Jurnal
Teknologi

Full Paper

SYNTHETIC MULTIVARIATE DATA GENERATION PROCEDURE WITH VARIOUS OUTLIER SCENARIOS USING R PROGRAMMING LANGUAGE

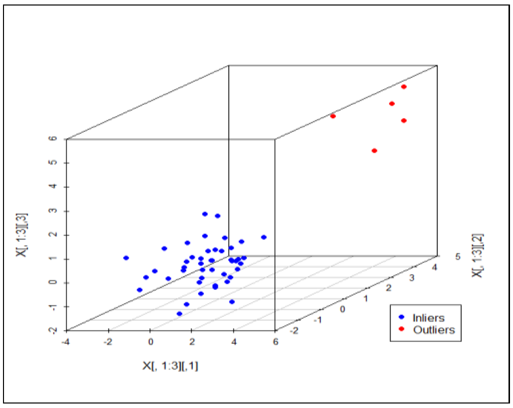
Sharifah Sakinah Syed Abd Mutalib[a,b], Siti Zanariah Satari[a*], Wan Nur Syahidah Wan Yusoff[a]

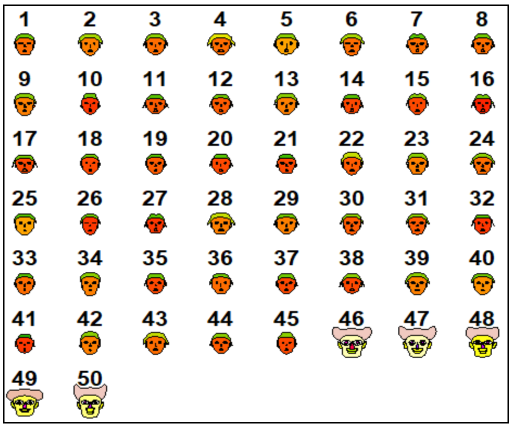
[a]Centre for Mathematical Sciences, College of Computing and Applied Sciences, Universiti Malaysia Pahang, 26300 Gambang, Kuantan, Pahang, Malaysia
[b]Faculty of Computer, Media and Technology Management, University College TATI, Jalan Panchur, Telok Kalong, 24000 Kemaman, Terengganu, Malaysia

Article history
Received
29 October 2021
Received in revised form
11 February 2022
Accepted
17 February 2022
Published Online
20 April 2022

*Corresponding author
zanariah@ump.edu.my

Graphical abstract





Abstract

A synthetic data generation procedure is a procedure to generate data from either a statistical or mathematical model. The data generation procedure has been used in simulation studies to compare statistical performance methods or propose a new statistical method with a specific distribution. A synthetic multivariate data generation procedure with various outlier scenarios using R is formulated in this study. An outlier generating model is used to generate multivariate data that contains outliers. Data generation procedures for various outlier scenarios by using R are explained. Three outlier scenarios are produced, and graphical representations using 3D scatterplot and Chernoff faces for these outlier scenarios are shown. The graphical representation shows that as the distance between outliers and inliers by shifting the mean, λ increases in Outlier Scenario 1, the outliers and inliers are completely separated. The same pattern can also be seen when the distance between outliers and inliers, by shifting the covariance, δ increase in Outlier Scenario 2. For Outlier Scenario 3, when both values λ and δ increase, the separation of outliers and inliers are more apparent. The data generation procedure in this study will be continually used in other applications, such as identifying outliers by using the clustering method.

Keywords: Data generation procedure, multivariate data, outlier generating model, Chernoff faces, scatterplot 3D, R

Abstrak

Prosedur penghasilan data sintetik ialah satu prosedur untuk menghasilkan data daripada model statistik atau matematik. Dengan taburan spesifik, prosedur penghasilan data telah digunakan dalam kajian simulasi untuk membandingkan prestasi kaedah statistik atau mencadangkan kaedah statistik baru. Dalam kajian ini, satu prosedur penghasilan data multivariat sintetik dengan pelbagai senario data terpencil menggunakan R diformulakan. Satu model penghasilan data terpencil digunakan untuk menghasilkan data multivariat yang mengandungi data terpencil. Langkah-langkah untuk prosedur penghasilan data dengan pelbagai senario data terpencil menggunakan R dijelaskan. Tiga senario data terpencil dihasilkan dan perwakilan grafik menggunakan *scatterplot* 3D

84:3 (2022) 89–101 | https://journals.utm.my/jurnalteknologi | eISSN 2180–3722 | DOI: https://doi.org/10.11113/jurnalteknologi.v84.17900 |

dan *Chernoff faces* ditunjukkan. Perwakilan grafik menunjukkan bahawa apabila jarak antara data terpencil dan bukan data terpencil dengan mengalih *mean*, λ meningkat dalam Senario Data Terpencil 1, data terpencil dan bukan data terpencil akan terpisah sepenuhnya. Keputusan yang sama juga boleh dilihat apabila jarak antara data terpencil dan bukan data terpencil dengan mengalih *covariance*, δ meningkat dalam Senario Data Terpencil 2. Bagi Senario Data Terpencil 3, apabila kedua-dua nilai λ dan δ meningkat, pemisahan antara data terpencil dan bukan terpencil semakin jelas. Prosedur penghasilan data dalam kajian ini akan terus digunakan dalam aplikasi yang lain seperti mengenalpasti data terpencil dengan menggunakan kaedah berkelompok.

Kata kunci: Prosedur penghasilan data, data multivariat, model penghasilan data terpencil, *Chernoff faces*, *scatterplot* 3D, R.

1.0 INTRODUCTION

The performance of new multivariate techniques is often tested on various data sets by using simulation studies [1]. These data sets are generated through a statistical model with specific data distribution. The procedure to generate the data is known as data generating procedure. A data generating procedure is a generic term to generate data using a data model, whether it is a statistical or mathematical model. The data produced by this procedure is known as synthetic data.

The limitation of real data had motivated most of the studies to use data generating procedures to produce synthetic data to test the proposed method's performance. For example, [1] used data generating procedure to propose a new algorithm for random data simulation where sample size and correlation can be set, and a new algorithm named *SimuleMV* was introduced. Data generating procedure was used by [2] to generate multivariate non-normal random numbers with given multivariate skewness and kurtosis via a simulation study. The study done by [2] is to overcome the problem in behavioural and social science as the data are rarely normally distributed in practice. Meanwhile, 500 data sets were generated by [3] to study the performance of the clinical model also via a simulation study. Synthetic data was also used in simulation studies such as in [1, 4–7]. Specific parameters can be set and tested throughout the studies by using synthetic data.

