Paper PH-89

Using SAS/STAT to implement a multivariate adaptive outlier detection approach to distinguish outliers from extreme values

Paulo Macedo, Ph.D., Integrity Management Services LLC

ABSTRACT

Hawkins (1980) defines an outlier as “an observation that deviates so much from other observations as to arouse the suspicion that it was generated by a different mechanism”. To identify data outliers, a classic multivariate outlier detection approach implements the Robust Mahalanobis Distance Method by splitting the distribution of distance values in two subsets (within-the-norm and out-of-the-norm), with the threshold value usually set to the 97.5% Quantile of the Chi-Square distribution with p (number of variables) degrees of freedom and items whose distance values are beyond it are labeled out-of-the-norm. This threshold value is an arbitrary number, however, and it may flag as out-of-the-norm a number of items that are actually extreme values of the baseline distribution rather than outliers. Therefore, it is desirable to identify an additional threshold, a cutoff point that divides the set of out-of-norm points in two subsets - extreme values and outliers.

One way to do this – in particular for larger databases – is to Increase the threshold value to another arbitrary number but this approach requires taking into consideration the size of the dataset since size will affect the threshold separating outliers from extreme values. A 2003 article by Gervini (Journal of Multivariate Statistics) proposes “an adaptive threshold that increases with the number of items **n** if the data is clean but it remains bounded if there are outliers in the data.” In 2005 Filzmoser, Garrett and Reimann (Computers & Geosciences) built on Gervini’s contribution to derive by simulation a relationship between the number of items **n,** the number of variables in the data **p** and a critical ancillary variable for the determination of outlier thresholds.

This paper implements the Gervini adaptive threshold value estimator using PROC ROBUSTREG and the SAS Chi-Square functions CINV and PROBCHI, available in the SAS/STAT environment. It also provides data simulations to illustrate the reliability and the flexibility of the method in distinguishing true outliers from extreme values.

.INTRODUCTION

The Hawkins (1980) reference to outliers as observations that may be generated by a “different mechanism” refers to the fact that most of the population from which the observed sample is drawn belongs to a prevalent distribution of values (background data), but this population is also expected to include a small subset of values associated with a distinct data generating process which is called a “contaminating distribution.”

In the context of fraud detection in Government Health Insurance Programs (GHIPs), for example, investigators frequently have to deal with the question whether a provider billing pattern is an indicator of an admissible large-scale health care operation - an extreme value - or a sign of a fraud scheme - an outlier in the relevant peer-comparison group. Properly defined peer-comparison groups of providers rendering services to GHIPs reveal that besides a number of acceptable “extreme values,” the presence of at least some outliers occurs with relative frequency - in particular in a few large metropolitan areas. For instance, The Health Care Fraud and Abuse Control Program Annual Report for Fiscal Year 2013 (Department of Health and Human Services and Department of Justice) finds that “Payments by Medicare for Durable Medical Equipment (DME) in Miami-Dade County13 alone hit an all-time high in the third-quarter of 2006, when payments exceeded $73 million; those payments have decreased over time, and in the first-quarter of 2013 payments were under $15 million.” The report credits law enforcement activity and better measures taken by the Center for Medicare and Medicaid Services for the decrease, which suggests the significant presence of outliers in the earlier data.

Reimann, Filzmoser and Garrett (2005) emphasize “the difference between extremes of a distribution and true outliers. Outliers are thought to be observations coming from one or more different distributions, and extremes are values that are far away from the center but which belong to the same distribution.” In exploratory data analysis it is convenient to start with simply identifying all out-of-the-norm observations as extreme and proceed then to select the subset of outliers.

It is important to take into account, also, that the way the center of the distribution is estimated affects the determination of the threshold(s) for outlier detection because the center is responsive to influential points that may “mask” the presence of some outliers if they are not properly dealt with in the estimation. For example, consider the identification of potential outliers among top values (right tail of the distribution) of a sample draw of 20 numbers, {2.0, 2.5, 3.0, 4.0, 5.0, 6.0, 7.0, 7.5, 8.0, 8.5, 8.7, 9.0, 9.5, 9.7, 10.0, 10.4, 10.5, 17.0, 17.5, and 19.0}. The often used outlier detection threshold computed as the sample mean plus twice the value of the standard deviation will identify in

this case only one right-tail outlier, the item valued as 19; it seems intuitively clear though that at least two other numbers are potential outliers, the two items valued as 17 and 17.5. A way to deal with the issue is to recalculate the sample mean with the exclusion of influential point(s), i.e., to estimate the sample mean in a "robust" fashion – which is by construction more resistant against the influence of outlying observations. In the example above, if the sample mean is re-estimated with the exclusion of the item valued as 19, the mean-plus-two-standard-deviations rule applied to the original 20 items will identify three of them as outliers – 17, 17.5 and 19, instead of one - 19.

Very often a robust estimation of the center (or location) and shape (or scatter) of the sample distribution sets thresholds which are able to detect more outliers than a non-robust estimation would identify, which makes even clearer the importance of distinguishing extreme values of a background distribution from outliers of a contaminating distribution. If the data structure involves more than one variable (multivariate case) an efficient method to accomplish this compares the values of the notional Chi-square distribution with the values of the empirical distribution, which are based on the squared distances of the sampled items to the center of the dataset. The adaptive outlier threshold suggested by Gervini (2003) implements the idea. His proposed threshold is flexible enough to increase with the number of items **n** if the data is clean – that means if there are no items from a contaminating distribution in the data - but it remains bounded if there are outliers in the data.

