Hacettepe Journal of Mathematics & Statistics

Hacet. J. Math. Stat.
Published online: 04.08.2025
DOI : 10.15672/hujms.1614841

Research Article

A robust probabilistic framework for principal component regression: optimizing parameter identification and outlier detection via approximate Bayesian computation

Aiman Tahir*, Maryam Ilyas

College of Statistical Sciences, University of the Punjab, Lahore, Pakistan

Abstract

Anomalies and ill-conditioned predictors present considerable obstacles to reliable parameter estimation in regression models. This paper presents an innovative approach that combines principal component regression with approximate Bayesian computation to address these issues. Principal component regression mitigates the effects of ill-conditioned variables by transforming highly correlated predictors into orthogonal components. Meanwhile, approximate Bayesian computation enhances robustness by approximating the posterior distribution of error variance (σ^2). This flexible framework models uncertainty and noise effectively. The integration of these methods improves both parameter estimation and anomaly detection. By assigning probabilistic scores to potential outliers, the method provides a more accurate and nuanced identification of anomalies. Extensive validation through simulated and real-world datasets demonstrates the favorable performance of the proposed technique over existing robust methods. These findings highlight the potential of approximate Bayesian computation as a powerful tool to improve the robustness and precision of regression analyzes in noisy and complex data environments.

Mathematics Subject Classification (2020). 65C20, 62G32

Keywords. Anomalies, approximate Bayesian computation, ill-conditioned predictors, principal component regression, robust estimation

1. Introduction

Anomalous data and ill-conditioned variables are common statistical challenges that are well-documented in fields such as finance [42], healthcare [3], economics [18], damage detection [56], and the social sciences [43, 44], especially in prediction tasks. These issues can severely distort the results of regression analysis. Anomalies are the data points that deviate substantially from the rest of the data. They often arise due to measurement errors or natural variability [7, 52]. On the other hand, ill-conditioned predictors occur when independent variables in regression are nearly collinear, leading to inflated standard errors and unreliable parameter estimates [1, 51]. This condition makes it difficult to isolate the individual contributions of predictors to the response variable [36]. Both anomalous data

*Corresponding Author.
Email addresses: aimantahir78@gmail.com (A. Tahir), maryam.stat@pu.edu.pk (M. Ilyas)
Received: 07.01.2025; Accepted: 30.07.2025

and ill-conditioned variables can mislead statistical inference and lead to flawed conclusions if not appropriately addressed.

Anomalies and ill-conditioned predictors present considerable obstacles to reliable parameter estimation in regression models. This paper presents an innovative approach that combines Principal Component Regression (PCR) with Approximate Bayesian Computation (ABC) to address these issues. PCR mitigates the effects of ill-conditioned variables by transforming highly correlated predictors into orthogonal components. Meanwhile, ABC enhances robustness by approximating the posterior distribution of error variance (σ^2). This flexible framework models uncertainty and noise effectively. The integration of these methods improves both parameter estimation and anomaly detection. By assigning probabilistic scores to potential outliers, the method provides a more accurate and nuanced identification of anomalies. Extensive validation through simulated and real-world datasets demonstrates the favourable performance of the proposed technique over existing robust methods. These findings highlight the potential of ABC as a powerful tool for improving the robustness and accuracy of regression analyses in noisy and complex data environments. The PCR offers a solution to an ill-conditioned predictor matrix by transforming correlated predictors into orthogonal principal components [15,36,40]. These components serve as new predictors, effectively reducing the multicollinearity problem. However, while PCR mitigates collinearity, it remains vulnerable to outliers, which can still skew results. To address this, various robust PCR techniques have been emerged. Walczak and Massart [54] combined robust Principal Component Analysis (PCA) using the least median of squares regression [47] with PCR to handle outliers. Pell [45] introduced a "resampling by halfmeans" method [20] that excludes outliers before PCA is conducted. Filzmoser [21] proposed a robust PCR using projection pursuit proposed by Li and Chen [38]. This technique obtains robust principal components, integrating them into the Least Trimmed Square (LTS) regression [47] for prediction. Hubert and Verboven [32] introduced two versions of robust PCR. One for low-dimensional data ($p < n$) using the Minimum Covariance Determinant (MCD) [47]. Another for high-dimensional data ($p > n$), incorporating the ROBPCA method [33]. Hubert et al. [33] proposed the ROBPCA technique to cope with high-dimensional data, including the $n = p$ setting by integrating projection pursuit and robust covariance estimation. Zhang et al. [62] suggested using principal sensitive vectors [46] to detect outliers before applying classical PCR. A comparative study between robust PCR and robust PLS was conducted by Engelen [19], focusing on efficiency, robustness, goodness of fit and predictive power. Further innovations include functional logistic regression proposed by Denhere and Billor [16] and the Bayesian framework proposed by Gagnon et al. [22]. Recently, several robust estimators for PCR have been developed to address multicollinearity, outliers, and high-dimensional data simultaneously. Notable contributions include the works of Tahir and Ilyas [53], Ebiwonjumi et al. [18], Arum et al. [5], and Dong et al. [17].

Within the Bayesian framework, various robust estimators have been developed to manage the influence of outliers and influential observations. One group of methods employs probabilistic measures such as γ-divergences (e.g., [28,58]), Kullback-Leibler divergences [34], conditional predictive density [13], and posterior probability density function of residuals [12,61], along with measures of surprise [8]. A second group relies on heavy-tailed likelihood functions to accommodate outliers, such as the t-distribution [9,37], mixtures of normal distributions [55], and robust error distributions [23], as well as more recent approaches like the revisited two-component mixture model [27]. However, these methods typically focus on robust parameter estimation without explicitly addressing outlier detection. A third category assumes outlier generation models, such as the mixture model [35], the mean-shift model [25,26], and the variance-inflation model [14]. Yuen and Mu [59] proposed a probabilistic approach for robust parameter estimation and outlier detection in linear regression, leveraging Bayes theorem. Recently, Xiao et al. [57] introduced a robust

regression model combining the Bayesian selection model with LTS to improve anomaly detection and predictive accuracy.

Despite the effectiveness of these Bayesian approaches, many rely on analytically tractable likelihood functions, which can limit their applicability when the model becomes complex. In cases where the likelihood function is difficult to express analytically or computationally expensive, simulation-based methods such as ABC offer a more flexible alternative. By simulating summary statistics from prior distributions, ABC can infer parameters without requiring an explicit likelihood, making it particularly valuable for addressing complex models accommodating anomalies and ill-conditioned predictors. Building on this framework, this paper introduces a novel approach integrating ABC with PCR to handle the dual challenges in regression models. The method estimates optimal parameter values while accounting for uncertainties in outlier detection, offering a robust and flexible solution to these pervasive issues. To evaluate its effectiveness, a comparative analysis is conducted against several existing robust techniques, including the Huber estimator [30, 31], LTS [47, 48], and Robust PCR (RobPCR). The proposed method leverages the strengths of these existing techniques while offering a more comprehensive framework by using ABC for robust parameter estimation and outlier identification. Specifically, RobPCR, which employs ROBPCA [33] for robust principal component extraction followed by regression, serves as a key benchmark for comparison.

The structure of the paper is as follows: Section 2 provides a detailed overview of Bayesian inference, ABC, maximum trimmed likelihood estimation, and PCR. Section 3 outlines the proposed method. Section 4 discusses the simulation settings, data generation process, and performance measures. Section 5 presents results that compare the proposed method with existing techniques. Section 6 applies the methods to real-life data, and Section 7 offers concluding remarks.

2. Preliminaries

2.1. Bayesian linear regression

In Bayesian linear regression [11], consider a linear regression model as described in Eq. 2.1. Here, $\mathbf{y}$ is the vector $(n \times 1)$ of the response variable, and $\mathbf{X}$ is a matrix of predictors of order $(n \times p)$. $\boldsymbol{\beta}$ is the $(p \times 1)$ vector of regression coefficients corresponding to the predictors $(\mathbf{X})$. The error term (ϵ) is $(n \times 1)$ vector that follows a normal distribution with zero mean and variance σ^2.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon \tag{2.1}$$

The likelihood function for the response variable is defined in Eq.(2.2). The posterior distribution of $\boldsymbol{\beta}$ and σ^2 is then determined using Eq. (2.3). In this context, $p(\boldsymbol{\beta}, \sigma^2)$ represents the prior distribution of $\boldsymbol{\beta}$ and σ^2, which encapsulates prior knowledge about the parameters, typically based on expert experience. The term $p(\mathbf{y}|\mathbf{X})^{-1}$ denotes a normalizing constant ensuring the integral of the posterior distribution is unity over the whole parametric space Θ.

$$p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \mathbf{X}) = (2\pi\sigma^2)^{-n/2}\ exp\ [\frac{-1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})] \tag{2.2}$$

$$p(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X}) = p(\mathbf{y}|\mathbf{X})^{-1}\ p(\boldsymbol{\beta}, \sigma^2)\ p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \mathbf{X}) \tag{2.3}$$