In multivariate data, one of the interests is to identify outliers by using robust estimation of mean and covariance matrix. Outliers' studies are closely related to robust estimators [8]. Most of the outliers' studies involved the process of proposing a new robust estimator [5, 9]. The performance of a new robust estimator can be tested via a simulation study, and a data generation procedure is used in a simulation study to generate data. Outlier is abnormal data that differs from most of the data [10]. Multivariate data is a set of data represented by $n \times p$ matrix where n is the sample size and p is the number of variables [11].

The presence of outliers in multivariate data are normally occurs and can be expected. However, the outliers can affect proper classical multivariate analysis, lead to incorrect conclusions, make modelling difficult and disrupt the mean and covariance matrix measures. Previous studies such as in [6, 9, 12] used synthetic data via simulation study to propose new robust estimators. Some application of outliers identification in multivariate data has been found, such as in geosciences [13, 14], financial data [15, 16] and medical [17, 18].

In the context of outliers for multivariate, the outlier generating model is used to generate multivariate data that contain outliers [19]. Most of the studies used multivariate normal distribution and called the outlier generating model as a mixture of *p*-variate normal distributions [4, 5, 9]. In this study, three outlier scenarios will be produced from the outlier generating model, which is the mean-shift model (Outlier Scenario 1), variance-inflation model (Outlier Scenario 2), and mean-shift and variance-inflation model (Outlier Scenario 3). Details for these outlier scenarios are explained in Section 2.1.

The objective of this study is to formulate a synthetic data generation procedure for multivariate data with various outlier scenarios using R. Most of the previous studies do not explain in detail about the data generation procedure for outlier scenarios in multivariate data. Studies about outliers in multivariate data only stated about the model that is used to generate data but not for the procedure such as in [4–6, 20, 21]. Procedure for new or proposed methods is explained in those studies with generated data is stated in general.

The rest of the paper is organized as follows. The following section explains about outlier generating model, which discusses three outlier scenarios in detail. Then, the data generation procedure for each outlier scenario using R is explained. A graphical representation for each outlier scenario using R is shown and discussed in the results and discussion section. The last section presents the conclusion of this study.

2.0 METHODOLOGY

2.1 Outlier Generating Model for Multivariate Data

Random data are generated from the following outlier generating model and is given as,

$$(1-\varepsilon)N_p(\vec{\mu}_0,\Sigma_0)+\varepsilon N_p(\lambda\vec{\mu}_1,\delta\Sigma_1) \qquad (1)$$

Where $\Sigma_0=\Sigma_1=I_p$, $\vec{\mu}_0=(0\,0\ldots0)'$ and $\vec{\mu}_1=(1\,1\ldots1)'$ is of dimension p. Many outlier studies have used these outlier generating model such as in [5,9,20].

Inliers are generated from $N_p(\vec{\mu}_0,\Sigma_0)$, whereas outliers are generated from $N_p(\lambda\vec{\mu}_1,\delta\Sigma_1)$ where λ and δ are the separation between outliers and inliers by shifting mean and covariance. The percentage of outliers, ε will be used to determine the number of outliers in the data. For example, if the data has a sample size, $n=50$ and percentage of outliers, $\varepsilon=0.1$, there will be five outliers in the data. In this study, the sample size of 50 and the percentage of outliers of 10% is used for illustration purposes. Various sample sizes and percentages of outliers were used by using the same R code in [6] and [7] studies. Sample size of 50 is chose since most of the historical multivariate data with outliers has sample sizes within 28 to 86 [22].

From the outlier generating model in Equation (1), λ and δ will determine the outliers' scenarios and the separation between outliers and inliers. There will be three outliers' scenarios. The λ values that used in this study are 1, 2, 4 and 10, and the δ values used in this study are 0.5, 2, 10 and 25. These values are based on the studies done by [4, 5, 8].

In the first scenario (Outlier Scenario 1), the separation between outliers and inliers is determined by the value of λ by shifting the mean, whereas the value of δ is fixed to 1. Outlier Scenario 1 also be named as the mean-shift model, the shifts of location, shift outliers or location outliers [23, 24]. Outlier Scenario 1 is given by Equation (2).

$$(1-\varepsilon)N_p(\vec{\mu}_0,\Sigma_0)+\varepsilon N_p(\lambda\vec{\mu}_1,\Sigma_1). \qquad (2)$$

The second outliers scenario (Outlier Scenario 2) has also been named the shift of scale or scatter outliers [5,24]. The separation of outliers and inliers in this scenario will be determined by the value of δ by shifting the covariance, whereas the value of λ is fixed to 0. Outlier Scenario 2 is given by Equation (3).

$$(1-\varepsilon)N_p(\vec{\mu}_0,\Sigma_0)+\varepsilon N_p(0,\delta\Sigma_1). \qquad (3)$$

Outlier Scenario 3 shifts both λ and δ simultaneously. Outliers and inliers are separated by shifting the mean and covariance simultaneously. Outlier Scenario 3 is given as follows,

$$(1-\varepsilon)N_p(\vec{\mu}_0,\Sigma_0)+\varepsilon N_p(\lambda\vec{\mu}_1,\delta\Sigma_1). \qquad (4)$$

Random data will be generated according to these outlier scenarios by using the formulated procedure in the next section.

2.2 Synthetic Data Generation Procedure using R

This section presents the procedure and coding to generate synthetic multivariate data using R version 4.0.3. R is an open-source statistical software for the users. The following are the general steps to generate synthetic multivariate data, and Table 1 shows R coding for the procedure for each outlier scenario.

Step 1. Load ‘`MASS`’, ‘`base`’, ‘`scatterplot3d`’ and ‘`aplpack`’ packages in R.

Step 2. Define inputs for the outlier generating model.
(i) n : sample size
(ii) p : number of variable
(iii) ε : percentage of outliers
(iv) Values for λ and δ,
(v) $\vec{\mu}_0=(0\,0\ldots0)'$
(vi) $\vec{\mu}_1=(1\,1\ldots1)'$

Step 3. Use ‘`mvrnorm`’ function to generate the multivariate data randomly.

