Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 1899225, 12 pages
https://doi.org/10.1155/2021/1899225

Research Article

A Novel Approach for Outlier Detection in Multivariate Data

Saima Afzal,[1] **Ayesha Afzal**,[2] **Muhammad Amin**,[3] **Sehar Saleem**,[4] **Nouman Ali**,[5] **and Muhammad Sajid**[6]

[1]*Department of Statistics, Bahauddin Zakariya University, Multan 60000, Pakistan*
[2]*Department of Computer Science, Air University Multan Campus, Multan 60000, Pakistan*
[3]*Department of Statistics, University of Sargodha, Sargodha, Pakistan*
[4]*Department of Statistics, Lahore College for Women University, Lahore, Pakistan*
[5]*Department of Software Engineering, Mirpur University of Science & Technology (MUST), Mirpur 10250, AJK, Pakistan*
[6]*Department of Electrical Engineering, Mirpur University of Science & Technology (MUST), Mirpur 10250, AJK, Pakistan*

Correspondence should be addressed to Muhammad Amin; muhammad.amin@uos.edu.pk

Received 15 July 2021; Revised 13 September 2021; Accepted 15 September 2021; Published 7 October 2021

Academic Editor: Ishfaq Ahmad

Outlier detection is a challenging task especially when outliers are defined by rare combinations of multiple variables. In this paper, we develop and evaluate a new method for the detection of outliers in multivariate data that relies on Principal Components Analysis (PCA) and three-sigma limits. The proposed approach employs PCA to effectively perform dimension reduction by regenerating variables, i.e., fitted points from the original observations. The observations lying outside the three-sigma limits are identified as the outliers. This proposed method has been successfully employed to two real life and several artificially generated datasets. The performance of the proposed method is compared with some of the existing methods using different performance evaluation criteria including the percentage of correct classification, precision, recall, and *F*-measure. The supremacy of the proposed method is confirmed by abovementioned criteria and datasets. The *F*-measure for the first real life dataset is the highest, i.e., 0.6667 for the proposed method and 0.3333 and 0.4000 for the two existing approaches. Similarly, for the second real dataset, this measure is 0.8000 for the proposed approach and 0.5263 and 0.6315 for the two existing approaches. It is also observed by the simulation experiments that the performance of the proposed approach got better with increasing sample size.

1. Introduction

In most real-life datasets, there exist data observations that do not conform to general model and/or behavior of the data. Such observations that are significantly inconsistent with the majority of the observations in the dataset are known as outliers. Outlier detection problem needs to be addressed in a wide range of applications in fraud detection (e.g., suspicious use of credit cards or other kinds of financial transactions), health data analysis (e.g., detecting unusual responses to treatment plans among patients), fault detection in production processes, and network intrusion detection, etc. Moreover, several data analysis tasks are influenced due to the presence of outliers and require minimizing the effect of outlier observations or eliminating them all together. The problem of detecting outliers in multivariate data is a nontrivial task that becomes even more problematic in case of high dimensional datasets.

Existing techniques for the general outlier detection problem can be broadly categorized in four key approaches including statistical distribution-based approaches, distance-based approaches, density-based approaches, and the subspace-learning based approaches [1–4].

The *statistical distribution-based approaches* consider a distribution or probability model (such as normal distribution or Poisson distribution) for the given dataset to find any outlier observations with reference to the selected model by employing a "discordance test" with respect to some known parameters of the dataset, e.g., the mean, variance, and/or an assumed data distribution [3]. Most approaches in this category are designed for univariate datasets, i.e., having a single attribute; however, several problems involve outlier

detection in multidimensional datasets. Zhao et al. [5] presented COPOD outlier detection method that was motivated from statistical methods to model multivariate data distribution. COPOD first builds the empirical copula and then makes use of the fitted model for the prediction of tail probabilities of each data observation to classify it as a regular or outlier observation. A key concern with approaches relying on the statistical distribution of the dataset is that the statistical distribution and related parameters regarding the dataset model may not always be known a priori. Moreover, the statistical parameters of the dataset can also influence the outlier detection to the masking or swamping effect.

Distance-based approaches rely on the distances between observations to detect outliers. Data observations that do not have enough neighboring observations within a distance threshold are considered outliers [3, 6]. The first effort towards outlier detection in multivariate functional data using graphical tools was the Functional Outlier Map (FOM) approach [7, 8]. These methods utilize statistical depth functions and distance measures derived from them for outlier detection. Prykhodko et al. addressed outlier detection in multivariate nonnormal data based upon univariate and multivariate normalizing transformations [9]. They used squared Mahalanobis distance and a quantile of the Chi-Square distribution for the purpose. In a recent work, Caberoa et al. presented an outlier detection method [10] that performs archetype analysis to combine projections into relevant subspaces with a nearest-neighbor algorithm. In addition to their reliance on statistical characteristics of the dataset (e.g., mean values), a key concern with distance-based approaches is due to their reliance on the global information of the dataset as their performance depends on the neighborhood size of observations.

Density-based approaches [11] rely on the local outlier factors of data points computed by considering the local density of their neighborhoods. While approaches in this category achieve good accuracy without making any assumptions about the dataset distribution, these approaches have high computational complexity especially for high-dimensional large datasets.

Among *subspace-learning based approaches* [11, 12] for outlier detection, Zhao et al. proposed LOMA [12], a local outlier detection approach for massive high-dimensional datasets. LOMA performs data reduction by employing attribute relevance analysis. Further, it employs particle swarm optimization for efficient searching of sparse subspace, where the data density, i.e., the number of observations in the dataset, is very small. Our proposed method is somewhat similar as it also performs subspace-learning; however, unlike attribute relevance analysis, we employ principal component analysis for dimension reduction.

Most of the existing outlier detection methods are either designed for univariate datasets or require a large number of data points to perform effectively. For example, in case of distance-based methods, it is difficult to identify outliers simply by computing distances from the few available data points to mean value. Moreover, existing approaches considering the entire variable set are computationally expensive when considered for high dimensional datasets. However, multivariate datasets with high dimensionality with varying sizes in terms of their number of instances are often encountered in real life data analytics situations.

We employ PCA with three-sigma limits for the identification of outliers. PCA is one of the most prevalent linear dimension reduction techniques. It reduces the dimensionality of high-dimension multivariate datasets, with minimum loss of information. It works by producing new uncorrelated variables that successively maximize variance. The new variables are the linear combinations of all the original variables. The methods based upon graphs are useful tools for identifying outliers in multivariate data, especially when we are working on PCs, but they may not be effective for applications of real time detections. The validity of the existing formal tests is based upon some assumptions like the dataset having a multivariate normal distribution. If these assumptions are not satisfied, the application of these methods is not possible. We propose an innovative outlier detection approach based upon the PCs and three-sigma limits. The proposed approach can be employed in real time and does not require any assumption or restriction related to the dataset.

The rest of the paper is organized as follows. Section 2 describes the multivariate outlier detection problem and introduces important notation. Section 3 presents our proposed outlier detection method. Section 4 explains the datasets and presents the performance evaluation results of our proposed method. Section 5 provides a discussion of evaluation results and finally concludes the paper.

2. Multivariate Outlier Detection Problem

Most datasets contain one or more unusual observations that are considered as the outlying observation, i.e., dissimilar from the majority of the observations in the dataset, or are doubtful under the expected probability model of the dataset. In a dataset consisting of single feature, either very large or very small observations as compared to the others are unusual observations. If the distribution of the dataset is assumed to be normal, then an observation whose standardized value is greater than the absolute value is usually considered as an outlying observation. The situation becomes complex for a dataset having numerous features. In high dimensional datasets, there can be outliers that cannot be identified when each dimension is independently considered and, hence, cannot be identified by using the univariate criterion. Therefore a multivariate approach is required, and all the dimensions should be considered together.

