FACTA UNIVERSITATIS (NIŠ)
Ser. Math. Inform. Vol. 34, No 3 (2019), 583–595
https://doi.org/10.22190/FUMI1903583G

# MAHALANOBIS DISTANCE AND ITS APPLICATION FOR DETECTING MULTIVARIATE OUTLIERS

**Hamid Ghorbani**

**Abstract.** While methods of detecting outliers is frequently implemented by statisticians when analyzing univariate data, identifying outliers in multivariate data pose challenges that univariate data do not. In this paper, after short reviewing some tools for univariate outliers detection, the Mahalanobis distance, as a famous multivariate statistical distances, and its ability to detect multivariate outliers are discussed. As an application the univariate and multivariate outliers of a real data set has been detected using *R* software environment for statistical computing.

**Keywords**: Mahalanobis distance, multivariate normal distribution, multivariate outliers, outlier detection.

## 1. Introduction

The role of statistical distances when dealing with problems such as hypothesis testing, goodness of fit tests, classification techniques, clustering analysis, outlier detection and density estimation methods is of great importance. Using distance measures (or similarities) enable us to quantify the closeness between two statistical objects. These objects can be two random variables, two probability distributions, moment generating functions, an individual sample point and a probability distributions or two individual samples. There exists many statistical distance measures [38], among them the Mahalanobis distance has the advantage of its ability to detect multivariate outliers.

Outliers are those data that deviate from global behavior of majority of data. Outliers or outlying observation have different definition in texts, for example “an outlier deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism”, see [12]. Outliers have major influence on the statistical inference. They increase error variance and reduce the power of statistical

Received March 14, 2019; accepted June 26, 2019
2010 *Mathematics Subject Classification.* Primary 62H99; Secondary 62F35, 62G35

tests and cause bias estimates that may be of substantive interest [22]. Therefore, the process of outlier detection is an interesting and important aspect in the data analysis, see [3] and [5]. Depending on application synonyms are often used for the outlier detection process, among them, one can mention anomaly detection, deviation detection, exception mining, fault detection in safety critical systems, fraud detection for credit cards, intrusion detection in cyber security (unauthorized access in computer networks), misuse detection, noise detection and novelty detection see [1], [9], [23] and [32].

All proximity-based techniques for identification of outliers such as k-Nearest Neighbor (k-NN) algorithm calculate the nearest neighbors of a record using a suitable distance calculation metric such as Euclidean distance, Mahalanobis distance or some other measure of dissimilarity. For large data set using the Mahalanobis distance is computationally more expensive than Euclidean distance as it require to pass through all variables in data set to calculate the underlying inter-correlation structure. An iterative Mahalanobis distance type of method for the detection of outliers in multivariate data has been proposed by [10]. Due to the masking effect, in which one outlier masks a second outlier, if the second outlier can be considered as an outlier only by itself, but not in the presence of the first outlier, detecting multiple outliers is more completed than the case where data consist of a single outlier, since masking effects might decrease the Mahalanobis distance of an outlier. This might happen because a small cluster of outliers attracts mean and inflate variance towards its direction [4]. In such cases using robust estimates of sample mean and variance, can often improve the performance of the detection procedure, see [24] and [30].