Table 1 R code for synthetic multivariate data generation for each outlier scenario $(n=50,\ \varepsilon=0.1)$

Outlier Scenarios	R code

1
(Eq. 2)
	`n<-50`
	`p<-p      #p=3,5`
	`e<-0.1`
	`mu0<-rep(0,p)`
	`mu1<-rep(λ,p)`
	`sigma<-diag(p)`
	`n1<-floor((1-e)*n)`
	`n2<-ceiling(e*n)`
	`X<-rbind(mvrnorm(n1,mu0,sigma),`
	`mvrnorm(n2,mu1,sigma))`

2
(Eq. 3)
	`n<-50`
	`p<-p      #p=3,5`
	`e<-0.1`
	`mu<-rep(0,p)`
	`sigma0<-diag(p)`
	`sigma1<-δ*diag(p)`
	`n1<-floor((1-e)*n)`
	`n2<-ceiling(e*n)`
	`X<-rbind(mvrnorm(n1,mu,sigma0),`
	`mvrnorm(n2,mu,sigma1))`

3
(Eq. 4)
	`n<-50`
	`p<-p     #p=3,5`
	`e<-0.1`
	`mu0<-rep(0,p)`
	`mu1<-rep(λ,p)`
	`sigma0<-diag(p)`
	`sigma1<-δ*diag(p)`
	`n1<-floor((1-e)*n)`
	`n2<-ceiling(e*n)`
	`X<-rbind(mvrnorm(n1,mu0,sigma0),`
	`mvrnorm(n2,mu1,sigma1))`

3.0 RESULTS AND DISCUSSION

For illustration purposes, sample size of $n=50$, number of variables of $p=3,5$ and percentage of outliers, $\varepsilon=0.1$ are chosen. For each outlier scenario, 10% outliers which equivalent to five outliers are planted. From Table 1, n2 is coded as outliers and are arranged as the last five observations (observations 46-50). For Outlier Scenario 1, $\lambda=1,2,4$ and 10 are used. While $\delta=0.5,2,10$ and 25 are used for Outlier Scenario 2.

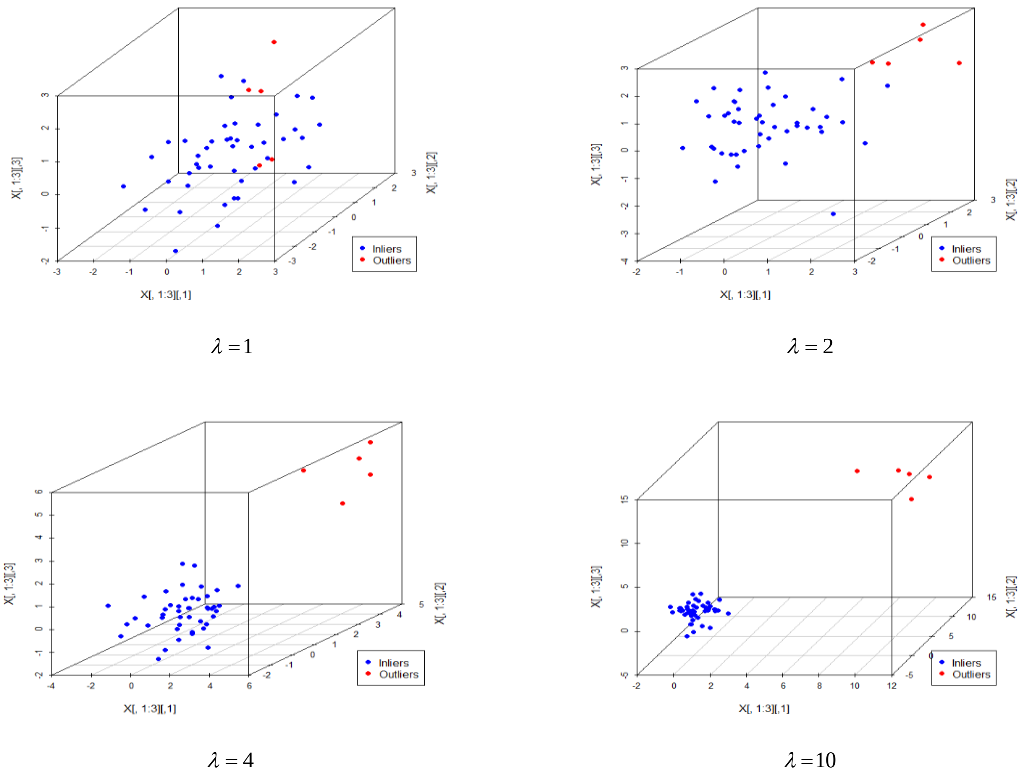
Outlier Scenario 3 used combination values of λ and δ.

3D scatterplot $(p=3)$ and Chernoff faces $(p=5)$ are used to illustrate the position of outliers in each outlier scenario. Chernoff faces is one of the methods to represent multivariate data graphically and was invented by Chernoff (1973). Each observation represented the properties of the face, such as the height of the face, the width of the eyes, and the styling of hair [25, 26]. According to [25], by using Chernoff faces, it is easy for the human mind to grasp the normality and abnormality in the data.

3.1 Outlier Scenario 1

Figures 1 and 2 show graphically the position of outliers for $\lambda = 1, 2, 4$ and 10 for $p=3$ and $p=5$. From the 3D scatterplot in Figure 1 with $p=3$, the outliers and inliers are mixed and difficult to distinguish when $\lambda = 1$ and $\lambda = 2$. Outliers also are far from each other for $\lambda = 1$ and $\lambda = 2$. Both of these conditions will make the detection of outliers difficult. As the λ values increase ($\lambda = 4$ and $\lambda = 10$), the outliers and inliers can be wholly separated. For $\lambda = 4$ and $\lambda = 10$, the detection of outliers is easier compared to $\lambda = 1$ and $\lambda = 2$. As the λ values increase, the success rate in detecting outliers increases. Outliers and inliers also form a clear separation as the λ values increase.

Figure 2 shows Chernoff faces for multivariate data with five variables. The same pattern can be seen for $p=5$. The last five observations (outliers) show different faces from inliers as the λ value increase. For $\lambda = 1$, the faces are mixed and difficult to classify whether it is outliers or inliers. However, observation 47 show a significant difference face from inliers. As the λ values increase, observations 46-50 can be wholly distinguished and can be classified as outliers.



