# IMPROVED ROBUST ESTIMATOR AND CLUSTERING PROCEDURES FOR MULTIVARIATE OUTLIERS DETECTION

SHARIFAH SAKINAH BT SYED ABD MUTALIB

DOCTOR OF PHILOSOPHY

UNIVERSITI MALAYSIA PAHANG

## SUPERVISOR'S DECLARATION

We hereby declare that We have checked this thesis and in our opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Doctor of Philosophy.

(Supervisor's Signature)

Full Name : DR. SITI ZANARIAH BINTI SATARI

Position : SENIOR LECTURER

Date : 17 July 2023

(Co-supervisor's Signature)

Full Name : DR. WAN NUR SYAHIDAH BINTI WAN YUSOFF

Position : SENIOR LECTURER

Date : 17 JULY 2023

## STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Universiti Malaysia Pahang or any other institutions.

_____

(Student's Signature)

Full Name : SHARIFAH SAKINAH BT SYED ABD MUTALIB

ID Number : PSS18002

Date : 17 JULY 2023

# IMPROVED ROBUST ESTIMATOR AND CLUSTERING PROCEDURES FOR MULTIVARIATE OUTLIERS DETECTION

SHARIFAH SAKINAH BT SYED ABD MUTALIB

Thesis submitted in fulfillment of the requirements
for the award of the degree of
Doctor of Philosophy

Centre for Mathematical Sciences

UNIVERSITI MALAYSIA PAHANG

JULY 2023

## ACKNOWLEDGEMENTS

BISMILLAHIRRAHMANIRRAHIM

In the name of Allah, the Most Gracious, the Most Merciful. Peace and blessings be upon the Messenger of Allah. Alhamdulillah, all praises and thanks to Allah, the Lord of the world, Most Gracious, Most Merciful, Master of the Day of Judgement, for blessing me with the will, determination, perseverance and strength to complete this thesis.

First and foremost, I would like to express my deepest gratitude to my supervisors, Dr. Siti Zanariah Satari and Dr. Wan Nur Syahidah Wan Yusoff for their valuable guidance, assistance, concern, motivation and continuous support throughout my study. I greatly appreciate my supervisor's patience in reviewing and providing constructive criticism for this thesis. A million thanks for their trust and confidence in my capability to complete this thesis.

I would also like to give special thanks to my parents, Syed Abd Mutalib bin Syed Abdullah and Tuan Saripah Arbiah bt Syed Muhammad for their continuous support and understanding when undertaking my research and writing my thesis. Your prayer for me was what sustained me this far. Last but not least, I thank my family, friends, colleagues and all those who have assisted me directly or indirectly in completing this thesis.

I hope this thesis will always remind me of my parents, my siblings, my close friend Azizah Mohd Rohni, and all the significant people who came into my life and my mind during the challenging times of this journey which leads me to the remembrance of Allah.

May Allah shower His blessings upon all of you.

## ABSTRACT

Outlier detection for multivariate data has been one of the areas that garnered attention to study due to the difficulty that arises as the number of variables, *p* increases. Visual inspection is insufficient to detect outliers in multivariate data, unlike univariate data. One of the methods to detect outliers in multivariate data is by using distance-based methods, which is Mahalanobis distance (MD). However, the sample mean and covariance matrix in MD is bound to masking and swamping problems. Therefore, many studies use robust estimators to replace the sample mean and covariance matrix. The development of robust estimators still continues until now. Although the robust estimator can overcome the problem of MD, it is still limited to detecting single point outliers only. Therefore, cluster-based methods have been proposed and developed in previous studies to overcome this problem. Hence, the main objective of this study is to propose a robust estimator in order to develop an improved procedure for detecting outliers in multivariate data using robust clustering-based methods. Firstly, an improved robust estimator based on the equality of covariance matrices that is less sensitive to the presence of outliers is proposed and named as Test on Covariance (TOC). TOC is developed by modified Concentration-Step (C-Step) in the Fast Minimum Covariance Determinant (FMCD) algorithm. In this step, the equality of covariance matrices test is done, and TOC is obtained. Secondly, an improved single linkage robust clustering procedure is developed. The similarity measure used in this procedure is the robust distance using TOC, named RDT. The improved single linkage robust clustering is robustified using RDT. Then, the performance of the proposed robust estimator and clustering procedure in detecting outliers for multivariate data are investigated using simulation studies and historical datasets. A data generation procedure is formulated in the simulation study to create synthetic data with three Outlier Scenarios using the R language. Three Outlier Scenarios used in this study are the Mean-shift (Outlier Scenario 1), Variance-inflation (Outlier Scenario 2), and Mean-shift and variance-inflation (Outlier Scenario 3). Three measurements are used to assess the effectiveness of the proposed robust estimator and clustering procedure, which are the probability that all the outliers are successfully detected (*pout*), the probability that the outliers are falsely detected as inliers (*pmask*), and the probability of inliers detected as outliers (*pswamp*). In particular, five historical datasets are used, which are Stackloss, Brain and Weight, Bushfire, Hawkins-Bradu Kass, and Milk. In this study, the performance of TOC in detecting outliers is compared with other existing robust estimators, which are Fast Minimum Covariance Determinant (FMCD), Minimum Vector Variance (MVV), Covariance Matrix Equality (CME) and Index Set Equality (ISE). Based on the simulation study, TOC shows good results in *pswamp* for all Outlier Scenarios, which indicates TOC has the lowest probability of misclassifying inliers as outliers compared to other robust estimators. TOC also shows similar performance as other robust estimators in most conditions. If the three measurements are considered simultaneously, TOC is the better estimator for the sample size, $n = 30, 50, 100, 200$, number of variables, $p = 3, 5, 10$ and all percentages of outliers, $1\% \le \varepsilon \le 25\%$. TOC also has proven able to detect outliers, does not have a masking effect, and performs similarly to other robust estimators in the historical datasets. Meanwhile, the performance of the improved single linkage robust clustering procedure is compared with single linkage by using Euclidean (ED), Mahalanobis distance (MD),