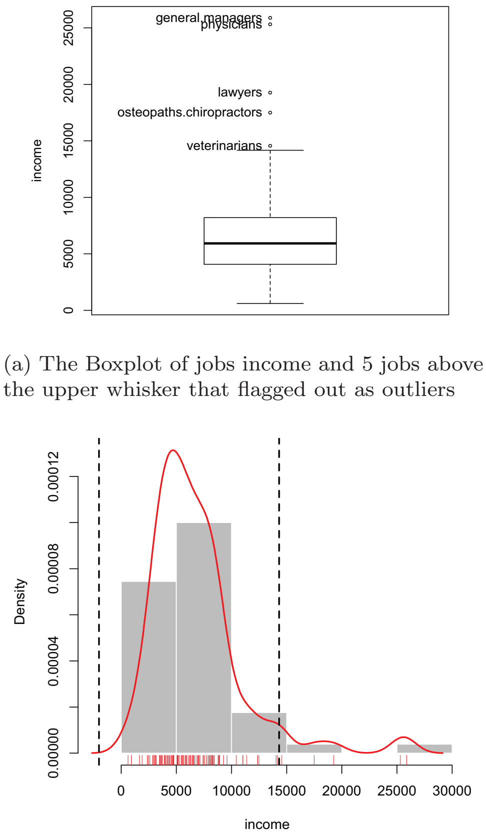
In this paper, the problems of the univariate and multivariate outlier detection has been addressed. For univariate outlier detection, the result of applying the classical visual method based on box-plot and Ven der Loo method [36] on a real data set has been compared. For multivariate outlier detection, usual and robust Mahalanobis distances has been used to find the outliers of a real data set using R software environment for statistical computing.

## 2. Univariate Outlier Detection

A simple visualization tools, such as scatter plot, box-and-whisker (boxplot), stem-and-leaf plot, QQ-plot, etc., can be used to discover the outliers. The box plots, first introduced by [35], are a standardized way of displaying the distribution of data based on a five number summary ("minimum", first quartile ($Q_1$), median, third quartile ($Q_3$), and "maximum"). In general, the box of a box plot shows the median and quartiles. The box plot rule declares observations as outliers if they lie outside the interval

$$Q_1 - k(Q_1 - Q_3), Q_3 + k(Q_3 - Q_1),$$

the common choices for $k$ is 1.5 for flagging (dubbed) outliers and 3.0 for flagging outliers, see Figure 2.1, in which the whiskers are shown for $k = 1.5$. This rule differs from standard outlier identification rules, since it is not sample-size dependent, the probability of declaring outliers when none exist changes with the number of observations [29]. Moreover, for data coming from a random normal sample of size 75, the probability of labeling at least one outlier is 0.5 [13]. Many other statistical tests have been used to detect outliers, as discussed in [3].



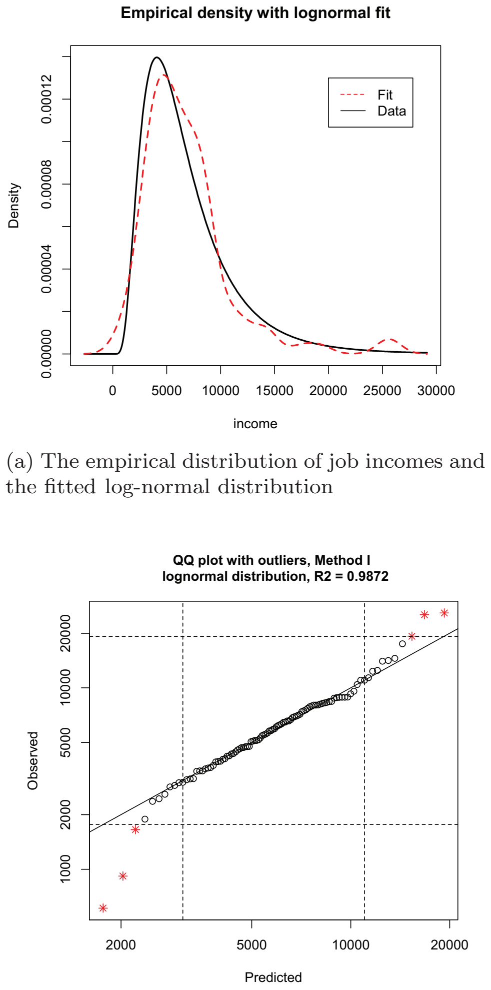
(a) The Boxplot of jobs income and 5 jobs above the upper whisker that flagged out as outliers

(b) The empirical density and the corresponding box plot whiskers. On the $x$ axis, five outliers are shown that exceed the upper whisker threshold

FIG. 2.1: Univariate outlier detection using the boxplot for job incomes in Prestige data set

Van der Loo [36] developed two methods to detect outliers in economic data, when an approximate data distribution is known. In the following, his first method is applied in order to detect the outliers of "income" variable (average income of incumbents, dollars, in 1971) from Prestige of Canadian Occupations data set in "car" package in R software environment [8]. The Prestige data set has 102 rows and 6 columns. This data consists of some measurment related to different occupations.