This paper sets up Gervini's adaptive threshold value estimator using PROC ROBUSTREG and the SAS Chi-Square functions CINV and PROBCHI, available in the SAS/STAT environment. The text includes five sections: **A Standard Multivariate Outlier Detection Method** describes briefly the Mahalanobis Distance, "a well-known criterion for multivariate outlier detection that depends on estimated parameters of the multivariate distribution"; **Robust Distance Outlier Detection Method** discusses concisely the essential features of a frequently used robust estimator of the location (center) and scatter (covariance matrix) of the data structure, the minimum covariance determinant estimator – implemented by the SAS PROC ROBUSTREG; **Adaptive Threshold Outlier Detection Method** presents the basic elements of Gervini's adaptive outlier threshold and discusses briefly the approximation proposed by Filzmoser Garrett and Reimann to estimate the threshold separating outliers and extreme values; **Adaptive Threshold Outlier Detection Method – A Simulation** illustrates by means of data simulation the reliability and the flexibility of the method in distinguishing true outliers from extreme values; and the **Conclusion**.

A STANDARD MULTIVARIATE OUTLIER DETECTION METHOD

When the data structure includes more than one variable it is essential that the outlier detection method takes into account the possible interdependence among variables. Ben-Gal (2005) summarizes the issue and provides a graphical two-dimensional example to make the point: "In many cases multivariable observations cannot be detected as outliers when each variable is considered independently. Outlier detection is possible only when multivariate analysis is performed, and the interactions among different variables are compared within the class of data. A simple example can be seen in the figure below, which presents data points having two measures on a two-dimensional space. The lower left observation is clearly a multivariate outlier but not a univariate one. When considering each measure separately with respect to the spread of values along the x and y axes, we can see that they fall close to the center of the univariate distributions. Thus, the test for outliers must take into account the relationships between the two variables, which in this case appear abnormal." Ben-Gal's Figure 1.1 illustrating the point is reproduced below.

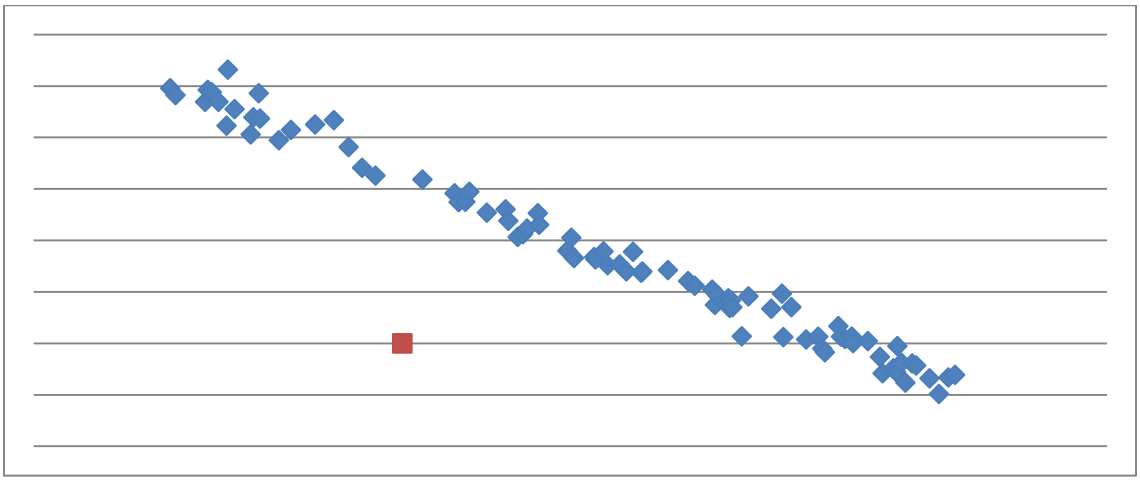

Figure 1. A Two-dimensional Space with one Outlying Observation (reproduced from Ben-Gal 2005)

The figure shows an isolated point far apart from a set of clustered points. By considering each measure separately with respect to the spread of values along the horizontal axis **x** and vertical axis **y** axis it is clear each coordinate of the isolated point falls close to the center of its respective univariate distribution. From this perspective then the

isolated point is not an outlier. However, if the interdependence between **x** and **y** is taken into account the isolated point emerges as an outlier because it is located far away from the two-dimensional center of the data structure.

In the multivariate case not only the distance of an observation from the center of the data but also the shape of the data has to be considered – much the same way as the distance of an observation to the central location of the data (arithmetic mean) and the dispersion of the data points (standard deviation) are frequently the elements taken into account in outlier detection in the univariate case. Multivariate outlier detection methods basically evaluate how far data points are from the center of the data distribution. Among a number of distance measures used to accomplish the task the Mahalanobis distance is a frequently used criterion - based on the estimated parameters (location and shape) of the multivariate distribution.

