., z_{ip})^T$: the vector of the transformed predictor variable (robust principal component scores) for the i-th observation
- p: the number of transformed predictor variables
- ε_i: error distributed normally with zero expectation and constant variance
- M: the number of non-constant basis functions
- α_0: constant coefficient of the basis function
- α_m: coefficient for the m-th basis function
- n: the number of observations
- q: the degree of the polynomial in each spline segment

- K_m: maximum interaction in the m-th basis function
- $z_{v(k,m)}$: the v-th transformed predictor variable selected in the k-th and m-th subregion
- $t_{k,m}$: knot point value of the transformed predictor variable $z_{v(k,m)}$
- $s_{k,m} \in \{+1, -1\}$: the sign indicator (positive or negative) for the k-th factor

Several parameters must be specified when constructing a MARS model [11]. These parameters collectively govern model complexity and generalization performance.

a. Knot
A knot is a point within a predictor variable where the slope of the regression line changes. Consequently, the knots function as boundaries that delineate specific segments within a dataset, marking the termination of one local region and the initiation of the next.
To regulate these segments, the Minimum Observation (MO) is used to define the minimum data span between knots, which controls model complexity. In the MARS framework, this minimum span near the data boundaries is mathematically determined based on the number of predictors (p) and a tuning parameter (α) via the following formula [2]:

$$L(\alpha) = 3 - \log_2(\frac{\alpha}{p}) \tag{2}$$

where $\alpha = 0.05$ is the significance level and p is the number of predictor variables. This formula ensures that the knots are not placed too close together, thereby helping to control the model's complexity and reduce the risk of overfitting

b. Basis Function
A basis Function (BF) represents the interval between successive knots. The recommended range for the maximum number of basis functions is 2 to 4 times the number of predictor variables, that is $2p \leq BF \leq 4p$
In this study, the BF parameter is calibrated within this range using p predictor variables derived from robust principal component scores, and the optimal value is selected based on the minimum GCV criterion

c. Maximum Interaction
Maximum Interaction (MI) is the product of the cross-products between correlated variables, corresponding to K_m in Eq. (1). It specifies the maximum order of interaction permitted within a single basis function, which is typically restricted to 1, 2, or 3. If the interaction exceeds 3, the resulting model becomes increasingly complex and difficult to interpret.
In this study, all three levels of MI are evaluated as part of the simulation scenarios (Table 1), where GCV is used to select the best basis functions within each run, and the results are aggregated into primary scenarios based on the minimum mean RMSE.

The selection of the best MARS model is based on the Generalized Cross Validation (GCV) criterion. The GCV criterion is used to eliminate redundant basis functions (or lack of fit) and to select the optimal basis functions in the MARS algorithm [10]. To account for model complexity, the number of effective parameters for a baseline model is first defined as:

$$P(M) = \text{trace}\left(B(B^T B)^{-1} B^T\right) + 1$$

To prevent overfitting during knot selection, an adjusted penalty term is introduced by incorporating a penalty cost (d) for each basis function (M) added to the model:

$$P(\tilde{M}) = P(M) + dM$$

By defining this adjusted penalty term prior, the GCV criterion for evaluating the MARS model is formulated as follows:

$$\text{GCV}(M) = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[y_i - \hat{f}_M(\mathbf{z}_i)\right]^2}{\left[1 - \frac{P(\tilde{M})}{n}\right]^2} \tag{3}$$

where:

- y_i: vector of the i-th observed response variable
- $\hat{f}_M(\mathbf{z}_i)$: MARS model estimate based on the basis function applied to the transformed variables
- $GCV(M)$: generalized cross-validation criteria for selecting MARS basis functions
- $\mathbf{z}_i$: vector of transformed predictor variables for the i-th observation $\mathbf{z}_i = (z_{i1}, z_{i2}, ..., z_{ip})^T$, $i = 1, 2, ..., n$
- n: the number of observations
- M: the number of nonconstant basis functions
- B: basis function of dimension $M \times n$
- $P(M)$: penalty measure for complex functions
- d: smoothing parameter, with range $2 \leq d \leq 4$

2.3. Outlier Detection

One of the problems frequently encountered in secondary data is the presence of outliers. An outlier is a data point that deviates from the general pattern of the relationship between the predictor variable and the response variable and does not follow the distribution pattern of the other data. There are two types of outliers: outliers in the predictor variable and outliers in the response variable [12]. Outliers in predictor variables can be detected using the Mahalanobis distance. In its calculation, the Mahalanobis distance uses a covariance matrix and a mean vector. The following is the formula for calculating the Mahalanobis distance [13].

$$d_i^2 = (x_i - \mu)' \Sigma^{-1} (x_i - \mu), \quad i = 1, 2, 3, \ldots, n$$

with,

- x_i: a random vector of predictor variables,
- μ: mean vector,
- Σ: covariance matrix.

Outlier detection and testing must be performed whenever outliers occur in the data as part of the analysis procedure and as the primary basis for using robust methods such as RPCA with MCD and MM estimators, which are robust against outliers.