Let $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ be a random sample of size n from a multivariate distribution, and we have m variables ($n \geq m$). Each $\mathbf{y}_j = (y_{j1}, y_{j2}, \ldots, y_{jm})^T$ is defined as a vector of observations, where $j = 1, 2, \ldots, n$.

Most commonly used approaches to identify outliers in multivariate data are based upon the measuring distances of observations from the central point of dataset. If $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ follow the multivariate normal distribution, then, for any forthcoming observation from the same multivariate normal distribution, a statistic T^2 that relies upon the Mahalanobis distance is defined as

$$T^2 = \frac{n}{(n+1)}(\mathbf{Y}-\overline{Y})^T\mathbf{S}^{-1}(\mathbf{Y}-\overline{Y}), \tag{1}$$

T^2 is distributed as $((n-1)m/(n-m))F_{m,n-m}$. $\overline{Y}$ and $\mathbf{S}$ are, respectively, the sample *mean vector* ($\overline{Y}$) and sample *covariance matrix* ($\mathbf{S}$), defined as follows:

$$\overline{Y} = \frac{1}{n}\sum_{j=1}^{n}\mathbf{Y}_j,$$
$$\mathbf{S} = \frac{1}{n-1}\sum_{j=1}^{n}\left(\mathbf{Y}_j-\overline{Y}\right)\left(\mathbf{Y}_j-\overline{Y}\right)^T. \tag{2}$$

$F_{m,n-m}$ follows an F-distribution with m and $n-m$ degrees of freedom [3]. A higher value of T^2 is an indication of a larger distance of the observations from the canter of the data. Other distance measures such as Euclidean distance or Canberra metric can also be used in place of Mahalanobis distance. An observation, which has a greater difference than a threshold value, is identified as an outlying observation. The threshold value is usually based upon the distribution of distance measure. The distribution of these distances is not easy to derive, even having the assumption of normality.

The PCA-based methods have a long history for the identification of outliers in multivariate data [13–15]. The largest cumulative proportion of the total sample variance is explained by the leading (first few) PCs that have large variances. These leading or major PCs have a tendency of strong relationship with the dimensions that have larger variances and covariances. As a consequence, the observations that are outlying cases with respect to the leading or major components typically relate to outlying observations on one or more of the original variables. In our proposed approach, we employ PCA with three-sigma limits on the error series to identify outlier observations as discussed in the following section.

Let $\xi_j = [\xi_{j1}\xi_{j2}\ldots\xi_{jm}]$ be the *principal components score vector* for observation vector $\mathbf{y}_j = [y_{1j}\, y_{2j}\ldots y_{mj}]$. The number of variables is "m" and "r" is the number of retained PCs and $r \le m$. The sum of squared principal components scores for j^{th} observation given as

$$\sum_{i=1}^{r}\frac{\xi_{ij}^2}{\lambda_{ij}} = \frac{\xi_{1j}^2}{\lambda_1} + \frac{\xi_{2j}^2}{\lambda_2} + \cdots + \frac{\xi_{rj}^2}{\lambda_r}, \quad r \le m \tag{3}$$

follows a chi-square distribution [16] having r degrees of freedom under the assumptions that $\lambda_1 > \lambda_2 > \ldots \lambda_m$ and all λ_i are distinct.

For a specified level of significance α, an observation is identified as outlier if $\sum_{j=1}^{r}\xi_j^2/\lambda_j > \chi_r^2(\alpha)$.

Here, $\chi_r^2(\alpha)$ is upper α percentage point of the chi-square distribution having r degrees of freedom. The value of α refers to false alarm rate in identifying a normal observation as an outlying observation.

3. Proposed Method

Our proposed method for outlier detection is based upon regenerating the variables using the major PCs by following [17, 18]. The step-by-step procedural details of the proposed method are presented below.

Step 1: Estimate the PCs of the original variables. In this step, we perform PCA by converting the original variables into a set of orthogonal variables, i.e., the principal components. These PCs are computed in a way such that the first PC is a maximum-variance linear combination of original variables, and the 2nd PC is a linear combination of original variables, which account for maximum remaining variation while considering a zero correlation between the 1st and 2nd PC. The remaining PCs are computed in a similar manner such that they all are uncorrelated with each other.

For computing the PCs, we first subtract the mean value of each variable from the dataset in order to center the original values in the dataset around the origin and compute pair-wise correlation among variables in the correlation matrix. Eigenvalues and eigenvectors of the correlation matrix are then computed. Scaled eigenvectors represent the PCs with corresponding eigenvalues representing the degree of variance among data observations in eigenvectors' direction.

Given the $m \times n$ multivariate data matrix $\mathbf{Y}$ where "m" is the number of variables and each of the "n" row values denotes data observations/values corresponding to these variables,

$$\mathbf{Y} = \left[y_{ij}\right] = \begin{bmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,n} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ y_{m,1} & y_{m,2} & \cdots & y_{m,n} \end{bmatrix} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y_m} \end{bmatrix}. \tag{4}$$

Let $\lambda_1, \ldots, \lambda_m$ denote the eigenvalues of the correlation matrix of $\mathbf{Y}$ such that $\lambda_1 > \lambda_2 > \ldots \lambda_m$, and all λ_i are distinct. Λ denotes the $m \times m$ matrix of eigenvectors corresponding to eigenvalues λ_i of the correlation matrix of $\mathbf{Y}$ given as

$$\mathbf{\Lambda} = \left[l_{ik}\right] = \begin{bmatrix} l_{1,1} & l_{1,2} & \cdots & l_{1,m} \\ l_{2,1} & l_{2,2} & \cdots & l_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ l_{m,1} & l_{m,2} & \cdots & l_{m,m} \end{bmatrix} = \begin{bmatrix} \mathbf{l}_1 \\ \mathbf{l}_2 \\ \vdots \\ \mathbf{l_m} \end{bmatrix} \tag{5}$$

where $\mathbf{l}_i$ are the eigenvectors.

The matrix of *estimated principal components scores* is then computed, which is as an $m \times n$ matrix defined as

$$\boldsymbol{\xi} = \mathbf{\Lambda Y}, \tag{6}$$

or

$$\boldsymbol{\xi} = \left[\xi_{ij}\right] = \begin{bmatrix} \xi_{1,1} & \xi_{1,2} & \cdots & \xi_{1,n} \\ \xi_{2,1} & \xi_{2,2} & \cdots & \xi_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \xi_{m,1} & \xi_{m,2} & \cdots & \xi_{m,n} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\xi}_1 \\ \boldsymbol{\xi}_2 \\ \vdots \\ \boldsymbol{\xi}_\mathbf{m} \end{bmatrix}. \tag{7}$$

Let $\mathbf{W} = \Lambda^{-1}$, and then

$$\mathbf{W} = \left[w_{ik}\right] = \begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,m} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m,1} & w_{m,2} & \cdots & w_{m,m} \end{bmatrix} = \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \\ \vdots \\ \mathbf{w_m} \end{bmatrix}. \tag{8}$$

The matrix of *weighted PCs for* i^{th} *variable* is then computed as

$$\boldsymbol{\xi}_{\mathbf{Wi}} = \begin{bmatrix} w_{i,1} \times \xi_{1,1} & w_{i,1} \times \xi_{1,2} & \cdots & w_{i,1} \times \xi_{1,n} \\ w_{i,2} \times \xi_{2,1} & w_{i,2} \times \xi_{2,2} & \cdots & w_{i,2} \times \xi_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{i,m} \times \xi_{m,1} & w_{i,m} \times \xi_{m,2} & \cdots & w_{i,m} \times \xi_{m,n} \end{bmatrix}. \quad (9)$$

Step 2: Regenerating the series. This step involves regenerating the original series with appropriate reduction (as suggested by any rule, e.g., scree plot) in dimensions. The original variables can be regenerated without any loss if we make use of all the PCs included in the process. This, however, will not contribute towards dimensionality reduction. In principle, the number of PCs involved to regenerate the original variables should essentially be lesser than original number of variables.