The Mahalanobis Distance is defined as follows:

Given **n** observations from a ***p-dimensional*** dataset (***p*** variables in the data), denote the sample mean vector by $\bar{\mathbf{x}}_\mathbf{n}$ and the sample covariance matrix by $\boldsymbol{V_n}$, where

$$\boldsymbol{V_n} = \frac{1}{n-1}\sum_1^n (\boldsymbol{x_i} - \bar{\boldsymbol{x}}_\boldsymbol{n})(\boldsymbol{x_i} - \bar{\boldsymbol{x}}_\boldsymbol{n})^T$$

The **Mahalanobis Distance** for each multivariate data point $\boldsymbol{i}, \boldsymbol{i} = \mathbf{1}, \ldots \boldsymbol{n}$, is denoted by $\boldsymbol{MD_i}$ and given by

$$\boldsymbol{MD_i} = \left(\sum_1^n (\boldsymbol{x_i} - \bar{\boldsymbol{x}}_\boldsymbol{n})^T \boldsymbol{V_n^{-1}} (\boldsymbol{x_i} - \bar{\boldsymbol{x}}_\boldsymbol{n})\right)^{1/2}$$

For multivariate normally distributed data the squared Mahalanobis distance $\boldsymbol{MD_i^2}$ follows a chi-square distribution $\boldsymbol{\chi_p^2}$ with degrees of freedom equal to the ***p*** number of variables included in the calculation. Typically the outlier detection threshold is set to a certain quantile of $\boldsymbol{\chi_p^2}$, for example 97.5% (or 98%). This value identifies an ellipsoid that includes the set of hypothetical boundary data points having the same squared Mahalanobis distances $\boldsymbol{MD_i^2}$ from the center of the data structure. Any point beyond the boundary is classified as an outlier.

ROBUST DISTANCE OUTLIER DETECTION METHOD

It is likely that the estimated location and scatter parameters based on the available sampled data points will not be the best representation of the data structure because the estimators may be sensitive to influential points, as in the example of a sample of 20 numeric values discussed in the Introduction. The standard measure used to evaluate the robustness of an estimator is the sample breakdown point, which is "the fraction of data that can be given arbitrary values without making the estimator arbitrarily bad" and "typically this number is some function of the sample size **n.**" In addition, "to get a single number, one uses the asymptotic breakdown point, which is the limit of the finite sample breakdown point as **n** goes to infinity" (quotes from Geyer, 2006). The closer the breakdown point is to zero the less robust the estimator is. If the estimator of the location of the data structure is the arithmetic mean then the finite sample breakdown point is (**1/n)** and the corresponding asymptotic breakdown point converges to 0. In contrast, if the estimator is the median of the data values then the finite sample breakdown point is [**(n−1)/(2n)]** and the related asymptotic breakdown point converges to 0.5 (50%). The median is a more robust estimator than the mean as illustrated in the case of the sample of 20 numeric values discussed above: no matter how large a new number is assigned to replace the sample original top value (first set to 19), the median remains equal to 8.6 while the arithmetic mean quickly drifts away from its first estimate, 8.74.

The standard Mahalonobis distance estimator of the location (center) of the data structure is actually the multivariate counterpart of the univariate arithmetic mean: it uses all the available observations in the data and it is likely to be subject to the pull of influential points that makes the estimate a bad representation of the data structure.

One of the most widely used distance estimators in the class of robust estimators is the minimum covariance determinant (MCD) estimator proposed by Rousseeuw (1984, 1985). The MCD estimator is determined by that subset of observations of size **h** which minimizes the determinant of the sample covariance matrix, computed from all the possible **h-subsets** of the original set of **n** data points.

The original MCD estimator required the evaluation of all possible $\binom{n}{h}$ subsets of size **h** of the **n** items in the reference sample making its computation very inefficient. Instead the fast algorithm developed by Rousseeuw and Van Driessen (1999) divides the data processing in computational steps including a rule to select a sequence of subsets of size **h** such that the value of the determinant of the covariance matrix in step **t** is less than the value in step **(t – 1).** The process stops when either the determinants of the last two sequential subsets have the same value or the determinant of the subset in the very last step is zero.

Given **n** observations from a ***p-dimensional*** dataset (***p*** variables in the data), the **MCD-based Robust Distance** for each multivariate data point $\boldsymbol{i}, \boldsymbol{i} = \mathbf{1}, \ldots \boldsymbol{n}$, is denoted by $\boldsymbol{RD_i}$ and given by

$$\boldsymbol{RD}(\boldsymbol{x_i}) = \left(\sum_i^n \left(\boldsymbol{x_i} - \boldsymbol{T}(\boldsymbol{X})\right)^{\boldsymbol{T}} \boldsymbol{C}(\boldsymbol{X})^{-1} \left(\boldsymbol{x_i} - \boldsymbol{T}(\boldsymbol{X})\right)\right)^{1/2}$$

where $\boldsymbol{T}(\boldsymbol{X})$ and $\boldsymbol{C}(\boldsymbol{X})$ are the robust multivariate location and scatter, respectively, obtained by MCD.

Very often ***h*** is set as ***h = 0.75n*** where ***n*** is the number of items in the sample. As the breakdown value of the MCD estimator is roughly equal to $(\boldsymbol{n} - \boldsymbol{h})/\boldsymbol{n}$ (Rousseeuw and Van Driessen, 1999) this choice of ***h*** amounts to set that value to approximately 25%, which means that the estimator does not generate biased estimates of the location and scatter of the data structure as long as the fraction of outliers in the data is below 25%.

