Pakistan Journal of Statistics and Operation Research

Detection of Outliers Method in Grouped Multivariate Data: A Method Based on Multiple Linear Regression

Suthat Phuttisen[1], Wuttichai Srisodaphol[2*]

* Corresponding Author

1. Department of Statistics, Faculty of Science, Khon Kaen University, Thailand, Suthad@kkumail.com
2. Department of Statistics, Faculty of Science, Khon Kaen University, Thailand, wuttsr@kku.ac.th

Abstract

Cluster analysis is applied to group data so that samples within the same group are similar. A common problem with multivariate data implementation is that the data differs significantly from most of the other data. Outliers can significantly impact data analysis and model performance, making their detection crucial in various domains. This study presents an investigation of the outlier detection method using multiple linear regression for grouped multivariate data. The research compares the performance of the proposed method with two existing approaches, namely the Caroni and Billor (2007) method and the Hardin and Rocke (2004) method. In the case of uncontaminated data, the proposed method still identifies inliers as outliers in uncontaminated data and exhibits an increased percentage of not detecting outliers as the number of variables and sample size increase. In the scenario of contaminated data, the results reveal that the proposed method consistently outperforms both the Caroni and Billor method and the Hardin and Rocke method in terms of accuracy and precision. These findings highlight the effectiveness of the proposed method for outlier detection in grouped multivariate data. The study contributes to the existing knowledge of outlier detection approaches and provides insights into their performance under different data conditions. Researchers and practitioners can benefit from these findings when selecting appropriate outlier detection methods for various applications.

Key Words: Multivariate data, Multiple linear regression, Outliers, Cluster analysis.

Mathematical Subject Classification: 62H30, 62J05.

1. Introduction

Nowadays, cluster analysis is applied to group data so that samples within the same group are similar, e.g., on the business side, customers are segmented according to their consumption behavior. Customers with similar consumption behaviors will be in the group. In medicine, patients are grouped according to their symptoms or severity of the disease to use different treatment methods according to the severity of the disease, etc. (Montgomery et al., 2012) Problems encountered in the analysis of the data may be outliers. Outliers can be caused by human error, the data collection tool, out-of-threshold simulation, or neither (Santoyo, 2017). Outliers, by definition, are not like normal data in a dataset. They are data points far away from normal data in each cluster.

When determining whether multivariate data that is multigroup may have outliers, a popular method for detecting outliers was DBSCAN (Density-Based Spatial Clustering of Applications with Noise) (Ester et al., 1996). The algorithm is designed to identify clusters in spatial databases, even in the presence of noise and outliers. DBSCAN operates based on the concept of density, gathering together densely packed data points while separating sparsely populated regions. It defines two important parameters: epsilon (ε), which specifies the neighborhood radius around each point, and minPts, which sets the minimum number of points required to form a dense region or cluster. In large spatial databases, the DBSCAN algorithm has proven effective at discovering clusters of arbitrary shape and managing noise. It has been widely adopted and functions as the basis for numerous density-based clustering methods. Nonetheless, Hardin and Rocke (2004) presented a method for outlier detection in datasets that contain multiple

clusters. They proposed using the Minimum Covariance Determinant (MCD) estimator to identify outliers within each cluster. They evaluated the performance of their approach using simulated data and real-world examples, demonstrating its effectiveness in detecting outliers within each cluster. They compared it with other outlier detection methods and showed that their proposed method performed well in various scenarios. Caroni and Billor (2007) proposed a method for detecting outliers in multivariate data that multigroup by improving the method of Billor et al. (2000) that proposed BACON (Blocked Adaptive Computationally Efficient Outlier Nominators), which the BACON method applies to data with a single group. The result showed in the simulated data that the data was uncontaminated. The result was that the modified outlier method from the BACON method found the percentage of outliers close to 0 percent according to the simulated data.