Let "r" be the *retention level*, i.e., the reduced number of PCs used for regenerating variables. Then, the initial r elements of the j^{th} column of PC scores are cumulated to construct cumulative PC's scores for observation j and variable i using the retention level r. Thus, the j^{th} observation of the i^{th} variable using the retention level r is regenerated as

$$\widehat{y}_{ij} = \sum_{k=1}^{r} w_{ik} \xi_{kj}. \quad (10)$$

Step 3: Compute the error series for each variable. In the PCA-based proposed procedure, we model the observations of original variables as original data points and the observations of regenerated variables as fitted points. In this step, we compute their difference as *error* (denoted by e_i).

Step 4: Employ three-sigma limits to detect outliers. Once the series of errors are computed by applying the abovementioned technique, we employ the three-sigma limits. Three-sigma limits are typically applied for identifying and/or removing anomalies or outliers in different datasets. Employing three-sigma limits implies that only a very small number of possible observations could fall outside specification limits of the corresponding dataset. Sigma is essentially a reference to the intervals under a normal or "Gaussian" curve. Each interval is equal to one standard deviation or sigma. Three-sigma limits hence refer to ± 3 sigma from the mean of the data under the curve. In the case of a normal distribution, 68.26% of the data points are within ± 1 from the mean, 95.46% are within ± 2, and 99.73% are within ± 3 sigma. A variation exceeding ± 3 sigma indicates room for improvement.

As discussed, the regenerated variables are based upon major PCs, which account for maximum of the variation of data and are essentially the linear combinations of all original variables. Considering the difference of regenerated and original series as error, three-sigma limits for each of the error series are computed, and the observations lying outside the limits are treated as outliers.

Algorithm 1 summarizes the step-wise details of the proposed method.

Note that we determine the retention level r by considering the scree plot.

4. Numerical Evaluation

In this section, we evaluate and compare the performance of proposed outlier detection method with two most commonly used available methods by considering two real world applications and a simulation study.

4.1. Real Applications. In this section, we present the performance evaluation results of our proposed outlier detection method using two real life applications.

4.1.1. Silicon Wafer Thickness Data. The first application is related to silicon wafer thickness data, and the data source is given in the data availability section. The thickness of a single wafer was measured at nine different locations $(Y_1, Y_2, \ldots, Y_9)$ for 184 consecutive lots. A single wafer from the tray of wafers was removed always at the same position for each lot of wafers after the completion of the chemical vapor decomposition process. All the observations of the dataset had been approximately cantered and scaled to disguise the original variables for privacy.

Figure 1 shows the matrix plot of each of nine variables of silicon wafer thickness dataset. All the variables are regenerated using the first two PCs, because the first two PCs account for almost 94% variation of the data.

4.1.2. Solvents Dataset. Second real life application is related to nine physical properties of 103 chemical solvents. The nine physical properties are Melting Point (Y_1), Boiling Point (Y_2), Dielectric (Y_3), Dipole Moment (Y_4), Refractive Index (Y_5), $E_T(30)$ (empirical solvent polarity parameter) (Y_6), Density (Y_7), log$P$ (partition coefficient of a molecule between an aqueous and lipophilic phases) (Y_8), and Solubility (Y_9). The data source is given in the data availability section.

Figure 2 presents the matrix dot-plot of nine variables of dataset. The reconstruction of all the nine variables is done using the first three PCs as suggested by scree plot. The first three PCs account for 77% of the total variation.

Table 1 presents the eigenvalues, proportion, and cumulative of variance accounted for by the respective components for both datasets. The elbows in scree plots presented in Figures 3(a) and 3(b) suggest retaining the first two PCs for silicon wafer thickness data and first three PCs for solvents data.

Outlier detection is done with the previously explained two existing methods based upon major PCs and Mahalanobis distance, and our proposed method. The error series, i.e., differences between the original and regenerated variables, are computed. The means of all these error series are approximately zero. This is an indication of how good our regenerated variables are. Table 2 presents the mean of the error series for both datasets ($\overline{e}_i$'s are the means of error series).

To gauge the performance of the existing and proposed methods, we use the confusion matrix [19], which is usually used for the performance evaluation of outlier detection methods (Table 3).

Input: $\mathbf{Y}$: $m \times n$ matrix of multivariate data, where m is the number of variables, and each of the n row values denotes data observations/values corresponding to these variables r: retention level, i.e., the reduced number of PCs to consider
Output: Identification of outlier data observations
Steps:
(1) $\Lambda = computeEigenvectors\ (\mathbf{Y})$/ * compute the matrix of eigenvectors using equation (5) * /
(2) $\mathbf{W} = \Lambda^{-1}$
(3) $\xi = estimatePCscores\ (\Lambda, \mathbf{Y})$/ * Calculate the principal component scores using equation (6) * /
(4) $\xi_{\mathbf{w}} = \text{weightedPCs}\ (\mathbf{W}, \xi)$/ * Compute weighted PCs of original variables using equation (9) * /
(5) $\widehat{Y} = \text{regenerateSeries}\ (\xi_{\mathbf{w}}, \mathbf{r})$/ * regenerate data series using equation (10) * // * Compute the error series by considering the difference of original variables and regenerated variables as errors * /
(6) for $i = 1$ to m
(7) $\mathbf{e_i} = \text{difference}\ (\mathbf{y_i}, \widehat{y}_\mathbf{i})$
(8) for each error series $\mathbf{e_i}$
(9) $\overline{X}_{ei} = \text{Mean}\ (\mathbf{e_i})$
(10) $s_{ei} = \text{Standard Deviation}\ (\mathbf{e_i})$
(11) For each data observations $\mathbf{y}$ in $\mathbf{Y}$
(12) Classify $\mathbf{y}$ as outlier if it lies outside the three-sigma limits i.e. $\overline{X}_{ei} \pm 3s_{ei}$

Algorithm 1: Outlier Detection Algorithm

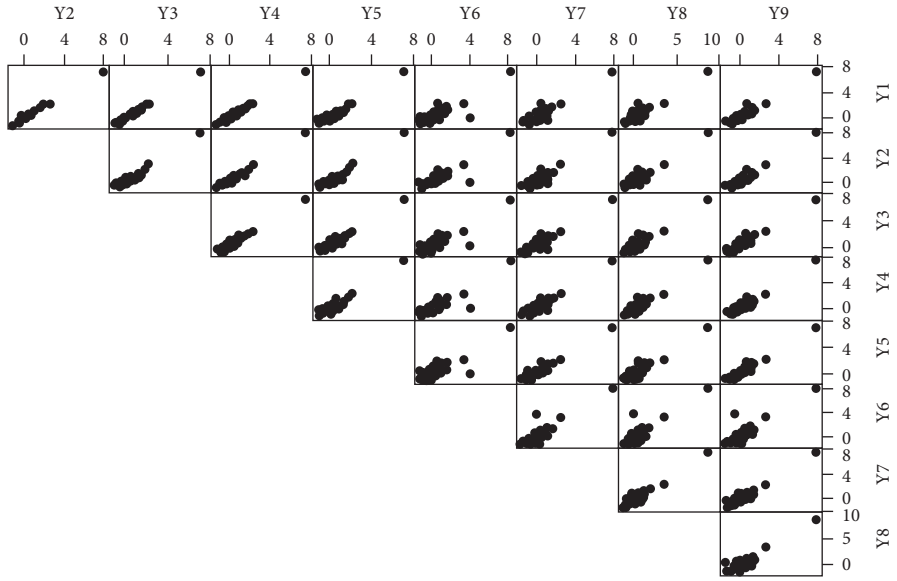

Figure 1: Matrix plot for the silicon wafer thickness dataset.