According to the Kolmogrov-Smirnov goodness-of fit test, the log-normal distribution fits well to income data (p-value=0.47), see the left panel of Figure 2.2. Therefore, the Var der Loo method was applied to detect possible outliers in this data using the plotting facilities developed in the “extremevalues” package in $R$ software environment [37].



(a) The empirical distribution of job incomes and the fitted log-normal distribution

(b) Outlier detected using the first Van der Loo method, which are indicated by $*$ sign

Fig. 2.2: Model based univariate outlier detection for job incomes in Prestige data set

As it is shown in the right panel of Figure 2.2, this method detects six outliers which are located on two sides of data. The Outliers on the left down part of the Figure are case numbers 53, 63, 68, and the rest are 2, 17, 24, whereas the upper outliers on the boxplot are case numbers 2, 17, 24, 25, 26.

The study of outliers in structured situations like regression models are based on the residuals and has been studied by several authors, see [29] and references therein. Five widely used test statistics for detecting outliers have been compared using Monte Carlo method by Balasooriya and Tse [2].

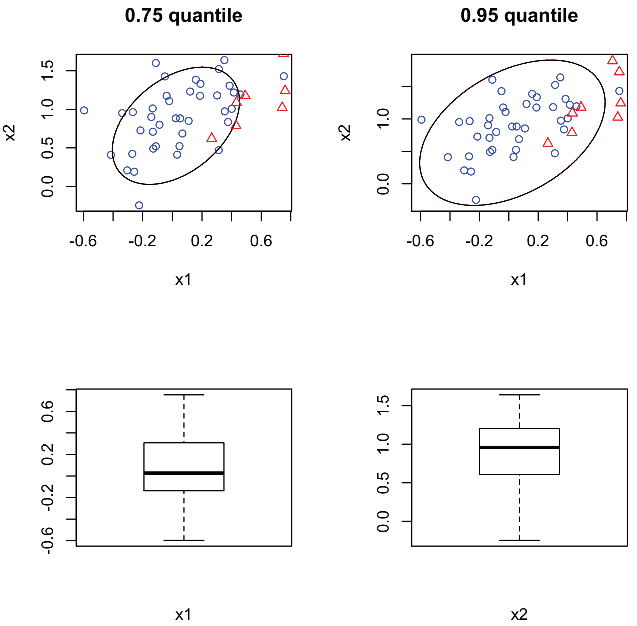


FIG. 2.3: (above) Scatter plot of two simulated samples from bivariate normal distributions, which show clear outliers out of 0.75 and 0.95 cutoffs corresponding to quantiles of the $\chi^2(2)$ distribution, (below) the box plot of margins of the same data with no points lying outside the whiskers

## 3. Multivariate Outliers Detection

Nowadays more and more observed data are multi-dimensional, which increase the chance of occurring unusual observations. The problem is that a few outliers is always enough to distort the results of data (by altering the mean performance, by increasing variability, etc.). Therefore, detecting outliers is a growing concern in many scientific areas, including but not limited to Psychology [18], Financial market [6] and Chemometrics [26].

In the field of multivariate statistics, the Mahalanobis distance has a major application for the detection of outliers [20]. The Mahalanobis distance is defined in the next section. Mahalanobis distance measures the number of standard deviations that an observation is from the mean of a distribution. Since outliers do not behave as normal as usuall observations at least in one dimension, this measure can be used to detect outliers. See [14] for a comparison of Mahalanobis distances with other proximity-based outlier detection techniques.