The entry "Robust Distance" in the SAS web link "The ROBUSTREG Procedure" (available at http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_rreg_sect021.htm) discusses the implementation of the MCD-based Robust Distance in the SAS STAT environment.

ADAPTIVE THRESHOLD OUTLIER DETECTION METHOD

Gervani in his 2003 article on robust estimators of location and scatter starts by noting that the threshold value frequently used to identify outliers in the data ($\boldsymbol{\chi^2_{p,0..975}}$) is an arbitrary number: "For large data sets a considerable number of observations will be discarded even if they do follow the normal model." He proceeds to state that "One way to avoid this problem (arbitrariness of the threshold choice) is to increase the threshold value to another arbitrary fix number, but this will affect the bias of the reweighted estimator. A better alternative is to use an adaptive threshold value that increases with **n** if the data is *clean* but remains bounded if there are outliers in the sample." The following paragraph draws from his 2003 paper and states its main idea (the original notation to represent the data location and scatter was replaced by the one used in the previous Section).

"We propose one method of constructing such adaptive threshold values. Let $\boldsymbol{G_n}(\boldsymbol{u}) = \frac{1}{n}\sum_1^n \boldsymbol{n}\mathbf{1}(\boldsymbol{d^2}(\boldsymbol{x_i}, \boldsymbol{T}(\boldsymbol{X}), \boldsymbol{C}(\boldsymbol{X})) \leq \boldsymbol{u})$ be the empirical distribution of the squared Mahalanobis distances. Let $\boldsymbol{G_p}(\boldsymbol{u})$ be the $\boldsymbol{\chi^2_p}$ distribution function. If $\boldsymbol{\eta} = \boldsymbol{\chi^2_{p,(1-\alpha)}}$ for a certain small $\boldsymbol{\alpha}$, say ***α = 0.025***, define $\boldsymbol{\alpha_n} = \boldsymbol{sup}_{(u \geq \eta)} \left\{\boldsymbol{G_p}(\boldsymbol{u}) - \boldsymbol{G_n}(\boldsymbol{u})\right\}^+$ where $\{.\}^+$ indicates the positive part. This $\boldsymbol{\alpha_n}$ can be regarded as a measure of outliers in the sample."

Filzmoser, Garrett and Reimann (FGR, 2005) build on Gervini (2003) by strongly emphasizing the way adaptive outlier thresholds may help to address the outliers/extreme values issue: "defining outliers by using a fixed threshold value, for example $\boldsymbol{\chi^2_{p,(1-\alpha)}}$ **-** where $(\mathbf{1} - \boldsymbol{\alpha}) = \mathbf{97.5\%}$ **or** $(\mathbf{1} - \boldsymbol{\alpha}) = \mathbf{98\%}$, is rather subjective because: 1) if the data should indeed come from a single multivariate normal distribution, the threshold would be infinity because there are no observations from a different distribution (only extremes); 2) there is no reason why this fixed threshold should be appropriate for every data set; and 3) the threshold has to be adjusted to the sample size."

FGR restate that the tails of the data distribution are the focus of the adaptive outlier detection approach, "The tails are defined by $\boldsymbol{\delta} = \boldsymbol{\chi^2_{p;(1-\alpha)}}$ for a certain small **α** (typically **α = 0.025** or **α = 0.02**) and

$$\boldsymbol{p_n}(\boldsymbol{\delta}) = \boldsymbol{sup}_{\boldsymbol{u}>\delta}(\boldsymbol{G}(\boldsymbol{u}) - \boldsymbol{G_n}(\boldsymbol{u}))^+,$$

where **+** indicates positive differences (because a negative difference would not be a sign of presence of outliers). $\boldsymbol{p_n}(\boldsymbol{\delta})$ measures the departure of the empirical from the theoretical distribution only in the tails, defined by the value of $\boldsymbol{\delta}$."

Data coming from a multivariate normal distribution have no observation that should be classified as an outlier and those observations having large robust distances deserve only the label of extreme values. However, more often than not data include both a background distribution and at least one contaminating distribution. Anyway $\boldsymbol{p_n}(\boldsymbol{\delta})$ is not used directly to distinguish extreme values from outliers; instead an ancillary critical value $\boldsymbol{p_{crit}}$ is introduced to do that. The measure of outliers is defined as

$$\alpha_n(\delta) = \begin{cases} \mathbf{0} & \boldsymbol{if}\ \boldsymbol{p_n}(\boldsymbol{\delta}) \leq \boldsymbol{p_{crit}}(\boldsymbol{\delta}, \boldsymbol{n}, \boldsymbol{p}) \\ \boldsymbol{p_n}(\boldsymbol{\delta}) & \boldsymbol{if}\ \boldsymbol{p_n}(\boldsymbol{\delta}) > \boldsymbol{p_{crit}}(\boldsymbol{\delta}, \boldsymbol{n}, \boldsymbol{p}) \end{cases}$$

The threshold value is defined as $\boldsymbol{c_n(\delta) = G_n^{-1}(1 - \alpha_n(\delta))}$.