The performance of outlier detection methods is also evaluated by its true detection rate.

Three other metrics, i.e., precision, recall [20], and F-measure [21], have also been used to evaluate the performance of proposed and existing approaches. Recall and precision are defined as follows:

$$\text{Precesion} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{11}$$

F-measure is the combination of precision and recall measures and defined as

$$F = \left(1 + \beta^2\right) * \frac{(\text{precision} * \text{recall})}{(\text{precision} + \text{recall})}. \tag{12}$$

The value of β is usually taken as 1 [22].

The analysis of the results for silicon wafer thickness data revealed that the proposed method detected observations 39, 111, and 155 as outliers. Outliers detected with the method based upon major PCs are 39, 72, 155, 161, and 174 observations. Similarly, observations 39, 61, and 145 are detected as outlier with the method based upon Mahalanobis distance.

For the solvents data, the proposed method detected observations 2, 5, 9, 15, 51, 70, 83, 97, and 101 as outliers. Outliers detected with method based upon major PCs are 2, 5, 9, 19, 61, 92, 97, and 101 observations. Similarly,

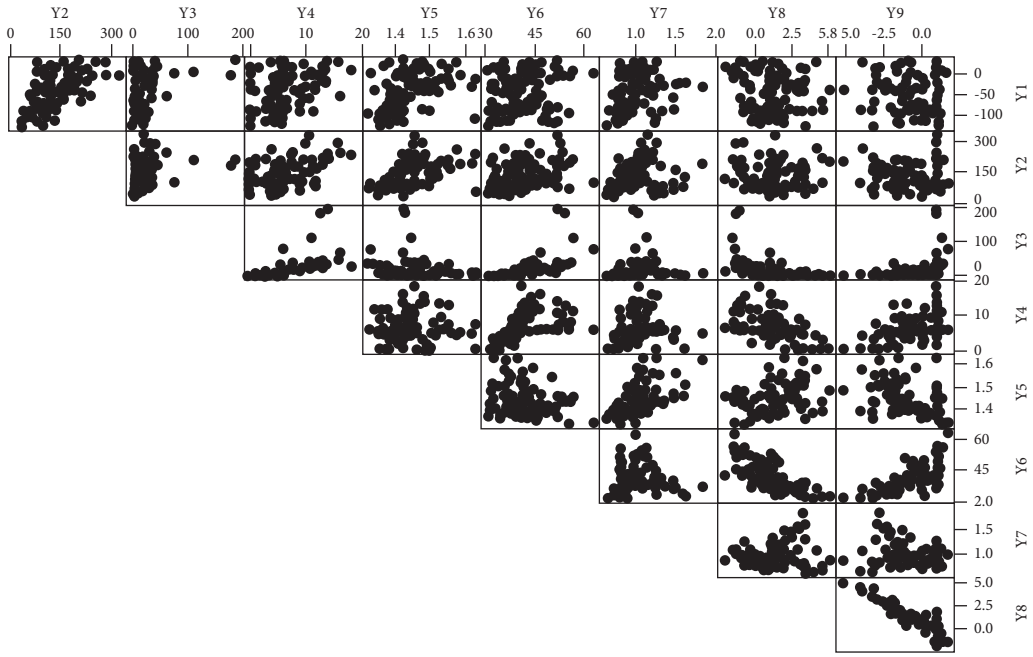

Figure 2: Matrix plot for the solvents dataset.

Table 1: Eigenvalues, proportion, and cumulative variance accounted for by the respective components for silicon wafer thickness and solvents dataset.

Component	Silicon wafer thickness dataset			Solvents dataset		
	Eigen value	Proportion	Cumulative	Eigen value	Proportion	Cumulative
1	8.1753	0.908	0.908	3.6587	0.407	0.407
2	0.254	0.028	0.937	2.499	0.278	0.684
3	0.2017	0.022	0.959	0.7692	0.085	0.77
4	0.1764	0.02	0.979	0.5893	0.065	0.835
5	0.0692	0.008	0.986	0.5154	0.057	0.892
6	0.0562	0.006	0.993	0.421	0.047	0.939
7	0.0304	0.003	0.996	0.286	0.032	0.971
8	0.0263	0.003	0.999	0.192	0.021	0.992
9	0.0104	0.001	1	0.0694	0.008	1

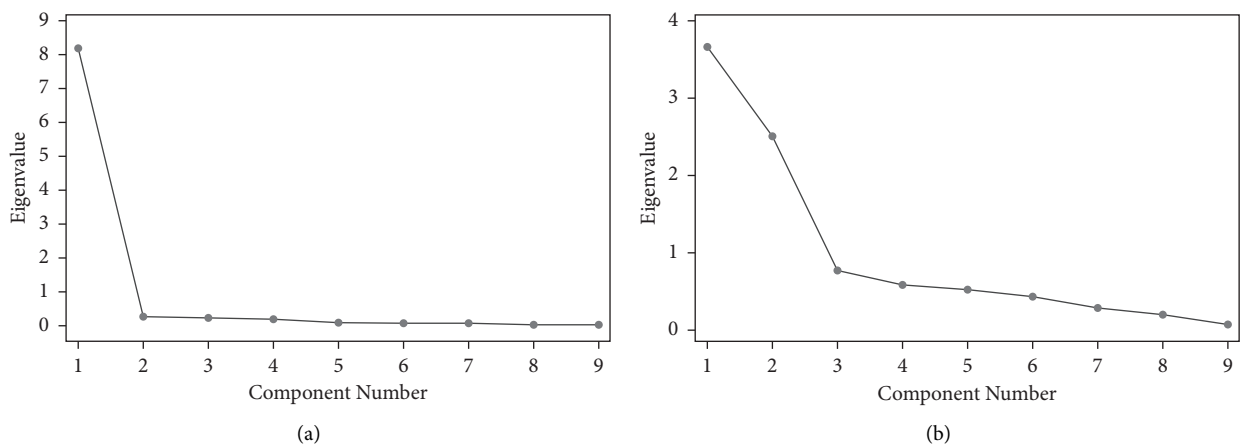

Figure 3: (a) Scree plot for silicon wafer thickness data, (b) scree plot for solvent data.

Table 2: Mean of the error series for real life datasets.

Data	$\bar{e}_1$	$\bar{e}_2$	$\bar{e}_3$	$\bar{e}_4$	$\bar{e}_5$	$\bar{e}_6$	$\bar{e}_7$	$\bar{e}_8$	$\bar{e}_9$
Silicon Wafer Thickness Data	0.0000	0.0000	−0.0488	0.0137	0.0071	0.0211	0.0069	0.0137	0.0135
Solvents Data	0.0039	0.0194	0.0000	−0.0116	0.0019	−0.0196	0.0186	−0.0088	0.0098

Table 3: Confusion matrix.