2.4. Classic Principal Component Analysis (CPCA)

PCA is a dimension-reduction method that transforms correlated predictor variables into a set of orthogonal (uncorrelated) components, called principal components. Suppose

$$X \in \mathbb{R}^{n \times p}$$

is data matrix with n observations on p variables. The sample mean vector $\bar{x}$ and the sample covariance matrix S are defined as [14] :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i;$$

and

$$S = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T, \quad i = 1, 2, \ldots, n$$

where:

- x_i: observation vector i-th sized p x 1

- $\bar{x}$: sample mean vector
- S: unbiased sample estimator of Σ

The covariance matrix S is decomposed into eigenvalues and eigenvectors as follows [15]:

$$S = A\Lambda A^T$$

where:

- λ_i: eigenvalue of the covariance matrix S
- a_i: eigenvector corresponding to λ_i
- A: matrix whose columns are eigenvectors, namely $(a_1, a_2, \ldots, a_p)$
- $\Lambda = diag(\lambda_1, \lambda_2, ...\lambda_p)$: diagonal matrix containing eigenvalues ordered as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$

Eigenvectors are obtained from the eigen decomposition of the covariance matrix by solving the equation:

$$|S - \lambda I| = 0$$

For each eigenvalue λ_i, the sample eigenvectors $\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_p$ are obtained by solving the equation:

$$(S - \lambda_i I)a_i = 0,$$

with the normalization condition

$$a_i^T a_i = 1$$

and the orthogonality property

$$a_i^T a_j = 0, \quad i \neq j$$

The principal component scores are defined as:

$$\mathrm{PC}_{ij} = a_j^T(x_i - \bar{x}), \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, p$$

with:

- x_i: i-th observation vector
- $\bar{x}$: sample mean vector
- a_j: j-th eigenvector
- PC_{ij}: j-th principal component score for the i-th observation

2.5. Robust Principal Component Analysis (RPCA)

RPCA is an extension of standard PCA that replaces traditional location and scatter estimates with robust estimators. In CPCA, principal components are obtained from the eigen-decomposition of the sample covariance matrix, which is sensitive to outliers. To overcome this limitation, this study employs RPCA, where the PCs are derived from a robust scatter matrix, such as MCD or MM-estimator, to mitigate the effect of outliers. Next, the principal component scores were extracted and incorporated into the proposed framework—MARS—as a new set of input variables (predictors). Therefore, RPCA uses robust location estimators and robust scatter matrices instead of classical ones. Suppose there are n observations on p variables, expressed as $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T, i = 1, 2, \ldots, n$

In RPCA, the initial step involves estimating the robust location vector $\hat{T}_{\text{rob}}$ $(p \times 1)$ and the robust scatter matrix $\hat{C}_{\text{rob}}$ $(p \times p)$. The robust location estimator is represented as

$$\hat{T}_{\text{rob}} = (\hat{T}_1, \hat{T}_2, \ldots, \hat{T}_p)^T$$

and the robust scatter matrix is represented as

$$\hat{C}_{\text{rob}} = \begin{bmatrix} \hat{c}_{11} & \hat{c}_{12} & \dots & \hat{c}_{1p} \\ \hat{c}_{21} & \hat{c}_{22} & \dots & \hat{c}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{c}_{p1} & \hat{c}_{p2} & \dots & \hat{c}_{pp} \end{bmatrix}$$

The estimation of the robust parameters $\hat{T}_{\text{rob}}$ and $\hat{C}_{\text{rob}}$ is performed using two methods: specifically, Minimum Covariance Determinant (MCD) and MM-estimator.

2.5.1. Minimum Covariance Determinant

The Minimum Covariance Determinant (MCD) is a robust covariance estimator that selects a subset of size h from n observations, then minimizes the determinant of the sample covariance matrix of that subset. The MCD estimator is defined as follows [6]:

$$H^* = \arg\min \det(S_H)$$

with

$$S_H = \frac{1}{h} \sum_{i \in H} (x_i - \bar{x}_H)(x_i - \bar{x}_H)^T$$

and

$$\bar{x}_H = \frac{1}{h} \sum_{i \in H} x_i$$

is the average of the subset. The value of h satisfies $\lfloor (n + p + 1)/2 \rfloor \leq h \leq n$

The optimal subset H^* is then used to estimate the robust location T and robust scatter C employing the following formula (Croux & Haesbroeck, 1999):

$$\hat{T}_{\text{MCD}} = \frac{1}{h} \sum_{i \in H^*} x_i$$

and

$$\hat{C}_{\text{MCD}} = c_{(p,h)} \frac{1}{h} \sum_{i \in H^*} (x_i - \hat{T}_{\text{MCD}})(x_i - \hat{T}_{\text{MCD}})^T$$

with

$$c_{(p,h)} = \frac{\alpha}{F_{\chi^2_{p+2}}(\chi^2_{p,\alpha})} \tag{4}$$

where:

- $\alpha = h/n$: represents the proportion of the selected subset of observations used to minimize the determinant
- h: subset size, $\lfloor (n + p + 1)/2 \rfloor \leq h \leq n$
- $F^2_{\chi^2_{p+2}}$: cumulative distribution function (CDF) of the chi-square with degrees of freedom $p + 2$
- $\chi^2_{p,\alpha}$: the α quantile of the χ^2_p distribution with p degrees of freedom
- i: observation index in a subset
- p: the number of variables

Note that α in Eq. (4) denotes the subset proportion h/n, which is distinct from the significance level $\alpha = 0.05$ used in Eq. (2).