### 3.1. The Mahalanobis distance

From geometric point of view, the Euclidean distance between two points is the shortest possible distance between them. One problem with the Euclidean distance measure is that it does not take the correlation between highly correlated variables into account. In this situation, Euclidean distance assigns equal weight to such variables, and since these variables measure essentially the same characteristic, therefore this single characteristic gets additional weight. In effect, correlated variables gets excess weight by Euclidean distance, see [16] and [21].

An alternative approach is to scale the contribution of individual variables to the distance value according to the variability of each variable. This approach is considered by the Mahalanobis distance, which has been developed as a statistical measure by PC Mahalanobis, an Indian statistician [19]. The Mahalanobis distance finds wide applications in the field of multivariate statistics. It differs from Euclidean distance in this way that it takes into account the correlations between variables. It is a scale invariant metric and provides a measure of distance between a point $\mathbf{x} \in R^p$ generated from a given $p-$variate (probability) distribution $f_{\mathbf{X}}(.)$ and the mean $\mu = E(\mathbf{X})$ of the distribution. Assume $f_{\mathbf{X}}(.)$ has finite second order moments and denote $\Sigma = E(\mathbf{X} - \mu)$ be the covariance matrix. Then the Mahalanobis distance is defined by

$$D(\mathbf{X}, \mu) = \sqrt{(\mathbf{X} - \mu)^T \Sigma^{-1} (\mathbf{X} - \mu)}. \tag{3.1}$$

If the covariance matrix is the identity matrix, the Mahalanobis distance reduces to the Euclidean distance. For the comparison of these two distances see Figure 3.1, in which the Euclidean and Mahalanobis distances of points located on the circles and ellipse are 1 and 2 unit far away from the center of data. The computation has been done on a data set, that are find under `geog.uoregon.edu/GeogR/data/csv/midwtf2.csv`. The observed difference stems from this fact that the Mahalanobis distance also accounts for the covariance (or correlation) structure of data.

Apart from usual application of the Mahalanobis distance in multivariate analysis techniques such as classification and clustering, discriminant analysis and pattern analysis, principal component analysis, there exists modern applications, among them financial applications [33], image processing [39], Neurocomputing [11] and Physics [31] might be mentioned.

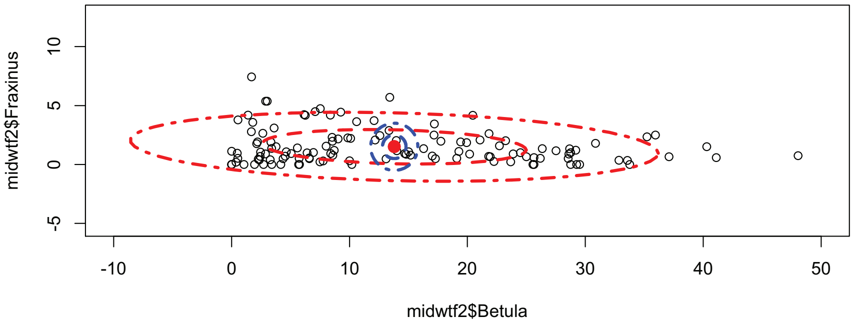


FIG. 3.1: Schematic comparison of the Mahalanobis (ellipse) and Euclidean (circle) distances calculated for a data set. The two lines, circles and ellipses, correspond to the Euclidean and the Mahalanobis distances, of one and two units apart from the center of data

### 3.2. Multivariate normal distribution

Recall the multivariate normal density function below, in which the parameters $\mu$ and $\Sigma$, are the mean and the covariance matrix of the distribution, respectively.