and TOC. Based on the simulation study, RDT only becomes the better similarity measure in a few conditions for *pout, pmask* and *pswamp* and performs similarly to other similarity measures in most conditions for all Outlier Scenarios. If the performance measurement of *pout*, *pmask* as well as *pswamp* are considered simultaneously for all Outlier Scenarios, RDT is the better similarity measure when $n = 50,100$, $p = 3,5$ and $\varepsilon = 5\%,10\%,15\%$. Moreover, RDT is the better similarity measure when the historical dataset contains 19% outliers, $p = 3$ and $n < 100$. From the findings of the simulation study and historical datasets, both TOC and RDT did not perform well for large sample size. It is also found that TOC outperforms RDT's ability to detect outliers in multivariate data. Therefore, this study concluded that TOC is a promising robust estimator and can be an alternative to other robust estimators for detecting outliers in multivariate data. RDT can also be used as an alternative similarity measure in clustering procedures and can also be used in other clustering methods. TOC can be further applied in other multivariate methods such as Principal Component Analysis, Factor Analysis and Discriminant Analysis. Furthermore, the improved single linkage robust clustering procedure in this study can be incorporated with Minimum Spanning Tree (MST).

# ABSTRAK

Pengesanan data terpencil untuk data multivariat telah mendapat perhatian kerana kesukarannya apabila bilangan pembolehubah, $p$ bertambah. Pemeriksaan visual sahaja tidak mencukupi untuk mengesan data terpencil dalam data multivariat, tidak seperti data univariat. Salah satu kaedah untuk mengesan data terpencil dalam data multivariat ialah dengan menggunakan kaedah berasaskan jarak iaitu Jarak Mahalanobis (MD). Walau bagaimanapun, sampel min dan matriks kovarians dalam MD mempunyai masalah *masking* dan *swamping*. Oleh itu, banyak kajian menggantikan min sampel dan matriks kovarians dengan penganggar teguh. Pembangunan dan kajian berkenaan penganggar teguh masih berterusan sehingga kini. Walaupun penganggar teguh boleh mengatasi masalah MD, ia hanya terhad untuk mengesan satu data terpencil pada satu masa. Oleh itu, kaedah berasaskan kluster dicadangkan dan dibangunkan dalam kajian lepas untuk mengatasi masalah ini. Justeru, objektif utama kajian ini ialah untuk mencadangkan penganggar teguh yang dipertingkatkan dan membangunkan prosedur yang dipertingkatkan untuk mengesan data terpencil bagi data multivariat menggunakan kaedah berasaskan pengelompokan teguh. Pertama, penganggar teguh yang dipertingkatkan berdasarkan kesamaan matriks kovarians yang kurang sensitif terhadap kehadiran data terpencil dicadangkan dan dinamakan Test on Covariance (TOC). TOC dibangunkan dengan mengubahsuai *C*-step dalam algorithm FMCD. Dalam *C*-step, ujian kesamaan antara matriks kovarians dilakukan dan TOC diperoleh. Kedua, prosedur pengelompokan rangkaian tunggal teguh yang dipertingkatkan dibangunkan. Ukuran kesamaan yang digunakan dalam prosedur ini ialah ukuran jarak teguh menggunakan TOC, dinamakan RDT. Pengelompokan rangkaian tunggal teguh yang dibangunkan telah diteguhkan dengan menggunakan RDT. Kemudian, prestasi penganggar teguh yang dipertingkatkan dan prosedur pengelompokan yang dicadangkan dalam mengesan data terpencil bagi data multivariat disiasat menggunakan simulasi dan set data lepas. Prosedur penjanaan data diformulasikan dalam simulasi bagi mencipta data dengan tiga senario data terpencil menggunakan R. Tiga scenario data terpencil yang digunakan ialah Anjakan Min (Senario Data Terpencil 1), Inflasi Varians (Senario Data Terpencil 2), dan Anjakan Min dan Inflasi Varians (Senario Data Terpencil 3). Tiga ukuran digunakan untuk menilai prestasi penganggar teguh yang dipertingkatkan dan prosedur pengelompokan yang dicadangkan iaitu kebarangkalian semua data terpencil berjaya dikesan (*pout*), kebarangkalian data terpencil dikesan sebagai bukan data terpencil (*pmask*), dan kebarangkalian bukan data terpencil dikesan sebagai data terpencil (*pswamp*). Secara khususnya, lima set data lepas digunakan iaitu Stackloss, Brain dan Weight, Bushfire, Hawkins-Bradu Kass, dan Milk. Prestasi TOC dalam mengesan data terpencil bagi data multivariat dibandingkan dengan penganggar teguh sedia ada iaitu Fast Minimum Covariance Determinant (FMCD), Minimum Vector Variance (MVV), Covariance Matrix Equality (CME), dan Index Set Equality (ISE). Berdasarkan simulasi, TOC menunjukkan hasil yang baik dalam *pswamp* untuk semua scenario data terpencil, yang bermaksud TOC mempunyai kebarangkalian paling rendah untuk salah mengklasifikasikan bukan data terpencil sebagai data terpencil berbanding dengan penganggar teguh yang lain. TOC juga menunjukkan prestasi yang sama seperti penganggar teguh lain dalam kebanyakan keadaan. Apabila ketiga-tiga ukuran dipertimbangkan, TOC ialah penganggar yang lebih baik jika saiz sampel,