The prior distributions of $\boldsymbol{\beta}$ and σ^2 are assumed to be independent of each other. Thus, the term $p(\boldsymbol{\beta}, \sigma^2)$ can be factored into $p(\boldsymbol{\beta})p(\sigma^2)$. The prior distribution for $\boldsymbol{\beta}$ is assumed to follow an independent uniform distribution with a specified range. Meanwhile, the prior for σ^2 is supposed to follow an inverse Gamma distribution, denoted as $IG(a^*, b^*)$, and is defined in Eq. (2.4). Here, $\Gamma(.)$ represents the Gamma function, and a^* and b^* are shape

and scale parameters, respectively. The posterior distribution in Eq. (2.3) is transformed into Eq. (2.5) by substituting the prior distribution as follows:

$$p(\sigma^2) = \frac{(b^*)^{a^*}}{\Gamma(a^*)}(\sigma^2)^{-a^*-1}\ exp\ (-\frac{b^*}{\sigma^2}) \tag{2.4}$$

$$p(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X}) \propto p(\boldsymbol{\beta})\ p(\sigma^2)\ p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \mathbf{X})$$

$$p(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X}) \propto p(\boldsymbol{\beta})\ (\sigma^2)^{-(\frac{n}{2}+a^*+1)}\ exp\ [\frac{-1}{2\sigma^2}\{2b^* + (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\}] \tag{2.5}$$

The marginal posterior distribution of σ^2 is calculated by integrating $p(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X})$ with respect to $\boldsymbol{\beta}$. The result is an inverse Gamma distribution with parameters $\hat{a}$ and $\hat{b}$ defined in Eq. (2.6). Similarly, integrating the posterior distribution with respect to σ^2 gives the marginal posterior distribution of $\boldsymbol{\beta}$. This posterior follows a Student-t distribution with mean $(\hat{\boldsymbol{\beta}})$, precision matrix $(\boldsymbol{\Lambda})$ and degrees of freedom (ν). Here, $\boldsymbol{\Lambda} = (\mathbf{X}^T\mathbf{X})(\frac{\hat{a}}{\hat{b}})^{-1}$ and $\nu = 2\hat{a}$.

$$\hat{a} = \frac{n-p}{2} + a^*, \hat{b} = \frac{1}{2}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}) + b^* \tag{2.6}$$

2.2. Approximate Bayesian Computation

ABC is a computational technique grounded in Bayesian statistics [10, 50]. In standard Bayesian inference, the likelihood function plays a central role in determining the parameter estimates. Although simple Bayesian models often allow for tractable analytical solutions, complex models can make likelihood integration difficult or computationally expensive. ABC provides a framework for likelihood-free inference, bypassing the need to compute the likelihood directly. ABC is particularly well-suited for models defined by a stochastic data-generating process. Instead of working with the full posterior distribution $p(\boldsymbol{\theta}|\mathbf{D})$, ABC focuses on the partial posterior distribution $p(\boldsymbol{\theta}|\mathbf{s}_{obs})$. Here, $\mathbf{s}_{obs}$ is a summary statistic vector derived from the observed data $(\mathbf{D})$. This partial posterior distribution is given in Eq. (2.7).

$$p(\boldsymbol{\theta}|\mathbf{s}_{obs}) = \frac{p(\mathbf{s}_{obs}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{p(\mathbf{s}_{obs})} \tag{2.7}$$

ABC works by simulating m values of θ_i from the prior distribution $\pi(\boldsymbol{\theta})$. For each θ_i, corresponding summary statistics s_i is computed based on the data model $p(\mathbf{s}|\theta_i)$. The partial posterior distribution is then approximated by conditional density estimation based on the simulated pairs (θ_i, s_i).

2.3. Maximum Trimmed Likelihood Estimation

In Maximum Trimmed Likelihood Estimation (MTLE), the likelihood function $p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \mathbf{X})$ is replaced by a likelihood based on a subset of the data, that is, $p(\mathbf{y}^H|\boldsymbol{\beta}, \sigma^2, \mathbf{y}^H)$. Specifically, the likelihood is computed for a trimmed dataset $\mathbf{U}^H = \mathbf{U}_i = (\mathbf{X}_i, \mathbf{y}_i), i \in H$. Here, H denotes an index set having h different data points from $1, 2, \ldots, n$. The modified likelihood is defined in Eq. (2.8).

$$p(\mathbf{y}^H|\boldsymbol{\beta}, \sigma^2, \mathbf{X}^H) = (2\pi\sigma^2)^{-h/2}\ exp\ [\frac{-1}{2\sigma^2}(\mathbf{y}^H - \mathbf{X}^H\boldsymbol{\beta})^T(\mathbf{y}^H - \mathbf{X}^H\boldsymbol{\beta})] \tag{2.8}$$

The MTLE $(\hat{\boldsymbol{\beta}}_{MTLE})$ and the corresponding MTLE data set $(\mathbf{U}^H_{MTLE})$ are obtained by maximizing the likelihood or, equivalently, by minimizing the trimmed error function $E(\boldsymbol{\beta}, \mathbf{U}^H)$. This error function is defined by

$$\hat{\boldsymbol{\beta}}_{MTLE} = \underset{\boldsymbol{\beta}, \mathbf{U}^H}{argmin}\ E(\boldsymbol{\beta}, \mathbf{U}^H) \tag{2.9}$$

$$E(\boldsymbol{\beta}, \mathbf{U}^H) = (\mathbf{y}^H - \mathbf{X}^H\boldsymbol{\beta})^T(\mathbf{y}^H - \mathbf{X}^H\boldsymbol{\beta}) \tag{2.10}$$

The efficient estimates of $\boldsymbol{\beta}_{MTLE}$ and $\mathbf{U}^{H}_{MTLE}$ are computed using the methodology proposed by Rousseeuw and Van Driessen, [49]. In the LTS approach, the efficiency and robustness of the estimation are controlled by the parameter h. Selecting the appropriate value for h is critical to achieve a balance between robustness and efficiency. A larger h brings the trimmed likelihood closer to the full likelihood, thereby reducing the robustness to outliers. In contrast, a smaller h increases robustness by excluding more data points, but this can result in a significant loss of information. Therefore, the choice of h requires careful consideration. Typically, the value of h depends on the expected number of outliers in the data set. It is set as a fixed fraction of the total sample size n, often assumed to be at least $n/2$. In this study, h is selected to be 70% of the total number of observations. This choice helps to ensure that any suspicious or anomalous entries can be effectively identified and removed from $\boldsymbol{\beta}_{MTLE}$, while maintaining a sufficient amount of data for reliable parameter estimation.

2.4. Detection of leverage points

A leverage point in linear regression, as defined by Rousseeuw and Leroy [47] and Hoaglin and Welsch [29], is a data point that deviates significantly from the rest of the data in the predictor space. These points can have a strong influence on Ordinary Least Squares (OLS) estimates, particularly when they are associated with large residuals. To identify leverage points, the Mahalanobis distance is commonly used [47] and is expressed in Eq. 2.11. Here, $\mathbf{x}_i^T$ denotes the i^{th} row of the data matrix ($\mathbf{X}$), $\bar{\mathbf{x}}$ is the mean row vector, and Σ_X represents the covariance matrix of the row vectors of the data matrix ($\mathbf{X}$) as follows:

$$MD(\mathbf{x}_i) = \sqrt{(\mathbf{x}_i - \bar{\mathbf{x}})\Sigma_X^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})^T} \tag{2.11}$$

Rousseeuw and Leroy [47] proposed a threshold to identify leverage points using the chi-square distribution quantile. The threshold is defined as $\sqrt{\Phi^{-1}_{\chi^2}(0.95)}$. Here, $\Phi^{-1}_{\chi^2}(.)$ is the quantile function of the chi-square distribution. The degrees of freedom correspond to the number of non-intercept terms in the model. Based on this threshold, the set of leverage points L is determined using the Eq. (2.12). This approach helps isolate points that may exert an undue influence on the regression model, ensuring more reliable parameter estimates as

$$L = \{\mathbf{U}_i = (\mathbf{x}_i, y_i) : MD(\mathbf{x}_i) > \sqrt{\Phi^{-1}_{\chi^2}(0.95)}, i = 1, 2, \ldots, n\} \tag{2.12}$$

2.5. Principal component regression

PCA, as described by Anderson [2], transforms the original correlated predictor variables into a set of uncorrelated variables called principal components (PCs). These PCs are linear combinations of the original predictors. Let $\mathbf{X}$ be the $n \times p$ matrix of predictors, where n is the sample size and p is the number of predictors. The principal components are represented as $\mathbf{z}_1 = \mathbf{e}_1^T\mathbf{x}_1, \mathbf{z}_2 = \mathbf{e}_2^T\mathbf{x}_2, \ldots, \mathbf{z}_p = \mathbf{e}_p^T\mathbf{x}_p$. Here, $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p$ are the eigenvectors of the covariance matrix ($\boldsymbol{\Sigma} = cov(\mathbf{X})$) , corresponding to the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$. The matrix $E = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p]$ contains these eigenvectors and is of size $p \times p$.