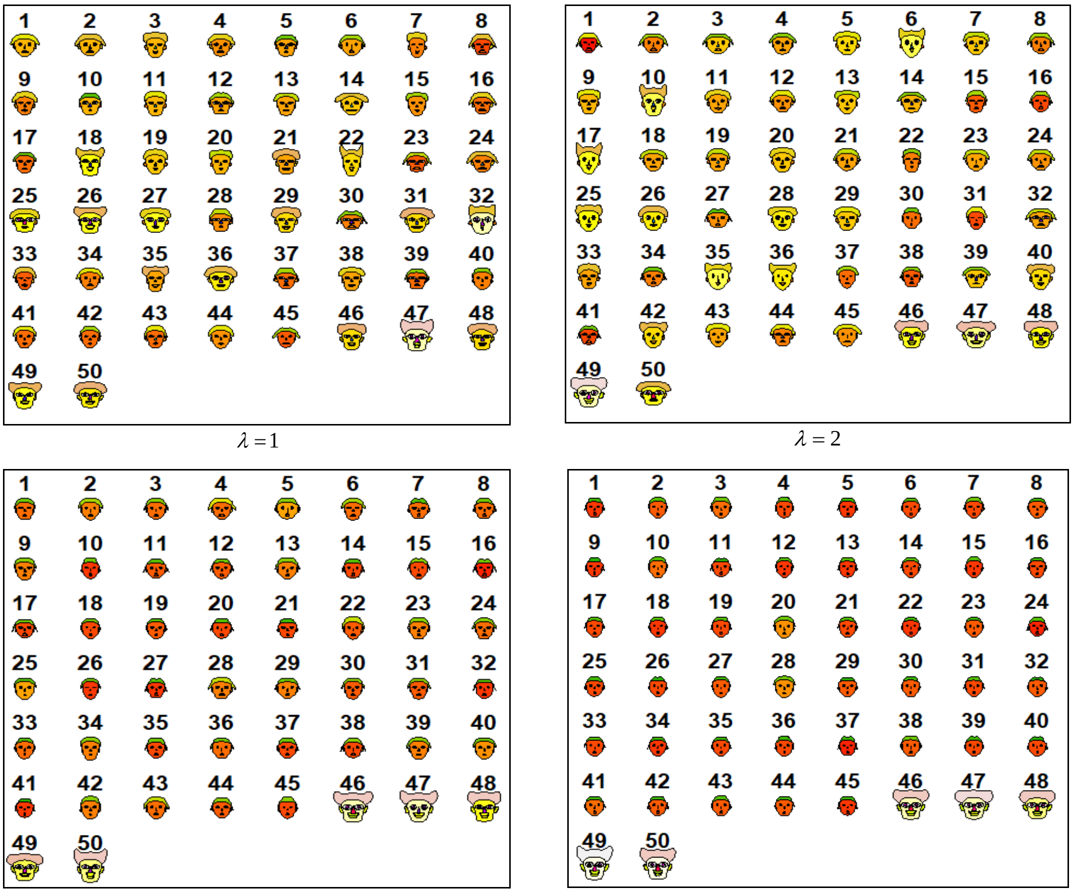
$\lambda = 1$

$\lambda = 2$

$\lambda = 4$

$\lambda = 10$

Figure 1 3D scatterplots illustration of Outlier Scenario 1 for *p*=3



$\lambda=1$

$\lambda=2$

$\lambda=4$

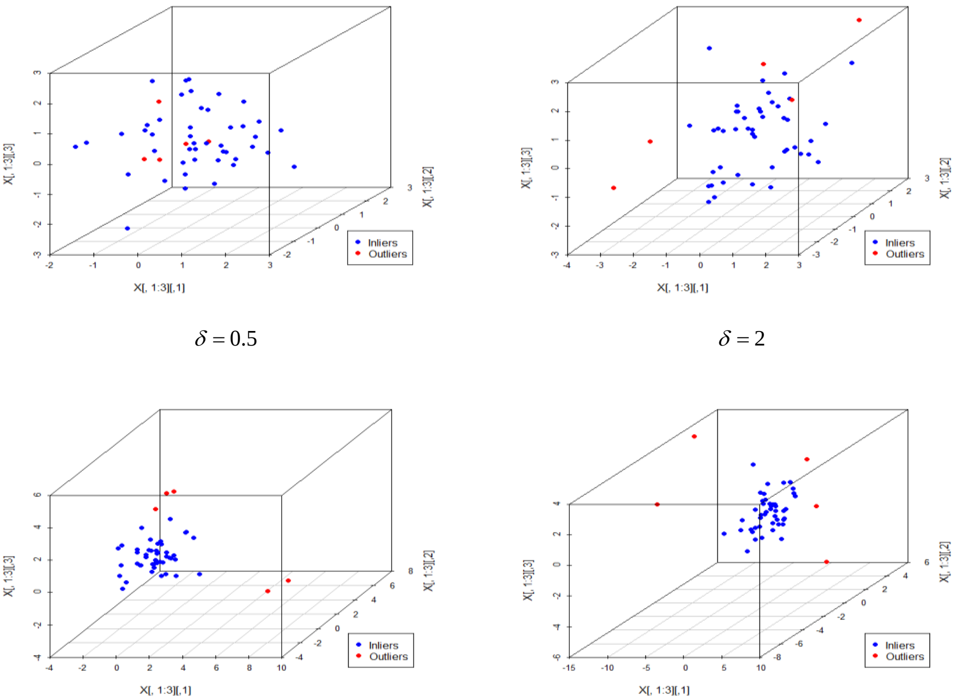
$\lambda=10$

Figure 2 Chernoff faces illustration of Outlier Scenario 1 for *p=5*

3.2 Outlier Scenario 2

Figures 3 and 4 show the position of outliers when covariance is shifted. From the 3D scatterplots Figure 3, the separation of outliers and inliers is getting more clearer as the values δ increase. For $\delta=0.5$, there is no separation between the outliers and inliers. For $\delta=2$, the separation between the outliers and inliers becomes more apparent, but there are still outliers mixed with inliers. For $\delta=10$, outliers and inliers are wholly separated, but a few outliers are still close to inliers. As the δ values increase, the outliers and inliers are entirely separated from each other and can be seen for $\delta=25$.