$n = 30, 50, 100, 200$ , bilangan pembolehubah, $p = 3, 5, 10$ dan bagi semua peratusan data terpencil, $1\% \le \varepsilon \le 25\%$. TOC telah terbukti mampu mengesan data terpencil, tidak mempunyai kesan penyamaran dan berprestasi serupa dengan penganggar teguh lain dalam set data lepas. Sementara itu, prestasi prosedur pengelompokan yang dipertingkatkan dibandingkan dengan pengelompokan rangkaian tunggal menggunakan ukuran jarak Euclidean (ED), Mahalanobis (MD) dan TOC. Berdasarkan simulasi, RDT menjadi ukuran kesamaan yang lebih baik dalam beberapa keadaan untuk *pout*, *pmask* dan *pswamp* dan berprestasi sama dengan ukuran kesamaan lain dalam kebanyakan keadaan bagi semua senario data terpencil. Apabila pengukuran prestasi *pout*, *pmask* dan juga *pswamp* dipertimbangkan secara serentak untuk semua senario data terpencil, RDT ialah ukuran kesamaan yang lebih baik apabila $n = 50, 100$, $p = 3, 5$ dan $\varepsilon = 5\%, 10\%, 15\%$. Tambahan pula, RDT ialah ukuran kesamaan yang lebih baik apabila set data lepas mengandungi 19% data terpencil, $p = 3$ dan $n < 100$. Daripada dapatan simulasi dan data lepas, kedua-dua TOC dan RDT tidak menunjukkan prestasi yang bagus apabila saiz sampel besar. Ianya juga didapati TOC mengatasi prestasi RDT dalam mengesan data terpencil bagi data multivariat. Oleh itu, kajian ini merumuskan TOC ialah penganggar teguh yang dipertingkatkan yang menjanjikan dan boleh digunakan sebagai alternatif kepada penganggar teguh lain bagi mengesan data terpencil dalam data multivariat. RDT juga boleh digunakan sebagai ukuran kesamaan alternatif dalam prosedur pengelompokan dan juga boleh digunakan bersama kaedah pengelompokan yang lain. TOC juga boleh digunakan dalam kaedah kaedah multivariat yang lain seperti Analisis Komponen Utama, Analisis Faktor dan Analisis Diskriminasi. Prosedur pengelompokan rangkaian tunggal teguh yang dipertingkatkan dalam kajian ini juga boleh digabungkan dengan Pokok Rentangan Minimum (MST).

## TABLE OF CONTENT

## REFERENCES


Affindi, A. N., Ahmad, S., & Mohamad, M. (2019). A comparative study between ridge MM and ridge least trimmed squares estimators in handling multicollinearity and outliers. *Journal of Physics: Conference Series*, *1366* (1), 012113. https://doi.org/10.1088/1742-6596/1366/1/012113

Afzal, S., Afzal, A., Amin, M., Saleem, S., Ali, N., & Sajid, M. (2021). A novel approach for outlier detection in multivariate data. *Mathematical Problems in Engineering*, *2021*. https://doi.org/10.1155/2021/1899225

Aggarwal, C. C. (2017). Outlier Analysis. In *Springer* (Second Edi).

Al-Zoubi, M. D. B., Ali, A. D., & Yahya, A. A. (2010). Fuzzy clustering-based approach for outlier detection. *Proceedings of the 9th WSEAS International Conference on Applications of Computer Engineering*, 192–197.

Almeida, J. A. S., Barbosa, L. M. S., Pais, A. A. C. C., & Formosinho, S. J. (2007). Improving hierarchical cluster analysis: A new method with outlier detection and automatic clustering. *Chemometrics and Intelligent Laboratory Systems*, *87*(2), 208–217. https://doi.org/10.1016/j.chemolab.2007.01.005

Amiri, S., Clarke, B. S., Clarke, J. L., & Koepke, H. (2019). A general hybrid clustering technique. *Journal of Computational and Graphical Statistics*, *28*(3), 540–551. https://doi.org/10.1080/10618600.2018.1546593

Atkinson, A. C. (1994). Fast very robust methods for the detection of multiple outliers. *Journal of the American Statistical Association*, *89*(428), 1329–1339. https://doi.org/10.1080/01621459.1994.10476872

Atkinson, A. C., & Mulira, H. M. (1993). The stalactite plot for the detection of multivariate outliers. *Statistics and Computing*, *3*(1), 27–35. https://doi.org/10.1007/BF00146951

Badaró, J. P. M., Campos, V. P., da Rocha, F. O. C., & Santos, C. L. (2021). Multivariate analysis of the distribution and formation of trihalomethanes in treated water for human consumption. *Food Chemistry*, *365*, 130469. https://doi.org/10.1016/j.foodchem.2021.130469

Balcan, M., Liang, Y., & Gupta, P. (2014). Robust hierarchical clustering. *Journal of Machine Learning*, *15*, 4011–4051. https://doi.org/10.1109/IMSCCS.2006.167

Banerjee, A., & Davé, R. N. (2012). Robust clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *2*(1), 29–59. https://doi.org/10.1002/widm.49

Barbosa, J. J., Duarte, A. R., & Martins, H. S. R. (2020). A performance evaluation in multivariate outliers identification methods. *Ciência e Natura*, *42*, e16. https://doi.org/10.5902/2179460x41662

Barnett, V., & Lewis, T. (1984). *Outliers in statistical data*. Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics

Becker, C., & Gather, U. (1999). The masking breakdown point of multivariate outlier identification rules. *Journal of the American Statistical Association*, *94*(447), 947–955. https://doi.org/10.1080/01621459.1999.10474199

Bondu, R., Cloutier, V., Rosa, E., & Roy, M. (2020). An exploratory data analysis approach for assessing the sources and distribution of naturally occurring contaminants (F, Ba, Mn, As) in groundwater from southern Quebec (Canada). *Applied Geochemistry*, *114*, 104500. https://doi.org/10.1016/j.apgeochem.2019.104500

Bongiorno, C., Miccichè, S., & Mantegna, R. N. (2022). Statistically validated hierarchical clustering: Nested partitions in hierarchical trees. *Physica A: Statistical Mechanics and Its Applications*, *593*, 126933. https://doi.org/10.1016/j.physa.2022.126933