In PCR, the response variable ($\mathbf{y}$) is regressed to a subset of the first q-PCs ($\mathbf{Z}_q$), where $q < p$. This subset captures most of the variance in the original data, while mitigating the problem of ill-conditioned predictors. The PCR model is expressed in Eq. 2.13. Here, $\boldsymbol{\gamma_q}$ is $(q \times 1)$ vector of regression coefficients for the q-PCs ($\mathbf{Z}_q$), and ϵ is the $(n \times 1)$ vector of error terms. The regression coefficients $\boldsymbol{\gamma_q}$ are estimated using Eq. 2.14 following the least squares estimation method. Once $\boldsymbol{\gamma_q}$ is estimated, the corresponding coefficients for the original predictors, $\hat{\boldsymbol{\beta}}$, are recovered by transforming back to the original predictor

space using the eigenvector matrix ($\mathbf{E}$) in Eq. 2.15. In PCR, only the first few principal components are retained for regression, thereby mitigating the effects of multicollinearity.

$$\mathbf{y} = \mathbf{Z}_q\boldsymbol{\gamma}_q + \epsilon \tag{2.13}$$

$$\hat{\boldsymbol{\gamma}}_q = (\mathbf{Z}_q^T\mathbf{Z}_q)^{-1}\mathbf{Z}_q^T\mathbf{y} \tag{2.14}$$

$$\hat{\boldsymbol{\beta}} = \mathbf{E}_{p\times q}\hat{\boldsymbol{\gamma}}_q \tag{2.15}$$

3. Proposed Probabilistic Robust Principal Component Regression (PRPCR)

This method is proposed to address robust parameter estimation and outlier identification in the presence of anomalies, leverage points, ill-conditioned predictors and high-dimensional data settings. It encompasses several distinct steps, which are outlined in the following subsections. Subsection 3.1 presents the algorithm for creating initial-suspicious sub-datasets. Subsection 3.2 elaborates on the computation of outlier probability. Subsection 3.3 discusses the ABC rejection sampler algorithm. Lastly, subsection 3.4 details the computational procedure of the proposed method.

3.1. Initial-suspicious sub-datasets

Parameter estimation is significantly affected by ill-conditioned regressors, outliers, and leverage points. The proposed technique aims to achieve robust parameter identification by overcoming the influence of outliers and mitigating multicollinearity. This subsection outlines the process of selecting the initial regular and suspicious subdatasets. Initially, the reliable subset of data, termed the initial regular dataset ($\mathbf{U}^R$), is identified from the complete dataset ($\mathbf{U}$). The selection of data points for $\mathbf{U}^R$ involves maximizing the trimmed likelihood while excluding leverage points, as defined in Eq. 2.12. The complement of the set of leverage points in $\mathbf{U}^H_{MTLE}$ constitutes $\mathbf{U}^R$. Furthermore, parameter identification is based on $\mathbf{U}^R$, excluding leverage points. In contrast, the initial suspicious subdataset, ($\mathbf{U}^S$), is derived by removing $\mathbf{U}^R$ from the entire dataset ($\mathbf{U}$).

3.2. Probability of outlier

Existing anomaly detection methods typically rely on a predefined threshold for the identification of outliers, often determined based on expert knowledge. Typically, a criterion such as $(|\epsilon|)/\sigma > 2.5$ is used to flag anomalies. Here, $\boldsymbol{\epsilon}$ denotes the $(n \times 1)$ vector of residual terms. However, the choice of this threshold can significantly affect outlier detection outcomes. In this study, we adopt an alternative criterion proposed by Yuen and Mu [59], which computes the probability of an outlier for each suspicious data point, providing a more nuanced approach to outlier identification.

It is important to note that the initial suspicious data set can include regular data points, particularly when a small value of h is selected in the trimmed likelihood estimation. To enhance outlier detection, each data point in the initial suspicious subset is assigned a probability of being an outlier. Data points in this subset with a probability less than 0.5 are reclassified as regular data points. This detecting criteria was adopted by Yuen and Mu (2012).

For instance, consider a suspicious data point (U_k^S) with its corresponding residual (ϵ_k^S), where $U_k^S = (\mathbf{X}_k^S, y_k^S)$ and $\epsilon_k^S = y_k^S - \mathbf{X}_k^S\boldsymbol{\beta}$. To determine whether this data point is an outlier, we calculate the probability of a data point relative to the total number of data points with a certain error. The probability of an outlier is described as the probability that the residuals of all n data entries lie within the interval $(-|\epsilon_k^S|, |\epsilon_k^S|)$ assuming that the prediction error follows the distribution, ie $G(\mathbf{0}, \sigma^2)$.

In simpler terms, this measure quantifies the probability of obtaining a data point under the assumed error distribution. The conditional probability is defined in Eq. 3.1,

representing that all n data entries fall in the interval $(-|\epsilon_k^S|, |\epsilon_k^S|)$ given ϵ_k^S and σ^2. Here, n is the total number of data points in the whole dataset ($\mathbf{U}$), and Φ is the cumulative distribution function of the standard normal distribution.

$$P_o(U_k^S|\epsilon_k^S, \sigma^2) = [1 - 2\Phi(-|\epsilon_k^S|/\sigma)]^n \tag{3.1}$$

Moreover, the uncertainty of ϵ_k^S is derived by the posterior uncertainty of the regression parameters ($\boldsymbol{\beta}$). The posterior uncertainty of $\boldsymbol{\beta}$ and σ^2 is accommodated by utilizing the theory of total probability. Consequently, each suspicious data point contains the probability of an outlier defined in Eq. 3.2. Here, the posterior pdf ($p(\epsilon_k^S, \sigma^2|U^R)$) serves as a weighting function, as Eq. 3.2 is defined by the weighted average of $P_o(U_k^S|\epsilon_k^S, \sigma^2)$.

$$P_o(U_k^S|\mathbf{U}^R) = \int_{-\infty}^{+\infty} [1 - 2\Phi(-|\epsilon_k^S|/\sigma)]^n \times p(\epsilon_k^S, \sigma^2|\mathbf{U}^R) d\boldsymbol{\beta} d\sigma^2 \tag{3.2}$$

Depending on the regular subdataset ($\mathbf{U}^R$), the regression coefficients ($\boldsymbol{\beta}^R$) are estimated using equation. 2.15 following least square estimation. The shape parameter (a^R) and the scale parameter (b^R) are updated using Eq. 2.6. Since the integral in Eq. 3.2 does not produce a closed form solution. Therefore, the probability of an outlier is computed using the Monte Carlo (MC) simulation technique.

To compute the probability of an outlier, the samples of σ^2 are first drawn from a prior distribution, and the posterior distribution of σ^2 is approximated using the ABC rejection sampling algorithm, as described in subsection 3.3. In this method, the ABC algorithm refines the prior samples by comparing simulated data with observed data and retaining only those samples that closely match. Once the posterior distribution of σ^2 has been approximated, the posterior distribution of the residual ϵ_k^S can be expressed as the product of the conditional posterior of ϵ_k^S given σ^2 and the posterior of σ^2. Specifically, the posterior pdf ($p(\epsilon_k^S, \sigma^2|\mathbf{U}^R)$) can be factorized as $p(\epsilon_k^S|\sigma^2, \mathbf{U}^R)\ p(\sigma^2|\mathbf{U}^R)$. Here, $p(\epsilon_k^S|\sigma^2, \mathbf{U}^R)$ represents a normal distribution with mean ($\hat{\epsilon}_k^S = y_k^S - \mathbf{X}_k^S \hat{\boldsymbol{\beta}}^R$) and variance ($\mathbf{P}_k^S = \sigma^2 \mathbf{X}_k^S [(X^R)^T X^R]^{-1} (\mathbf{X}_k^S)^T$). Using this approximated posterior distribution of σ^2, the corresponding samples of ϵ_k^S are generated from a normal distribution with estimated mean $\hat{\epsilon}_k^S$ and variance $\mathbf{P}_k^S$. Finally, the probability of an outlier for each suspicious data point is estimated by using Eq. (3.3) as follows:

$$P_o(U_k^S|\mathbf{U}^R) = E_(p(\epsilon_k^S, \sigma^2|\mathbf{U}^R))[1 - 2\Phi(-|\epsilon_k^S|/\sigma)]^n \approx \frac{1}{M} \Sigma_i^M [1 - 2\Phi(-|\epsilon_k^{S(i)}|/\sigma^{2(i)})]^n \tag{3.3}$$

Here, M denotes the number of MC runs. After computing the probability of an outlier ($P_o(U_k^S|\mathbf{U}^R)$) for each suspicious point (U_k^S), it is compared to a threshold of 0.5. If $P_o(U_k^S|\mathbf{U}^R) > 0.5$, the data point U_k^S is considered an outlier. Otherwise, it is reclassified as a regular data point.