		Predicted status	
		Normal	Attack
Actual status	Normal	True negative (TN)	False positive (FP)
	Attack	False negative (FN)	True positive (TP)

Table 4: True positive, false negative, false positive, and true negative in silicon wafer thickness dataset with three methods.

Data	Method	TP	FN	FP	TN
Silicon Wafer Thickness Data	Proposed	2	1	1	180
	Major PCs	1	1	3	179
	Mahalanobis	1	1	2	180
Solvents Data	Proposed	8	3	1	92
	Major PCs	5	6	3	89
	Mahalanobis	6	5	2	90

Table 5: Precision and recall for solvents dataset with three methods.

Data	Method	Precision	Recall	F-measure
Silicon Wafer Thickness Data	Proposed	0.6667	0.6667	0.6667
	Major PCs	0.2500	0.5000	0.3333
	Mahalanobis	0.3333	0.5000	0.4000
Solvents Data	Proposed	0.8889	0.7273	0.8000
	Major PCs	0.6250	0.4545	0.5263
	Mahalanobis	0.7500	0.5454	0.6315

Table 6: Mean of the error series for simulated datasets.

n	ρ	Contamination (%)	$\bar{e}_1$	$\bar{e}_2$	$\bar{e}_3$	$\bar{e}_4$	$\bar{e}_5$	$\bar{e}_6$	$\bar{e}_7$	$\bar{e}_8$	$\bar{e}_9$	$\bar{e}_{10}$
200	0.90	2	−0.0242	0.0145	0.0560	−0.0137	0.0000	−0.0139	−0.0212	−0.0144	0.0000	0.0405
		5	−0.0508	0.0587	−0.0129	−0.0050	−0.0336	0.0072	0.0112	0.0000	0.0081	0.0050
		10	0.0572	0.0723	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	−0.0315	0.0000
	0.95	2	−0.0800	0.0541	0.1001	0.0601	0.1143	−0.0834	−0.0445	−0.0230	0.1142	−0.0381
		5	−0.0481	0.0275	0.0247	−0.0164	−0.0083	−0.0513	−0.0132	0.0000	0.0220	−0.1398
		10	−0.0324	−0.0055	−0.0055	−0.0501	−0.0213	−0.0036	−0.0428	−0.0513	0.0260	0.0365
	0.975	2	−0.0309	0.0226	−0.0226	0.0296	0.0000	0.0011	−0.0512	0.0003	−0.0029	0.0231
		5	−0.0507	0.0121	0.0172	−0.0511	−0.0509	−0.0506	−0.0509	−0.0269	0.0285	0.0347
		10	0.0000	0.0000	0.0000	0.0000	0.0000	−0.0588	0.0000	−0.0843	0.0000	0.0000
500	0.90	2	−0.0503	0.0244	0.0065	−0.0065	0.0065	−0.0445	−0.0046	−0.0491	0.0446	−0.0139
		5	−0.0510	0.0000	0.0125	0.0000	−0.0066	0.0000	0.0000	0.0000	0.0000	0.0000
		10	0.0000	0.0000	−0.0296	0.0071	0.0163	0.0000	0.0000	0.0000	−0.0676	0.0000
	0.95	2	−0.0426	0.0162	0.0138	−0.0192	−0.0081	−0.0027	−0.0292	−0.0152	0.0335	−0.0279
		5	−0.0348	0.0359	−0.0118	−0.0048	−0.0205	−0.0260	−0.0383	−0.0237	0.0392	0.0241
		10	−0.0526	0.0178	0.0000	0.0070	−0.0285	−0.0519	−0.0192	−0.0117	0.0194	−0.0077
	0.975	2	−0.0647	0.0895	0.0122	−0.0091	−0.0123	−0.0637	−0.0506	−0.0497	0.0937	−0.0235
		5	0.0000	0.0000	0.0000	−0.0289	−0.0339	0.0000	0.0706	0.0000	−0.0234	−0.0513
		10	−0.0507	0.0362	−0.0233	−0.0032	0.0420	−0.0431	0.0054	0.0190	0.0228	−0.0103
1000	0.90	2	−0.0499	−0.0027	0.0000	0.0526	0.0000	0.0000	−0.0635	−0.0447	0.0000	0.0000
		5	−0.0504	0.0089	0.0709	−0.0231	−0.0434	−0.0500	−0.0502	0.0098	0.0497	−0.0116
		10	0.0000	0.0000	−0.0082	0.0000	−0.0066	−0.0185	−0.0171	0.0000	0.0000	0.0000
	0.95	2	−0.0485	0.0000	0.0272	0.0000	0.0000	0.0000	0.0000	0.0000	−0.0507	0.0290
		5	−0.0091	−0.0595	0.0223	0.0000	0.0094	−0.0157	0.0189	−0.0031	0.0308	−0.0214
		10	−0.0510	−0.0073	0.0066	−0.0303	−0.0103	−0.0509	0.0015	−0.0513	0.0367	−0.0513
	0.975	2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		5	0.0183	−0.0953	0.0583	0.1324	−0.0167	−0.1252	−0.0096	−0.0905	0.1096	0.0000
		10	0.0000	0.0000	0.0000	0.0000	0.0000	−0.0019	−0.0513	0.0000	0.0000	0.0439

Table 7: Percentage of true positive, false negative, false positive, and true negative in simulated datasets.

Method	n	Contamination (%)	$\rho = 0.90$				$\rho = 0.95$				$\rho = 0.975$			
			TP	FN	FP	TN	TP	FN	FP	TN	TP	FN	FP	TN
Proposed			75	25	1.0204	98.9796	75	25	1.5306	98.4694	75	25	1.5306	98.4694
Major PCs		2	50	50	2.0408	97.9592	50	50	2.0408	97.9592	50	50	2.0408	97.9592
Mahalanobis			50	50	2.0408	97.9592	50	50	2.0408	97.9592	50	50	2.5510	97.4490
Proposed			90	10	1.0526	98.9474	90	10	0.5263	99.4737	90	10	0.5263	99.4737
Major PCs	200	5	80	20	1.5789	98.4211	70	30	1.5789	98.4211	70	30	1.5789	98.4211
Mahalanobis			80	20	2.1053	97.8947	80	20	2.1053	97.8947	70	30	1.5789	98.4211
Proposed			90	10	2.7778	97.2222	95	5	2.2222	97.7778	95	5	1.1111	98.8889
Major PCs		10	80	20	3.8889	96.1111	85	15	4.4444	95.5556	90	10	2.7778	97.2222
Mahalanobis			85	15	4.4444	95.5556	85	15	4.4444	95.5556	90	10	2.7778	97.2222
Proposed			80	20	0.4082	99.5918	90	10	0.4082	99.5918	90	10	0.2041	99.7959
Major PCs		2	70	30	0.8163	99.1837	80	20	0.8163	99.1837	70	30	0.6122	99.3878
Mahalanobis			70	30	1.2245	98.7755	70	30	0.6122	99.3878	70	30	0.6122	99.3878
Proposed			92	8	0.4211	99.5789	96	4	0.4211	99.5789	96	4	0.2105	99.7895
Major PCs	500	5	84	16	1.4737	98.5263	84	16	1.0526	98.9474	88	12	0.8421	99.1579
Mahalanobis			84	16	1.2632	98.7368	84	16	1.0526	98.9474	84	16	0.6316	99.3684
Proposed			94	6	0.4444	99.5556	96	4	0.4444	99.5556	98	2	0.2222	99.7778
Major PCs		10	90	10	1.5556	98.4444	92	8	1.5556	98.4444	90	10	1.5556	98.4444
Mahalanobis			86	14	0.8889	99.1111	92	8	0.8889	99.1111	92	8	0.8889	99.1111
Proposed			85	15	0.4082	99.5918	85	15	0.4082	99.5918	100	0	0.3061	99.6939
Major PCs		2	75	25	0.4082	99.5918	80	20	0.3061	99.6939	85	15	0.3061	99.6939
Mahalanobis			70	30	0.4082	99.5918	80	20	0.3061	99.6939	90	10	0.2041	99.7959
Proposed			94	6	0.5263	99.4737	96	4	0.7368	99.2632	100	0	0.6316	99.3684
Major PCs	1000	5	90	10	0.6316	99.3684	90	10	0.6316	99.3684	92	8	0.6316	99.3684
Mahalanobis			86	14	0.7368	99.2632	84	16	0.8421	99.1579	86	14	0.7368	99.2632
Proposed			96	4	0.5556	99.4444	98	2	0.4444	99.5556	100	0	0.2222	99.7778
Major PCs		10	89	11	1.2222	98.7778	90	10	0.3333	99.6667	93	7	0.4444	99.5556
Mahalanobis			88	12	1.1111	98.8889	90	10	0.3333	99.6667	93	7	0.5556	99.4444