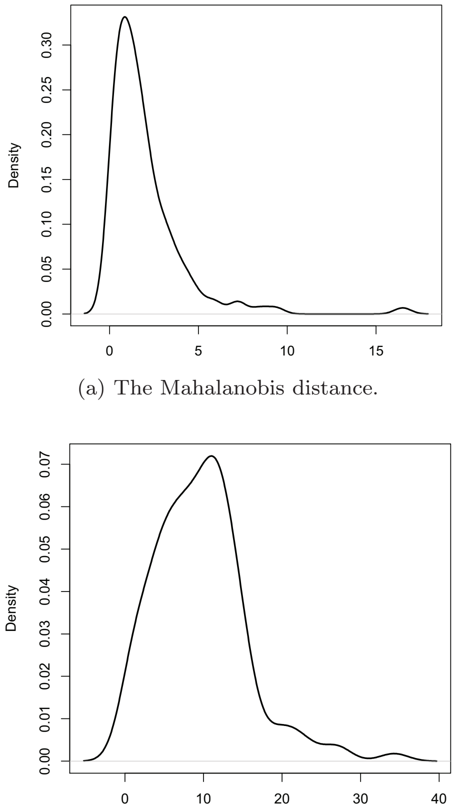
$$\phi(\mathbf{x}) = \left(\frac{1}{2\pi}\right)^{p/2} |\Sigma|^{-1/2} \exp\{-\frac{1}{2}(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)\},$$

note that this density function, $\phi(x)$, only depends on $x$ through the following squared Mahalanobis distance in the exponent:

$$(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu).$$

There are some important facts about this exponent:

- All values of $\mathbf{x}$ such that $(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu) = c$ for any specified constant value $c$ have the same value of the density f(x) and thus have equal likelihood. The paths of these $\mathbf{x}$ values yielding a constant height for the density are ellipsoids. That is, the multivariate normal density is constant on surfaces where the square of the distance $(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)$ is constant. These paths are called contours, which can be constructed from the eigenvalues and eigenvectors of the covariance matrix, meaning that the direction of the ellipse axes are in the direction of the eigenvalues and the length of the ellipse axes are proportional to the constant times the eigenvectors [15].
- As the value of $(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)$ increases, the value of the density function decreases.
- The value of $(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)$ increases as the distance between $\mathbf{x}$ and $\mu$ increases.



(a) The Mahalanobis distance.

(b) The Eucleadn distance

Fig. 3.2: Emperical densities

- The Mahalanobis distance $d^2 = (\mathbf{x} - \mu)'\Sigma^{-1}(\mathbf{x} - \mu)$ has a chi-square distribution with $p$ degrees of freedom, see Figure 3.1.

Suppose that $X$, is a $p$-dimensional vector having multivariate normal distribution, $X \sim N_p(\mu, \Sigma)$, the Mahananobis squared distance $D^2(\mathbf{X}, \mu)$ is then distributed as a $\chi^2$ random variable with $p$ degrees of freedom. The classical approach of outlier detection uses the estimates of the Mahalanobis distance, by plugging in multivariate sample mean $\bar{X}$ and covariance matrix $S$ estimates for unknown mean $\mu$ and covariance matrix $\Sigma$, and tags as outlier any observation which has a Mahalanobis squared distance $d^2(\mathbf{X}, \bar{X})$ lying above a predefined quantile of the $\chi^2$ distribution with p degrees of freedom [7].

This method is problematic, because all relies on normality assumption and the parameters estimates are particularly sensitive to outliers. Therefore, it is important to consider robust alternatives to these estimators for calculating robust Mahalanobis distances. The most widely used estimator of this type is the minimum covariance determinant (MCD) estimator defined in [25] for which also a fast computing algorithm was constructed [27].

In the next section, a sample data has been subjected to find its multivariate outliers by calculating the robust version of the Mahalanobis distances using the $R$ as a modern statistical software for heavy computations involved.

