Statistical Outlier Detection in Large Multivariate Datasets

Pradipto Das
Netaji Subhash Engineering College, Computer Applications, Kolkata - 700152.[1]

Dr. Deba Prasad Mandal
Indian Statistical Institute, Machine Intelligence Unit, Kolkata – 700108.

Abstract

This work focuses on detecting outliers within large and very large datasets using a computationally efficient procedure. The algorithm uses Tukey's biweight function applied on the dataset to filter out the effects of extreme values for obtaining appropriate location and scale estimates. Robust Mahalanobis distances for all data points are calculated using these location and scale estimates. A suitable rejection point for the outliers is determined by a separation boundary obtained using non-parametric density estimation by Parzen window where the probability density curve of the robust Mahalanobis distances descends and then again ascends for outlying distances. This procedure demonstrates good success at identifying outliers even in cases where data is highly skewed and overlapping, compared to established statistical outlier detection methods for both univariate and multivariate data where the underlying distribution needs to be known.

Key words: Outlier detection, robust estimation, parzen windows, biweight function, Mahalanobis distances, large datasets.

1 Introduction

Very often, there exist data objects that do not comply with the general behavior or model of the data. Such data objects, which are grossly different or inconsistent from the rest of the data set, are termed outliers. In the univariate case, we perform discordancy tests on an extreme (upper or lower) value of the data set or k extreme values to determine whether the extreme value is truly an outlier in relation to the underlying distribution of the dataset. Discordancy tests for univariate data require test statistics be set up and then its distribution be determined to evaluate the significance probability. However many of these statistics are prone to, to differing extents, *masking*, specially the Dixon statistic – an excess/spread statistic of the form $(x_{(n)} - x_{(n-1)})/(x_{(n)} - x_{(2)})$ for examining upper outlier $x_{(n)}$ avoiding $x_{(1)}$. Masking occurs when a group of outlying points skews the mean and covariance estimates towards it. Swamping occurs when a group of outlying points skews the mean and covariance estimates towards it and away from other inlying points, and the resulting distance from inlying points to the mean is large. In many of statistical outlier identification methods for univariate data that involve test statistics[1], in addition to assuming a distribution function, the number of outliers that can be tested needs to be fixed.

Identifying outliers in multivariate data pose challenges that univariate data do not. A multivariate outlier need not be an extreme in any of its components. The idea of extremeness arises inevitably from some form of 'ordering' of the data. Barnett categorizes sub-ordering principles in four types: marginal, reduced (or aggregate), partial and conditional. For multivariate outlier study, reduced sub-ordering is almost the only principle that has been employed. With reduced sub-ordering we transform any multivariate observation x, of dimension p, to a scalar quantity R(x). That observation x_i which yields the maximum value $R_{(n)}$ will be adjudged discordant if $R_{(n)}$ is unreasonably large in relation to the distribution of $R_{(n)}$ under the basic model F. There are two problems with this form of approach:

§ We may loose useful information on multivariate structure by employing reduced (or any other form of) sub-ordering. The special cases where R(x) singles out some marginal component of x, or

[1] *Research carried out as Final Semester Project work at ISI for Netaji Subhash Engineering College.*

identifies, say, the first principal component of the multivariate sample, intuitively demonstrate this risk.

§ Even if we have chosen R(x), the distributional form of $R_{(n)}$ under F may not be easy to determine or be employed as a basis for employing a test of discordancy.

Let us consider the case where we choose to represent a multivariate observation x, by means of a **distance measure**, $R(x; x_0, \Gamma) = (x - x_0)'\Gamma^{-1} (x - x_0)$, where x_0 reflects the location of a data set or underlying distribution and Γ^{-1} applies a differential weighting to the components of the multivariate observation inversely related to their scatter or to the population variability. For e.g. $\mathbf{x_0}$ might be the zero vector **0**, or the true mean m or the sample mean `**x**, and Γ might be the variance-covariance matrix **V** or its sample equivalent **S**, depending on the state of our knowledge about m and **V**.