Boudt, K., Rousseeuw, P. J., Vanduffel, S., & Verdonck, T. (2020). The minimum regularized covariance determinant estimator. *Statistics and Computing*, *30*(1), 113–128. https://doi.org/10.1007/s11222-019-09869-x

Bouguettaya, A., Yu, Q., Liu, X., Zhou, X., & Song, A. (2015). Efficient agglomerative hierarchical clustering. *Expert Systems with Applications*, *42*(5), 2785–2797. https://doi.org/10.1016/j.eswa.2014.09.054

Bulut, H. (2020). Mahalanobis distance based on minimum regularized covariance determinant estimators for high dimensional data. *Communications in Statistics - Theory and Methods*, *49*(24), 5897–5907. https://doi.org/10.1080/03610926.2020.1719420

Cabana, E., Lillo, R. E., & Laniado, H. (2021). Multivariate outlier detection based on a robust Mahalanobis distance with shrinkage estimators. *Statistical Papers*, *62*(4), 1583–1609. https://doi.org/10.1007/s00362-019-01148-1

Cafaro, M., Melle, C., Pulimeno, M., & Epicoco, I. (2021). Fast online computation of the Qn estimator with applications to the detection of outliers in data streams. *Expert Systems with Applications*, *164*, 113831. https://doi.org/10.1016/j.eswa.2020.113831

Camacho, J. (2017). On the generation of random multivariate data. *Chemometrics and Intelligent Laboratory Systems*, *160*, 40–51. https://doi.org/10.1016/j.chemolab.2016.11.013

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

Cerioli, A., Riani, M., & Torti, F. (2011). Accurate and powerful multivariate outlier detection. *Int. Statistical Inst.: Proc. 58th World Statistical Congress*, 5608–5613.

Cha, S. H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *City*, *1*(2), 1. https://doi.org/10.1007/s00167-009-0884-z

Chatzinakos, C., Pitsoulis, L., & Zioutas, G. (2016). Optimization techniques for robust multivariate location and scatter estimation. *Journal of Combinatorial Optimization*, *31*(4), 1443–1460. https://doi.org/10.1007/s10878-015-9833-6

D'Urso, P., De Giovanni, L., & Massari, R. (2020). Smoothed self-organizing map for robust clustering. *Information Sciences*, *512*, 381–401. https://doi.org/10.1016/j.ins.2019.06.038

Da Costa, J. F. P., & Cabral, M. (2022). Statistical methods with applications in data mining: A review of the most recent works. *Mathematics*, *10*(6), 1–22. https://doi.org/10.3390/math10060993

Daudin, J. J., Duby, C. D., & Trecourt, P. (1988). Stability of principal component analysis studied by the bootstrap method. *Statistics: A Journal of Theoretical Applied Statistics*, *19*(2), 241–258. https://doi.org/10.1080/02331888808802095

De Ketelaere, B., Hubert, M., Raymaekers, J., Rousseeuw, P. J., Vranckx, I., (2020). Real-time outlier detection for large datasets by RT-DetMCD. *Chemometrics and Intelligent Laboratory Systems*, *199*, 103957. https://doi.org/10.1016/j.chemolab.2020.103957

De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. L. (2000). The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, *50*, 1–18. https://doi.org/10.1016/S0169-7439(99)00047-7

Deb, A. B., & Dey, L. (2017). Outlier detection and removal algorithm in k-means and hierarchical clustering. *World Journal of Computer Application and Technology*, *5*(2), 24–29. https://doi.org/10.13189/wjcat.2017.050202

Djauhari, M. A. (2008). Highly robust estimation of location and scatter when data sets are of high dimension: An open problem. *The 3rd International Conference on Mathematics and Statistics (ICoMS-3)*, 1–8.

Djauhari, M. A. (2007). A measure of multivariate data concentration. *Journal of Applied Probability and Statistics*, *2*(2), 139–155.

Djauhari, M. A. (2011). Geometric Interpretation of Vector Variance. *Matematika*, *27*(1), 51–57.

Domino, K. (2020). Multivariate cumulants in outlier detection for financial data analysis. *Physica A: Statistical Mechanics and Its Applications*, *558*, 124995. https://doi.org/10.1016/j.physa.2020.124995

Dotto, F., Farcomeni, A., García-Escudero, L. A., & Mayo-Iscar, A. (2018). A reweighting approach to robust clustering. *Statistics and Computing*, *28*(2), 477–493. https://doi.org/10.1007/s11222-017-9742-x

Duan, L., Xu, L., Liu, Y., & Lee, J. (2009). Cluster-based outlier detection. *Annals of Operations Research*, *168*(1), 151–168. https://doi.org/10.1007/s10479-008-0371-9

Evans, K., Love, T., & Thurston, S. W. (2015). Outlier identification in model-based cluster analysis. *Journal of Classification*, *32*(1), 63–84. https://doi.org/10.1007/s00357-015-9171-5

Everitt, B. S., & Dunn, G. (2001). *Applied multivariate data analysis* (Second Edi). London: Arnold.

Fauconnier, C., & Haesbroeck, G. (2009). Outliers detection with the minimum covariance determinant estimator in practice. *Statistical Methodology*, *6*(4), 363–379. https://doi.org/10.1016/j.stamet.2008.12.005

Fawzy, A., Mokhtar, H. M. O., & Hegazy, O. (2013). Outliers detection and classification in wireless sensor networks. *Egyptian Informatics Journal*, *14*(2), 157–164. https://doi.org/10.1016/j.eij.2013.06.001

Ferreira-Dias, S., Gominho, J., Baptista, I., & Pereira, H. (2018). Pattern recognition of cardoon oil from different large-scale field trials. *Industrial Crops and Products*, *118*, 236–245. https://doi.org/10.1016/j.indcrop.2018.03.038

Filzmoser, P., & Gregorich, M. (2020). Multivariate outlier detection in applied data analysis: global, local, compositional and cellwise outliers. *Mathematical Geosciences*, *52*(8), 1049–1066. https://doi.org/10.1007/s11004-020-09861-6