3.3. ABC rejection sampler

Techniques based on the approximate likelihood function of ABC through simulations, comparing the results of these simulated samples with observed data. ABC rejection sampling is a basic form of this approach. In this method, we first simulate the candidate parameter θ^* from a prior distribution $\pi(\theta)$. This candidate parameter (θ^*) is used to simulate the synthetic dataset (W) from the assumed model, which matches the dimensions of the observed data (Q).

Comparisons between the synthetic and observed data sets are made using a distance metric, such as the squared distance between the summary statistics of the data sets. The summary statistic $S(.)$ is defined as the sample mean. The distance function $d(W, Q)$ is then represented as $(\bar{W} - \bar{Q})^2$. If $d(W, Q) \leq \epsilon_o$, the candidate parameter (θ^*) is retained as a posterior sample. If $d(W, Q) > \epsilon_o$, the candidate parameter (θ^*) is discarded. Here, ϵ_o

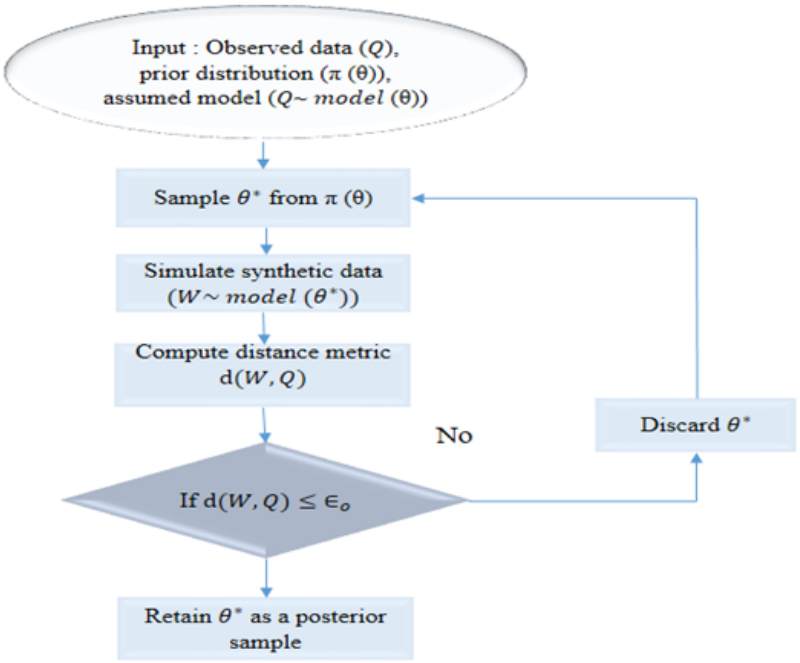

Figure 1. The flowchart of the algorithm of the ABC rejection sampler.

denotes the tolerance value for accepting simulated parameters based on the discrepancy between the simulated and observed data. A lower value of ϵ_o leads to a better posterior approximation but a lower acceptance rate and high computational cost. However, the larger value of ϵ_o yields computational efficiency but may compromise the posterior approximation. An appropriate value is chosen to balance these trade-offs, following the standard practice in ABC literature. The ABC rejection sampler algorithm is discussed in Figure 1.

3.4. Algorithm of the proposed technique

The proposed technique aims to obtain the uncertainty of the parameters after eliminating the anomalies and the problem of ill-conditioned explanatory variables. Firstly, it chooses the reliable portion of the data, termed the regular subset ($\mathbf{U}^R$) from the complete dataset ($\mathbf{U}$). Subsequently, the suspicious subset ($\mathbf{U}^S$) is investigated by considering the probability of outliers. Data points with a probability of outliers less than 0.5 are returned to regular data points ($\mathbf{U}^R$). The algorithm of this technique can be found below and its flow chart can be seen in Figure 2.

1. Obtain the initial regular data set ($\mathbf{U}^R$) in two steps.
   (a) Identify the subset ($\mathbf{U}^H_{MTLE}$) of the entire data ($\mathbf{U}$) by choosing a conservatively small value of h in MTLE following the procedure described in sub-section 2.3.
   (b) Subsequently, remove the leverage points (L) from $\mathbf{U}^H_{MTLE}$ using Eq. 2.12 to obtain $\mathbf{U}^R$.
2. Determine the initial suspicious subset ($\mathbf{U}^S$) by excluding $\mathbf{U}^R$ from the entire data matrix ($\mathbf{U}$) i.e., $\mathbf{U}^S = \mathbf{U} \backslash \mathbf{U}^R$.
3. Estimate the regression parameter vector ($\boldsymbol{\beta}^R$) using PCR.
   (a) Perform PCR on $\mathbf{U}^R$ to address multicollinearity.
   (b) Retain principal components that explain at least 80% of the data variability.

   (c) Estimate $\boldsymbol{\beta}^R$ using the least squares method taking the principal components as explanatory variables (See, Eq. 2.14 and Eq. 2.15).
4. Approximate the posterior distribution of σ^2 using the ABC rejection sampler.
   (a) Draw samples of σ^2 from the prior distribution $\pi(\sigma^2)$ (e.g., inverse Gamma distribution with parameters $\hat{a}$ and $\hat{b}$ defined in Eq. 2.6, taking $a^* = b^* = 1$).
   (b) For each σ^2, simulate the synthetic response variable ($\mathbf{y}_{sn}$).
   (c) Compute the squared difference between the summary statistics (e.g., sample mean) of the synthetic response ($\mathbf{y}_{sn}$) and the observed response ($\mathbf{y}$).
   (d) Retain that candidate of σ^2 if the distance $d(\mathbf{y}_{sn}, \mathbf{y})$ is less than the predefined threshold ϵ_o i.e., $(\bar{\mathbf{y}}_{sn} - \bar{\mathbf{y}})^2 \leq \epsilon_o$.
   (e) Repeat this process to approximate the posterior distribution of σ^2.
5. For each suspicious data point (U_k^S), generate corresponding residual ϵ_k^S from a Gaussian distribution with mean ($\hat{\epsilon}_k^S = y_k^S - \mathbf{X}_k^S \hat{\boldsymbol{\beta}}^R$) and variance ($\mathbf{P}_k^S = \sigma^2 \mathbf{X}_k^S [(\mathbf{X}^R)^T \mathbf{X}^R]^{-1} (\mathbf{X}_k^S)^T$). Here, σ^2 is sampled from the approximated posterior distribution defined in step 4.
6. Calculate the probability of outliers $P_o(U_k^S | \mathbf{U}^R)$ for each suspicious observation (U_k^S) considering Eq. 3.3, which incorporates the approximated posterior distribution of σ^2 and its corresponding residual ϵ_k^S.
7. If $P_o(U_k^S | \mathbf{U}^R) \geq 0.5$, retain U_k^S in the suspicious data set $\mathbf{U}^S$. Otherwise, reclassify $\mathbf{U}^S$ as a regular data point and move it to $\mathbf{U}^R$.
8. Update the regular and suspicious data sets $\mathbf{U}^R$ and $\mathbf{U}^S$. Repeat steps 3–7 until all suspicious data points satisfy $P_o(U_k^S | \mathbf{U}^R) \geq 0.5$. Convergence is typically achieved in two or three iterations.
9. The final regression parameters ($\boldsymbol{\beta}^R$) are based on the final regular data set ($\mathbf{U}^R$). Any data point that remains in the final suspicious dataset $\mathbf{U}^S$ is considered a potential outlier, with an outlier probability $P_o(U_k^S | \mathbf{U}^R)$.

4. Simulation study

4.1. Data generation

This subsection presents various data simulation scenarios designed to assess the performance of the proposed technique against existing methods, including the Huber estimator, LTS, and RobPCR. The data generation process is structured to simulate challenging conditions that include high levels of contamination, high correlation fractions, and different types of outliers. The goal is to demonstrate that the proposed technique outperforms baseline counterparts under extreme outlier conditions and in the presence of collinear predictors.

Two types of outliers are introduced in the simulations. The first type arises from a high level of measurement noise, and the other one from modelling errors. Both types of outliers lead to abnormally large errors, but the second type introduces more substantial bias compared to the first. To model these outliers, two sub-datasets are generated. The contaminated set ($\mathbf{U}^1$) consists of the regular sub-dataset mixed with the outliers of the first type, while the disordered dataset ($\mathbf{U}^2$) comprise outliers of the second type. The total data set $\mathbf{U}$ combines both $\mathbf{U}^1$ and $\mathbf{U}^2$, simulating the coexistence of both types of outliers.