Intuitively, we label a point an outlier because it is sufficiently “far away” from the majority of the data. An important tool for quantifying “far away”, is the Mahalanobis distance, defined as $MD^2 =$ R(x;`**x**, Γ) for each point $x_i$, where`**x** is the sample mean of the data set, and Γ is the sample covariance matrix. Clearly, the Mahalanobis distance relies on classical location and scale estimators. As such, it is subjected to the masking effect, and is not suitable for general use in contaminated data.

The conventional location (sample mean) and scale (sample variance) estimates are not robust to outliers and thus render the use of MD^2 useless. An estimator should possess several qualities. It should have a high breakdown point, but this could be outweighed by other criteria. Intuitively, the breakdown point of an estimator is the least amount of contamination in the data that could change the estimator so that it fails to meaningfully describe the data set. For realistic applications, a breakdown point of 20% is usually satisfactory.

A large attention has focused on use of Huber’s M-estimators for estimating **V**(often simultaneously with m, since we are unlikely to know the location parameter of the distribution.) Maronna (1976) specifically examines robust M-estimators for m and **V**, by concentrating on affine-invariance and a basic model F that is assumed to have an elliptically symmetric density. Huber (1981) identifies different possible interests in robust estimations of **V** and the correlation matrix **P** and estimation of the shape matrix for some relevant elliptical distribution with density of the form $f(x) = |\mathbf{V}|.h((x\text{-}m)\mathbf{V}^{-1}(x\text{-}m))$ where h(y) is a spherically symmetric density in p-dimensional space. A tangible form for outlier-robust M-estimators, relevant to an assumed elliptically basic model and an associated normal contamination distribution, is exhibited by Maronna and Campbell (1980) [1]. The estimators m and **V** are obtained as iteratively derived simultaneous solutions to the equations:

$m = \Sigma_{i=1}{}^{n}\, w_i x_i \,/\, \Sigma_{i=1}{}^{n}\, w_i$ and $\mathbf{V} = \Sigma_{i=1}{}^{n}\, w_i^2 (x_i - m)'(x_i - m)/(\Sigma_{i=1}{}^{n}\, w_i^2 - 1)$

$w_i = w(R_i)/R_i$ and R_i is the sample value of the reduced measure relevant to the normal distribution: that is

$R_i = (x_i\text{-}m)\mathbf{V}^{-1}(x_i\text{-}m)$

The Minimum Covariance Determinant (MCD) location and shape estimates are resistant to outliers. However, finding the exact MCD sample can be difficult and time consuming. The only known method for finding the exact MCD is to search every half sample and calculate the determinant of the covariance matrix of the sample. Finding outlying distances based on an assumed distribution of the MD^2 may not yield good results. The distance distribution is based on the knowledge of the data distribution and its parameters. A new method for detecting outliers in a multivariate normal sample had been derived by Hardin and Rocke [2] which are superior to the commonly used Chi-Square cutoff.

We are concerned here with the fact that the data may follow any distribution not necessarily normal and may contain samples obtained from a mixture of distributions. Gnanadesikan (1977) presents a comprehensive review of the robust estimation of m and **V** from robust estimators of the *individual elements* of m and V as well as for directly ‘multivariate’ estimators of m and **V.** Mosteller and Tukey (1977) propose a robust estimator of **V** based on robust regression estimation: regressing x_j on $x_1, x_2, ..., x_{j-1}$ j = 2, 3, ..., p (where x_j is the jth component of the typical observation vector x).[1] To this end, we use Tukey’s biweight function iteratively on the individual components of the features or components of the data for obtaining a robust estimate of location.

2 Background of the method

The new method focuses on the development of an outlier detection method suitable for large data sets. Since the basic assumption was that the underlying distribution of the data remains unknown, we chose to focus on distance-based methods. A key ingredient of a successful distance-based method is a robust estimate of the covariance matrix.

Despite being more sensitive to outliers (i.e. more difficult to estimate robustly) than the location estimate, the Mahalanobis distance is also extremely sensitive to an incorrect scale estimate and could be rendered incapable of separating the outliers and inliers. We therefore devoted considerable time to obtaining a fast, robust estimate of the covariance matrix.