Based on the statement about the problems of detecting outliers, the researcher was interested in the outliers that were not caused by various errors. This extremely significant outlier that was not the result of multiple errors is significant because it was directly generated by the sample unit, such as when people with diabetes have higher blood pressure and blood sugar than normal people without diabetes (Kelleher, 2022).

Therefore, it is essential to account for outlier values that are not the consequence of errors. From the method to detect outliers in the case of multivariate data that multigroup mentioned above, there is a difficult application requirement to determine the distance between two points; the number of outliers depends on the cut-off point caused by the given distribution and level of significance. If the level of significance is too high, the outlier value will not be found, but if the level of significance is too small, it will cause many outliers to be found. It is not necessary that all data sets have outlier values. To be more accurate in identifying outliers as true outliers for detecting outliers in multivariate data that is multigroup, this research would like to present a method using the principles of Euclidean distance and multiple linear regression. The remainder of the paper is organized as follows. In Section 2, we introduce material and methodology. Section 3 describes the results and discussion. Finally, we describe the conclusion in Section 4.

2. Material & Methodology

2.1 The proposed algorithm

The purpose of this research is to present a method to detect outliers in multivariate data that is multigroup using the principles of Euclidean distance and multiple linear regression. We have the following method for conducting research:

Algorithm. Detection Outliers Method in Grouped Multivariate Data
Inputs: X and k (k is the number of groups of data)
Outputs: $\mathbf{X}_{inliers}$ and $\mathbf{X}_{outliers}$

01: **for** each $j = 1$ **to** k **do**
02: **for** each $i = 1$ **to** n_j **do**
03: Compute Euclidean distance: $d_{ij} = \sqrt{\sum_{m=1}^{p} (x_{imj} - \bar{x}_{mj})^2}$ when $\bar{x}_{mj} = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{imj}$, p is the number of independent variables and sample size $n = n_1 + n_2 + \cdots + n_k$
04: **end for**
05: Create matrix $\mathbf{H}_j$ with a size $h_j \times p$ and $\mathbf{H}_j^{'}$ with a size $(n_j - h_j) \times p$
06: **if** $d_{ij} \leq \frac{\min(d_{ij}) + \max(d_{ij})}{2}$
07: **Then send X to** $\mathbf{H}_j$
08: **else send X to** $\mathbf{H}_j^{'}$
09: **end if**
10: **for** each dimension $m = 1$ **to** p **do**
11: Compute multiple linear regression equations in matrix $\mathbf{H}_j$
12: Compute R_m^2

if $R_m^2 = \max(R_m^2)$
then select this m is dependent variable
end for
Compute $|\varepsilon_{lj}|$ form $\mathbf{H}_j$
for each $l = 1$ **to** h_j **do**
$|\varepsilon_{lj}| = |y_{lj} - \hat{y}_{lj}|$;
end for
Compute $C_j = \max(|\varepsilon_{lj}|)$
Compute $|\varepsilon_{vj}|$ form $\mathbf{H}_j^{'}$
for each $v = 1$ **to** $n_j - h_j$ **do**
$|\varepsilon_{vj}| = |y_{vj} - \hat{y}_{vj}|$
end for
end for
for each $j = 1$ **to** k **do**
for each $i = 1$ **to** n **do**
if $|\varepsilon_{ij}| > C_j$ all j
Then send X to $\mathbf{X}_{outliers}$
else send X to $\mathbf{X}_{inliers}$
end if
end for
end for
Return $\mathbf{X}_{inliers}$ and $\mathbf{X}_{outliers}$

The above algorithm of our proposed method is depicted in Figure 1.