The contaminated set ($\mathbf{U}^1$) is generated using a weighted mixture model, as defined in Eq. (4.1) to simulate the error terms of data points in $\mathbf{U}^1$. Here, the contamination level α controls the proportion of outliers in the data set. The error terms in $\mathbf{U}^1$ follow a mixture distribution defined in Eq. 4.1. Here, regular data points are generated from a Gaussian distribution ($G(\tilde{e}|\mathbf{0}, \tilde{\sigma}^2)$), while the outliers are drawn from a mixture of triangular distribution ($f(\tilde{e})$, as defined in Eq. (4.2). The triangular distribution $T(\tilde{e}|a, b, c)$ is

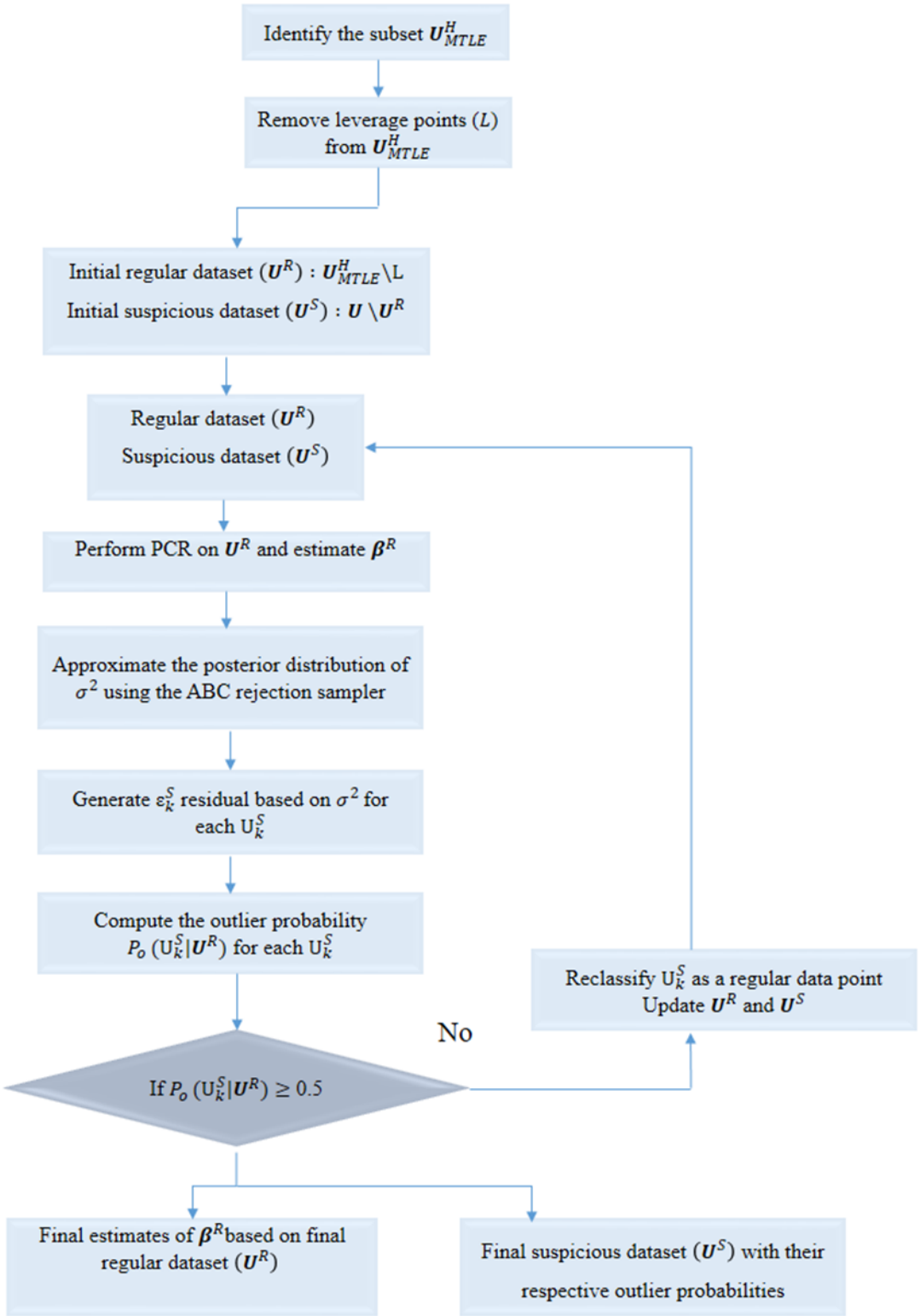

Figure 2. The flowchart of the algorithm of the proposed PRPCR.

defined in Eq. 4.3 and approaches its maximum at $\tilde{e} = b$.

$$p(\tilde{e}) = (1-\alpha)\ G(\tilde{e}|\mathbf{0}, \tilde{\sigma}^2) + \alpha f(\tilde{e}) \tag{4.1}$$

$$f(\tilde{e}) = 0.5\ T(\tilde{e}| - 5\tilde{\sigma}, -4\tilde{\sigma}, -3\tilde{\sigma}) + 0.5\ T(\tilde{e}|5\tilde{\sigma}, 4\tilde{\sigma}, 3\tilde{\sigma}) \tag{4.2}$$

$$T(\tilde{e}|a,b,c) = \begin{cases} 2(\tilde{e}-a)/(c-a)(b-a) & a \le \tilde{e} \le b \\ 2(c-\tilde{e})/(c-a)(c-b) & b \le \tilde{e} \le c \end{cases} \tag{4.3}$$

To simulate the disordered data set ($\mathbf{U}^2$), a substantial bias is introduced into the data. The response variable (y_i) and explanatory variables ($\mathbf{x}_i^T$) in $\mathbf{U}^2$ are generated from a Gaussian distribution with mean (μ_i) and covariance ($\sigma^2\mathbf{I}$). The total number of outliers (N_o) in the whole data set is defined in Eq. (4.4). Here, N_1 and N_2 denotes number of observations in $\mathbf{U}^1$ and $\mathbf{U}^2$, respectively. Then, we have

$$N_o = \alpha N_1 + N_2 \tag{4.4}$$

The explanatory variables in $\mathbf{U}^1$ are simulated using Eq.4.5 following [5]. Here, ρ represents the correlation between two explanatory variables and D_{ij} are independent pseudo-random numbers derived from the standard normal distribution. This introduces control multicollinearity, which varies between scenarios to assess its effect on the robustness of the proposed method. The response variable (y_i) for the observations in $\mathbf{U}^1$ is generated according to Eq. (4.6). Here, $\tilde{e}_i$ has the weighting mixture distribution, as defined in Eq. (4.1). Then, we obtain

$$x_{ij} = (1-\rho^2)^{1/2} D_{ij} + \rho D_{(i,p+1)},\ i = 1,2,3,\ldots,N_1 \text{ and } j = 1,2,3,\ldots,p \tag{4.5}$$

$$y_i = \Sigma_{j=1}^{p} \beta_j x_{ij} + \tilde{e}_i,\ i = 1,2,3,\ldots,N_1 \tag{4.6}$$

To assess the performance of the proposed technique over baseline counterparts, four simulation scenarios are designed, each of which tests specific aspects of robustness against anomalous data and ill-conditioned predictors.

Scenario I involves a sample size $n = 40$, with 33 observations in $\mathbf{U}^1$ (including 4 outliers of the first type with contamination fraction $\alpha = 0.1$) and 7 outliers of the second type in $\mathbf{U}^2$. The number of predictors is $p = 3$, and the correlation between the predictors is set to $\rho = 0.8$. The regression coefficients for $\mathbf{U}^1$ are $\beta_1 = 1, \beta_2 = 2$ and $\beta_3 = 3$. The disordered data set $\mathbf{U}^2 = (\mathbf{X}, \mathbf{y})$ is normally distributed with mean $\mu = [70, 50, 40, -10]$ and covariance matrix $\mathbf{\Sigma} = 25\mathbf{I}$. This scenario tests the ability of the method to handle small sample sizes and strong multicollinearity. Scenario II is similar to scenario I but with an increased level of multicollinearity, setting the correlation between predictors at $\rho = 0.9$. This scenario evaluates how the method compares with baseline techniques under high multicollinearity. Scenario III modifies scenario I by increasing the sample size to $n = 70$, with $N_1 = 63$ (including 4 outliers of the first type with $\alpha = 0.06$) and $N_2 = 7$. This scenario examines the scalability of the proposed method as the sample size increases, with a focus on robustness to outliers and leverage points. Scenario IV introduces five predictors, with regression coefficients $\beta = (1,2,3,1,2)^T$. The mean vector for $\mathbf{U}^2 = (\mathbf{X}, \mathbf{y})$ is $\mu = [70, 50, 40, 30, 60, -10]$, while the other settings remain the same as in scenario I. This scenario tests the performance of the proposed method with an increased number of predictors, compared to the baseline methods. Scenarios V and VI are modified versions of Scenario IV, which incorporate correlation levels of $\rho = 0.9$ and $\rho = 0.95$, respectively.

4.2. Performance evaluation measures

We use the Mean Square Error (MSE) of the estimated regression parameters to evaluate the effectiveness of the proposed technique in terms of robust parameter identification. The MSE is defined in Eq. (4.7) and has been widely used by researchers ([24, 39]) as an

Table 1. The estimates of regression parameters with their respective MSE for different techniques regarding scenario I.