The first thing to be done was to devise some mechanism to assign weights to each of the observations. The weights were to have the property that potential outliers would receive very low weights and thus would not influence the scale estimate much. Thus in accordance with conventional statistical outlier detection methods we would like to assign a zero weight to the extremes in the sample and then test for discordancy of the extremes relative to the estimated distribution of the data set.

M-estimators minimize functions of the deviations of the observations that are more general than the sum of squared deviations or the sum of absolute deviations. A suitably chosen M-estimator will have good robustness of efficiency in large samples. The bi-weight, or bisquare, estimator of location is the solution T_n of $\Sigma_{i=1}{}^{n}\ \Psi(u_i) = 0$, where

$$\Psi(u) = \begin{cases} u(1 - u^2) & |u| \leq 1 \\ 0 & |u| > 1 \end{cases} \qquad \text{... 2.1}$$

and,

$$u = (x_i - T_n)/(cS_n), \qquad \text{... 2.2}$$

where,

Ψ is the derivative of the objective function $\rho(x, t)$, which is minimized for estimation of T_n.

Outliers do not affect the biweight because $\Psi(u) = 0$ if $|u|$ is sufficiently large. Moreover, the estimator is not as sensitive to small changes in data values as is the median. Since $\Psi(u) \approx u$ for small u, near the center of the sample the biweight behaves like the mean. An estimator is *resistant* if it is affected only to a limited extent either by a small number of gross errors or by any number of small rounding and grouping errors. An estimator has *robustness of efficiency* over a range of distributions if its variance (or, for biased estimators, its mean squared error) is close to the minimum for each distribution.

The gross-error sensitivity (g.e.s) expresses, asymptotically, the maximum effect a contaminated observation can have on the estimator $T_{n.}$ It is the maximum absolute value of the Influence Curve (IC). The IC for the biweight when T = 0, shows that g.e.s of the biweight is finite. An estimator has *robustness of efficiency* and is *resistant* only if the g.e.s is bounded.[3]

The Ψ-function of the biweight is 0 for |u| larger than 1. Thus IC(x) = 0 for |x| > cS. We say the rejection point is r = cS. Observations beyond the rejection point do not contribute to the value of the estimate except possibly through the auxiliary scale estimate. Estimators whose Ψ-function have finite rejection point (e.g, the biweight) are particularly well protected against sufficiently large outliers.[3]

The biweight can be computed as an iteratively reweighted sample mean, where the weights w_i are defined in terms of the estimate obtained in the previous iteration,

T_{k-1}:

$T_k = \Sigma_{i=1}{}^{n}\ w_i x_i / \Sigma_{i=1}{}^{n}\ w_i$ 2.3

$w_i = w(u_i)$, $u_i = (x_i - T_{k-1})/(cS_{k-1})$, ...2.4

$w(u) = (1 - u^2)^2 I_{[-1,1]}(u)$, I is uniform on [-1,1], ...2.5

where *c* is a "tuning constant" (usually in the range 4–6) and *S* is an estimate of scale (usually made to be unbiased if the x_i 's came from a Gaussian population). Beaton and Tukey (1974, pages 151–152) offered a rationale for the name *biweight*: "the 'bi-' referring to the outer exponent (whose value ensures continuity for both *w(u)* and *w'(u)*)." The biweight has been shown to be highly efficient in many diverse contexts, including robust analysis of variance, time series and even in control charts used to monitor product quality[4].

3 The proposed method

The procedure starts by fitting Tukey's biweight function to each component (feature) separately, keeping record of the weights assigned to each data value. Let there be *n* observations and *p* components.
Analyzing each component separately, weights were assigned to observations by the biweight function. Considering the *Median Absolute Deviation (MAD)* to be more robust than the conventional sample mean, we denote the univariate observations be $x_1, \ldots, x_n$ with corresponding weights $w_1, \ldots, w_n$. Let $\bar{x}_{bw}$ the location estimate, which we initially take to be the *median*, S a scale estimate based on the median absolute deviation from $\bar{x}_{bw}$,