## 4. Analyzing a Sample Data

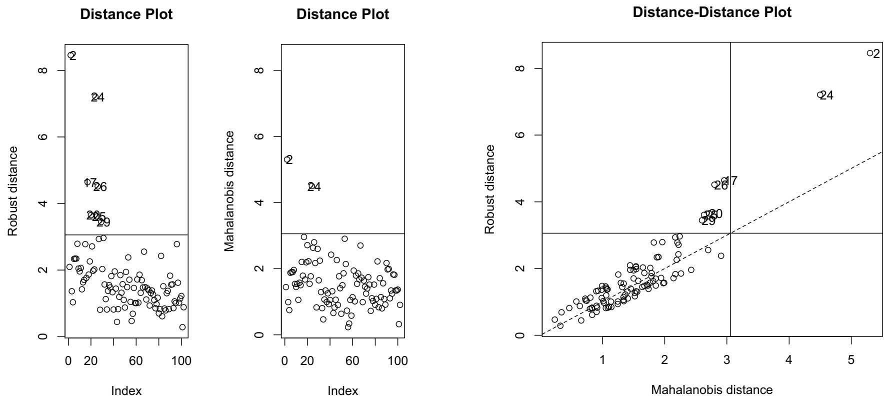
In the following, the vector of three variables of Prestige data set are considered as a multivariate observation. These variables are “education” (average education of occupational incumbents), “income” (average income of incumbents) and “prestige” (Pineo-Porter prestige score for occupation). The aim is to detect multivariate outliers in this data set using robust version of the Mahalanobis distance, the (MCD) estimator, which has been implemented in “rrcov” package in R [34]. First the mean vector and usual (classic) covariance matrix of the observation and the robust version of them are calculated. The results are:

```
-> Method:  Classical Estimator.
Estimate of Location:
education     income   prestige
    10.74    6797.90      46.83

Estimate of Covariance:
           education  income     prestige
education  7.444e+00  6.691e+03  3.991e+01
income     6.691e+03  1.803e+07  5.222e+04
prestige   3.991e+01  5.222e+04  2.960e+02

-> Method:  Robust Estimator.
Robust Estimate of Location:
education     income   prestige
     9.97    5833.96      41.64

Robust Estimate of Covariance:
           education  income     prestige
education  7.156e+00  4.355e+03  3.192e+01
income     4.355e+03  9.695e+06  3.923e+04
prestige   3.192e+01  3.923e+04  2.559e+02
```

Comparing classical and robust estimators of mean vector $\mu$ and the covariance matrix $\Sigma$, shows clear differences. These robust estimators are relatively insensitive to small changes in the bulk of the observations (inliers) or large changes in small number of observations (outliers).

In two left panels of Figure 4.1, the robust and classical Mahalanobis distances are shown in parallel. In most right panel of this figure, the distance-distance plot defined by [28] is shown, which plots the classical Mahalanobis versus robust distances and enable us to classify the observations and identify the potential outliers. The dashed line represents the points for which the robust and classical distances are equal. The horizontal and vertical lines are drawn at values $x = y = \sqrt{\chi^2_{(3,0.975)}}$. Points beyond these lines can be considered as outliers and are identified by their labels. In all panels, the outliers have large robust distances and are identified by their labels, for more details see [34].

Looking at the non-robust Mahalanobis distances at right panel of Figure 4.1 flagged out the observation number 2 and 24 as outliers, whereas robust Mahalanobis at the same panel flagged out the observation number 2, 7, 24, 25, 26 and 29 as outliers. In other words, applying the robust method enabled us to detect hidden outliers which has been masked by each other.



(a) Distance plot.

(b) Chi-square Q-Q plot.

FIG. 4.1: Multivariate outlier detection using the robust Mahalanobis distances

## 5. Conclusion

In this paper, the Mahalanobis distance as a multivariate distance and its advantages relative to the Euclidean distance was reviewed. It made clear when dealing with correlated multivariate data the Mahalanobis distance is more suitable than the Euclidean distance because it takes the correlation into account. Moreover, It was shown how the Mahalanobis distances can be used as a tool for identifying multivariate outliers. When calculating the Mahalanobis distances one needs to estimate the theoretical mean vector and covariance matrix. Estimating these parameters using their usual empirical counterparts especially when data contain outliers yields misleading results, since these estimators are affected seriously by outliers. One reasonable solution is to use robust statistical techniques. There are different robust estimates, but distance-based methods, such as MCD are based on robust estimates of the mean and covariance matrix so that a robust Mahalanobis distance can be computed for each point. In this paper, the above mentioned methods have been applied to detect multivariate outliers in a real data set, using R software environment for statistical computing.