Filzmoser, P., Maronna, R., & Werner, M. (2008). Outlier identification in high dimensions. *Computational Statistics and Data Analysis*, *52*(3), 1694–1711. https://doi.org/10.1016/j.csda.2007.05.018

Filzmoser, P., & Todorov, V. (2013). Robust tools for the imperfect world. *Information Sciences*, *245*, 4–20. https://doi.org/10.1016/j.ins.2012.10.017

Filzmoser, P. (2005). Identification of multivariate outliers: a performance study. *Austrian Journal of Statistics*, *34*(2), 127–138. https://doi.org/10.17713/ajs.v34i2.406

Fischer, D., Berro, A., Nordhausen, K., & Ruiz-Gazen, A. (2021). REPPlab: An R package for detecting clusters and outliers using exploratory projection pursuit. *Communications in Statistics: Simulation and Computation*, *50*(11), 3397–3419. https://doi.org/10.1080/03610918.2019.1626880

Fung, W. K. (1993). Unmasking outliers and leverage points: A confirmation. *Journal of the American Statistical Association*, *88*(422), 515–519. https://doi.org/10.1080/01621459.1993.10476302

Gan, G., Ma, C., & Wu, J. (2007). *Data clustering: theory, algorithms, and applications* (SIAM, Soci).

García-Escudero, L. A., Gordaliza, A., Matrán, C., & Mayo-Iscar, A. (2008). A general trimming approach to robust cluster analysis. *The Annals of Statistics*, *36*(3), 1324–1345. https://doi.org/10.1214/07-AOS515

Garcia-Escudero, L. A., Gordaliza, A., Matran, C. & Mayo-Iscar, A. (2010). A review of robust clustering methods. *Advances in Data Analysis and Classification*, *4*(2), 89–109. https://doi.org/10.1007/s11634-010-0064-5

Giorgi, F. M., Ceraolo, C., & Mercatelli, D. (2022). The R Language: An Engine for Bioinformatics and Data Science. *Life*, *12*(5), 648. https://doi.org/10.3390/life12050648

Gupta, P. (2011). *Robust Clustering Algorithms*. Unpublished PhD Thesis. Georgia Institute of Technology.

Grübel, R. (1988). A minimal characterization of the covariance matrix. *Metrika*, *35*(1), 49–52. https://doi.org/10.1007/BF02613285

Hadi, A. S. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *54*(3), 761–771.

Hadi, A. S., Imon, A. R., & Werner, M. (2009). Detection of outliers. *Wiley Interdisciplinary Reviews: Computational Statistics*, *1*(1), 57–70. https://doi.org/10.1002/wics.6

Hardin, J., & Rocke, D. M. (2004). Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. *Computational Statistics & Data Analysis*, *44*(4), 625–638. https://doi.org/10.1016/S0167-9473(02)00280-3

Hawkins, D. M., Bradu, D., & Kass, G. V. (1984). Location of several outliers in multiple-regression data using elemental sets. *Technometrics*, *26*(3), 197–208. https://doi.org/10.1080/00401706.1984.10487956

Herdiani, E. T., Sari, P. P., & Sunusi, N. (2019). Detection of outliers in multivariate data using Minimum Vector Variance Method. *Journal of Physics: Conference Series*, *1341*(9), 092004. https://doi.org/10.1088/1742-6596/1341/9/092004

Herwindiati, D. E., Djauhari, M. A., & Mashuri, M. (2007). Robust multivariate outlier labeling. *Communications in Statistics-Simulation and Computation*, *36*(6), 1287–1294. https://doi.org/10.1080/03610910701569044

Herwindiati, D. E., Hendryli, J., & Mulyono, S. (2018). Robust kurtosis projection approach for mangrove classification. *International Conference on Computing and Information Technology*. *Springer*, 93–103. https://doi.org/10.1007/978-3-319-93692-5_10

Hubert, M. (2020). Robust multivariate statistical methods. In *Comprehensive Chemometrics*, Second Edi, 107–122. https://doi.org/10.1016/b978-0-12-409547-2.14879-6

Hubert, M., & Debruyne, M. (2010). Minimum covariance determinant. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(1), 36–43. https://doi.org/10.1002/wics.61

Hubert, M., Debruyne, M., & Rousseeuw, P. J. (2018). Minimum covariance determinant and extensions. *Wiley Interdisciplinary Reviews: Computational Statistics*, *10*(3), e1421. https://doi.org/10.1002/wics.1421

Hubert, M., Rousseeuw, P., & Vakili, K. (2014). Shape bias of robust covariance estimators: an empirical study. *Statistical Papers*, *55*(1), 15–28. https://doi.org/10.1007/s00362-013-0544-8

Hubert, M., Rousseeuw, P., Vanpaemel, D., & Verdonck, T. (2015). The DetS and DetMM estimators for multivariate location and scatter. *Computational Statistics and Data Analysis*, *81*, 64–75. https://doi.org/10.1016/j.csda.2014.07.013

Iqbal, M. Z., Riaz, M., & Nasir, W. (2017). Multivariate outlier detection: a comparison among two clustering techniques. *Pakistan Journal of Agricultural Sciences*, *54*(1), 227–231. https://doi.org/10.21162/PAKJAS/17.4743

Jerison, H. J. (1973). *Evolution of the brain and intelligence* (A. P. Inc. (ed.))

Jiang, M. F., Tseng, S. S., & Su, C. M. (2001). Two-phase clustering process for outliers detection. *Pattern Recognition Letters*, *22*(6–7), 691–700. https://doi.org/10.1016/S0167-8655(00)00131-8

Johnson, R. A., & Wichern, D. W. (2002). *Applied multivariate statistical analysis* (Fifth Edit). New Jersey: Hall.