$$S = \text{median}\{| x_i - \bar{x}_{bw} |\}; \qquad \ldots.2.6$$

and c a tuning constant. The weights w_i are then iteratively calculated (until there is no difference in weight values) according to:

$$w_i = \begin{cases} (1 - ((x_i - \bar{x}_{bw})/cS)^2)^2, \text{ when } ((x_i - \bar{x}_{bw})/cS)^2 < 1 \\ 0, \text{ otherwise.} \end{cases} \qquad \ldots.2.7$$

$$\bar{x}_{bw} = \Sigma^n w_i x_i / \Sigma^n w_i,\ i = 1, 2, \ldots, n \qquad \ldots.2.8$$

and $S = \text{median}\{| x_i - \bar{x}_{bw} |\};$ 2.9

Mosteller and Tukey recommended c = 6 or c = 9; The biweight has a higher efficiency at c = 9 because less data is rejected, but better (i.e., lower) gross error sensitivity at c = 6.[3] Kafadar also conducted a study investigating the biweight's efficiency for different values of c, sample sizes and distributions and found that its efficiency increases with both sample size and tuning constant c at the Gaussian.

After fitting the biweight to each of the p components, we used the weights thus obtained to calculate a weighted mean and covariance matrix. For the covariance, we used the product of the weights for the different coordinates as follows:

$$m = \Sigma_{i=1}{}^n w_i x_i / \Sigma_{i=1}{}^n w_i , \qquad \ldots.2.10$$

$$C = (\Sigma_{i=1}{}^n w_i x_i' x_i)/(\Sigma_{i=1}{}^n w_i) - m'm , \qquad \ldots.2.11$$

where the subscript indicates the observation number (1, …, n), m is an [1 x p] matrix and C is a [p x p] matrix.

Investigation into the values of the estimated parameters revealed that most of the inliers (data points that are not outliers) received weights in the range [0.89, 0.99], so that their full contributions to the sample covariance were not counted. This was noteworthy because points near the tail of the primary distribution received reduced weights, yet their full weights were necessary for an accurate estimate. From equations 2.1 – 2.2 we find that, 0<|u|<1 and so is w(u). We needed a mapping from a continuous w(u) to a 0-1 step function that would give a 0 weight to the outliers so that their influence while calculating the location and scale estimates was not present. Empirical studies show that better outlier detection rate occurs for cut off values between 0.75 and 0.90. For each sample x_i, we assign weights w_{ij} for all of its j components and select the minimum of them as a measure of the weight for the sample x_i. In practice, this can be done by calculating the weights of the first components of the entire sample and then moving onto the next component. This way we can reduce the space complexity for the weights to O(n).

At this point, it is ambiguous as to what the Robust Mahalanobis Distances are that determines whether a data point is an outlier. If the data are normally distributed, the distribution of the Mahalanobis Distances will follow a χ^2 distribution. More studies in this regard have been presented by Hardin and Rocke[2] . Our assumption is that the data may follow any distribution not necessarily normal. We therefore attempt to estimate non-parametrically, the density curve of the Mahalanobis Distances using parzen windows and kernel estimators. The primary distribution of the data is described by a central peak, but we look for the rejection boundary where the density curve flattens and rises again to form small secondary peaks after the central peak. Multiple peaks of significant size could mean dense clusters in the data.

To calculate the sample density, we make use of the kernel density estimate of Silverman (1986) [5]. For robust Mahalanobis distances *RMDi*, the sample density at point *d*, $f_n(d)$ is given by

$$f_n(d) = (1/nh)\, \Sigma_{i=1}^{\;n} K((d - RMDi)/h) \qquad ....2.12$$

where *K* is the kernel function and *h* the window width. *K* is commonly taken to be the normal density, which returns to zero fairly quickly and limits the effect of distant observations. To estimate the density at d, form a sequence of regions $R_1, R_2, ..$,containing d – such that, the first region is to be used with one sample, the second with two, and so on. We assume R_n is a p-dimensional hypercube and h_n be the length of an edge of the hypercube. Let V_n be the volume of $R_n = h_n^{\,d}$. k_n be the number of samples falling in R_n, and $f_n(\mathbf{d})$ be the n^{th} estimate of $f(\mathbf{d})$**:** then the probability of falling in the window is $p_n = k_n/n$. and the density estimated is $f_n(d) = (k_n/n)/V_n$
Goal: to design a sequence of windows $V_1, V_2, \ldots, V_n$ at point d, so that, as $n \to \propto$, $f_n(d) \to f(d)$, where f(d) is the true density.