The factor $c_{(p,h)}$ is a consistency factor that ensures the $\hat{C}_{\text{MCD}}$ estimator is consistent with the multivariate normal distribution, correcting for the bias arising from the use of only a subset of h observations. The T_{MCD} and C_{MCD} estimators are then used in RPCA to extract robust principal components by decomposing the robust scatter matrix.

2.5.2. MM-Estimator

The MM estimator was introduced by Yohai (1987) as a robust estimator that combines the high breakdown point of the S-estimator with the high asymptotic efficiency of the M-estimator. For multivariate location and scatter estimation, the MM procedure is mathematically defined through three stages [7].

Stage 1: Initial S-estimate and Robust Scale The first stage computes an initial S-estimator of location ($\hat{T}_S$) and scatter ($\hat{C}_S$) that guarantees 50% breakdown point. This is achieved by minimizing the determinant of the candidate scatter matrix C:

$$(\hat{T}_S, \hat{C}_S) = \arg\min_{(T,C)} \det(C_S)$$

subject to the constraint on the average loss function ρ_0

$$\frac{1}{n}\sum_{i=1}^{n} \rho_0(d_i) = \delta$$

where $d_i = \sqrt{(x_i - T)^T C^{-1}(x_i - T)}$ represents the robust Mahalanobis distance calculated for any candidate location T and the scatter matrix C. Here, ρ_0 is a bounded symmetric resistance function, and $\delta = 0.5$ to ensure the maximum breakdown point. Once the optimal full scatter matrix C_S is obtained under this constraint, the robust scalar scale $\hat{\sigma}$ can be geometrically derived from its volume element, which is defined as $\hat{\sigma} = |\hat{C}_S|^{1/2p}$

Stage 2: M-estimate of Location and Shape Using the fixed scale $\hat{\sigma}$ from stage 1, the MM-estimator of location ($\hat{T}_{MM}$) and the shape matrix ($\hat{\Gamma}$) are obtained by minimizing a second loss function ρ_1:

$$\min_{T,\Gamma} \frac{1}{n}\sum_{i=1}^{n} \rho_1 \left(\frac{\left[(x_i - T)^T \Gamma^{-1}(x_i - T)\right]^{1/2}}{\hat{\sigma}} \right)$$

subject to the absolute constraint $|\Gamma| = 1$. Computationally, the optimal location $\hat{T}_{MM}$ is derived using an Iteratively Reweighted Least Squares (IRLS) algorithm expressed as a weighted mean:

$$\hat{T}_{\mathrm{MM}} = \frac{\sum_{i=1}^{n} u(\tilde{d}_i/\hat{\sigma}) x_i}{\sum_{i=1}^{n} u(\tilde{d}_i/\hat{\sigma})}$$

where $\tilde{d}_i = \sqrt{(x_i - \hat{T}_{MM})^T \hat{\Gamma}^{-1}(x_i - \hat{T}_{MM})}$ and $u(t) = \rho_1'(t)/t$ is the weight function.

Stage 3: Final MM Scatter Matrix The robust scatter matrix ($\hat{C}_{MM}$) is constructed by multiplying the squared initial robust scale by the optimal shape matrix from Stage 2

$$\hat{C}_{\mathrm{MM}} = \hat{\sigma}^2 \hat{\Gamma}$$

Loss Function and Computational Tuning Parameters This study employs Tukey's bisquare function for both ρ_0 and ρ_1, defined as:

$$\rho(t; c) = \begin{cases} \frac{t^2}{2} - \frac{t^4}{2c^2} + \frac{t^6}{6c^4} & \text{if } |t| \le c \\ \frac{c^2}{6} & \text{if } |t| > c \end{cases}$$

The MM estimator has a breakdown point of up to 50% and an asymptotic efficiency of 95% under a normal model, making it well-suited for small sample sizes with mild contamination. The derived $\hat{C}_{\mathrm{MM}}$ is then passed to the RPCA procedure for eigen decomposition.