FGR derive the way the critical value $\boldsymbol{p_{crit}}$ relates to different sample sizes ***n*** and different number of variables ***p*** by simulation. The resulting definition of the critical value as a function of ***n*** and ***p*** is:

$$\boldsymbol{p_{crit}(\delta, n, p) = \frac{0.24 - 0.003p}{\sqrt{n}} \; for \; p \leq 10,}$$

$$\boldsymbol{p_{crit}(\delta, n, p) = \frac{0.252 - 0.0018p}{\sqrt{n}} \; for \; p > 10.}$$

ADAPTIVE THRESHOLD OUTLIER DETECTION METHOD – A SIMULATION

To evaluate the performance of the three outlier detection methods discussed above - Mahalanobis Distance with Fixed Threshold, Robust Distance with Fixed Threshold and Robust Distance with Adaptive Threshold – this study draws items from a two-dimensional bi-normal distribution and including three levels of contaminating distribution – 5%, 10% and 15%. The correlation between the two variables was set to zero - meaning that these variables are independent by design, although the relaxation of this constraint would not change the main results of the study.

The samples have size of 1,000 data points corresponding to the sum of items drawn from the background distribution and the items drawn from the contaminating distribution. For example, 5% of contamination means that there are 50 true outliers in the sample and 950 within-the-norm values, and so on. The background distribution of items classified as within-the-norm is $\boldsymbol{N_2(0, I)}$ and the contaminating distribution is $\boldsymbol{N_2(3.5 \times 1, I)}$, where ***0*** and ***1*** stand for the 2 x 1 vectors $\begin{bmatrix}0\\0\end{bmatrix}$ and $\begin{bmatrix}1\\1\end{bmatrix}$ and ***I*** is the two-dimensional identity matrix $\boldsymbol{I} = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$. The implementation of both the Mahalanobis Distance and Robust Distance used the **SAS PROC ROBUSTREG** (SAS/STAT environment) with the method option = LTS and the model option = Leverage. The implementation of the Adaptive Threshold method used parts of the texts of Gervini (2003) and Filzmoser, Garrett and Reimann (2005) as pseudo-code sources for the SAS code provided below.

Tables 1 to 3 below summarize the results obtained by using the three methods in the analysis of sets of data points including different contamination levels.

The results of the simulation with the lowest level of contamination, displayed in Table 1, confirm that even at that relatively low level the Robust Distance with Fixed Threshold (RDFT) estimator does a better job in identifying outliers than the standard Mahalanobis Distance with Fixed Threshold (MDFT) method. This better performance though comes at the cost of larger number of identified outliers that are actually false positives. In contrast, the Adaptive Threshold (AT) method does also a good job in classifying true outliers as outliers with the advantage of misclassifying as outliers a much lesser number of items than the RDTF does.

The results corresponding to the intermediary and the highest levels of contamination, displayed in Tables 2 and 3, reinforce the pattern – the AT method is always efficient in identifying true outliers but parsimonious in misclassifying items as outliers when they are not. As the contamination level increase the RDTF shares of misclassified items decrease but it never outperforms the AT method in that indicator; in fact, the misclassification of items by the AT method at the highest level of contamination is remarkably low. As expected, the ability of the MDTF method in detecting outliers deteriorates quickly as the contamination level increases.

Table 1 Comparing Three Outlier Detection Methods, Sample Size = 1,000 and Contamination = 50

METHOD	SAMPLE NUMBER	NUMBER OF TRUE POSITIVES	NUMBER OF FALSE POSITIVES	%(NUMBER OF IDENTIFIED TRUE OUTLIERS/TRUE OUTLIERS)	%(NUMBER OF IDENTIFIED FALSE OUTLIERS/TRUE OUTLIERS)
MAHALANOBIS	1	36	4	72.00%	8.00%
MAHALANOBIS	2	43	6	86.00%	12.00%
MAHALANOBIS	3	39	7	78.00%	14.00%
MAHALANOBIS	4	43	10	86.00%	20.00%
MAHALANOBIS	5	43	6	86.00%	12.00%
ROBUST	1	50	30	100.00%	60.00%
ROBUST	2	50	25	100.00%	50.00%
ROBUST	3	49	28	98.00%	56.00%
ROBUST	4	50	25	100.00%	50.00%
ROBUST	5	50	23	100.00%	46.00%
ADAPTIVE	1	48	9	96.00%	18.00%
ADAPTIVE	2	48	3	96.00%	6.00%
ADAPTIVE	3	48	7	96.00%	14.00%
ADAPTIVE	4	49	3	98.00%	6.00%
ADAPTIVE	5	49	2	98.00%	4.00%

Table 2 Comparing Three Outlier Detection Methods, Sample Size = 1,000 and Contamination = 100					
METHOD	SAMPLE NUMBER	NUMBER OF TRUE POSITIVES	NUMBER OF FALSE POSITIVES	%(NUMBER OF IDENTIFIED TRUE OUTLIERS/TRUE OUTLIERS)	%(NUMBER OF IDENTIFIED FALSE OUTLIERS/TRUE OUTLIERS)
MAHALANOBIS	1	47	10	47.00%	10.00%
MAHALANOBIS	2	47	8	47.00%	8.00%
MAHALANOBIS	3	46	11	46.00%	11.00%
MAHALANOBIS	4	43	7	43.00%	7.00%
MAHALANOBIS	5	49	9	49.00%	9.00%
ROBUST	1	100	30	100.00%	30.00%
ROBUST	2	98	30	98.00%	30.00%
ROBUST	3	100	43	100.00%	43.00%
ROBUST	4	100	27	100.00%	27.00%
ROBUST	5	99	40	99.00%	40.00%
ADAPTIVE	1	96	10	96.00%	10.00%
ADAPTIVE	2	95	8	95.00%	8.00%
ADAPTIVE	3	100	18	100.00%	18.00%
ADAPTIVE	4	99	6	99.00%	6.00%
ADAPTIVE	5	96	20	96.00%	20.00%