Conditions for the window design:

§ $\mathrm{Lim}_{n \to \propto} V_n = 0$, Increasing spatial resolution.

§ $\mathrm{Lim}_{n \to \propto} k_n = \propto$, Assuming $p(x) \neq 0$ and large samples at each point.

§ $\mathrm{Lim}_{n \to \propto} k_n /n = 0$, k_n grows in an order smaller than n.

Duda and Hart[7] mention the use of Parzen window for non-parametric density estimate that we use here.
Parzen window considers V_n as a function of n, e.g. $V_n = 1/\sqrt{n}$. Here focus is to shrink the size of V_n
It also considers a window function φ, such that, $\varphi(u) >= 0$ with $\int_\Omega \varphi(u)du = 1$, where Ω is the domain of u - for example: A Gaussian function $\varphi(u) = (1/(2\pi)^d)e-^{(u2)/2}$.

For the Parzen window, we can estimate the density in the following way: $\varphi((u\text{-}d_i)/h_n) = 1$ if d_i falls within the hypercube of volume V_n centered at d and is 0 otherwise. The number of samples falling in the hypercube is given by

$$k_n = \Sigma_{i=1}^{\;n}\, \varphi((d\text{-}d_i)/h_n) \qquad ....2.13$$

$$f_n(d) = 1/n\Sigma_{i=1}^{\;n}\, \varphi((d\text{-}d_i)/h_n)/V_n = (1/nh_n^{\,p})\Sigma_{i=1}^{\;n}\, \varphi((d\text{-}d_i)/h_n) \qquad ....2.14$$

Substituting p = 1 and $\varphi = K$ and letting $h_n = h$ (a constant), gives us equation 2.12.

Silverman (1978)[6] discusses the amount of smoothing necessary when using the kernel method to estimate a probability density from independent identically distributed observations, by using the test graph method. For data containing outliers, Silverman (1986) recommends $h = 0.9An^{-1/5}$, where *A* = min{standard deviation, interquartile range/1.34}. For computational purposes, Silverman (1982) presents a fast method of calculating $f_n(d)$ based on the Fast Fourier Transform. We have not implemented the FFT procedure but literature reveals significant savings in computational time using the FFT procedure from O(nm) to O(nlogn) where m denotes the number of divisions into which the maximum Mahalanobis Distance is divided which we assigned to be 100. Having chosen K to be the Gaussian kernel and h as discussed above, we next focus our attention on how to find out a suitable rejection boundary.

It would not in general be a good idea to select the rejection point as the value at which the central peak returns to zero because it could taper down very gradually and return zero long after the main peak has ended. A better option would be to choose the rejection point as that value where the *slope* is sufficiently close to zero. We estimate the slope by first differences; In practice, the rejection line should be taken as the first data value where the slope enters a confidence interval containing zero (after the central peak has ended). Precise computation of an appropriate confidence interval requires knowledge of the distribution of the Mahalanobis distances, which is known only when employing the sample mean and covariance from normally distributed data.

We used a simple technique to obtain the rejection point. We find the index of the slope of the estimated density curve where the slope enters a confidence interval [–0.0001, 0.0001] after attaining the central peak i.e., the slope has a large negative value and increases towards zero. Find the value of Mahalanobis distance where this negative slope first enters a confidence interval containing zero. A good value for the half-length of this interval is 0.0001. Let the value of Mahalanobis distance where the slope is sufficiently close to zero (and also where the density is less than 0.3 of its maximum value), be called the rejection point. Finally, we classify all points as outliers that have robust Mahalanobis distance greater than this rejection point.