2.5.3. Formation of Robust Principal Components

After obtaining the robust scatter matrix with each estimator, the robust principal components are computed through eigen decomposition

$$\hat{C}_{\text{rob}} = \hat{A}_{\text{rob}} \hat{\Lambda}_{\text{rob}} \hat{A}_{\text{rob}}^{T}$$

where:

- $\hat{C}_{\text{rob}}$: robust scatter estimator, namely $\hat{C}_{\text{MCD}}$ or $\hat{C}_{\text{MM}}$
- $\hat{A}_{\text{rob}} = (\hat{a}_1^{\text{rob}}, \hat{a}_2^{\text{rob}}, \ldots, \hat{a}_p^{\text{rob}})$: matrix whose columns are robust eigenvectors
- $\hat{\Lambda}_{\text{rob}} = \text{diag}(\hat{\lambda}_1^{\text{rob}}, \hat{\lambda}_2^{\text{rob}}, \ldots, \hat{\lambda}_p^{\text{rob}})$: diagonal matrix containing robust eigenvalues in the order $\hat{\lambda}_1^{\text{rob}} \geq \hat{\lambda}_2^{\text{rob}} \geq \cdots \geq \hat{\lambda}_p^{\text{rob}} \geq 0$

The eigenvalue estimator is derived by solving the equation

$$|\hat{C}_{\text{rob}} - \lambda I| = 0$$

Meanwhile, the eigenvector estimator $\hat{a}_i^{\text{rob}}$ is obtained by solving the equation:

$$(\hat{C}_{\text{rob}} - \hat{\lambda}_i^{\text{rob}} I)\hat{a}_i^{\text{rob}} = 0$$

with normalization conditions;

$$(\hat{a}_i^{\text{rob}})^T \hat{a}_i^{\text{rob}} = 1$$

and orthogonal properties;

$$(\hat{a}_i^{\text{rob}})^T \hat{a}_j^{\text{rob}} = 0, \quad i \neq j$$

The principal component score of the sample is defined as:

$$\hat{\text{PC}}_{ij}^{\text{rob}} = (\hat{a}_j^{\text{rob}})^T (x_i - \hat{T}_{\text{rob}}), \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, p$$

with

$$\hat{T}_{rob} \in \{\hat{T}_{\text{MCD}}, \hat{T}_{\text{MM}}\}$$

where:

- i: index of the i-th observation, where i=1,2,...,n
- j: index of the j-th principal component, where j=1,2,...,p
- x_i: the p x 1 vector of the i-th observation
- $\hat{a}_j^{rob}$: the robust eigenvector j-th
- $\hat{T}_{rob}$: robust location estimator ($\hat{T}_{MCD}$ or $\hat{T}_{MM}$)

After extracting the robust principal components via the eigen decomposition of the robust scatter matrix C_{rob}, the original data centered by the robust location T_{MM} is projected onto these components to generate the robust principal component scores. In this study, RPCA is used exclusively as a robust orthogonal transformation of the original predictor space, rather than as a dimension reduction technique. Consequently, all p robust principal components are retained, yielding k=p. Thus, the computational input vector for the MARS model is:

$$z_i = \left(PC_{i1}^{\text{rob}}, PC_{i2}^{\text{rob}}, \ldots, PC_{ip}^{\text{rob}}\right)^{\mathsf{T}}, \quad i = 1, 2, \ldots, n$$

This approach is justified for two reasons. First, the main goal of integrating RPCA into the MARS framework is to reduce the effect of outliers in the predictor variables through robust estimation of location and scatter, rather than to perform variable reduction. Second, keeping all components ensures that no potentially useful information is removed before the MARS algorithm selects the adaptive basis function through its GCV-based forward-backward procedure. In this way, variable selection is fully delegated to MARS, which identifies the relevant structure through its stepwise knot placement mechanism

2.6. Criteria Evaluation

To evaluate model performance, Root Mean Square Error (RMSE) metrics are used. RMSE measures how well models perform by calculating the difference between the observed value and the predicted value y_i and the estimated value $\hat{f}_i$ on the response variable using the following equation.

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

where:

- y_i: the actual observed value of the response variable for the i-th observation
- $\hat{f}_i$: the predicted value of the response variable for the i-th observation, estimated by the MARS model
- n: the total number of observations in a single dataset

The model with the lowest RMSE indicated the best predictive performance. To assess performance robustness across the simulation study, the mean prediction RMSE over a set of R replications (where $R = 100$) is calculated as follows:

$$\overline{\text{RMSE}} = \frac{1}{R}\sum_{i=1}^{R}\text{RMSE}_i$$

where R is the number of simulation replications ($R = 100$)

The standard deviation (SD) of the RMSE across 100 replications is given by the square root of the variance of the prediction RMSE, formulated as:

$$\text{SD}_{\text{RMSE}} = \sqrt{\frac{1}{R-1}\sum_{i=1}^{R}(\text{RMSE}_i - \overline{\text{RMSE}})^2}$$

The 95% confidence interval (CI) of the RMSE across 100 replications is given by:

$$\text{CI}_{95\%} = \overline{\text{RMSE}} \pm t_{\alpha/2,R-1}\left(\frac{\text{SD}_{RMSE}}{\sqrt{R}}\right)$$

where:

- $CI_{95\%}$: the 95% confidence interval for the mean RMSE
- $t_{\alpha/2,R-1}$: the critical value from the Student's t-distribution with significance level $\alpha = 0.05$ and degrees of freedom $R - 1$