Kosinski, A. S. (1999). A procedure for the detection of multivariate outliers. *Computational Statistics and Data Analysis*, *29*(2), 145–161. https://doi.org/10.1016/S0167-9473(98)00073-5

Kumar, M., & Orlin, J. B. (2008). Scale-invariant clustering with minimum volume ellipsoids. *Computers & Operations Research*, *35*(4), 1017–1029. https://doi.org/10.1016/j.cor.2006.07.001

Kunjunni, S. O., & Abraham, S. T. (2020a). Multidimensional outlier detection and robust estimation using Sn covariance. *Communications in Statistics- Simulation and Computation*, 1-11. https://doi.org/10.1080/03610918.2020.1725820

Kunjunni, S. O., & Abraham, S. T. (2020b). Sn covariance. *Communications in Statistics - Theory and Methods*, *49*(24), 6133–6138. https://doi.org/10.1080/03610926.2019.1628275

Kuwil, F. H., Shaar, F., Topcu, A. E., & Murtagh, F. (2019). A new data clustering algorithm based on critical distance methodology. *Expert Systems with Applications*, *129*, 296–310. https://doi.org/10.1016/j.eswa.2019.03.051

Leys, C., Klein, O., Dominicy, Y., & Ley, C. (2018). Detecting multivariate outliers: use a robust variant of the Mahalanobis distance. *Journal of Experimental Social Psychology*, *74*, 150–156. https://doi.org/10.1016/j.jesp.2017.09.011

Lim, H. A., & Midi, H. (2016). Diagnostic Robust Generalized Potential based on Index Set Equality (DRGP (ISE)) for the identification of high leverage points in linear model. *Computational Statistics*, *31*(3), 859–877. https://doi.org/10.1007/s00180-016-0662-6

Liu, X., Gao, F., Wu, Y., & Zhao, Z. (2018). Detecting outliers and influential points: an indirect classical Mahalanobis distance-based method. *Journal of Statistical Computation and Simulation*, *88*(11), 2013–2033. https://doi.org/10.1080/00949655.2018.1448981

Loperfido, N. (2018). Skewness-based projection pursuit: A computational approach. *Computational Statistics and Data Analysis*, *120*, 42–57. https://doi.org/10.1016/j.csda.2017.11.001

Maronna, R. A., & Yohai, V. J. (2017). Robust and efficient estimation of multivariate scatter and location. *Computational Statistics and Data Analysis*, *109*, 64–75. https://doi.org/10.1016/j.csda.2016.11.006

Maronna, R. A., & Yohai, V. J. (1995). The behavior of the Stahel-Donoho robust multivariate estimator. *Journal of the American Statistical Association*, *90*(429), 330–341. https://doi.org/10.1080/01621459.1995.10476517

Melendez-Melendez, G., Cruz-Paz, D., Carrasco-Ochoa, J. A., & Martínez-Trinidad, J. F. (2019). An improved algorithm for partial clustering. *Expert Systems with Applications*, *121*, 282–291. https://doi.org/10.1016/j.eswa.2018.12.027

Meropi, P., Bikos, C., & George, Z. (2018). Outlier detection in skewed data. *Simulation Modelling Practice and Theory*, *87*, 191–209. https://doi.org/10.1016/j.simpat.2018.05.010

Midi, H., Hendi, H. T., Arasan, J., & Uraibi, H. (2020). Fast and robust diagnostic technique for the detection of high leverage points. *Pertanika Journal of Science and Technology*, *28*(4), 1203–1220. https://doi.org/10.47836/pjst.28.4.05

Mikulec, A., & Kupis-Fijalkowska, A. (2012). An empirical analysis of the effectiveness of Wishart and Mojena criteria in cluster analysis. *Statistics in Transition*, *13*(3), 569–580.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, *50*(2), 159–179. https://doi.org/10.1007/BF02294245

Mojena, R. (1977). Hierarchical grouping methods and stopping rules: An evaluation. *The Computer Journal*, *20*(4), 359–363.

Möller, S. F., Frese, J. Von, & Bro, R. (2005). Robust Methods for Multivariate Data Analysis. *Journal of Chemometrics*, *19*(10), 549–563. https://doi.org/10.1002/cem.962

Olukanmi, P. O. O., & Twala, B. (2017). K-means-sharp: modified centroid update for outlier-robust k-means clustering. *2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech)*, 14–19. https://doi.org/10.1109/RoboMech.2017.8261116

Ott, L., Pang, L., Ramos, F., & Chawla, S. (2014). On integrated clustering and outlier detection. *Advances in Neural Information Processing Systems*, *27*.

Pan, J.-X., Fung, W.-K., & Fang, K.-T. (2000). Multiple outlier detection in multivariate data using projection pursuit techniques. *Journal of Statistical Planning and Inference*, *83*(1), 153–167. https://doi.org/10.1016/s0378-3758(99)00091-9

Pasillas-Díaz, J. R., & Ratté, S. (2016). An unsupervised approach for combining scores of outlier detection techniques, based on similarity measures. *Electronic Notes in Theoretical Computer Science*, *329*, 61–77. https://doi.org/10.1016/j.entcs.2016.12.005

Peña, M. (2018). Robust clustering methodology for multi-frequency acoustic data: A review of standardization, initialization and cluster geometry. *Fisheries Research*, *200*, 49–60. https://doi.org/10.1016/j.fishres.2017.12.013

Pokojovy, M., & Jobe, J. M. (2022). A robust deterministic affine-equivariant algorithm for multivariate location and scatter. *Computational Statistics & Data Analysis*, *172*, 107475. https://doi.org/10.1016/j.csda.2022.107475

Popat, S. K., & Emmanuel, M. (2014). Review and comparative study of clustering techniques. *International Journal of Computer Science and Information Technologies*, *5*(1), 805–812.