4 Algorithmic steps of the method

We summarize the preceding discussion by presenting this algorithm in steps (the order of computation follows each step in parentheses):
(1) Calculate the biweight weights for each data point using Tukey's biweight function, fitting each component separately. O(npt) [t is the number of iterations required for convergence of the weights for each individual component of all data points. Typically t << n.]

(2) If an observation's weight is less than the cutoff value, reassign that weight to zero, otherwise let it be one. We select the minimum weight across each observation's components as the weight for the other *p • 1* components too. O(np)

(3) Compute the weighted mean and covariance of all the observations, which amounts to the sample mean and covariance of those observations with weight equal to one. O(np + np^2)

(4) Calculate a robust Mahalanobis distance for all n observations, using the mean and covariance matrix computed in the previous step. O(p^3 + n)

(5) Calculate the density of these robust Mahalanobis distances, using Silverman's kernel density method using Parzen Windows and approximate the slope O(nm), where m denotes the number of divisions into which the maximum Mahalanobis Distance is divided. We assigned m to be 100. If we use Silverman's FFT method, the order of computation becomes O(nlog(n)).

(6) Determine the rejection point as that value of Mahalanobis distance where the slope following the main density peak first enters a confidence interval containing zero. O(n)

(7) All points are considered outliers if their robust Mahalanobis distances are larger than this rejection point. O(n)

5 Results on Simulated Data

We ran 100's of simulation on overlapped correlated data coming from a 10000 x 10 randomly chosen data belonging to Poisson distributions as follows:
The first component of the first 8000 observations was randomly generated from a Poisson distribution with λ = 12. The first component of the next 2000 observations was randomly generated from a Poisson

distribution with $\lambda = 42$. Components 2-10 of the rest of the data were randomly generated from Poisson distribution with $\lambda = 12$. Finally, to all data was added a sample of size 10000 x 1 generated from a random Poisson distribution with $\lambda = 12$.

The robust Mahalanobis distances and the estimated density are shown below with bi-weight cut-off value of 0.97.

The percentage error is calculated to be:

$$\frac{100 \text{ x } (|\#\text{true outliers - }\#\text{observed outliers}|)}{\#\text{true outliers}}$$

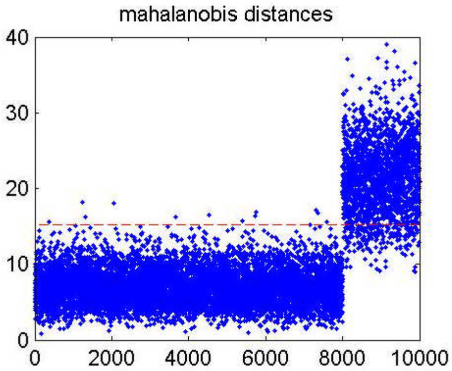


Fig 5.1

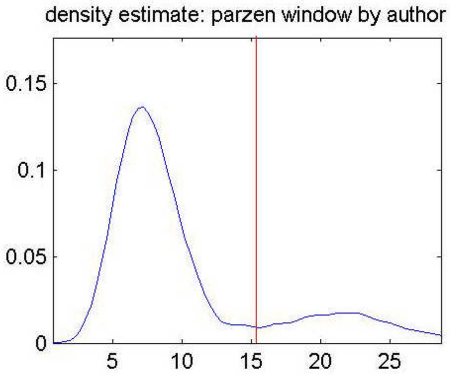
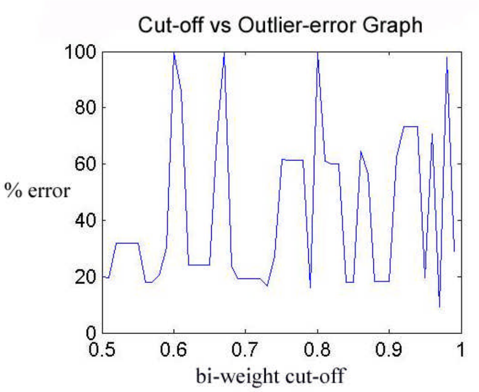


Fig 5.2

Averaging the results of all simulations of datasets generated by this process for cutoff values 0.85, 0.90 and 0.97, we find the outlier error rates to be 17.85%, 18.2% and 9.25% respectively. Normally, the higher cutoff value should result in lower outlier error rates. The graph to the right shows the outlier detection errors when the bi-weight cutoff value is varied from 0.5 to 0.99 in steps of 0.01. For this data set, a cut-off value of 0.97 yields the lowest outlier-detection error rates. Repeated trials have shown that for data coming from symmetrical distributions a cut-off value of 0.75 to 0.85 suffices to guarantee near perfect outlier detection.