2.7. Stages of RPCA-MCD-MARS and RPCA-MM-MARS

Data analysis was conducted in multiple steps, as outlined below:

1. Generate simulated data from a uniform distribution with outliers, covering sample sizes of 50, 100, 200, 500, and 1000; outlier percentages of 5%, 10%, and 25%; and maximum MARS interaction levels of 1, 2, and 3.
2. Compute the robust location T and robust scatter matrix C from the predictor matrix X using the MCD-estimator and MM-estimators for each generated simulation data.
3. Perform an eigenvalue decomposition on the estimated robust scatter matrix to obtain the robust load matrix (eigenvectors).
4. Calculate the robust principal component scores $\hat{PC}_{ij}^{rob}$ for $i = 1, 2, \ldots, n; j = 1, 2, \ldots, p$. In this framework, the complete set of transformed predictors ($p = 8$) is fully retained. Rather than serving as a dimensionality reduction technique, RPCA is strictly used as a robust orthogonal transformation. This step maps the original predictors into an orthogonal space. The goal is to clearly identify and reduce the leverage effects of outliers in the predictor space before building the model.

5. Construct the MARS model using the full matrix of robust principal component scores PC^{rob} as the new input variables. The model parameters are configured as follows: BF is set within the range $[2p, 4p]$, MO is determined using the formula in the Eq. (2), and MI is evaluated across levels 1,2,3 as defined in Table 1. The optimal combination of these parameters is selected based on the minimum GCV using the formula in Eq. (3) through the forward-backward stepwise procedure.
6. Evaluate and compare the performance of RPCA-MCD-MARS and RPCA-MM-MARS using the mean, standard deviation (SD), and confidence intervals (CI) of the RMSE metrics.

3. Results and Discussion

The development of the two methods in this study was performed by integrating Robust Principal Component Analysis (RPCA) with two robust estimators, specifically the MCD-estimator and the MM-estimator, into the Multivariate Adaptive Regression Spline (MARS) framework. This integration aims to provide an outlier-resistant transformation of predictor variables before MARS modeling. In the proposed approach, the original predictor variables are first transformed into robust principal component scores using RPCA. These scores are then used as input variables in the MARS model to evaluate the relative effectiveness of the MCD-estimator and MM-estimators in managing the influence of outliers on the model structure.

To ensure methodological clarity, it is important to distinguish between the two selection criteria used at different stages of this study. Within each simulation replication (the model tuning stage), the MARS algorithm uses GCV (Eq. (3)) as an internal criterion throughout the forward-backward stepwise procedure to adaptively select basis functions under a specific maximum interaction bound ($MI \in \{1, 2, 3\}$). Across replications (the performance evaluation stage), the mean RMSE is computed for each of the 45 experimental conditions. The original 45 experimental runs are then consolidated into 15 primary scenarios (five sample sizes and three levels of outlier contamination) by retaining, for each (n, r) combination, the MI level that yields the minimum mean RMSE.

Among the 15 primary scenarios evaluated, the MM estimator exhibited superior predictive performance in 8 cases, while the MCD estimator performed better in the other 7 cases. Table 2 shows the best RMSE values and the corresponding optimal interaction level (MI) chosen by each estimator. The following subsections analyze performance across different contamination levels.

3.1. Effect of 5% Outlier Proportion

At a 5% outlier level, the MCD estimator demonstrates lower mean RMSE values than the MM estimator across most sample sizes, specifically at $n = 50, n = 100, n = 200$, and $n = 1000$. However, analysis of the 95% Confidence Intervals (CI) shows that this dominance is not consistently statistically significant. At $n = 50$, although MCD attains a smaller mean RMSE (0.5256 versus 0.6021), its 95% confidence interval $[0.4194, 0.6318]$ substantially overlaps with MM's confidence interval $[0.5379, 0.6663]$, and MCD exhibits a considerably larger standard deviation ($SD = 0.5416$ compared with $SD = 0.3278$ for MM). The same pattern of overlapping intervals and negligible performance margin appears at $n = 200$, where the 95% CIs for MCD $[0.4144, 0.5390]$ and MM $[0.4557, 0.5263]$ share a broad intersection. At $n = 500$, the MM estimator marginally outperforms MCD with a lower mean RMSE (0.4685) and a tighter CI $[0.4438, 0.4932]$, though their intervals still overlap. These trends demonstrate that under lower contamination, both estimators exhibit a high asymptotic efficiency, evidenced by the consistent contraction of SD values as n increases, rendering their predictive performance statistically comparable in intermediate samples.