Puig, X., & Ginebra, J. (2018). Outlier detection for multivariate categorical data. *Quality and Reliability Engineering International*, *34*(7), 1400–1412. https://doi.org/10.1002/qre.2339

Qu, W., Liu, H., & 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. https://doi.org/10.3758/s13428-019-01291-5

Rampado, O., Gianusso, L., Nava, C. R., & Ropolo, R. (2019). Analysis of a CT patient dose database with an unsupervised clustering approach. *Physica Medica*, *60*, 91–99. https://doi.org/10.1016/j.ejmp.2019.03.015

Rencher, A. C. (2002). Methods of Multivariate Analysis. In *A John Wiley & Sons, Inc. Publication*.

Riley, R. D., Snell, K. I. E., Martin, G. P., Whittle, R., Archer, L., Sperrin, M., & 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. https://doi.org/10.1016/j.jclinepi.2020.12.005

Ro, K., Zou, C., Wang, Z., & Yin, G. (2015). Outlier detection for high-dimensional data. *Biometrika*, *102*(3), 589–599. https://doi.org/10.1093/biomet/asv021

Rocke, D. M., & Woodruff, D. L. (2002). Computational connections between robust multivariate analysis and clustering. In *Compstat* (pp. 255–260). Physica, Heidelberg. https://doi.org/10.1007/978-3-642-57489-4_35

Rocke, D. M., & Woodruff, D. L. (1996). Identification of outliers in multivariate data. *Journal of the American Statistical Association*, *91*(435), 1047–1061. https://doi.org/10.1080/01621459.1996.10476975

Roelant, E., Van Aelst, S., & Willems, G. (2009). The minimum weighted covariance determinant estimator. *Metrika*, *70*(2), 177–204. https://doi.org/10.1007/s00184-008-0186-3

Roizman, V., Jonckheere, M., & Pascal, F. (2021). Robust clustering and outlier rejection using the Mahalanobis distance distribution. *2020 28th European Signal Processing Conference*, 2448–2452. https://doi.org/10.23919/Eusipco47968.2020.9287356

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, *79*(388), 871–880.

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications*, *8*(283-297), 37.

Rousseeuw, P. J., & Van Driessen, K. (1999). A fast algorithm for the Minimum Covariance Determinant estimator. *Technometrics*, *41*(3), 212–223.

Rousseeuw, P. J., & Hubert, M. (2011). Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(1), 73–79. https://doi.org/10.1002/widm.2

Rousseeuw, P., & Yohai, V. (1984). Robust regression by means of S-estimators. In *Robust and Nonlinear Time Series Analysis* (pp. 256–272). https://doi.org/10.1007/978-1-4615-7821-5_15

Rousseeuw, P. J., & Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, *85*(411), 633–639. https://doi.org/10.2307/2289999

Sad, C. M. S., da Silva, M., dos Santos, F. D., Pereira, L. B., Corona, R. R. B., Silva, S. R. C., Portela, N. A., Castro, E. V. R., Filgueiras, P. R., & Jr, V. L. (2019). Multivariate data analysis applied in the evaluation of crude oil blends. *Fuel*, *239*, 421–428. https://doi.org/10.1016/j.fuel.2018.11.045

Salleh, R. M. (2013). *A robust estimation method of location and scale with application in monitoring process variability*. Unpublished PhD Thesis. Universiti Teknologi Malaysia, Malaysia.

Salleh, R. M., & Djauhari, M. A. (2010). Robust start up stage for beltline moulding process variability monitoring using vector variance. *Malaysian Journal of Fundamental and Applied Sciences*, *6*(1), 67–71. https://doi.org/10.11113/mjfas.v6n1.179

Salleh, R. M., & Djauhari, M. A. (2011). Robust hotelling's $T^2$ control charting in spike production process. *International Seminar on the Application of Science & Mathematics 2011*, 1–8.

Santos-Pereira, C. M., & Pires, A. M. (2002). Detection of outlier in multivariate data: a method based on clustering and robust estimators. In *Compstat* (pp. 291–296). Physica, Heidelberg. https://doi.org/10.1007/978-3-642-57489-4_41

Santos-Pereira, C. M. M., & Pires, A. M. A. M. (2013). Robust clustering method for the detection of outliers: using AIC to select the number of clusters. In *Advances in Regression, Survival Analysis, Extreme Values, Markov Processes and Other Statistical Applications.* (pp. 409–415). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-34904-1_43

Satari, S. Z. (2015). *Parameter estimation and outlier detection for some types of circular model*. Unpublished PhD Thesis. University of Malaya, Malaysia.

Satari, S. Z., Di, N. F. M., & Zakaria, R. (2019). Single-linkage method to detect multiple outliers with different outlier scenarios in circular regression model. *AIP Conference Proceedings*, *2059*, 020003. https://doi.org/10.1063/1.5085946

Satari, S. Z., Di, N. F. M., & Zakaria, R. (2017). The multiple outliers detection using agglomerative hierarchical methods in circular regression model. *Journal of Physics: Conference Series*, *890*(1), 012152. https://doi.org/10.1088/1742-6596/890/1/012152

Savić, M., Atanasijević, J., Jakovetić, D., & Krejić, N. (2022). Tax evasion risk management using a hybrid unsupervised outlier detection method. *Expert Systems with Applications*, *193*, 116409. https://doi.org/10.1016/j.eswa.2021.116409

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P. P., Tiwari, A., Er, M. J. J., Ding, W., & Lin, C.-T. T. (2017). A review of clustering techniques and developments. *Neurocomputing*, *267*, 664–681. https://doi.org/10.1016/j.neucom.2017.06.053

Sebert, D. M., Montgomery, D. C., & Rollier, D. A. (1998). A clustering algorithm for identifying multiple outliers in linear regression. *Computational Statistics & Data Analysis*, *27*(4), 461–484. https://doi.org/10.1016/S0167-9473(98)00021-8

Serfling, R., & Mazumder, S. (2013). Computationally easy outlier detection via projection pursuit with finitely many directions. *Journal of Nonparametric Statistics*, *25*(2), 447–461. https://doi.org/10.1080/10485252.2013.766335