Method	PRPCR	Huber	LTS	RobPCR
β_1	1.41604	-0.53815	-0.63383	-0.20220
β_2	2.19267	0.12232	1.36812	-0.00760
β_3	2.60138	0.55802	1.97541	0.12904
MSE($\hat{\boldsymbol{\beta}}$)	2.20821	4.29534	6.08967	4.65631

evaluation criterion. A lower MSE value indicates a more accurate and reliable estimator. Then, we have

$$MSE(\hat{\boldsymbol{\beta}}) = \frac{1}{p}\ \Sigma_{l=1}^{p}(\hat{\beta}_l - \beta_l)^2 \tag{4.7}$$

In addition, three metrics are used to compare the performance of the proposed technique in outlier detection. These metrics are masking percentage, swamping percentage, and percentage of correctly identified outliers. Masking occurs when an actual outlier is not identified correctly, leading to it being missed. The masking percentage is calculated by dividing the number of masked outliers by the total number of actual outliers. Swamping, on the other hand, refers to cases in which a regular observation is mistakenly classified as an outlier. The percentage of swamping is computed by dividing the number of swamped observations by the total number of regular observations. Finally, the percentage of correctly identified outliers is determined by dividing the number of correctly detected outliers by the total number of actual outliers. Lower values of the masking and swamping percentages are desirable. This is because they indicate fewer errors in missing outliers or incorrectly identifying regular observations as outliers. On the other hand, a higher percentage of correctly identified outliers is preferred. This reflects a more accurate and reliable outlier detection process. Conventional techniques, such as the Huber estimator, LTS, and RobPCR, generally use a simple threshold to detect outliers. Specifically, they are based on a rule in which an observation is flagged as an outlier if $\frac{|\epsilon|}{\sigma} > 2.5$. However, the proposed method takes a more refined approach. It employs a sophisticated algorithm, described in subsection , which enhances both the accuracy and robustness of outlier identification and parameter estimation.

5. Results and discussion

This section presents the results for scenarios I-VI as discussed earlier. The performance of the proposed technique, PRPCR, is compared to classical methods such as Huber estimator, LTS, and RobPCR. For parameter identification, the comparison is made using the MSE of the estimated regression coefficients, denoted as MSE($\hat{\boldsymbol{\beta}}$). In addition, the performance of outlier detection is evaluated using metrics such as masking, swamping, and the percentage of correct outlier identification. All results are based on 100 Monte Carlo simulations, with average MSE values ($\hat{\boldsymbol{\beta}}$) presented in Tables 1 to 6 and Figure 3. All computations are performed using the R programming language.

Tables 1 to 6 provide estimates of regression parameters along with their respective MSE values for the different techniques. In all scenarios, the proposed method (PRPCR) consistently demonstrates favorable performance compared to the baseline methods. It achieves the lowest MSE values, reflecting more accurate parameter estimation. In scenario I in Table 1, PRPCR exhibits significantly lower MSE than the Huber estimator, LTS, and RobPCR. This pattern continues in Scenarios II (Table 2), III (Table 3), IV (Table 4), V (Table 5) and VI (Table 6). These results suggest that PRPCR is particularly robust across a range of challenging conditions, including multicollinearity, outliers and small sample sizes.

Table 2. The estimates of regression parameters with their respective MSE for different techniques regarding scenario II.

Method	PRPCR	Huber	LTS	RobPCR
β_1	1.62498	-0.50272	-1.47959	-0.05938
β_2	2.02728	0.11117	1.51853	-0.06612
β_3	2.41077	0.49982	2.86190	-0.05546
MSE($\hat{\boldsymbol{\beta}}$)	2.78721	4.45396	9.95846	4.91042

Table 3. The estimates of regression parameters with their respective MSE for different techniques regarding scenario III.

Method	PRPCR	Huber	LTS	RobPCR
β_1	1.09499	-0.74040	-0.21586	-0.26198
β_2	2.01070	0.14153	1.41100	0.00055
β_3	2.80065	0.88914	2.63472	0.23407
MSE($\hat{\boldsymbol{\beta}}$)	0.88033	4.07584	4.32620	4.58448

As the sample size increases in scenario III, the MSE decreases for all methods. However, PRPCR shows the most significant improvement. This sharp reduction in MSE indicates the scalability of PRPCR, which becomes more efficient as the size of the dataset grows. In contrast, the Huber estimator, LTS, and RobPCR exhibit more modest improvements, as their MSE values remain relatively high despite the increase in sample size.

In scenario IV-VI, where the number of predictors is increased and the levels of correlation between predictors are high, PRPCR continues to perform well. It maintains a low MSE even in this more complex, high-dimensional setting and ill-conditioned predictors. However, the baseline methods struggle to cope with the increased complexity, as evidenced by their much higher MSE values. The ability of PRPCR to handle high-dimensional and collinear data effectively highlights its robustness and versatility.

Figure 3 reports the results for the outlier detection performance. The metrics evaluated include the percentages of swamping, masking, and correct outlier identification. PRPCR consistently outperforms baseline methods in outlier detection in all scenarios. In Scenario I (see Figure 3 (a)), PRPCR correctly identifies the majority of true outliers, performing much better than the Huber estimator and RobPCR. These methods detect only a small fraction of the outliers. LTS performs slightly better than the other baseline methods, but is still less effective compared to PRPCR.

As multicollinearity increases in scenario II, PRPCR continues to perform robustly. It maintains a high percentage of correct outlier identification, despite the stronger correlation between predictors (See, Figure 3 (b)). In contrast, the Huber estimator and RobPCR continue to struggle with masking and misidentification. LTS shows a slight improvement in scenario II but remains less effective than PRPCR in detecting outliers. The resilience of PRPCR to multicollinearity is particularly noteworthy, as it maintains its accuracy under conditions that typically challenge other techniques.

In scenario III, the increase in sample size further highlights the scalability of PRPCR (See, Figure 3 (c)). As the data set grows, PRPCR continues to correctly identify most outliers. In contrast, baseline methods show limited improvement, with the Huber estimator and RobPCR missing a significant number of outliers. LTS also improves slightly but remains less effective than PRPCR in both detecting outliers and minimizing masking.

One notable observation is that PRPCR exhibits slightly higher flood percentages than some of the other techniques in specific scenarios. For instance, in scenario I, PRPCR incorrectly classifies a small portion of regular observations as outliers (See, Figure 3 (a)).

Table 4. The estimates of regression parameters with their respective MSE for different techniques regarding scenario IV.

Method	PRPCR	Huber	LTS	RobPCR
β_1	1.02538	-0.90464	-1.31541	-0.19590
β_2	1.91603	0.13745	1.15565	0.03341
β_3	2.61959	0.84009	2.80794	-0.11644
β_4	1.12684	0.98245	2.18452	0.12609
β_5	1.97761	-0.24281	0.57275	-0.03875
MSE($\hat{\boldsymbol{\beta}}$)	1.92190	4.53691	8.66796	3.88487

Table 5. The estimates of regression parameters with their respective MSE for different techniques regarding scenario V.

Method	PRPCR	Huber	LTS	RobPCR
β_1	1.44945	-1.13963	-1.81763	-0.05728
β_2	1.70711	0.21365	1.30130	-0.02774
β_3	2.48359	0.90165	3.61647	-0.00615
β_4	1.53601	1.43184	2.62697	-0.00997
β_5	1.82448	-0.31031	0.67037	-0.05544
MSE($\hat{\boldsymbol{\beta}}$)	2.01708	5.49700	16.86565	3.91835

Table 6. The estimates of regression parameters with their respective MSE for different techniques regarding scenario VI.

Method	PRPCR	Huber	LTS	RobPCR
β_1	1.39977	-1.16495	-3.80404	-0.03744
β_2	1.53073	0.49822	1.69527	-0.03752
β_3	2.43001	0.96838	4.77208	-0.03821
β_4	1.38702	1.462437	4.10974	-0.03785
β_5	2.11703	-0.57633	0.51420	-0.03787
MSE($\hat{\boldsymbol{\beta}}$)	3.64084	5.57154	36.87591	3.93777

However, this slight increase in swamping is outweighed by PRPCR's much lower masking rates and its consistently high rate of correctly identifying actual outliers. In Scenario IV-IV, where the number of predictors and correlation levels increases, PRPCR reduces its swamping percentage to its lowest level (see Figure 3 (d-f)). This shows that PRPCR adapts effectively to more complex models with additional predictors.

6. Application

The performance of the proposed method is demonstrated using the pollution dataset, which reflects the issues discussed in this study. This data set has been used by various researchers (e.g. [5, 6]) and consists of fifteen predictors. The goal is to predict the age-adjusted mortality rate per 100,000 population ($\mathbf{y}$). Detailed descriptions of the covariates can be found in previous studies (e.g.[41, 60]). Previous studies have identified severe multicollinearity and outliers in this data set [5], evidenced by variance inflation factors that reach 98.6 for $\mathbf{x}_{12}$ and 104.9 for $\mathbf{x}_{13}$. The correlation concentration between the predictors is presented in Figure 4.