Regarding the optimal interaction level, first-order interaction ($MI = 1$) dominates in most

Table 2: Comprehensive Performance Metrics over 100 Simulation Replications for Different Sample Sizes and Outliers in RPCA Using MCD and MM Estimators on MARS

n	Outlier	MCD				MM				Decision
		Opt. Int.	RMSE	SD	95% CI	Opt. Int.	RMSE	SD	95% CI	
50	5%	1	0.5256	0.5416	[0.4194, 0.6318]	1	0.6021	0.3278	[0.5379, 0.6663]	MCD
50	10%	2	0.6064	0.7206	[0.4652, 0.7476]	1	0.7131	0.3921	[0.6362, 0.7900]	MCD
50	25%	1	1.1347	0.6720	[1.0030, 1.2664]	1	0.8874	0.3697	[0.8149, 0.9600]	MM*
100	5%	3	0.3962	0.3569	[0.3262, 0.4662]	1	0.5670	0.2928	[0.5096, 0.6244]	MCD*
100	10%	3	0.7165	0.9314	[0.5339, 0.8991]	1	0.6470	0.2657	[0.5949, 0.6991]	MM
100	25%	1	0.9549	0.4111	[0.8743, 1.0355]	1	0.7879	0.2319	[0.7424, 0.8334]	MM*
200	5%	1	0.4767	0.3180	[0.4144, 0.5390]	1	0.4910	0.1801	[0.4557, 0.5263]	MCD
200	10%	1	0.6779	0.3356	[0.6121, 0.7437]	1	0.5841	0.1857	[0.5477, 0.6205]	MM
200	25%	1	0.8882	0.3372	[0.8221, 0.9543]	1	0.6862	0.2122	[0.6446, 0.7278]	MM*
500	5%	1	0.4723	0.1900	[0.4351, 0.5095]	1	0.4685	0.1259	[0.4438, 0.4932]	MM
500	10%	1	0.5399	0.1602	[0.5085, 0.5713]	1	0.4962	0.1202	[0.4726, 0.5198]	MM
500	25%	1	0.6529	0.0986	[0.6336, 0.6722]	1	0.6224	0.0923	[0.6043, 0.6405]	MM
1000	5%	1	0.3918	0.1224	[0.3678, 0.4158]	1	0.4480	0.1106	[0.4263, 0.4697]	MCD*
1000	10%	1	0.4587	0.0806	[0.4429, 0.4745]	1	0.4700	0.0596	[0.4583, 0.4817]	MCD
1000	25%	2	0.5639	0.0745	[0.5493, 0.5785]	1	0.5821	0.0535	[0.5716, 0.5926]	MCD

*Statistically significant dominance (the 95% confidence intervals of the two estimators do not overlap). Decisions without an asterisk (*) indicate numerical superiority in mean RMSE, but the estimators remain statistically comparable.

Note: Opt. Int. represents the maximum interaction level (MI) that minimizes the mean RMSE across 100 simulation replications for each specific (n, r) condition, whereas the internal basis functions within each individual replication are optimized via the Minimum GCV criterion.

scenarios at this contamination level. The only exception is observed for MCD at $n = 100$, which achieves the lowest RMSE at third-order interaction ($MI = 3$). This suggests that, under low contamination (5%) and moderate sample sizes ($n = 100$ to 500), MCD's clean subset selection can successfully preserve complex underlying interaction structures without inflating prediction errors. For most cases in an outlier proportion 5%, additive models ($MI = 1$) are sufficient to capture the data structure, which is consistent with the principle of parsimony under minimal contamination.

3.2. Effect of 10% Outlier Proportion

At a 10% outlier level, MCD's overall dominance diminishes, while MM's superiority becomes increasingly evident. At $n = 50$, MCD remains superior with an RMSE of 0.6064 compared to MM's 0.7131, but it suffers from extreme volatility, as indicated by a high standard deviation and an expansive CI ($SD = 0.7206, CI[0.4652, 0.7476]$). The high standard deviation and wide CI reflect the instability of the MCD estimate in small samples with moderate contamination; therefore, any claim of MCD's superiority in this context should be treated with caution. Conversely, MM provides a more stable and predictable solution.

For $n = 100, n = 200$, and $n = 500$, MM consistently outperforms MCD with lower mean RMSE values and markedly narrower confidence intervals; e.g., at $n = 100$, MM's $SD = 0.2657$ versus MCD's $SD = 0.9314$. However, at $n = 1000$, MCD achieves a lower mean RMSE (0.4587) compared with MM's 0.4700, but their confidence intervals (CI) overlap substantially, indicating that their asymptotic performances converge under large samples.

In terms of interaction structure, MM consistently selects first-order interaction ($MI = 1$) across all sample sizes. In contrast, MCD favors higher-order interactions at smaller sample sizes ($MI = 2$ at $n = 50$ and $MI = 3$ at $n = 100$) before converging to $MI = 1$ for $n \geq 200$. This indicates that in smaller samples, MCD's hard-rejection approach retains enough clean observations to support complex interactions modelling. However, as the sample size increases at

10% contamination, both estimators align toward a simpler additive structure ($MI = 1$), which provides greater stability and is less sensitive to outlying observations.