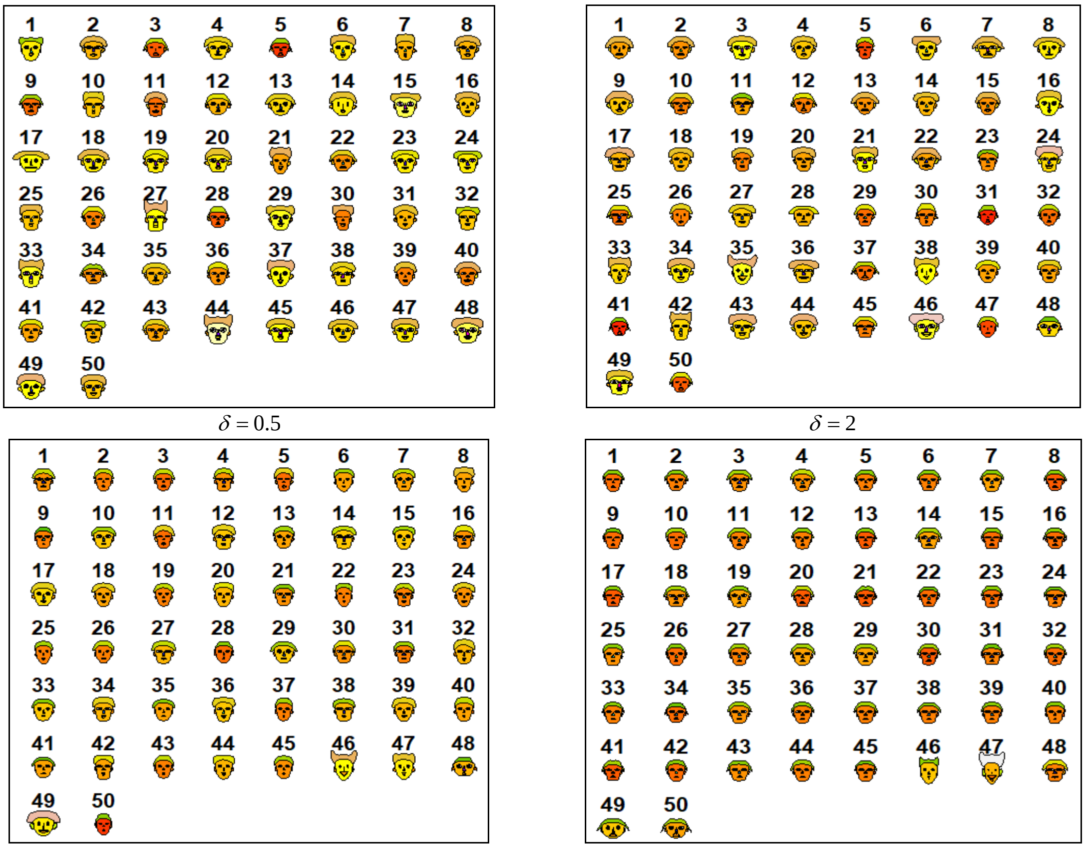
From the Chernoff faces in Figure 4, the separation of outliers and inliers are unclear between faces for $\delta=0.5$ and $\delta=2$. For $\delta=0.5$, face for observation 50th (outliers) are pretty similar to observation 25th (inliers). Outliers and inliers are still mixed for $\delta=2$ and can be seen from the face of observation 50th (outliers) and 29th (inliers). For $\delta=10$ and $\delta=25$, faces for all outliers are different from inliers. As the value δ increases, it can be seen that the faces of outliers show different features from inliers .



$\delta = 0.5$ $\delta = 2$

$\delta = 10$ $\delta = 25$

Figure 3 3D scatterplots illustration of Outlier Scenario 2 for *p*=3



$\delta = 0.5$ $\delta = 2$

$\delta = 10$ $\delta = 25$

Figure 4 Chernoff faces illustration of Outlier Scenario 2 for *p=5*

3.3 Outlier Scenario 3

Outlier Scenario 3 used combinations values of λ and δ. Figures 5 and 6 illustrate the 3D scatterplot of Outlier Scenario 3 for the number of variables, $p=3$. Both figures show as the values λ increase, the separation between outliers and inliers becomes clearer. The same condition also can be seen for the values δ. Two different clusters can be seen formed when the values of λ and δ increase. For $\lambda=1$, outliers and inliers are still mixed together, and a few outliers are separated from inliers as the δ values increase. For $\lambda=2$, the separation of outliers and inliers becomes more apparent as the δ value increases. Inliers can be seen forming one compact cluster and separated far from the outliers for $\lambda=4$ and $\lambda=10$ for all δ values.

Figures 7 and 8 show the Chernoiff face illustrations for Outlier Scenario 3 for the number of variables, $p=5$. From Figure 7, the faces of outliers are not too different from the inliers for $\lambda=1$. As the δ values increase, the faces of outliers show different features from inliers. From Figures 7 and 8, the faces of outliers show different features for all δ values for $\lambda=2$. As the λ values increase, the faces of outliers can be seen showing different features clearly from inliers for all δ values.

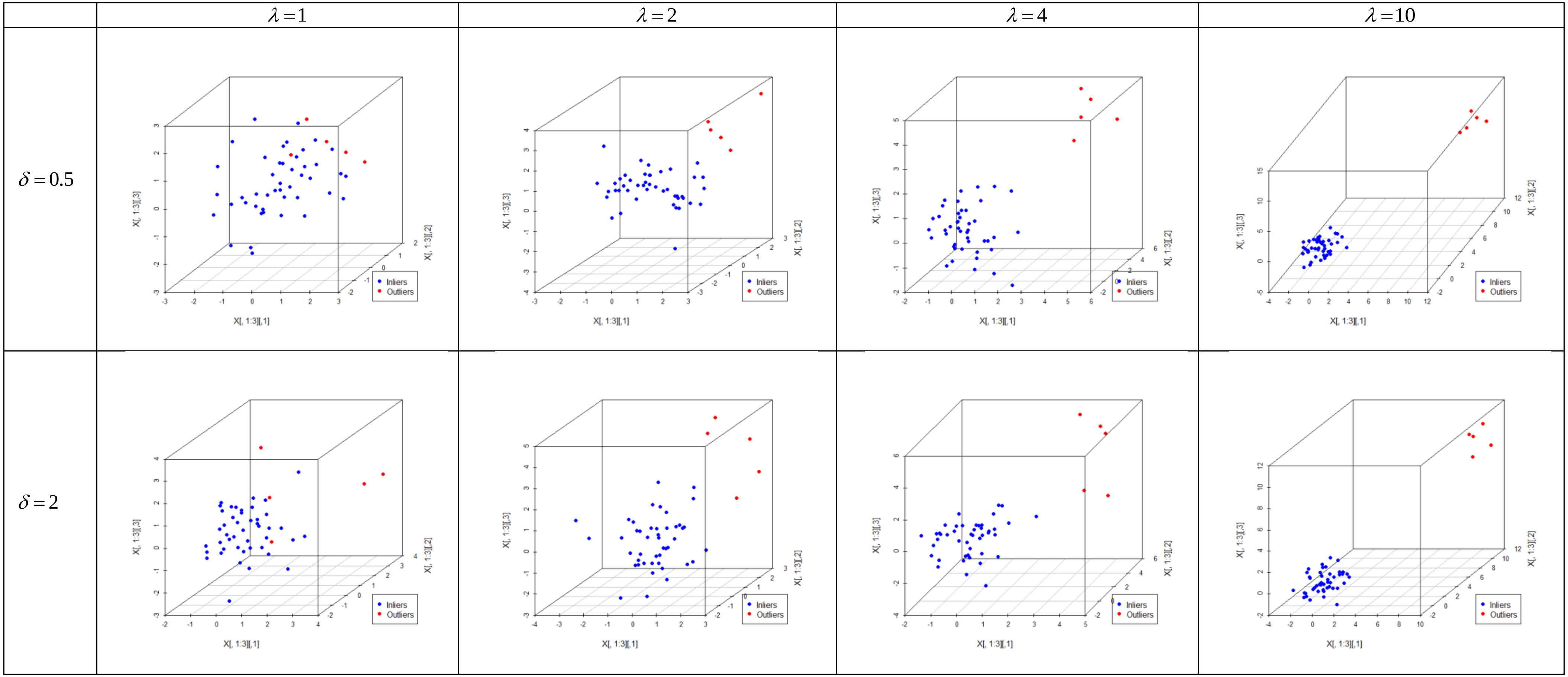


Figure 5 3D scatterplots illustration of Outlier Scenario 3 for *p*=3 and $\delta = 0.5, 2$.