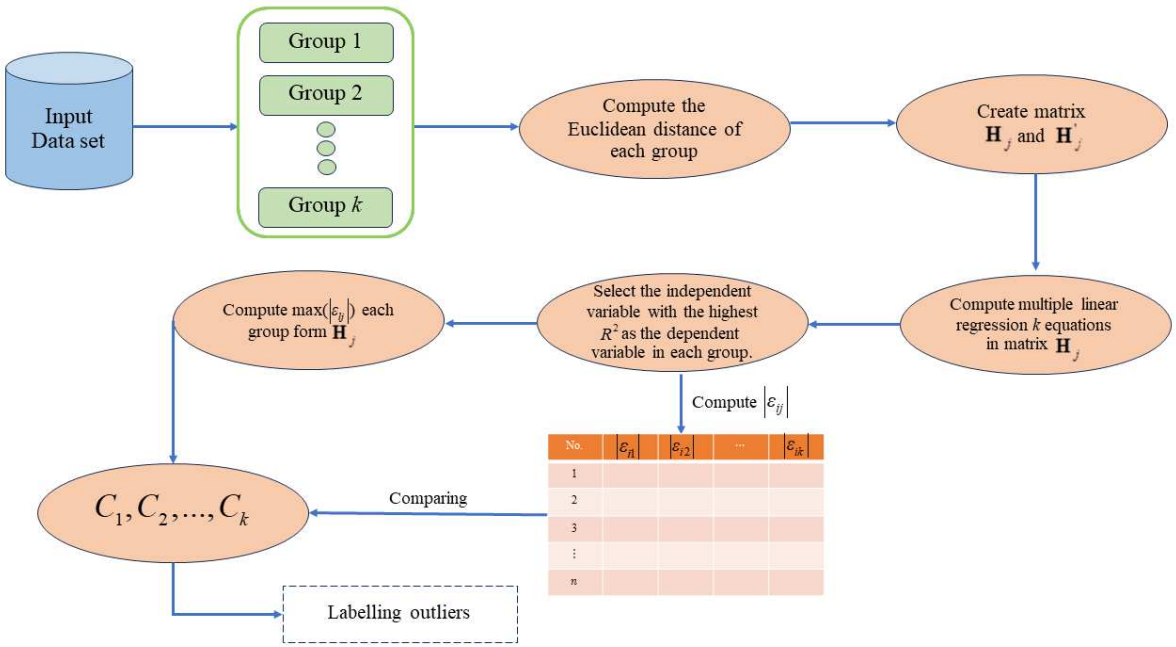

Figure 1: The operating procedure of the proposed method.

2.2. Performance comparison

In this section, we compare the performance of the proposed method with the previous two methods, Hardin and Rocke (2004) and Caroni and Billor (2007). We simulate the data using a multivariate normal distribution with and without contaminated data for 1000 iterations. The criterion for uncontaminated case is the proportion of outliers detected. Accuracy, precision, and recall are criteria for contaminated case. These methods are evaluated the performance for 2 or 3 groups of data under conditions as follows:

1. We simulate the uncontaminated and contaminated data in each group k from a multivariate normal distribution.

$$f(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma},\tau,\eta)=\tau\Phi(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})+(1-\tau)\Phi(\mathbf{x};\boldsymbol{\mu},\eta\boldsymbol{\Sigma})$$

where τ is contamination proportion, $\boldsymbol{\mu}$ is mean, $\boldsymbol{\Sigma}$ is variance – covariance matrix, η is degree of contamination, $\Phi(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$ is p-variate probability density function of multivariate normal distributions (Punzo and McNicholas, 2016).

2. For 2 groups data, we denote

$$\boldsymbol{\mu}_1=[5,0,...,0]',\ \boldsymbol{\mu}_2=[7.5,0,...,0]',\ \boldsymbol{\Sigma}=\begin{bmatrix}1 & 0.8 & \cdots & 0.8\\ 0.8 & 1 & \cdots & 0.8\\ \vdots & \vdots & \ddots & 0.8\\ 0.8 & 0.8 & \cdots & 1\end{bmatrix}_{p\times p}.$$

For unequal variance case, the variance – covariance matrix of the second group is placed by the $\boldsymbol{\Sigma}=\begin{bmatrix}1 & 0.6 & \cdots & 0.6\\ 0.6 & 1 & \cdots & 0.6\\ \vdots & \vdots & \ddots & 0.6\\ 0.6 & 0.6 & \cdots & 1\end{bmatrix}_{p\times p}$, τ = 0, 0.06, p = 2, 5, 10 and η = 1.5. Let n_j is sample size on group j; j = 1, 2 when (n_1,n_2) = (100,100) and (500,700).