Table 8: Precision and recall in simulated datasets.

Method	n	Contamination	$\rho = 0.90$			$\rho = 0.95$			$\rho = 0.975$		
			Precision	Recall	F-Measure	Precision	Recall	F-Measure	Precision	Recall	F-Measure
Proposed			0.6000	0.7500	0.6667	0.5000	0.7500	0.6000	0.5000	0.7500	0.60000
Major PCs		2	0.3333	0.5000	0.4000	0.3333	0.5000	0.4000	0.3333	0.5000	0.39998
Mahalanobis			0.3333	0.5000	0.4000	0.3333	0.5000	0.4000	0.2857	0.5000	0.36362
Proposed			0.8182	0.9000	0.8572	0.9000	0.9000	0.9000	0.9000	0.9000	0.90000
Major PCs	200	5	0.7273	0.8000	0.7619	0.7000	0.7000	0.7000	0.7000	0.7000	0.70000
Mahalanobis			0.6667	0.8000	0.7273	0.6667	0.8000	0.7273	0.7000	0.7000	0.70000
Proposed			0.7826	0.9000	0.8372	0.8261	0.9500	0.8837	0.9048	0.9500	0.92685
Major PCs		10	0.6957	0.8000	0.7442	0.6800	0.8500	0.7556	0.7826	0.9000	0.83720
Mahalanobis			0.6800	0.8500	0.7556	0.6800	0.8500	0.7556	0.7826	0.9000	0.83720
Proposed			0.8000	0.8000	0.8000	0.8182	0.9000	0.8572	0.9000	0.9000	0.90000
Major PCs		2	0.6364	0.7000	0.6667	0.6667	0.8000	0.7273	0.7000	0.7000	0.70000
Mahalanobis			0.5385	0.7000	0.6087	0.7000	0.7000	0.7000	0.7000	0.7000	0.70000
Proposed			0.9200	0.9200	0.9200	0.9231	0.9600	0.9412	0.9600	0.9600	0.96000
Major PCs	500	5	0.7500	0.8400	0.7925	0.8077	0.8400	0.8235	0.8462	0.8800	0.86277
Mahalanobis			0.7778	0.8400	0.8077	0.8077	0.8400	0.8235	0.8750	0.8400	0.85714
Proposed			0.9592	0.9400	0.9495	0.9600	0.9600	0.9600	0.9800	0.9800	0.98000
Major PCs		10	0.8654	0.9000	0.8824	0.8679	0.9200	0.8932	0.8654	0.9000	0.88236
Mahalanobis			0.9149	0.8600	0.8866	0.9200	0.9200	0.9200	0.9200	0.9200	0.92000
Proposed			0.8095	0.8500	0.8293	0.8095	0.8500	0.8293	0.8696	1.0000	0.93025
Major PCs		2	0.7895	0.7500	0.7692	0.8421	0.8000	0.8205	0.8500	0.8500	0.85000
Mahalanobis			0.7778	0.7000	0.7369	0.8421	0.8000	0.8205	0.9000	0.9000	0.90000
Proposed			0.9038	0.9400	0.9215	0.8727	0.9600	0.9143	0.8929	1.0000	0.94342
Major PCs	1000	5	0.8824	0.9000	0.8911	0.8824	0.9000	0.8911	0.8846	0.9200	0.90195
Mahalanobis			0.8600	0.8600	0.8600	0.8400	0.8400	0.8400	0.8600	0.8600	0.86000
Proposed			0.9505	0.9600	0.9552	0.9608	0.9800	0.9703	0.9804	1.0000	0.99010
Major PCs		10	0.8900	0.8900	0.8900	0.9677	0.9000	0.9326	0.9588	0.9300	0.94418
Mahalanobis			0.8980	0.8800	0.8889	0.9677	0.9000	0.9326	0.9490	0.9300	0.93940