Sharma, K. K., & Seal, A. (2021). Outlier-robust multi-view clustering for uncertain data. *Knowledge-Based Systems*, *211*, 106567. https://doi.org/10.1016/j.knosys.2020.106567

Sitio, A., Sinaga, A. S., Haikal, A., & Dewi, S. (2022). Multivariate analysis of commodity availability of staple foods using complete linkage hierarchical clustering method. *JITK (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer)*, *7*(2), 61–66. https://doi.org/10.33480/jitk.v7i2.2830

Stimolo, M. I., & Ortiz, P. A. (2020). Projection pursuit algorithms to detect outliers. *Cuadernos de Administración*, *33*. https://doi.org/10.11144/javeriana.cao33.ppado

Stromberg, A. J. (1997). Robust covariance estimates based on resampling. *Journal of Statistical Planning and Inference*, *57*(2), 321–334. https://doi.org/10.1016/S0378-3758(96)00051-1

Su, X., & Tsai, C.-L. L. (2011). Outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(3), 261–268. https://doi.org/10.1002/widm.19

Tarr, G., Müller, S., & Weber, N. C. (2016). Robust estimation of precision matrices under cellwise contamination. *Computational Statistics and Data Analysis*, *93*, 404–420. https://doi.org/10.1016/j.csda.2015.02.005

Tavares, A. H., Raymaekers, J., Rousseeuw, P. J., Brito, P., & Afreixo, V. (2020). Clustering genomic words in human DNA using peaks and trends of distributions. *Advances in Data Analysis and Classification*, *14*(1), 57–76. https://doi.org/10.1007/s11634-019-00362-x

Team, R. C. (2000). R Language Definition. In *R foundation for statistical computing* (Vol. 3, Issue 1). https://doi.org/10.1145/1061414.1061416

Uzabaci, E., Ercan, I., & Alpu, O. (2020). Evaluation of outlier detection method performance in symmetric multivariate distributions. *Communications in Statistics- Simulation and Computation*, *49*(2), 516–531. https://doi.org/10.1080/03610918.2018.1487068

Van Aelst, S., & Rousseeuw, P. (2009). Minimum volume ellipsoid. *Wiley Interdisciplinary Reviews: Computational Statistics*, *1*(1), 71–82. https://doi.org/10.1002/wics.19

Vijayarani, S., & Nithya, S. (2011). An efficient clustering algorithm for outlier detection. *International Journal of Computer Applications*, *32*(7), 22–27.

Wada, K., Kawano, M., & Tsubaki, H. (2020). Comparison of multivariate outlier detection methods for nearly elliptical distributions. *Austrian Journal of Statistics*, *49*(2), 1–17. https://doi.org/10.17713/ajs.v49i2.87

Wang, H., Bah, M. J., & Hammad, M. (2019). Progress in outlier detection techniques: A survey. *IEEE Access*, *7*, 107964–108000. https://doi.org/10.1109/ACCESS.2019.2932769

Wang, K., & Lan, H. (2020). Robust support vector data description for novelty detection with contaminated data. *Engineering Applications of Artificial Intelligence*, *91*, 103554. https://doi.org/10.1016/j.engappai.2020.103554

Werner, M. (2003). *Identification of multivariate Outliers in large data sets*. Unpublished PhD Thesis. University of Colorado.

Wu, G., Chen, C., & Yan, X. (2011). Modified minimum covariance determinant estimator and its application to outlier detection of chemical process data. *Journal of Applied Statistics*, *38*(5), 1007–1020. https://doi.org/10.1080/02664761003692456

Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, *2*(2), 165–193. https://doi.org/10.1007/s40745-015-0040-1

Yesilbudak, M. (2016). Partitional clustering-based outlier detection for power curve optimization of wind turbines. *5th International Conference on Renewable Energy Research and Applications (ICRERA)*, 1080–1084. https://doi.org/10.1109/ICRERA.2016.7884500

Yim, O., & Ramdeen, K. T. (2015). Hierarchical cluster analysis: comparison of three linkage measures and application to psychological data. *The Quantitative Methods for Psychology*, *11*(1), 8–21. https://doi.org/10.20982/tqmp.11.1.p008

Yoon, K. A., Kwon, O. S., & Bae, D. H. (2007). An approach to outlier detection of software measurement data using the k-means clustering method. *First International Symposium on Empirical Software Engineering and Measurement (ESEM 2007)*, 443–445. https://doi.org/10.1109/ESEM.2007.49

Zhang, J. (2013). Advancements of outlier detection: A survey. *ICST Transactions on Scalable Information Systems*, *13* (1), 1–26. https://doi.org/10.4108/trans.sis.2013.01-03.e2

Zheng, S., Zhu, Y. X., Li, D. Q., Cao, Z. J., Deng, Q. X., & Phoon, K. K. (2021). Probabilistic outlier detection for sparse multivariate geotechnical site investigation data using Bayesian learning. *Geoscience Frontiers*, *12*(1), 425–439. http://dx.doi.org/10.1016/j.gsf.2020.03.017

Zhou, Z., Ye, Z., Yu, J., & Chen, W. (2018). Cluster-aware arrangement of the parallel coordinate plots. *Journal of Visual Languages and Computing*, *46*, 43–52. https://doi.org/10.1016/j.jvlc.2017.10.003

Zulkipli, N. S., Satari, S. Z., & Wan Yusoff, W. N. S. (2021). A synthetic data generation procedure for univariate circular data with various outliers scenarios using Python programming language. *Journal of Physics: Conference Series*, *1988*(1). https://doi.org/10.1088/1742-6596/1988/1/012111

Zulkipli, N. S., Satari, S. Z., & Wan Yusoff, W. S. (2022). The effect of different similarity distance Measures in Detecting Outliers Using Single-Linkage Clustering Algorithm for Univariate Circular Biological Data. *Pakistan Journal of Statistics and Operation Research*, *18*(3), 561–573.