3.3. Effect of 25% Outlier Proportion

At a 25% outlier level, the MM-estimator demonstrates clear and statistically significant superiority across almost all sample sizes. For $n = 50, n = 100, n = 200$, MM yields substantially lower mean RMSE values along with tighter 95% confidence intervals that exhibit absolutely no overlap with those of MCD. For instance, at $n = 100$, MM achieves an RMSE of 0.7879 ($95\%CI : [0.7424, 0.8334]$) compared to MCD's 0.9549 ($95\%CI : [0.8743, 1.0355]$). At $n = 500$, the MM-estimator maintains its advantage with a lower RMSE and smaller standard deviation compared to the MCD-estimator, though their 95% CIs display a minor overlap within the [0.6336, 0.6405] range. These narrow margins underscore the advantage of MM's smooth iterative downweighting function over MCD's binary subset selection when the data space is heavily corrupted. The only exception is $n = 1000$, where MCD attains a lower RMSE of 0.5639 versus MM's 0.5821. However, this difference must be interpreted with caution, as their 95% CIs overlap ($CI[0.5716, 0.5785]$), implying that the distinction in predictive capability between the two estimators becomes statistically indistinguishable in very large samples.

Across all sample sizes in this contamination, first-order interactions ($MI = 1$) dominate for both estimators, except MCD at $n = 1000$, which selects a second-order interaction ($MI = 2$). This pattern suggests that, under heavy contamination, simpler additive models are generally more resistant to overfitting, regardless of the estimator used. Overall, the results indicate that the iterative downweighting of MM is more effective than the subset selection approach of MCD in reducing the influence of outliers at this contamination level, except when a sufficiently large sample size is given ($n = 1000$).

3.4. Performance Evaluation of MCD and MM Estimators

Based on Table 2, there is no universally superior estimator within the robust PCA-MARS framework. The optimal choice between MCD and MM estimators is heavily dependent on the level of outlier contamination, the sample size, and the balance between bias and estimator variance. Out of the 15 evaluated scenarios, 10 scenarios exhibit overlapping 95% CIs, indicating that the estimators remain statistically comparable in many settings despite numerical differences in mean RMSE.

Specifically, MCD is recommended for datasets with low outlier proportions ($\leq 5\%$) in most sample sizes (except $n = 500$) and for larger datasets ($n \geq 1000$) at all contamination levels. However, at $n \leq 200$, under moderate-to-high contamination (10% and 25%), MCD exhibits high sample-to-sample variability (inflated SD values), making it a riskier choice for practical applications. Conversely, the MM estimator displays greater stability–manifested by consistently narrower confidence intervals and lower standard deviations–across moderate-to-high outlier settings (10% and 25%), making it the more reliable option for intermediate sample sizes ($n = 100$ to 500).

Statistically, additive models ($MI = 1$) are overwhelmingly selected in 26 of the 30 combinations of estimator-scenario tested. MM strictly favors $MI = 1$ across all scenarios. In contrast, MCD captures higher-order interactions ($MI > 1$) under specific profiles: $n = 50$ (10% outliers), $n = 100$ (5% and 10% outliers), and $n = 1000$ (25% outliers). This indicates that the MCD estimator approach preserves complex data structures better than the MM estimator under these specific conditions.

4. Conclusion

This study developed and evaluated two robust nonparametric regression approaches, namely RPCA-MCD-MARS and RPCA-MM-MARS, which integrate Robust Principal Component

Analysis (RPCA) with Multivariate Adaptive Regression Spline (MARS) as a robust predictor transformation strategy. The framework transforms predictor variables into robust principal component scores using either the MCD or MM estimator, and then uses those scores as inputs to MARS. Since no baseline models such as standard MARS or CPCA-MARS were included, the extent to which the framework mitigates MARS sensitivity to outlier-contaminated predictors cannot be directly assessed. Instead, this study evaluated the relative predictive performance of two robust estimators, MCD and MM, within the RPCA-MARS framework across varying outlier proportions and sample sizes.

The simulation results show that the predictive performance of the RPCA-MARS framework depends heavily on the choice of the robust estimator and on data characteristics. The MCD estimator is highly efficient and recommended for data with minimal outlier proportions ($\leq 5\%$) or for very large sample sizes ($n \geq 1000$) across all contamination levels, although its advantage in large samples is often characterized by narrow margins and overlapping 95% confidence intervals with MM.

On the other hand, the MM estimator provides a more stable and robust alternative under moderate-to-high contamination (10%–25%) for practically relevant sample sizes ($n = 100$–500), consistently yielding lower standard deviations and narrower, highly concentrated confidence intervals. Furthermore, regardless of the estimator, severe contamination favors first-order additive models ($MI = 1$) to protect the MARS framework from outlier propagation and overfitting.

CRediT Authorship Contribution Statement

Uswatun Hasanah: Conceptualization, Methodology, Software, Formal Analysis, Investigation, Writing–Original Draft Preparation, Writing–Review & Editing, Project Administration. **Solimun:** Supervision, Methodology, Formal Analysis, Writing–Review & Editing. **Atiek Iriany:** Supervision, Methodology, Validation, Writing–Review & Editing.