## REFERENCES

1. C. C. Aggarwal: *Outlier Analysis, 2th Edition*. Springer, New York, 2017.
2. U. Balasooriya and Y. K. TSE: *Outlier detection in linear models: A comparative study in simple linear regression*. Communications in Statistics: Theory and Methods **15(12)** 1986, 3589–3598.
3. V. Barnett and T. Lewis: *Outliers in Statistical Data*. John Wiley and Sons, Chichester, England, 1994.
4. C. Becker and U. Gather: *The masking breakdown point of multivariate outlier identification rules*. Journal of the American Statistical Association **94(447)** (1999), 947–955.
5. V. Chandola, A. Banerjee and V. Kumar: *Anomaly detection: a survey*. ACM Comput. Surv. **41(3)** (2009), 1–58.
6. W. Dai and M. G. Genton: *Multivariate functional data visualization and outlier detection*. Journal of Computational and Graphical Statistics **27(4)** (2018), 923–934.
7. C. Fauconnier and G. Haesbroeck: *Outliers detection with the minimum covariance determinant estimator in practice*. Statistical Methodology **6(4)** (2009), 363–379.
8. J. Fox and S. Weisberg: *An R Companion to Applied Regression, 3th Edition*. SAGE Publications, Los Angeles, 2019.
9. M. Goldstein and S. Uchida: *A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data*. PLoS ONE **11(4)** (2016), 1–31.
10. A. S. Hadi: *Identifying multiple outliers in multivariate data*. Journal of the Royal Statistical Society, Series B, **54** (1992), 761-771.
11. N. Haldar K. Farrukh A. Aftab and H. Abbas: *Arrhythmia classification using Mahalanobis distance based improved fuzzy C-Means clustering for mobile health monitoring systems*. Neurocomputing, **220** (2016), 221–235.
12. D. M. Hawkins: *Identification of Outliers*. Chapman and Hall, London, 1980.
13. D. C. Hoaglin, B. Iglewicz and J. W. Tukey: *Performance of some resistant rules for outlier labeling*. Journal of the American Statistical Association **81** (1986), 991–999.
14. V. J. Hodge and J. Austin: *A survey of outlier detection methodologies*. Artif. Intell. Rev. **22** (2004), 85126.
15. R. A. Johnson and D. Wichern: *Applied Multivariate Statistical Analysis*. Prentice Hall, 2007.
16. I. T. Jolliffe: *Principal Component Analysis*. Springer-Verlag (1986).

17. W. J. Krzanowski: *Principles of Multivariate Analysis: A Users Perspective*, Oxford Science Publications, 1988.

18. C. Leys, O. Klein, Y. Dominicy and C. Ley: *Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance*. Journal of Experimental Social Psychology **7**4 (2018), 150–156.

19. P. C. Mahalanobis: *On the generalized distance in statistics*. Proceedings of the National Institute of Sciences (Calcutta), 1936, **2**, pp. 49–55.

20. J. Majewska: *Identification of multivariate outliers problems and challenges of visualization methods*. Informatyka i Ekonometria **4** (2015), 69–83.

21. G. M. Mimmack, S. Mason and J. Galpin: *Choice of distance matrices in cluster analysis: defining regions*. Journal of Climate **14** (2001), 2790–2797.

22. J. W. Osborne and A. Overbay: *The power of outliers (and why researchers should always check for them)*. Pract. Assess. Res. Eval. **9(6)** (2004), 1–9.