Table 3 Comparing Three Outlier Detection Methods, Sample Size = 1,000 and Contamination = 150					
METHOD	SAMPLE NUMBER	NUMBER OF TRUE POSITIVES	NUMBER OF FALSE POSITIVES	%(NUMBER OF IDENTIFIED TRUE OUTLIERS/TRUE OUTLIERS)	%(NUMBER OF IDENTIFIED FALSE OUTLIERS/TRUE OUTLIERS)
MAHALANOBIS	1	27	4	18.00%	2.67%
MAHALANOBIS	2	31	8	20.67%	5.33%
MAHALANOBIS	3	30	9	20.00%	6.00%
MAHALANOBIS	4	30	9	20.00%	6.00%
MAHALANOBIS	5	32	2	21.33%	1.33%
ROBUST	1	146	13	97.33%	8.67%
ROBUST	2	149	17	99.33%	11.33%
ROBUST	3	148	28	98.67%	18.67%
ROBUST	4	148	20	98.67%	13.33%
ROBUST	5	150	16	100.00%	10.67%
ADAPTIVE	1	142	1	94.67%	0.67%
ADAPTIVE	2	144	0	96.00%	0.00%
ADAPTIVE	3	146	8	97.33%	5.33%
ADAPTIVE	4	143	0	95.33%	0.00%
ADAPTIVE	5	142	1	94.67%	0.67%

Table 4 shows that the performance pattern of the three methods remains consistent as the number of samples analyzed increases. The table displays the mean and the standard deviations (STD) computed for 100 samples for the three levels of contamination included in the sets of data points. As the contamination levels increase the MDFT method has a significant reduction in its average ability to identify true outliers; on the other hand, this ability remains high for both the RDTF and the AT methods; however, the AT method has a clear edge over the RDTF method because it consistently identifies a much lower average share of false positives across all contamination levels.

Table 4 Statistics from 100 Samples of Size = 1,000 and Three Levels of Contamination					
METHOD	CONTAMINATION LEVEL	MEAN OF SHARES OF TRUE POSITIVES	STD OF SHARES OF TRUE POSITIVES	MEAN OF SHARES OF FALSE POSITIVES	STD OF SHARES OF FALSE POSITIVES
MAHALANOBIS	50	81.78%	4.70%	18.90%	5.59%
MAHALANOBIS	100	46.69%	3.59%	8.16%	2.51%
MAHALANOBIS	150	20.51%	2.29%	4.85%	1.50%
ROBUST	50	99.36%	1.30%	62.42%	12.10%
ROBUST	100	99.29%	0.86%	27.33%	5.29%
ROBUST	150	99.12%	0.81%	15.96%	3.09%
ADAPTIVE	50	97.34%	2.22%	17.80%	9.95%
ADAPTIVE	100	96.88%	1.67%	6.71%	4.25%
ADAPTIVE	150	96.47%	1.59%	3.58%	2.32%

CONCLUSION

The two main contributions of this paper are a concise review of key papers on location and scatter estimators used in multivariate outlier detection which emphasize the advantage of the Adaptive Threshold over Standard Fixed Threshold frameworks to identify data outliers and, based on this review, the development of SAS code to implement an efficient and reliable Adaptive Threshold Outlier Detection method in the SAS/STAT environment.

The Adaptive Threshold method uses information of the tails of the distributions of values - theoretical and empirical - and a computed critical value to identify a set of items that are the strongest candidates to be true outliers. This critical value depends both on the number of items in the sample ***n*** and on the number of variables ***p*** in the set of data points, and allows for the distinction between outliers and extreme values.

The implementation of the Adaptive Threshold method in the SAS/STAT environment used elements of the literature reviewed as pseudo-code sources for the development of the SAS code. This code processed samples of items drawn from a two-dimensional bi-normal distribution including three levels of contaminating distributions in a simulation study that evaluated the performance of the Adaptive Threshold method as compared to that of Fixed Threshold approaches. The simulation showed that the Adaptive Threshold method had by far the best overall performance as it was able to identify most of the true outliers in the data and had the least number of false positives for all contamination levels.