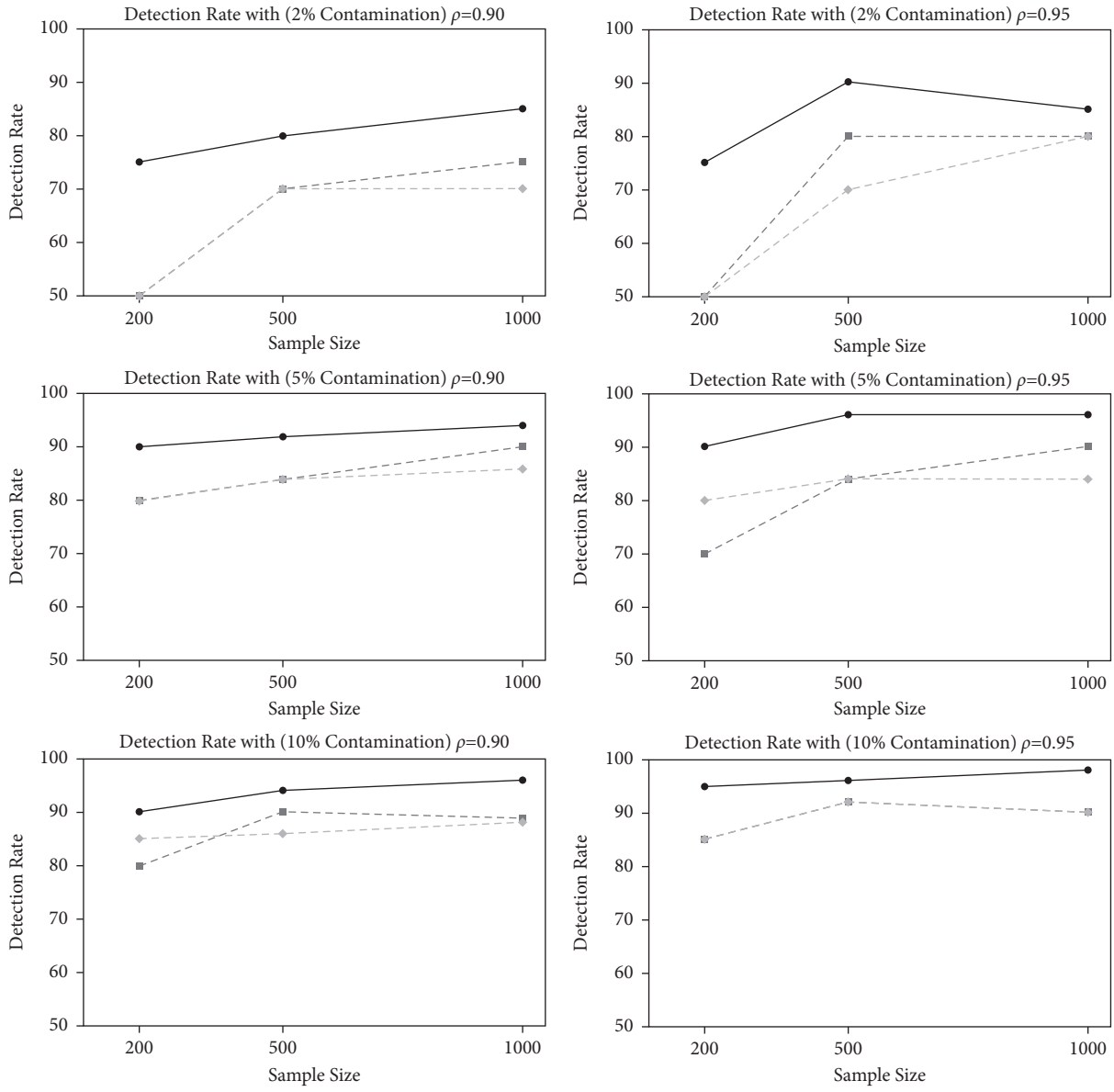

Figure 4: Continued.

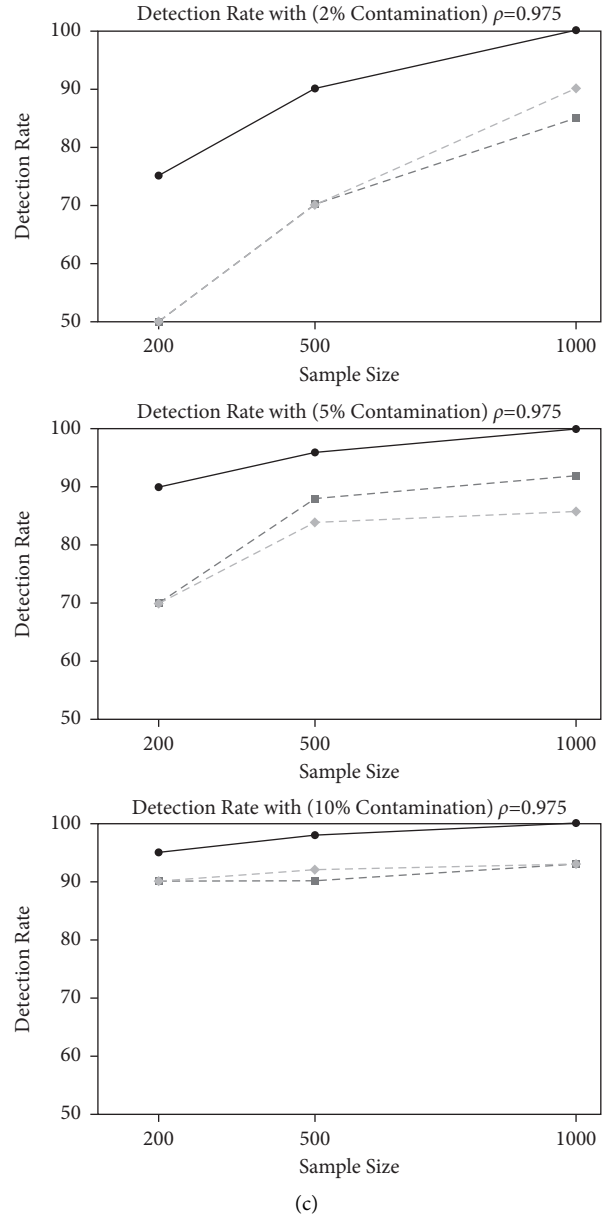

Figure 4: (a) Sample size versus true detection rate ($\rho = 0.9$) with the proposed (Black), major PCs (green), and Mahalanobis (red) methods. (b): Sample size versus true detection rate ($\rho = 0.95$) with the proposed (black), major PCs (green), and Mahalanobis (red) methods. (c): Sample size versus true detection rate ($\rho = 0.975$) with the proposed (black), major PCs (green), and Mahalanobis (red) methods.

observations 2, 5, 9, 15, 34, 83, 97, and 102 are detected as outlier with the method based upon Mahalanobis distance.

Table 4 presents the True Positive, False Negative, False Positive, and True Negative detected in both datasets with proposed and existing methods. The precision and recall computed using all three approaches are given in Table 4.

The results presented in Tables 4 and 5 indicate the supremacy of the proposed method. All the three evaluation criteria, i.e., precision, recall, and F-measure, are higher for the proposed method. The same results can be observed for solvents dataset. TPs and TNs are highest, and FNs and FPs are lowest with the proposed method for Solvents dataset. Precision, recall, and F-measure are found to be 0.8889, 0.7273, and 0.8000, respectively, (highest) for the proposed method.

4.2. Simulated Datasets. In this subsection, we present the comparative performance evaluation results of the proposed and two existing methods with the help of simulated datasets. For this purpose, ten variables are generated from a multivariate standard normal distribution with three different levels of correlation, i.e., 0.90, 0.95, and 0.975. Three different sample sizes are used, i.e., 200, 500, and 1000 observations for each of the three sets of variables. Three

levels of contamination, i.e., $c = 2\%$, 5%, and 10% are used to insert outliers in each of the ten variables. A total of c (contamination or number of outliers) random numbers between 1 and 1000 are produced to select the "c" rows to insert outliers in the dataset. The mean and standard deviation of the data are calculated. Each observation of the certain row is multiplied by

$$\text{Mean} + 10\,(\text{standard deviation}). \tag{13}$$

Hence, a total of twenty-seven datasets are generated in this way. The simulation experiments are replicated 1000 times to compute the percentage of true detection rate, precision, and recall.

The error series for all the simulated datasets are computed, and their means are presented in Table 6. It can be observed that the mean of all these error series are approximately zero.

Tables 7 and 8 present the percentage of true positive, false negative, false positive, and true negative, precision, recall, and F-measure for the simulated datasets. The percentage of true detection is highest, and false detection is the lowest for the proposed method as compared to the existing methods. The size of sample has a direct effect on any research findings. It can undermine or strengthen the internal and external validation of any study. The same can be observed from the results of our study. A substantial improvement in results can be seen with increasing sample sizes. As the sample size has increased from 200 to 1000, the percentage of true positives has increased from 75% to 85% when the contamination level is 2%, from 90% to 94% when the contamination level is 5%, and similarly from 90% to 96% when the contamination level is 2%. It can also be observed that the correlation level has no effect on the simulated results. Similarly, the change in contamination levels also has not any effect on the results. The same can be confirmed from precision, recall, and F-measure. A set of figures presented as Figures 4(a)–4(c) show the percentages of true detection rates versus sample sizes for the three methods. True detection rate is also higher for the proposed method and is getting more improved with the increase in sample size.

5. Discussion and Conclusion

This paper suggests a novel approach based upon PCA and three-sigma limits for outlier detection. The predictive model is developed using the major principal components suggested by the scree plots. The main advantage of the proposed approach is that it does not require any distributional assumptions. We performed the outlier detection with our proposed method as well as with two existing classical approaches to gauge the performance of the proposed method. The performance comparison is made using two real life and several simulated datasets. The examples from real life data and simulation experiments confirm the better performance of our proposed technique as compared to the two existing approaches. First, the three outlier detection methods were applied to silicon wafer thickness data. The computed values of precision, recall, and F-Statistic were highest with the proposed method, i.e., 0.6667, 0.6667, and 0.6667, respectively, while using major PCs, the three measures were 0.2500, 0.5000, and 0.3333. The method based upon Mahalanobis distance produced the three measures as 0.3333, 0.5000, and 0.4000, respectively. Major PCs based method produced the worst results. The same scenario can be observed from the application of these three outlier detection methods to solvents data. Precision, recall, and F-Statistic were computed as 0.8889, 0.7273, and 0.8000 with the proposed method, 0.6250, 0.4545, and 0.5263 by using major PCs, and 0.7500, 0.5454, and 0.6315 with Mahalanobis distance method. Simulation experiments also confirmed the same situation. For all the sample sizes, correlation, and contamination levels, the proposed method performed best among the three.