3. For 3 groups data, we denote

$$\boldsymbol{\mu}_1=[5,0,...,0]',\ \boldsymbol{\mu}_2=[7.5,0,...,0]',\ \boldsymbol{\mu}_3=[2.5,4.66,...,0]',\ \boldsymbol{\Sigma}=\begin{bmatrix}1 & 0.8 & \cdots & 0.8\\ 0.8 & 1 & \cdots & 0.8\\ \vdots & \vdots & \ddots & 0.8\\ 0.8 & 0.8 & \cdots & 1\end{bmatrix}_{p\times p}.$$

For unequal variance case, the variance – covariance matrix of the third group is placed by the $\boldsymbol{\Sigma}=\begin{bmatrix}1 & 0.6 & \cdots & 0.6\\ 0.6 & 1 & \cdots & 0.6\\ \vdots & \vdots & \ddots & 0.6\\ 0.6 & 0.6 & \cdots & 1\end{bmatrix}_{p\times p}$, τ = 0, 0.06, p = 2, 5, 10 and η = 1.5. Let n_j is sample size on group j; j = 1, 2, 3 when (n_1,n_2,n_3) = (100,100,100) and (500,700,500).

3. Results and Discussion

3.1 Result

In this section, we compare the performance of the proposed methods with the previous two methods, Hardin and Rocke (2004) and Caroni and Billor (2007). Tables 1 and 2 depict the percentage of detected outliers without contamination data when assuming equal and unequal covariances, respectively. The accuracy, precision, and recall with 6% contamination data are depicted in Tables 3 and 4, assuming equal and unequal covariances, respectively.

Table 1. The percentage of detected outliers without contamination data: equal covariances assumed.

			Hardin and Rocke			Caroni and Billor			Proposed method		
k	*p*	*n*	No. of outliers detected			No. of outliers detected			No. of outliers detected		
			0	1	2+	0	1	2+	0	1	2+
2	2	(100) (100)	0	0	100	97.80	1.20	1.00	10.72	27.03	62.25
		(500) (700)	0	0	100	51.78	24.47	23.75	20.72	36.03	43.25
	5	(100) (100)	0	0	100	97.90	1.80	0.30	47.48	32.69	19.83
		(500) (700)	0	0	100	52.49	23.88	23.63	43.16	50.22	6.62
	10	(100) (100)	0	0	100	98.01	1.32	0.67	58.53	26.27	15.20
		(500) (700)	0	0	100	52.77	24.61	22.62	84.49	14.83	0.68
3	2	(100) (100) (100)	0	0	100	98.90	1.00	0.01	16.22	22.57	61.21
		(500) (700) (500)	0	0	100	56.69	27.17	16.14	18.90	27.17	46.07
	5	(100) (100) (100)	0	0	100	98.10	1.40	0.50	51.24	30.66	18.10
		(500) (700) (500)	0	0	100	51.22	24.91	23.87	69.77	24.88	5.35
	10	(100) (100) (100)	0	0	100	98.10	1.30	0.60	63.72	23.84	12.44
		(500) (700) (500)	0	0	100	54.62	26.74	18.64	91.23	8.66	1.10

Table 2. The percentage of detected outliers without contamination data: unequal covariances assumed.