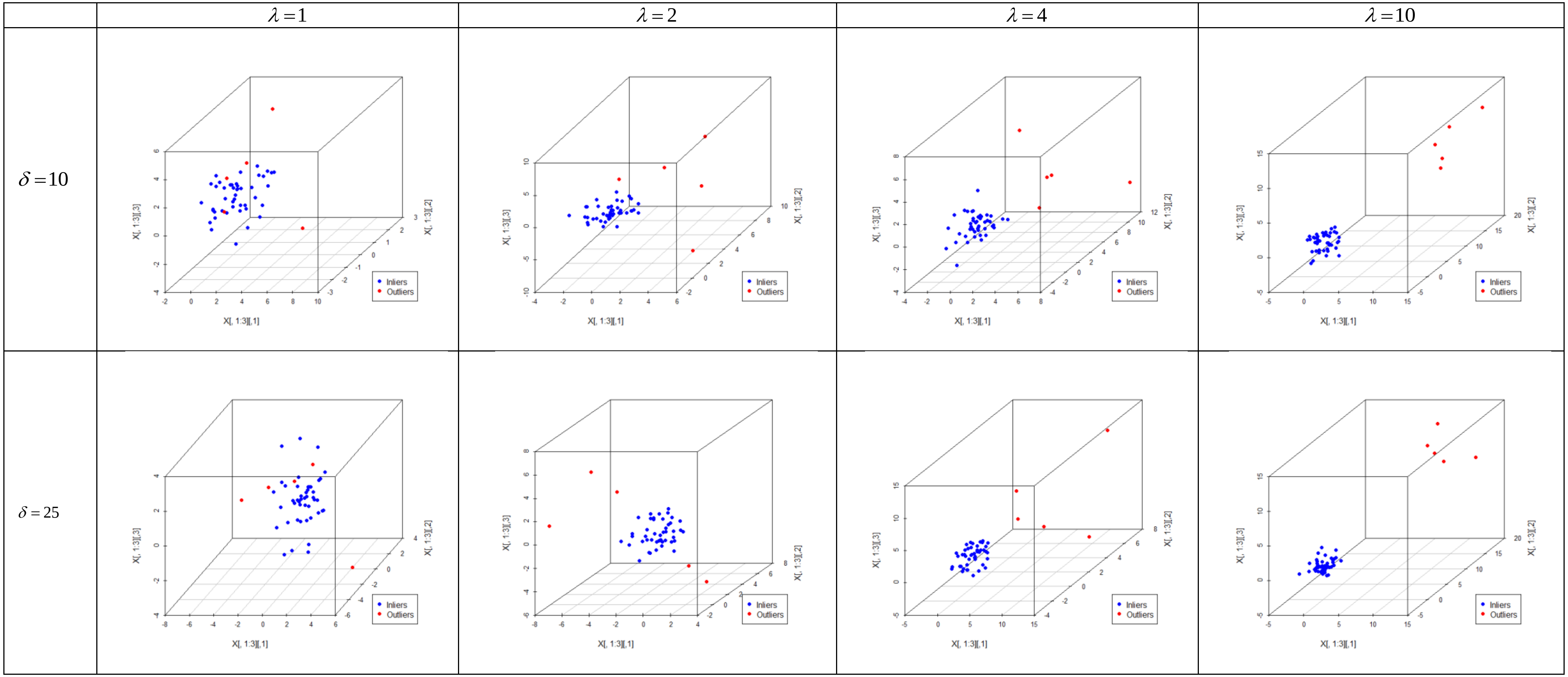


Figure 6 3D scatterplots illustration of Outlier Scenario 3 for *p*=3 and $\delta=10, 25.$

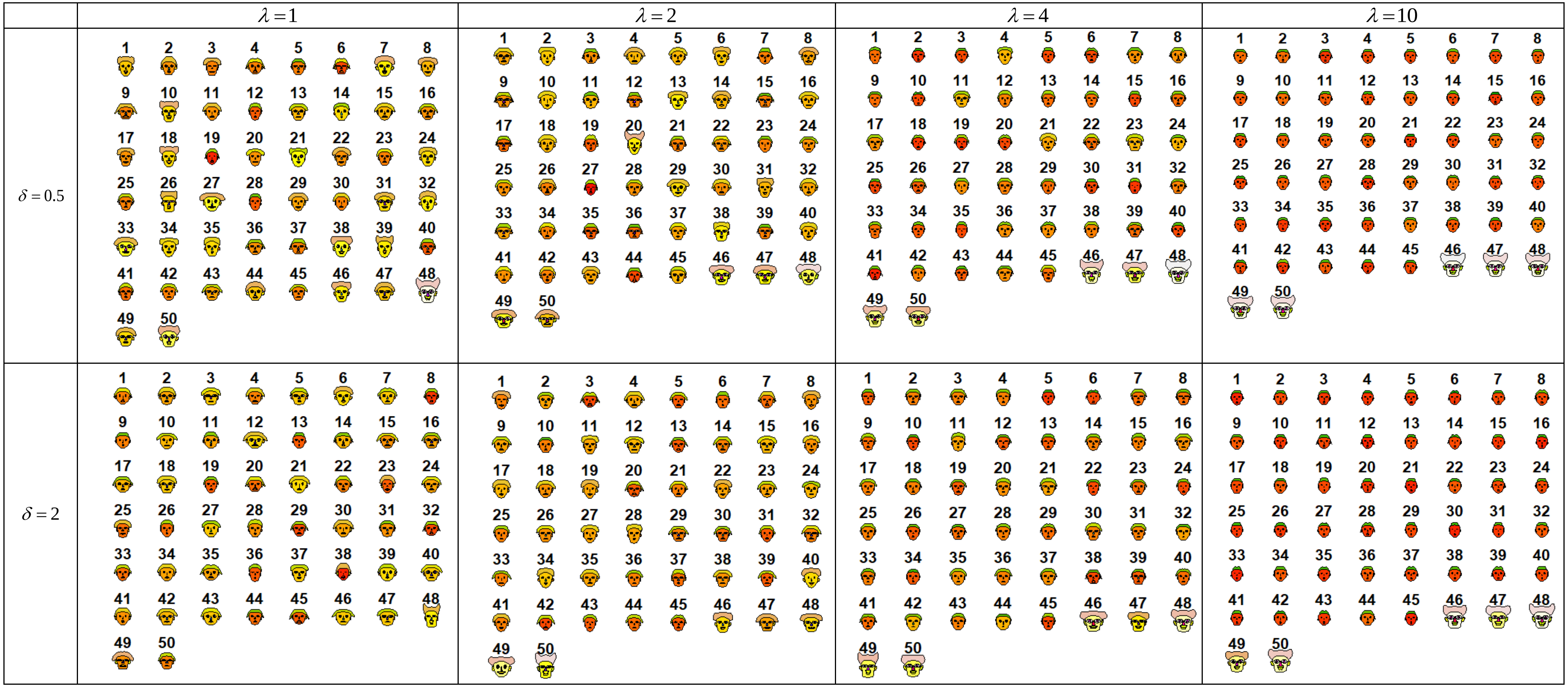


Figure 7 Chernoff faces illustration of Outlier Scenario 3 for *p*=5 and $\delta = 0.5, 2$.