23. M. A. F. Pimentel, D. A. Clifton, L. Clifton and L. Tarassenko: *A review of novelty detection*. Signal Processing **99** (2014), 215-249.

24. D. M. Rocke and D. L. Woodruff: *Identification of outliers in multivariate data*. Journal of the American Statistical Association **91(435)** (1996), 1047–1061.

25. P. J. Rousseeuw: *Multivariate estimation with high breakdown point*. In: Mathematical Statistics and Applications (W. Grossmann, G. Pflug, I. Vincze, W. Wertz, eds.), Reidel, Dordrecht, **B**, 1985, pp 283–297.

26. P. Rousseeuw, M. Debruyne, S. Engelen and M. Hubert: *Robustness and outlier detection in Chemometrics*. Critical Reviews in Analytical Chemistry **36(3)**, (2006), 221–242.

27. P. J. Rousseeuw and K. Van Driessen: *A fast algorithm for the minimum covariance determinant estimator*. Technometrics, **41** (1999), 212–223.

28. P. J. Rousseeuw and B. C. van Zomeren: *Robust distances: simulation and cutoff Values*. In: Directions in Robust Statistics and Diagnostics, Part II. (W. Stahel, S. Weisberg, eds.), Springer-Verlag, New York, 1991.

29. T. A. Sajesh and M. R. Srinivasan: *An overview of multiple outliers in multidimensional data*. Sri Lankan Journal of Applied Statistics **14** (2013), 86–120.

30. C. J. Santos-Pereira and A. M. Pires: *Detection of outliers in multivariate data: a method based on clustering and robust estimators*. In: Compstat (W. Hrdle, B. Rnz, eds.), Physica, Heidelberg, 2002, pp 291–296.

31. N. G. Sharma, M. Silarski, T. Bednarski, P. Biaas, E. Czerwiski, A. Gajos, M. Gorgol, B. Jasiska, D. Kamiska, . Kapon, G. Korcyl, P. Kowalski, T. Kozik, W. Krzemie, E. Kubicz, S. Niedwiecki, M. Paka, L. Raczyski, Z. Rudy, O. Rundel, A. Somski, A. Strzelecki, A. Wieczorek, W. Wilicki, M. Zieliski, B. Zgardziska and P. Moskal: *Reconstruction of hit time and hit position of annihilation quanta in the J-PET detector using the Mahalanobis distance*. Nukleonika **4** (2015), 765–769.

32. K. Singh and D. S. Upadhyaya: *Outlier detection: Applications and techniques*. International Journal of Computer Applications **89(6)** (2014) 307–323.

33. S. Stckl and M. Hanke: *Financial applications of the Mahalanobis distance*, SSRN Electronic Journal **1(2)** (2014), 78–84.

34. V. Todorov and P. Filzmoser: *An object-oriented framework for robust multivariate analysis*. Journal of Statistical Software **32(3)** (2009), 1–47.

35. J. W. Tukey: *Exploratory Data Analysis*. Addison-Wesley, New York, USA, 1977.

36. M. P. J. van der Loo: *Distribution based outlier detection for univariate data*. Discussion paper **1**0003 Statistics Netherlands (2010), 3–14.

37. M. P. J. van der Loo: *Extremevalues, an R package for outlier detection in univariate data*. R package version 2.3 (2010), url = http://www.github.com/markvanderloo/extremevalues.

38. G. M. Venturini: *Statistical Distances and Probability Metrics for Multivariate Data*. Ph. D. Thesis, Charles III University of Madrid, 2015.

39. Y. Zhang, B. Du, L. Zhang and S. Wang: *A low-rank and sparse matrix decomposition-based Mahalanobis distance method for hyperspectral anomaly detection*. IEEE Transactions on Geoscience and Remote Sensing **220** (2016), 1376–1389.

Hamid Ghorbani
Faculty of Mathematical Sciences
Department of Statistics
University of Kashan
Kashan 87317-53153, I. R. Iran
hamidghorbani@kashanu.ac.ir