			Hardin and Rocke			Caroni and Billor			Proposed method		
k	*p*	*n*	No. of outliers detected			No. of outliers detected			No. of outliers detected		
			0	1	2+	0	1	2+	0	1	2+
2	2	(100) (100)	0	0	100	98.90	0.30	0.80	2.90	7.40	89.70
		(500) (700)	0	0	100	53.21	14.55	32.24	2.12	5.81	92.07
	5	(100) (100)	0	0	100	98.20	0.20	1.60	42.27	33.76	23.97
		(500) (700)	0	0	100	49.22	14.23	36.55	62.12	24.94	12.94
	10	(100) (100)	0	0	100	97.90	0.90	1.20	53.77	26.13	20.10
		(500) (700)	0	0	100	51.68	14.92	33.40	80.56	17.32	2.12
3	2	(100) (100) (100)	0	0	100	98.80	0.00	1.20	4.59	19.74	75.67
		(500) (700) (500)	0	0	100	49.77	14.21	36.02	11.89	15.12	72.99
	5	(100) (100) (100)	0	0	100	97.12	1.20	1.68	53.12	30.44	16.44
		(500) (700) (500)	0	0	100	48.89	13.62	37.49	68.46	22.47	9.07
	10	(100) (100) (100)	0	0	100	97.60	1.70	0.70	59.82	24.12	16.06
		(500) (700) (500)	0	0	100	48.77	13.55	37.68	90.57	2.11	7.32

In Tables 1 and 2, we present the results of the outlier detection method for grouped multivariate data obtained using equal and unequal covariances. In the case of uncontaminated data, the results demonstrate that as the number of variables and sample size increase, the proposed method exhibits an increased percentage of not detecting outliers (No. of outliers detected are 0 and 1) based on the sample size and number of variables. Conversely, the Caroni and Billor method displays a high percentage of detected outliers (No. of outliers detected is 0) when the sample size is small, but this percentage decreases as the sample size increases. On the other hand, the Hardin and Rocke method detects outliers in all cases (No. of outliers detected is 2+).

Table 3. Accuracy, precision, and recall with 6% contamination: equal covariances assumed.

k	*p*	*n*	Hardin and Rocke			Caroni and Billor			Proposed method		
			Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
2	2	(100) (100)	0.9706	0.7023	1.0000	0.9324	0.4842	0.5024	0.9976	0.9625	1.0000
		(500) (700)	0.9896	0.6929	1.0000	0.9137	0.4769	0.4994	0.9996	0.9645	1.0000
	5	(100) (100)	0.9555	0.7033	1.0000	0.9045	0.4854	0.4868	0.9903	0.9625	0.9792
		(500) (700)	0.9891	0.6756	1.0000	0.8669	0.4611	0.4900	0.9879	0.9646	0.9879
	10	(100) (100)	0.9517	0.7193	1.0000	0.9256	0.4967	0.4875	0.9756	0.9618	0.9678
		(500) (700)	0.987	0.7129	1.0000	0.9513	0.4913	0.4849	0.9726	0.9648	0.9469
3	2	(100) (100) (100)	0.9167	0.4211	1.0000	0.9800	1.0000	0.6667	0.9987	0.9883	0.9967
		(500) (700) (500)	0.9526	0.4388	1.0000	0.9849	1.0000	0.6701	0.9997	0.9897	0.9970
	5	(100) (100) (100)	0.8908	0.4084	1.0000	0.9382	0.9703	0.6567	0.9788	0.9880	0.9761
		(500) (700) (500)	0.9502	0.4376	1.0000	0.9393	0.9597	0.6642	0.9859	0.9829	0.9832
	10	(100) (100) (100)	0.8567	0.3916	1.0000	0.9400	0.9787	0.6524	0.9733	0.9826	0.9754
		(500) (700) (500)	0.9498	0.4370	1.0000	0.9182	0.9580	0.6502	0.9511	0.9998	0.9613

Table 4. Accuracy, precision, and recall with 6% contamination: unequal covariances assumed.