As this paper proposes a robust approach for parameter identification and outlier detection, the outliers in this dataset are identified using the proposed method and also their competing techniques. The proposed method also provides the probability that each

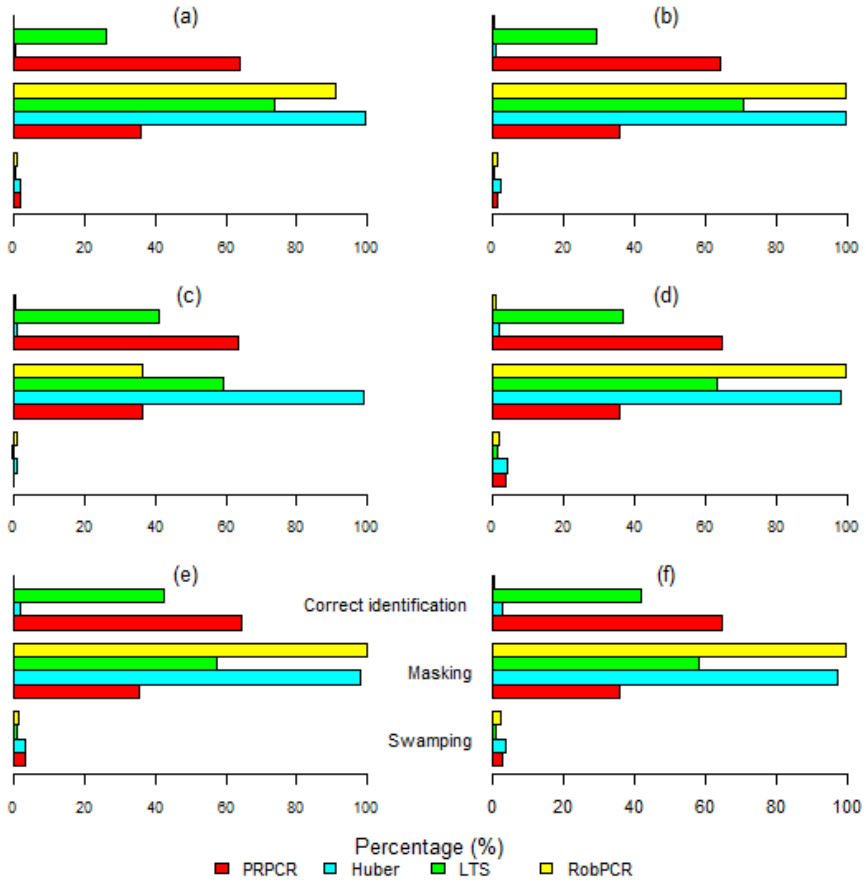

Figure 3. The percentages of correct identification, masking and swamping of outliers for the proposed technique and existing ones regarding scenario I (a), scenario II (b), scenario III (c), scenario IV (d), scenario V (e) and scenario VI (f).

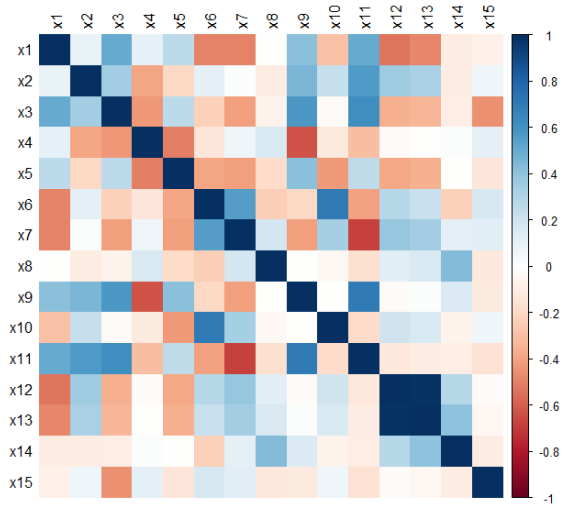

Figure 4. The graphical representation of the correlation matrix of fifteen predictors of pollution data.

Table 7. The identified outliers through several studied techniques.

Method	outliers
PRPCR	29 (0.8339903)
Huber	2, 28, 37, 59
LTS	32, 59
RobPCR	No outlier

Table 8. The estimates of regression parameters with their respective MVB for different techniques regarding the pollution dataset.

Method	PRPCR	Huber	LTS	RobPCR
β_1	0.10609	0.29701	0.36944	0.03579
β_2	0.08648	-0.24200	-0.31981	0.07411
β_3	0.04368	-0.2230519	-0.23401	0.02691
β_4	-0.01592	-0.20577	-0.28303	-0.07177
β_5	0.02457	-0.19382	-0.31532	0.08479
β_6	-0.19678	-0.20987	-0.23938	-0.20654
β_7	-0.09023	-0.11700	0.034223	-0.08505
β_8	0.20569	0.14191	0.28674	0.12361
β_9	0.18117	0.57831	0.44908	0.12018
β_{10}	-0.07083	-0.04725	-0.03868	-0.14046
β_{11}	0.13785	-0.08816	0.24593	0.11929
β_{12}	0.04174	-0.78242	-1.14211	0.02727
β_{13}	0.07745	0.76737	1.12630	0.05206
β_{14}	0.25770	0.12681	0.02428	0.21179
β_{15}	0.04765	-0.02347	0.02395	-0.02112
MVB($\hat{\boldsymbol{\beta}}$)	0.00370	0.04575	0.43073	0.00403

suspicious data point is an outlier. It can be seen from Table 7, that RobPCR does not detect outliers, despite their presence in the data. LTS identifies observations 32 and 59 as outliers, while the Huber estimator identifies observations 2, 28, 37 and 59 as outliers. The proposed method detects an outlier in observation 29, with its respective probability reported in parentheses.

Additionally, the parameter identification calculated using the proposed method and existing techniques are presented in Table 8. The bootstrap variance of each regression coefficient is also calculated. The median bootstrap variations of all regression parameters MBV($\hat{\boldsymbol{\beta}}$) for all studied estimators are discussed in Table 8. The proposed method outperforms competing techniques, achieving the lowest value of MBV($\hat{\boldsymbol{\beta}}$).

7. Conclusion

In conclusion, this paper proposes a probabilistic approach that combines principal component regression and approximate Bayesian computation to address outliers, leverage points, and multicollinearity issues in linear regression problems. The proposed method, PRPCR, enhances the efficiency of parameter estimation and the capability of outlier detection. Extensive simulations demonstrate PRPCR's favorable performance over baseline counterparts in both parameter identification and outlier detection. The technique proves to be robust even in challenging scenarios characterized by high correlation, a large number of predictors, and the presence of significant leverage points with large residuals. The application of PRPCR to the pollution data set further validates its effectiveness,

consistent with the results of the empirical study.

Acknowledgements

The author(s) would like to sincerely thank the reviewers for their valuable and constructive feedback.

Author contributions. Aiman Tahir was responsible for conducting the formal analysis and played a key role in developing the methodology. Additionally, Aiman contributed to the writing of the manuscript. Maryam Ilyas led the conceptualization of the study, contributed to the development of the methodology, and was involved in validating the results.

Conflict of interest statement. The authors declare no competing interests.

Funding. This research received no external funding.

Data availability. The published data has been used and complete reference has been provided in this manuscript.

References

[1] A. Alin, *Multicollinearity*, Wiley Interdisciplinary Reviews: Computational Statistics **2** (3), 370-374, 2010.

[2] T.W. Anderson, *An Introduction to Multivariate Statistical Analysis*, Wiley, 2003.

[3] M. Alshakhs, P.J. Goedecke, J.E. Bailey and C. Madlock-Brown, *Racial differences in healthcare expenditures for prevalent multimorbidity combinations in the usa: A cross sectional study*, BMC Medicine **21**(1), 399, 2023.

[4] K.C. Arum and F.I. Ugwuowo, *Combining principal component and robust ridge estimators in linear regression model with multicollinearity and outlier*, Concurr. Comput. Pract. Exp. **34**(10), 2022.

[5] K.C. Arum, F.I. Ugwuowo, H.E. Oranye, T.O. Alakija, T.E. Ugah and O. C. Asogwa, *Combating outliers and multicollinearity in linear regression model using robust Kibria-Lukman mixed with principal component estimator, simulation and computation*, Sci. Afr. **19**, 2023.

[6] F.A. Awwad, I. Dawoud and M.R. Abonazel, *Development of robust Özkale–Kaçiranlar and Yang–Chang estimators for regression models in the presence of multicollinearity and outliers.* Concurr. Comput. Pract. Exp. **34**(6), 2022.

[7] V. Barnett, and T. Lewis, *Outliers in statistical data*, John Wiley and Sons, New York, 1994.

[8] M.J. Bayarri and J. Morales, *Bayesian measures of surprise for outlier detection.* J. Stat. Plan. Inference **111**(1-2), 3-22, 2003.

[9] J.O. Berger, E. Moreno, L.R. Pericchi, M.J. Bayarri, J.M. Bernardo, J.A. Cano and S. Sivaganesan, *An overview of robust Bayesian analysis.* Test **3**(1), 5-124, 1994.

[10] M.A. Beaumont, W. Zhang and D.J. Balding, *Approximate Bayesian computation in population genetics.* Genetics **162**(4), 2025-2035, 2002.