With three increasing levels of sample sizes, i.e., 200, 500, and 1000, the percentages of true detections are increased, and false detection rates are decreased. The performance is getting better with increasing sample sizes regardless the level of correlation between variables and contamination level. The results showed that, with $\rho = 0.9$ when the contamination level is 2%, the precision is increased from 0.6000 to 0.8095, recall is increased from 0.7500 to 0.8500, and F-measure is increased from 0.6667 to 0.8293 when the sample size is increased from 200 to 1000. When the contamination level is 5%, these three measures are 0.8182, 0.9000, and 0.8572 for sample size 200 and increased to 0.9038, 0.9400, and 0.9215 with the sample size 1000. The same can be observed with $\rho = 0.95$. The three measures with sample size 200 and 2% contamination are 0.5000, 0.7500, and 0.6000 and increased to 0.8095, 0.8500, and 0.8293 with increasing sample size of 1000. A similar situation persists with 5% and 10% contamination. The results with datasets having maximum correlation level, i.e., $\rho = 0.975$, gave the same scenario. For 10% contamination level with sample size = 200, the three measures were 0.9048, 0.9500, and 0.92685. When sample size was increased to 500 and 1000, these measures were increased to 0.9800, 0.9800, 0.9800, and 0.9804, 1.0000, and 0.99010, respectively.

Not only is the proposed method useful for datasets with variables having interdependence relationship, but it can also be applied to data having variable with dependence relationship, i.e., variables categorized as response and explanatory variables. The outlying observations in the set of explanatory variables can be detected by using the step by step approach of the proposed method. After taking care of outlying observations in explanatory variables, response variable can be checked for outlying observations by using studentized deleted residuals, or a formal test can be conducted by means of the Bonferroni test procedure. In future work, investigating the proposed method for variables having such relationship might prove important.

Data Availability

Previously reported data were used to support this study and are available at https://openmv.net/info/silicon-wafer-thickness https://openmv.net/info/solvents.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Saima Afzal conceived, supervised, and designed the study. Ayesha Afzal performed simulation experiments and computational work. Muhammad Amin made the analysis of results. Sehar Saleem wrote the manuscript. Nouman Ali helped in conducting simulation studies and analysis of results. Muhammad Sajid reviewed and edited the manuscript. All the authors discussed the results and contributed to the final manuscript.

References

[1] K. I. Penny and I. T. Jolliffe, "A comparison of multivariate outlier detection methods for clinical laboratory safety data," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 50, no. 3, pp. 295–307, 2001.

[2] A. Lazarevic, L. Ertoz, V. Kumar, A. Ozgur, and J. Srivastava, "A comparative study of anomaly detection schemes in network intrusion detection," in *Proceedings of the Third SIAM Conference on Data Mining*, University of Illinois, Chicago, IL, USA, May 2003.

[3] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*, Elsevier, Amsterdam, Netherlands, 2011.

[4] C. C. Aggarwal, "An introduction to outlier analysis," in *Outlier Analysis*Springer, Cham, Switzerland, 2017.

[5] Z. Li, Y. Zhao, N. Botta, C. Ionescu, and X. Hu, "COPOD: copula-based outlier detection," in *Proceedings of the 2020 IEEE International Conference on Data Mining (ICDM)*, pp. 1118–1123, IEEE, Sorrento, Italy, November 2020.

[6] P. J. Rousseeuw and B. C. van Zomeren, "Unmasking multivariate outliers and leverage points," *Journal of the American Statistical Association*, vol. 85, no. 411, pp. 633–639, 1990.

[7] M. Hubert, P. J. Rousseeuw, and P. Segaert, "Multivariate functional outlier detection," *Statistical Methods and Applications*, vol. 24, no. 2, pp. 177–202, 2015.

[8] P. J. Rousseeuw, J. Raymaekers, and M. Hubert, "A measure of directional outlyingness with applications to image data and video," *Journal of Computational & Graphical Statistics*, vol. 27, no. 22, pp. 345–359, 2018.

[9] S. Prykhodko, N. Prykhodko, L. Makarova, and A. Pukhalevych, "Application of the squared mahalanobis distance for detecting outliers in multivariate non-Gaussian data," in *Proceedings of the 14th International Conference on Advanced Trends in Radioelecrtronics, Telecommunications and Computer Engineering (TCSET)*, University of Illinois, Chicago, IL, USA, pp. 962–965, 2018.

[10] I. Cabero, I. Epifanio, A. Piérola, and A. Ballester, "Archetype analysis: a new subspace outlier detection approach," *Knowledge-Based Systems*, vol. 217, p. 106830, 2021.

[11] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. . Sander, "LOF: identifying density-based local outliers," in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pp. 93–104, Dallas, TX, USA, May 2000.

[12] X. Zhao, J. Zhang, and X. Qin, "LOMA: a local outlier mining algorithm based on attribute relevance analysis," *Expert Systems with Applications*, vol. 84, pp. 272–280, 2017.

[13] R. Gnanadesikan and J. R. Kettenring, "Robust estimates, residuals, and outlier detection with multiresponse data," *Biometrics*, vol. 28, no. 1, pp. 81–124, 1972.

[14] I. T. Jolliffe, *Principal Component Analysis*, Springer-Verlag, New York, NY, USA, 2nd edition, 2002.

[15] T. Chen, E. Martin, and G. Montague, "Robust probabilistic PCA with missing data and contribution analysis for outlier detection," *Computational Statistics & Data Analysis*, vol. 53, no. 10, pp. 3706–3716, 2009.

[16] M. L. Shyu, S. C. Chen, K. Sarinnapakor, and L. Chang, "A novel anomaly detection scheme-based on principal component classifier," in *Proceedings of the 3rd IEEE International Conference on Data Mining*, pp. 353–365, Melbourne, FL, USA, December 2003.

[17] A. D. Back and A. S. Weigend, "A first application of independent component analysis to extracting structure from stock returns," *International Journal of Neural Systems*, vol. 8, no. 4, pp. 473–484, 1997.

[18] S. Afzal and M. M. Iqbal, "A new way to order independent components," *Journal of Applied Statistics*, vol. 43, no. 9, pp. 1753–1764, 2016.

[19] P. Dokas, L. Ertoz, V. Kumar, A. Lazarervic, J. Srivastava, and P.-N. Tan, "Data mining for network intrusion detection," *Proceedings of National Science Foundation Workshop on Next Generation Data Mining*, 2002.

[20] Y. Yang, *An Evaluation of Statistical Approaches to Text Categorization*, Kluwer Academic Publishers, Dordrecht, Netherlands, 1999.

[21] N. Chinchor, "MUC-4 evaluation metrics," in *Proceedings of the 4th Conf. Message Understand (MUC4)*, pp. 22–29, Association for Computational Linguistics, Stroudsburg, PA, USA, 1992.

[22] Y. Sasaki, "The truth of the *F*-measure," *Teach Tutor Mater*, vol. 1, no. 5, pp. 1–5, 2007.