k	*p*	*n*	Hardin and Rocke			Caroni and Billor			Proposed method		
			Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
2	2	(100) (100)	0.9714	0.7029	1.0000	0.9322	0.4897	0.5009	0.9972	0.9568	0.9998
		(500) (700)	0.9892	0.6926	1.0000	0.9124	0.4816	0.4979	0.9995	0.9592	1.0000
	5	(100) (100)	0.9563	0.6922	1.0000	0.8885	0.4823	0.4848	0.9867	0.9576	0.9995
		(500) (700)	0.9886	0.6731	1.0000	0.8589	0.4662	0.4837	0.9694	0.9594	0.9871
	10	(100) (100)	0.9556	0.7211	1.0000	0.9589	0.5029	0.5028	0.9757	0.9563	0.9797
		(500) (700)	0.9872	0.7105	1.0000	0.9546	0.4937	0.5087	0.9827	0.9594	0.9643
3	2	(100) (100) (100)	0.9145	0.4232	1.0000	0.9804	1.0000	0.6678	0.9908	0.9783	0.9444
		(500) (700) (500)	0.9527	0.4409	1.0000	0.9852	1.0000	0.6712	0.9924	0.9797	0.9445
	5	(100) (100) (100)	0.8986	0.4104	1.0000	0.9117	0.9592	0.6578	0.9751	0.9780	0.9278
		(500) (700) (500)	0.9469	0.4397	1.0000	0.9163	0.9320	0.6653	0.9672	0.9899	0.9324
	10	(100) (100) (100)	0.8678	0.3935	1.0000	0.9387	0.9729	0.6535	0.9767	0.9729	0.9273
		(500) (700) (500)	0.9453	0.4392	1.0000	0.9247	0.9472	0.6513	0.9676	0.9898	0.9036

In Tables 3 and 4, we present the results of the outlier detection method for grouped multivariate data obtained using equal and unequal covariances. In the case of 6% contaminated data, the results indicate that the proposed method consistently outperforms both the Caroni and Billor method and the Hardin and Rocke method in terms of accuracy and precision. Specifically, in all cases, the proposed method demonstrates higher accuracy and precision compared to the Caroni and Billor method. Moreover, when $k = 2$, the accuracy and precision of the Caroni and Billor method are lower than those of the Hardin and Rocke method. However, when $k = 3$, the precision of the Caroni and Billor method surpasses that of the Hardin and Rocke method. Notably, the Hardin and Rocke method consistently achieves a recall value of 1 in all cases.

Furthermore, we show the maximum number of detected outliers in 6% contaminated data for equal and unequal covariances assumed in Table 5. The results show that the proposed method identically detects outliers of about 6% for all scenarios.

Table 5. The maximum number of detected outliers of 6% contaminated data.

k	*p*	*n*	Hardin and Rocke		Caroni and Billor		Proposed method	
			equal covariance	unequal covariance	equal covariance	unequal covariance	equal covariance	unequal covariance
2	2	(100) (100)	47(23.5%)	49(24.5%)	39(19.5%)	35(17.5%)	14(7.0%)	14(7.0%)
		(500) (700)	211(17.6%)	210(17.5%)	116(9.7%)	102(8.5%)	82(6.8%)	80(6.7%)
	5	(100) (100)	44(22.0%)	46(23.0%)	46(23.0%)	43(21.5%)	13(6.5%)	13(6.5%)
		(500) (700)	199(16.6%)	198(16.5%)	107(8.9%)	117(9.8%)	82(6.8%)	79(6.6%)
	10	(100) (100)	38(19.0%)	41(20.5%)	41(20.5%)	40(20.0%)	12(6.0%)	13(6.5%)
		(500) (700)	194(16.2%)	198(16.5%)	122(10.2%)	118(9.8%)	78(6.5%)	80(6.7%)
3	2	(100) (100) (100)	74(24.7%)	77(25.7%)	21(7.0%)	23(7.7%)	20(6.7%)	21(7.0%)
		(500) (700) (500)	397(23.4%)	389(22.9%)	101(5.9%)	103(6.1%)	111(6.5%)	107(6.3%)
	5	(100) (100) (100)	71(23.7%)	86(28.7%)	20(6.7%)	22(7.3%)	22(7.3%)	22(7.3%)
		(500) (700) (500)	389(22.9%)	396(23.3%)	111(6.5%)	108(6.4%)	106(6.2%)	115(6.8%)
	10	(100) (100) (100)	75(25.0%)	76(25.3%)	25(8.3%)	23(7.7%)	21(7.0%)	21(7.0%)
		(500) (700) (500)	402(23.6%)	417(24.5%)	108(6.4%)	121(7.1%)	104(6.1%)	108(6.4%)