Code to compare three multivariate outlier detection methods

```
/*****************************************************
THE DATA STEPS TO GENERATE THE INPUT DATASETS WITH DATA
DRAWN FROM BI-NORMAL DISTRIBUTIONS ARE NOT SHOWN HERE
*****************************************************/

%macro ODSOff(); /* Call prior to BY-group processing */
ods graphics off;
ods exclude all;
ods noresults;
%mend;

%macro ODSOn(); /* Call after BY-group processing */
ods graphics on;
ods exclude none;
ods results;
%mend;
/*************************************************************************
Macro parameters: "p" is the number of variables; "n_samples" is the number
of samples; "bgrd" is the number of items from the background distribution;
and "cont" is the number of items from the contaminating distribution
*************************************************************************/
%macro out_meth(p, n_samples, bgrd, cont);
%let size = %eval(&bgrd + &cont);

%do i = 1 %to &n_samples;
%ODSoff;
proc robustreg data=Points_&size._out_&cont._smpl_&i method=LTS ;
model z = x  y /leverage /* (mcdinfo opc)*/;
output out = hat_robust_m RD = RD_DIST MD = Mahal ;
run;
%ODSon;

/*************************************************
Steps below identify sets of outliers using both the
Mahalanobis Distance and the Robust Distance methods
*************************************************/

Proc sort data = hat_robust_m; by descending mahal; run;
```

```
Data Mh_distance_-detection_&i;
set Hat_robust_m;
Format DISTRIBUTION_STATUS $15.;
DISTRIBUTION_STATUS = "WITHIN THE NORM";
if PROBCHI((Mahal)**2, &p) > 0.975 then DISTRIBUTION_STATUS = "OUTLIER";
run;

Proc sort data = hat_robust_m out = hat_robust_m_v2;
by descending RD_DIST; run;

Data Rb_distance_detection_&i;
set Hat_robust_m_v2;
Format DISTRIBUTION_STATUS $15.;
DISTRIBUTION_STATUS = "WITHIN THE NORM";
if PROBCHI((RD_DIST)**2, &p) > 0.975 then DISTRIBUTION_STATUS = "OUTLIER";
run;

proc sort data = Rb_distance_detection_&i; by descending RD_DIST; run;

/*********************************************************************************
Steps below identify sets of outliers using the adaptive outlier detection method,
which was proposed by Gervini (2002) and Reinmann, Filzmoser and Garret (2005);
the method uses a critical threshold to separate extreme values of the background
distribution from outlier values coming from contaminating distribution(s).

The identification of the critical threshold requires the computation of an
intermediary indicator (called here p_crit) which is a function of the number of
variables "p" and the number of items analyzed "n". Reinmann, Filzmoser and Garret
developed the formulas linking these variables by means of data simulation
*********************************************************************************/

%global p_crit;

Data Gervini_p_crit;
if &p <= 10 then p_crit = (0.24 - 0.003*&p.)/SQRT(&size.);
else p_crit = (0.252 - 0.0018*&p.)/SQRT(&size.);
run;

Data _NULL_;
set Gervini_p_crit;
if _N_ = 1;
call symput('p_crit', p_crit);
run;

%put &p_crit;

/*************************************************************************
To distinguish potential outliers from extreme values of the data generating
process the threshold value obtained above is compared to another variable
(called here p_n_delta) constructed by comparing the chi-square distribution
with the empirical distribution of the squared Mahalanobis robust distances
*************************************************************************/

Data Rb_distance_detection_&i._v2;
set Rb_distance_detection_&i;
if DISTRIBUTION_STATUS = "OUTLIER" then DISTRIBUTION_STATUS = "EXTREME VALUE";
run;

Proc sort data = Rb_distance_detection_&i._v2;
by RD_DIST; run;
data hat_robust_m_v3;
set Rb_distance_detection_&i._v2;
prb=(_n_ - 1)/&size.;
```

```
chiquant=cinv(prb,&p.);
if PROBCHI((RD_DIST)**2, &p) - prb > 0 then P_N_DELTA = PROBCHI((RD_DIST)**2, &p) -
prb;
else P_N_DELTA = 0;
/*************************************************************************
Note: this computation takes into account only positive differences
because negative differences would not indicate the presence of outliers
*************************************************************************/
run;

data p_n_delta;
set hat_robust_m_v3;
if PROBCHI((RD_DIST)**2, &p) >= 0.975;
run;

%global Supremum_P_N_DELTA;

Proc SQL Noprint;
select MAX(P_N_DELTA) into :Supremum_P_N_DELTA
from p_n_delta;
quit;

%if &Supremum_P_N_DELTA > &p_crit %then %do;

Data Ad_distance_detection_&i;
set Hat_robust_m_v3;
if prb /* empirical distribution */ > (1 - &Supremum_P_N_DELTA) then
DISTRIBUTION_STATUS = "OUTLIER";
run;

%end;

%else %do;

Proc SQL;
create table Ad_distance_detection_&i as
select *
from Rb_distance_detection_&i
order by RD_DIST desc;
run;

%end;

Proc Sort data = Ad_distance_detection_&i; by descending RD_DIST; run;
Proc sort data = Rb_distance_detection_&i; by descending RD_DIST; run;

Proc SQL; Drop table Rb_distance_detection_&i._v2; quit;

%do j = 1 %to 3;
%if &j = 1 %then  %let prefix = Mh;
%if &j = 2 %then  %let prefix = Rb;
%if &j = 3 %then  %let prefix = Ad;

Data &prefix._perf_&i._v0;
set &prefix._distance_detection_&i;
Keep ID DISTRIBUTION_STATUS IND_TRUE_POS IND_FALSE_POS METHOD SAMPLE_NUMBER;
Format IND_TRUE_POS comma8. IND_FALSE_POS comma8. METHOD $2. SAMPLE_NUMBER comma8.;
IND_TRUE_POS = 0; IND_FALSE_POS = 0; METHOD = UPCASE("&prefix"); SAMPLE_NUMBER = &i;
If ID >= &bgrd and DISTRIBUTION_STATUS = "OUTLIER" then IND_TRUE_POS = 1;
If ID < &bgrd and DISTRIBUTION_STATUS = "OUTLIER" then IND_FALSE_POS = 1;
run;
```

```
Proc SQL;
Create table &prefix._perf_&i as
Select SAMPLE_NUMBER, METHOD, SUM(IND_TRUE_POS) as NUMBER_TRUE_POS_&prefix format
comma8.,
SUM(IND_FALSE_POS) as NUMBER_FALSE_POS_&prefix format comma8.,
((Calculated NUMBER_TRUE_POS_&prefix)/&cont) as SHARE_TRUE_POS_&prefix format
PERCENT7.4,
((Calculated NUMBER_FALSE_POS_&prefix)/&cont) as SHARE_FALSE_POS_&prefix format
PERCENT7.4
from &prefix._perf_&i._v0
group by SAMPLE_NUMBER, METHOD;
quit;

Proc SQL;
create table sample_&i as
select a.SAMPLE_NUMBER, a.NUMBER_TRUE_POS_Mh, a.NUMBER_FALSE_POS_Mh,
a.SHARE_TRUE_POS_Mh, a.SHARE_FALSE_POS_Mh,  b.NUMBER_TRUE_POS_Rb,
b.NUMBER_FALSE_POS_Rb, b.SHARE_TRUE_POS_Rb, b.SHARE_FALSE_POS_Rb,
c.NUMBER_TRUE_POS_Ad, c.NUMBER_FALSE_POS_Ad, c.SHARE_TRUE_POS_Ad,
c.SHARE_FALSE_POS_AD
from Mh_perf_&i a, Rb_perf_&i b, Ad_perf_&i c
where a.SAMPLE_NUMBER = b.SAMPLE_NUMBER and b.SAMPLE_NUMBER = c.SAMPLE_NUMBER;
quit;

%end;

%if &i = 1 %then %do;
Data samples_cont_&cont;
set sample_&i;
run;
%end;
%else %do;
Data samples_cont_&cont;
set samples_cont_&cont sample_&i;
run;
%end;

Proc SQL; drop table Sample_&i table Mh_perf_&i._v0
table Rb_perf_&i._v0 table Ad_perf_&i._v0 table Mh_perf_&i table Rb_perf_&i
table Ad_perf_&i; quit;
%end; /* clause ends the sanples loop */

%mend out_meth;

%out_meth(2, 100, 950, 50);

%out_meth(2, 100, 900, 100);

%out_meth(2, 100, 850, 150);
```