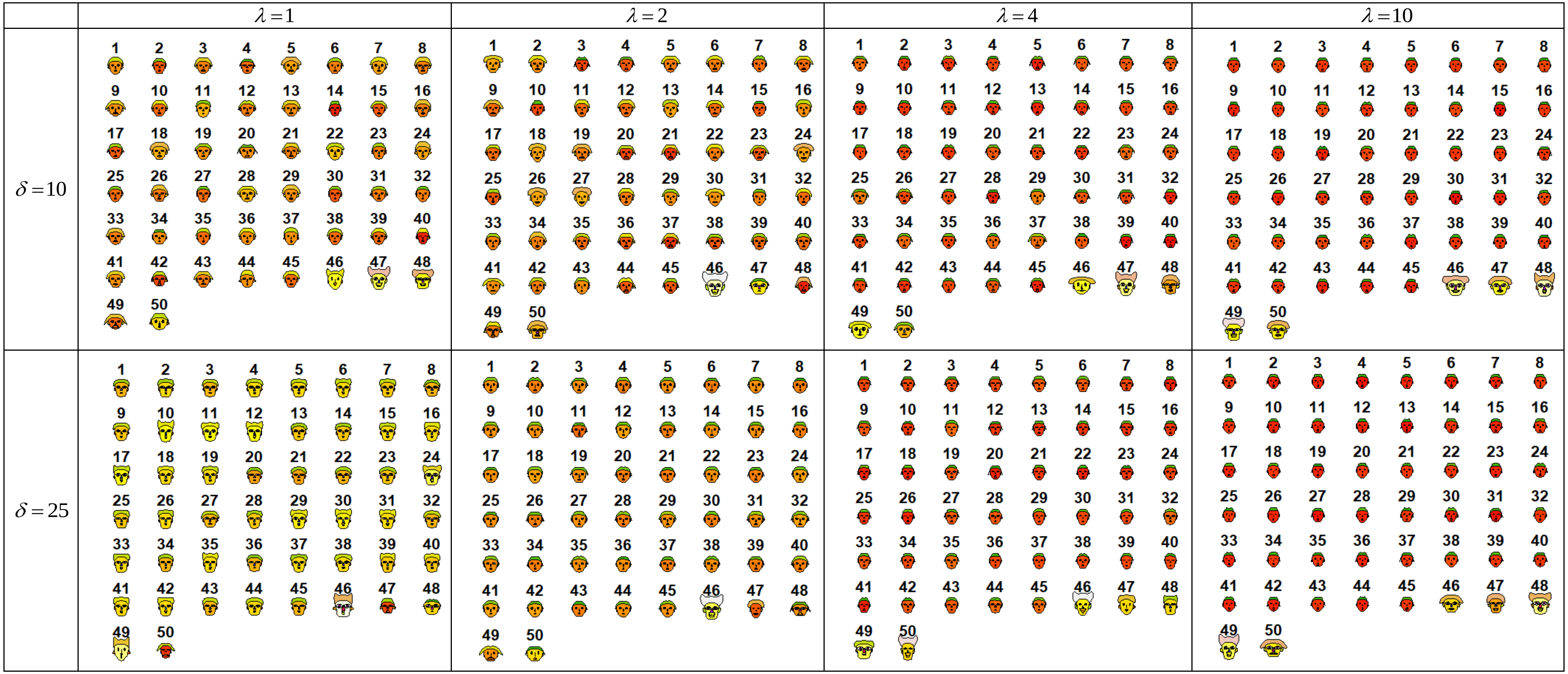


Figure 8 Chernoff faces illustration of Outlier Scenario 3 for *p*=5 and $\delta = 10, 25$.

Table 2 Findings from three outlier scenarios

Outlier Scenario	Number of variables, p	
	3	5
1	Outliers and inliers form a clear separation as the λ values increase.	As the λ values increase, features of faces for outliers are different and can be distinguished from inliers.
2	As the δ values increase, outliers and inliers are completely separated.	As the δ values increase, it can be seen that the faces of outliers show different features from inliers.
3	As the values of λ and δ increase, the separation between outliers and inliers becomes clearer.	As the values of λ and δ increase, the faces of outliers can be seen showing different features clearly from inliers.

Table 2 shows the findings from the three outlier scenarios of multivariate data. The results show a similar pattern between the two variables of $p=3$ and $p=5$. Overall, the results show that as the value for shift mean and covariance increase, the separation between the outliers and inliers becomes clearer.

4.0 CONCLUSION

In this study, a synthetic multivariate data generation procedure for three outlier scenarios using R is formulated and presented. Three outliers scenarios and steps to generate multivariate data that contain outliers are explained. Graphical representation using scatter plot 3D $(p=3)$ and Chernoff faces $(p=5)$ for the three outlier scenarios are also provided. The graphical representation is used to show the position of outliers and inliers in the three outlier scenarios.

For both number of variables, it can be concluded that as the λ value increases in Outlier Scenario 1, the outliers and inliers are entirely separated. The same pattern also can be seen when the δ value increases in Outlier Scenario 2. For Outlier Scenario 3, when both values of λ and δ increase, the separation of outliers and inliers are more apparent.

Previous studies only stated a model to generate outliers but did not explain in detail the procedure to generate the multivariate data that contained outliers. Hence, this study presented the data generation procedure for multivariate data that contained outliers in detail. This data generation procedure will be continually used and applied in outlier detection methods such as clustering for future studies.

Acknowledgement

The authors would like to thank Universiti Malaysia Pahang (UMP) for the financial support under the internal grant (RDU190363) and University College TATI (UCTATI).

References

[1] Camacho, J. 2017. On the Generation of Random Multivariate Data. *Chemometrics and Intelligent Laboratory Systems*. 160: 40-51. DOI: 10.1016/j.chemolab.2016.11.013.