We utilize the proposed method on a real dataset comprising measurements of seven characteristics from 76 young bulls representing three different breeds (Johnson and Wichern, 2002). In Figure 2, we have plotted the data in the space defined by the first two principal components. Notably, the 2 black points in the plot indicate potential outliers. These observations are visible in Figure 2. The 2 black points mean 2 young bulls are outliers because the first young bull has back fat and sale weight more than the other young bulls, and the second young bull has a fat-free body, percent fat-free body, and sale weight more than the other young bulls.

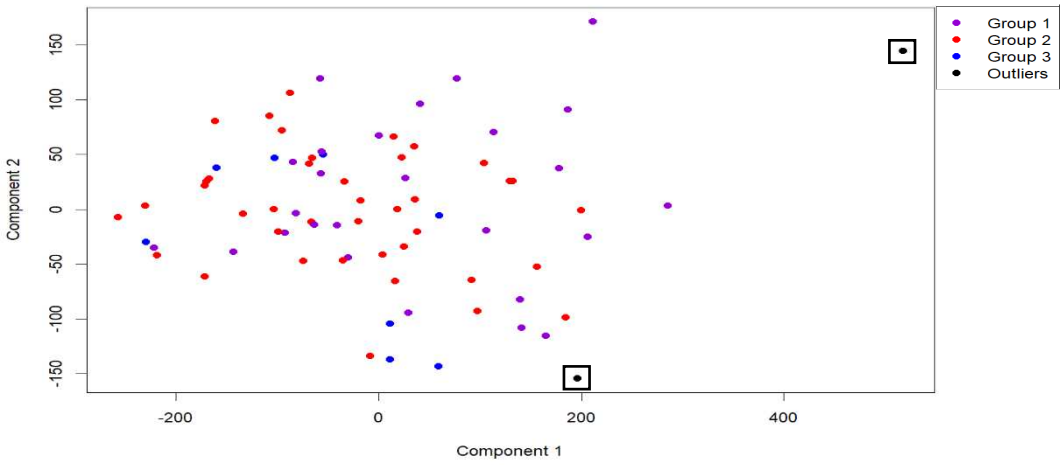

Figure 2: Outlier detection results of the proposed method.

From Figure 2, the proposed method for the detection of outliers can detect two outliers that may be outliers, the same as the Caroni and Billor method. These methods demonstrate a similarity in their outcomes for outlier detection.

3.2 Discussion

Uncontaminated data: the proposed method demonstrates the proposed method exhibits an increased percentage of not detecting outliers (0 and 1) as the number of variables and sample size increase. This suggests that the proposed method still identifies inliers as outliers in uncontaminated data and exhibits an increased percentage of not detecting outliers when dealing with larger sample sizes and more variables. The Caroni and Billor method exhibits a high percentage of detected outliers (0) when the sample size is small, but this percentage decreases as the sample size

increases. This implies that the Caroni and Billor method may struggle to detect outliers effectively in larger sample sizes. The Hardin and Rocke method detects outliers in all cases (2+). This suggests that the Hardin and Rocke method is sensitive enough to identify outliers even in small sample sizes.