Declaration of Generative AI and AI-assisted technologies

There are generative AI or AI-assisted technologies that were used during the preparation of the manuscript, such as DeepL and Grammarly, which were employed for translation from Indonesian to English and for proofreading/clarity edits. No generative AI or AI-assisted technologies were used for data analysis, statistical modeling, or result generation.

Declaration of Competing Interest

The authors declare no competing interests

Funding and Acknowledgments

This research received no external funding."

Data and Code Availability

The simulation datasets presented in this study are available upon reasonable request from the corresponding author. The custom code for data generation and analysis is available from the corresponding author upon reasonable request. Due to the simulation study's computational nature, the raw datasets can be regenerated using the provided code.

References

[1] A. Araveeporn. "The Estimating Parameter and Number of Knots for Nonparametric Regression Methods in Modelling Time Series Data". In: *Modelling* 5.4 (2024), pp. 1413–1434. DOI: 10.3390/modelling5040073.

[2] Jerome H. Friedman. "Multivariate Adaptive Regression Splines". In: *The Annals of Statistics* 19.1 (Mar. 1991), pp. 1–67. DOI: 10.1214/aos/1176347963.

[3] N. Murat. "Outlier detection in statistical modeling via multivariate adaptive regression splines". In: *Communications in Statistics-Simulation and Computation* 52.7 (2023), pp. 3379–3390. DOI: 10.1080/03610918.2021.2007400.

[4] M. Hubert, P. J. Rousseeuw, and K. Vanden Branden. "ROBPCA: A new approach to robust principal component analysis". In: *Technometrics* 47.1 (2005), pp. 64–79. DOI: 10.1198/004017004000000563.

[5] A. M. Gad and M. E. Qura. "Regression estimation in the presence of outliers: A comparative study". In: *International Journal of Probability and Statistics* 5.3 (2016), pp. 65–72. DOI: 10.5923/j.ijps.20160503.01.

[6] P. J. Rousseeuw and K. V. Driessen. "A Fast Algorithm for the Minimum Covariance Determinant Estimator". In: *Technometrics* 41.3 (1999), pp. 212–223. DOI: 10.1080/00401706.1999.10485670.

[7] Peter Rousseeuw and Mia Hubert. "High-Breakdown Estimators of Multivariate Location and Scatter". In: *Robustness and Complex Data Structures: Festschrift in Honour of Ursula Gather*. Berlin, Heidelberg: Springer, 2013. Chap. 4, pp. 49–66. DOI: 10.1007/978-3-642-35344-6_4.

[8] M. Mohammadi, M. K. Khorrami, A. Rezaei, H. Vatanparast, and M. M. K. Khorrami. "Robust principal component analysis-multivariate adaptive regression splines (rPCA-MARS) model for determining total acid number (TAN) and total base number (TBN) of crude oil samples using attenuated total reflectance fourier transform infrared (ATR-FTIR) spectroscopy". In: *Vibrational Spectroscopy* 129 (2023), p. 103579. DOI: 10.1016/j.vibspec.2023.103579.

[9] Christophe Croux and Anne Ruiz-Gazen. "High breakdown estimators for principal components: the projection-pursuit approach revisited". In: *Journal of Multivariate Analysis* 95.1 (2005), pp. 206–226. DOI: 10.1016/j.jmva.2004.08.002.

[10] M. B. Adiguzel and M. A. Cengiz. "Model selection in multivariate adaptive regressions splines (MARS) using alternative information criteria". In: *Heliyon* 9.9 (2023), e19964. DOI: 10.1016/j.heliyon.2023.e19964.

[11] N. Shanty and M. K. Aidid. "Application of Multivariate Adaptive Regression Splines (MARS) to Model the Factors Affecting the Percentage of Poor Population in Indonesia". In: *VARIANSI: Journal of Statistics and Its Application on Teaching and Research* 7.03 (2025), pp. 228–240. DOI: 10.35580/variansiunm518.

[12] P. Filzmoser and K. Nordhausen. "Robust Linear Regression for High-Dimensional Data: An Overview". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 13.4 (2021), e1524. DOI: 10.1002/wics.1524.

[13] H. Bulut. "Mahalanobis Distance Based on Minimum Regularized Covariance Determinant Estimators for High Dimensional Data". In: *Communications in Statistics—Theory and Methods* 49.24 (2020), pp. 5897–5907. DOI: 10.1080/03610926.2020.1719420.

[14] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. 3rd. Wiley Series in Probability and Statistics. Hoboken, New Jersey: John Wiley & Sons, Inc., 2003, p. 742. https://www.scribd.com/document/747066403/T-W-Anderson-An-Introduction-to-Multivariate-Statistical-Analysis-Wiley-Series-in-Probability-and-Statistics-3rd-Edition-2003.

[15] Richard A. Johnson and Dean W. Wichern. *Applied Multivariate Statistical Analysis*. 6th. Upper Saddle River, NJ: Pearson Prentice Hall, 2007. https://mathematics.foi.hr/Applied%20Multivariate%20Statistical%20Analysis%20by%20Johnson%20and%20Wichern.pdf.