[2] Qu, W., Liu, H., and Zhang, Z. 2020. A Method of Generating Multivariate Non-normal Random Numbers with Desired Multivariate Skewness and Kurtosis. *Behavior Research Methods*. 52(3): 939-946. DOI: 10.3758/s13428-019-01291-5.

[3] Riley, R. D., Snell, K. I. E., Martin, G. P., Whittle, R., Archer, L., Sperrin, M., and Collins, G. S. 2021. Penalization and Shrinkage Methods Produced Unreliable Clinical Prediction Models Especially when Sample Size was Small. *Journal of Clinical Epidemiology*. 132: 88-96. DOI: 10.1016/j.jclinepi.2020.12.005.

[4] Cerioli, A., Marco, R., and Francesca, T. 2011. Accurate and Powerful Multivariate Outlier Detection. *Int Statistical Inst: Proc. 58th World Statistical Congress 2011, Dublin*. 5608-5613. https://2011.isiproceedings.org/papers/950478.pdf.

[5] Filzmoser, P., Maronna, R., and Werner, M. 2008. Outlier Identification in High Dimensions. *Computational Statistics & Data Analysis*. 52: 1694-1711. DOI: 10.1016/j.csda.2007.05.018.

[6] Abd Mutalib, S. S. S., Satari, S. Z., and Wan Yusoff, W. N. S. 2019. A New Robust Estimator to Detect Outliers for Multivariate Data. *Journal of Physics: Conference Series*. 1366: 1-9. DOI: 10.1088/1742-6596/1366/1/012104.

[7] Abd Mutalib, S. S. S., Satari, S. Z., and Wan Yusoff, W. N. S. 2021. Comparison of Robust Estimators' for Detecting Outliers in Multivariate Data. *Journal of Statistical Modeling and Analysis*. 3(2): 36-64. DOI: 10.1088/1742-6596/1988/1/012095.

[8] Werner, M. 2003. Identification of Multivariate Outliers in Large Data Sets. Doctoral Thesis, University of Colorado. http://math.ucdenver.edu/graduate/thesis/werner_thesis.pdf.

[9] Herwindiati, D. E., Djauhari, M. A., and Mashuri, M. 2007. Robust Multivariate Outlier Labeling. *Communications in Statistics - Simulation and Computation*. 36(6): 1287-1294. DOI: 10.1080/03610910701569044.

[10] Su, X., and Tsai, C-L. 2011. Outlier Detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 1(3): 261-268. DOI: 10.1002/widm.19.

[11] Johnson, R. A., and Wichern, D. W. 2002. *Applied Multivariate Statistical Analysis*. Fifth Edition. Prentice Hall, Inc.

[12] Rousseeuw, P. J., and Van Driessen, K. 1999. A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*. 41(3): 212-223. DOI: 10.1080/00401706.1999.10485670.

[13] Filzmoser, P., and Gregorich, M. 2020. Multivariate Outlier Detection in Applied Data Analysis: Global, Local, Compositional and Cellwise Outliers. *Mathematical Geosciences*. 52 (8): 1049-1066. DOI: 10.1007/s11004-020-09861-6.

[14] Zheng, S., Zhu, Y. X., Li, D. Q., Cao, Z. J., Deng, Q. X., and Phoon, K. K. 2021. Probabilistic Outlier Detection for Sparse Multivariate Geotechnical Site Investigation Data using Bayesian Learning. *Geoscience Frontiers*. 12(1): 425–439. DOI: 10.1016/j.gsf.2020.03.017.

[15] Ismail, M. T., and Mohd Nasir, I. N. 2019. Outliers in Islamic and Conventional Stock Indices: An Empirical Analysis using Impulse Saturation Indicator. *ASM Science Journal*. 12 (Special Issue 5): 130-136. https://www.akademisains.gov.my/asmsj/article/outliers

-in-islamic-and-conventional-stock-indices-an-empirical-analysis-using-impulse-saturation-indicator/.

[16] Domino, K. 2020. Multivariate Cumulants in Outlier Detection for Financial Data Analysis. *Physica A: Statistical Mechanics and its Applications*. 558: 1-13. DOI: 10.1016/j.physa.2020.124995.

[17] Estiri, H., and Murphy, S. N. 2019. Semi-supervised Encoding for Outlier Detection in Clinical Observation Data. *Computer Methods and Programs in Biomedicine*. 181: 1-16. DOI: 10.1016/j.cmpb.2019.01.002.

[18] Abuzaid, A. H. 2020. Identifying Density-based Local Outliers in Medical Multivariate Circular Data. *Statistics in Medicine*. 39(21): 2793-2798. DOI: 10.1002/sim.8576.

[19] Barnett. V., and Lewis, T. 1984. *Outliers in Statistical Data*. Second Edition. John Wiley and Sons.

[20] Wada, K., Kawano, M., and Tsubaki, H. 2020. Comparison of Multivariate Outlier Detection Methods for Nearly Elliptical Distributions. *Austrian Journal of Statistics*. 49: 1-17. DOI: 10.17713/ajs.v49i2.872.

[21] Djauhari, M. A., Mashuri, M., and Herwindiati, D. E. 2008. Multivariate Process Variability Monitoring. *Communication in Statistics - Theory and Methods*. 37(11): 1742-1754. DOI: 10.1080/03610920701826286.

[22] Abd Mutalib, S. S. S., Satari, S. Z., and Wan Yusoff, W. N. S. 2021. Comparison of Robust Estimators for Detecting Outliers in Multivariate Datasets. *Journal of Physics: Conference Series*. 1988: 1-9. DOI: 10.1088/1742-6596/1988/1/012095.

[23] Filzmoser, P. 2005. Identification of Multivariate Outliers: A Performance Study. *Austrian Journal of Statistics*. 34(2): 127-138. DOI: 10.17713/ajs.v34i2.406.

[24] Pan, J-X., Fung, W-K., and Fang K-T. 2000. Multiple Outlier Detection in Multivariate Data using Projection Pursuit Techniques. *Journal of Statistical Planning and Inference*. 83(1): 153-167. DOI: 10.1016/s0378-3758(99)00091-9.

[25] Chernoff, H. 1973. The Use of Faces to Represent Points in k-Dimensional Space Graphically. *Journal of the American Statistical Association*. 68(342): 361-368. DOI: 10.1080/01621459.1973.10482434.

[26] Zuziak, J., Moskal, G., and Jakubowska, M. 2017. Effective Multivariate Data Presentation and Modeling in Distinction of the Tea Infusions. *Journal of Electroanalytical Chemistry*. 806: 97-106. DOI: 10.1016/j.jelechem.2017.10.059.