6% contaminated data: the results indicate that the proposed method consistently outperforms both the Caroni and Billor method and the Hardin and Rocke method in terms of accuracy and precision. The proposed method demonstrates higher accuracy and precision compared to the Caroni and Billor method in all cases. When $k = 2$ (referring to the number of outliers), the accuracy and precision of the Caroni and Billor method are lower than those of the Hardin and Rocke method. However, when $k = 3$, the precision of the Caroni and Billor method surpasses that of the Hardin and Rocke method. Notably, the Hardin and Rocke method consistently achieves a recall value of 1 in all cases. This indicates that the method correctly identifies all the true outliers.

These findings suggest that the proposed method shows promising performance in outlier detection for both uncontaminated and contaminated data, outperforming the other two methods in terms of accuracy and precision. However, it's important to consider the specific characteristics and limitations of each method and their applicability to different data scenarios.

In real data applications, the proposed outlier detection method can detect two outliers identical to those of Caroni and Billor, but our simulation results are better; our method is better than Hardin and Rocke and Caroni and Billor.
In this sect

4. Conclusion

The proposed outlier detection method demonstrates strong performance in both uncontaminated and 6% contaminated grouped multivariate data that is our proposed method identifies outliers with completeness and accuracy. It exhibits an increased percentage of not detecting outliers in uncontaminated data, especially with larger sample sizes and more variables. Moreover, when compared to the Caroni and Billor method and the Hardin and Rocke method, the proposed method consistently achieves higher accuracy and precision in 6% of contaminated data. However, the specific characteristics and limitations of each method should be considered when choosing the appropriate outlier detection approach for different data scenarios. These findings contribute to the field of outlier detection for grouped multivariate data and provide valuable insights for researchers and practitioners seeking effective methods to identify outliers in various data settings. Further research can explore the performance of these methods in different levels of contamination and evaluate their robustness in real-world applications.

Acknowledgment

This research is fully supported by the Research Fund for Supporting Lecturer to Admit High Potential Student to Study and Research on His Expert Program Year 2022, Graduate School, Khon Kaen University.

References

1. Billor, N., Hadi, A. S. & Velleman, P. F. (2000). BACON: blocked adaptive computationally efficient outlier nominators. *Computational Statistics and Data Analysis*, 34: 279–298. https://doi.org/10.1016/S0167-9473(99)00101-2
2. Caroni, C. & Billor, N. (2007). Robust Detection of Multiple Outliers in Grouped Multivariate Data, *Journal of Applied Statistics*, 34(10): 1241-1250. https://doi.org/10.1080/02664760701592877
3. Ester, M., Kriegel, H. P., Sander, J. & Xu, X. (1966). A density-based algorithm for discovering clusters in large spatial databases with noise, Proceedings of 2nd International Knowledge Discovery and Data Mining, 226–231.
4. Hardin, J. & Rocke, D. M. (2004). Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator, *Computational Statistics and Data Analysis*, 44: 625–638. https://doi.org/10.1016/S0167-9473(02)00280-3
5. Johnson, R. A. & Wichern, D. W., 2002. *Applied Multivariate Statistical Analysis*, fifth ed. Prentice Hall, New Jersey.
6. Kelleher, S. (2022). Can Diabetes Cause High Blood Pressure?, https://www.endocrineweb.com/condition/diabetes/diabetes-and-high-blood-pressure, (accessed 11 May 2023).

7. Montgomery, D. C., Peck, E. A. & Vining, G. G., 2012. *Introduction to Linear Regression Analysis*, third ed. John Wiley & Sons, New York.
8. Punzo, A. & McNicholas, P. D. (2016). Parsimonious Mixtures of Multivariate Contaminated Normal Distributions, *Biometrical Journal*, 58(6): 1506–1537. https://doi.org/10.1002/bimj.201500144
9. Santoyo, S. (2017). A Brief Overview of Outlier Detection Techniques. http://engdashboard.blogspot.com/2017/09/a-brief-overview-of-outlier-detection.html, (accessed 30 October 2021).