[11] L.D. Broemeling, *Bayesian analysis of linear models.* CRC Press, 2017.

[12] K. Chaloner and R. Brant, *A Bayesian approach to outlier detection and residual analysis.* Biometrika **75**(4), 651-659, 1988.

[13] G.E. Box, *Sampling and Bayes' inference in scientific modelling and robustness.* J. R. Stat. Soc.,A: Stat.Soc.**143**(4), 383-404, 1980.

[14] G.E. Box and G.C. Tiao, *A Bayesian approach to some outlier problems*. Biometrika **55**(1), 119-129, 1968.
[15] R.D. Cook and L. Forzani, *Partial Least Squares Regression: and Related Dimension Reduction Methods*, CRC Press, 2024.
[16] M. Denhere and N. Billor, *Robust principal component functional logistic regression*, Commun.Stat.–Simul.Comput. **45** (1) 264-281, 2016.
[17] H. Dong, T. Tong, C. Ma and Y. Chi, *Fast and provable tensor robust principal component analysis via scaled gradient descent*, Inf. Inference: A J. of the IMA **12**(3), 1716-1758, 2023.
[18] A. Ebiwonjumi, R. Chifurira and K. Chinhamu, *A robust principal component analysis for Estimating economic growth in Nigeria in the presence of multicollinearity and outlier*, J. Stat. Appl. Probab. **12**(2), 611–627, 2023.
[19] S. Engelen, M. Hubert, K. V. Branden, and S. Verboven, *Robust PCR and Robust PLSR: a comparative study*, In Theory and applications of recent robust methods 105-117, Birkhäuser, Basel, 2004.
[20] W.J. Egan, and S.L. Morgan, *Outlier detection in multivariate analytical chemical data*, Anal.Chem. **70** (11), 2372-2379, 1998.
[21] P. Filzmoser, *Robust principal component regression. Computer data analysis and modeling. Robust and computer intensive methods*, Belarusian State University, Minsk 132-137, 2001.
[22] P. Gagnon, M. Bédard, and A. Desgagné, *An automatic robust Bayesian approach to principal component regression*, J. Appl. Stat. **48** (1), 84-104, 2021.
[23] P. Gagnon, A. Desgagné and M. Bédard, *A new Bayesian approach to robustness against outliers in linear regression*, Bayesian Anal. **15**(2), 389-414, 2020.
[24] D.G. Gibbons, *A simulation study of some ridge estimators*, J.Am.Stat.Assoc. 76(373), pp.131-139, 1981.
[25] I. Guttman, *Care and handling of univariate or multivariate outliners in detecting spuriosity-a Bayesian approach*, Technometrics **15**(4), 723-738, 1973.
[26] I. Guttman, R. Dutter and P.R. Freeman, *Care and handling of univariate outliers in the general linear model to detect spuriosity—A Bayesian approach* , Technometrics **20**(2), 187-193, 1978.
[27] Y. Hamura, K. Irie, and S. Sugasawa, *Posterior robustness with milder conditions: Contamination models revisited*, Stat.Probab.Lett. 210, 2024.
[28] S. Hashimoto, S. Sugasawa, *Robust Bayesian regression with synthetic posterior distributions*, Entropy **22**(6), 661, 2020.
[29] D. C. Hoaglin and R. E. Welsch, *The hat matrix in regression and ANOVA*, Am. Stat. **32** (1), 17-22, 1987.
[30] P.J. Huber, *Robust regression: asymptotics, conjectures and Monte Carlo*, Ann. Stat., 799-821, 1973.
[31] P. Huber, *Robust statistics*, New York: John wiley and son, 1981.
[32] M. Hubert and S. Verboven, *A robust PCR method for high-dimensional regressors*, Journal of Chemometrics: A Journal of the Chemometrics Society **17** (8-9), 438-452, 2003.
[33] M. Hubert, P. J. Rousseeuw and K. V. Branden, ROBPCA: a new approach to robust principal component analysis, Technometrics **47** (1), 64-79, 2005.
[34] W. Johnson and S. Geisser. *A predictive view of the detection and characterization of influential observations in regression analysis*, J. Am. Stat. Assoc. **78**(381), 137-144, 1983.
[35] E.T. Jaynes, *Probability theory: The logic of science*. Cambridge university press, 2003.
[36] I. T. Jolliffe, Principal components in regression analysis. In Principal Component Analysis, 129-155, Springer, New York, 1986.

[37] K.L. Lange, R.J. Little and J.M. Taylor, *Robust statistical modeling using the t distribution*, J. Am. Stat. Assoc. **84**(408), 881-896, 1989.

[38] G. Li and Z. Chen, *Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and Monte Carlo*, J. Am. Stat. Assoc. **80**(391), 759-766, 1985.

[39] A.F. Lukman, R.A. Farghali, B.G. Kibria and O.A. Oluyemi, *Robust-stein estimator for overcoming outliers and multicollinearity*, Sci. Rep. **13**(1), 2023.

[40] W.F. Massy, *Principal components regression in exploratory statistical research*, J. Am. Stat. Assoc. **60** (309), 234–256, 1965.

[41] G.C. McDonald and R.C. Schwing, *Instabilities of regression estimates relating air pollution to mortality*, Technometrics **15** (3), 463-81, 1973.

[42] M. T. Molebatsi, *Handling of multicollinearity problem in modelling non- performing loans in africa's portfolio data [Doctoral dissertation]*, 2023.

[43] E.J. Montenegro, J.E. Pitti and B.O. Olivares, *Identification of the main subsistence crops of teribe: A case study based on multivariate techniques*, Idesia **39**(3), 83–94, 2021.

[44] B.O. Olivares, J.E. Pitti and E.J. Montenegro, *Socioeconomic characterization of bocas del toro in panama: An application of multivariate techniques*, Rev. Bras. Gest. Desenvolv. Reg. **16**(3), 59–71, 2020.

[45] R. J. Pell, *Multiple outlier detection for multivariate calibration using robust statistical techniques*, Chemom. Intell. Lab. Syst. **52** (1), 87-104, 2000.

[46] D. PEna, and V. Yohai, *A fast procedure for outlier diagnostics in large regression problems*, J. Am. Stat. Assoc. **94** (446), 434-445, 1999.

[47] P. J. Rousseeuw and A. M. Leroy, *Robust regression and outlier detection*, John Wiley & Sons, 1987.

[48] P. J. Rousseeuw, *Least median of squares regression*, J. Am. Stat. Assoc. **79** (388), 871-880, 1984.

[49] P.J. Rousseeuw, and K. Van Driessen, *Computing lts regression for large datasets*, Data Min. Knowl. Discov. **12**, 29–45, 2006.

[50] D.B. Rubin, *Bayesianly justifiable and relevant frequency calculations for the applied statistician*, Ann. Stat. 1151-1172, 1984.

[51] N. Shrestha, *Detecting multicollinearity in regression analysis*, Am. J. Appl. Math. Stat. **8** (2), 39-42, 2020.

[52] G.N. Singh, D. Bhattacharyya and A. Bandyopadhyay, *Robust estimation strategy for handling outliers*, Commun. Statis.-Theor. Meth. **53**(15), 5311-5330, 2024.

[53] A. Tahir and M. Ilyas, *Robust correlation scaled principal component regression*, Hacet. J. Math. Stat. **52** (2), 459–486, 2023.

[54] B. Walczak and D. L. Massart, *Robust principal components regression as a detection tool for outliers*, Chemom. Intell. Lab. Syst. **27** (1), 41-54, 1995.

[55] M. West, *Outlier models and prior distributions in Bayesian linear regression*, J. R. Stat. Soc., B: Stat. Methodol. **46**(3), 431-439, 1984.

[56] K. Worden, G. Manson and N.R. Fieller, *Damage detection using outlier analysis*, J. Sound Vib., 229(3), 647-667, 2000.

[57] S. Xiao, L. Cheng, C. Ma, J. Yang, X. Xu and J. Chen, *An adaptive identification method for outliers in dam deformation monitoring data based on Bayesian model selection and least trimmed squares estimation*, J. Civ. Struct. Health Monit. 1-17, 2024.

[58] S. Yonekura, and S. Sugasawa, *Adaptation of the tuning parameter in general Bayesian inference with robust divergence*, Stat. Comput. **33**(2), 39, 2023.

[59] K.V. Yuen and H.Q. Mu, *A novel probabilistic method for robust parametric identification and outlier detection*, Probabilistic Eng. Mech. **30**, 48–59, 2012.

[60] B. Yüzbaşı, M. Arashi and S. Ejaz Ahmed, *Shrinkage Estimation Strategies in Generalised Ridge Regression Models: Low/High-Dimension Regime*, Int. Stat. Rev. **88**(1), 229-51, 2020.

[61] A. Zellner, *Bayesian analysis of regression error terms*, J. Am. Stat. Assoc. **70**(349), 138-144, 1975.

[62] M. H. Zhang, Q. S. Xu and D. L. Massart, *Robust principal components regression based on principal sensitivity vectors*, Chemom. Intell. Lab. Syst. **67** (2), 175-185, 2003.