REFERENCES

Ben-Gal I. “Outlier Detection,” In: Maimon O. and Rockach L. (Eds.) *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*, Chapter 1, Kluwer Academic Publishers, 2005, ISBN 0-387-24435-2

http://www.eng.tau.ac.il/~bengal/outlier.pdf

Filzmoser P., Garrett R.G. and C. Reimann, “Multivariate outlier detection in exploration geochemistry,” Computers & Geosciences 31 (2005) 579–587

http://www.statistik.tuwien.ac.at/public/filz/papers/ArticleFGR05.pdf

Filzmoser P. “Identification of Multivariate Outliers: A Performance Study,” Austrian Journal of Statistics Volume 34 (2005), Number 2, 127–138

http://stat.ethz.ch/education/semesters/ss2012/ams/paper/outlierDetection.pdf

Geyer, C. J., Breakdown Point Theory Notes, Class Notes on Nonparametric Statistics (2006), University of Minnesota

http://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&cad=rja&uact=8&ved=0CCUQFjAB&url=http%3A%2F%2Fwww.stat.umn.edu%2Fgeyer%2F5601%2Fnotes%2Fbreak.pdf&ei=n6TdU-6xBsylyATE2ILIBw&usg=AFQjCNFtrlrFrBANu4zSJSVW1i34_sP0wA

Gervini D. “A robust and efficient adaptive reweighted estimator of multivariate location and scatter,” Journal of Multivariate Analysis 84 (2003) 116–144

https://pantherfile.uwm.edu/gervini/www/papers/JMVA2080.pdf

Hubert, M. and M. Debruyne, Minimum Covariance Determinant, John Wiley & Sons, Inc. WIREs Computational Statistics 2 (2010), pp. 36–43

http://www.researchgate.net/publication/229588284_Minimum_covariance_determinant

Reimann C., Filzmoser P. and R. G. Garrett, “Background and threshold: critical comparison of methods of determination.” Science of the Total Environment, Vol. 346 (2005), pp. 1-16

http://www.statistik.tuwien.ac.at/public/filz/papers/Stoten05.pdf

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

Rousseeuw P.J.,”Multivariate Estimation With High Breakdown Point,” in *Mathematical Statistics and Applications, Vol B*, eds. W. Grossmann, G. Pflug, I.Vincze and W.Wertz, (1985) Dordrecht: Reidel, pp. 283-297

Rousseeuw PJ, Van Driessen K. “A fast algorithm for the Minimum Covariance Determinant estimator.” Technometrics 41 (1999), pp. 212–223.

The Health Care Fraud and Abuse Control Program Annual Report for Fiscal Year 2013, Department of Health and Human Services and The Department of Justice

https://oig.hhs.gov/publications/docs/hcfac/FY2013-hcfac.pdf

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Paulo Macedo
Enterprise: Integrity Management Services
Address: 5911 Kingstowne Village Parkway
City, State ZIP: 22315
Work Phone: (703) 683-9600 Ext. 408
E-mail: pmacedo@integritym.com