Although for these highly overlapped data sets the algorithm looks less promising, however, it will be hard to detect even this amount of outliers efficiently using conventional statistical methods. Better success rates were obtained for synthetic data generated from symmetric distributions contaminated with data generated from the same distribution but with parameters "slipped" for e.g., data from N(μ, V) contaminated with data from N(μ+**a**, V). With skewed and uncorrelated data belonging to asymmetric distributions where there is a clear separation between outliers and inliers, the algorithm detects outliers with almost 0% outlier error rates.

6 Conclusions and Future Directions

We consider this algorithm to be a success against the established statistical outlier detection methods where the underlying distribution of the data needs to be known. We believe that further necessary optimizations to this algorithm are necessary to accurately determine outliers. Although no algorithm is a solution for all data set scenarios for outlier detection, this algorithm dominated by the term max{O(np), O(nlogn)} is a sufficient improvement over combinatorial algorithms like the Minimum Covariance Determinant (MCD) or over methods that assume normally generated data where the Mahalanobis distances follow χ^2, F or Beta distributions. However, in cases where p is very large this method proves to be computationally expensive.

An interesting extension to this work leads to the question whether simultaneous outlier detection and a "good" clustering of the data is possible? In this case, only two possible clusters are generated – one is the outlier cluster and the other is the inlier cluster. One may say that several peaks of the density curve formed before the density curve lies in the confidence interval corresponding to the rejection point can represent distinct density clusters. For an illustration, we ran the above algorithm on an 800 x 2 dataset generated randomly from a uniform distribution with varying parameters to obtain 5 major classes of data and few outlying data. However, as in the figure, the 2 major peaks in the density curve do not represent the natural clustering of the data set. In addition, this algorithm does not model the case where data contain imprecise information, as does the popular fuzzy c-means clustering. The results of clustering and outlier detection (based on distance from cluster centers) by a variant of the popular fuzzy c-means algorithm in which the number of clusters is unknown (I am currently pursuing research in this direction where outliers are to be identified amongst data which represent uncertainty) is given in Fig 6.4 below.

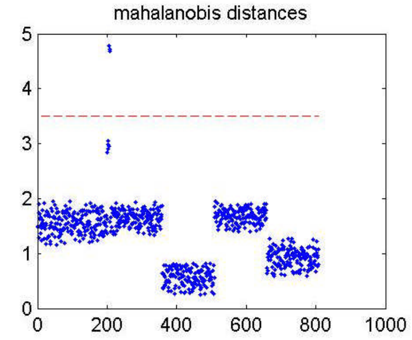


Fig 6.1

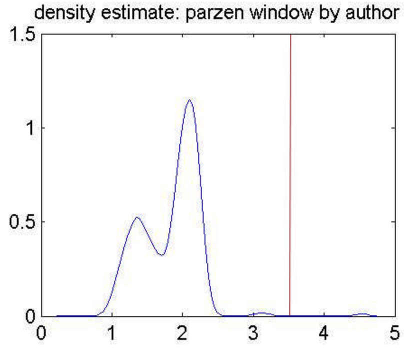


Fig 6.2

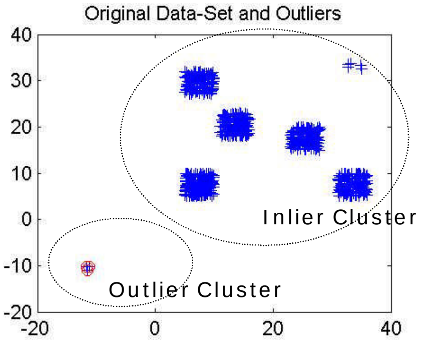


Fig 6.3

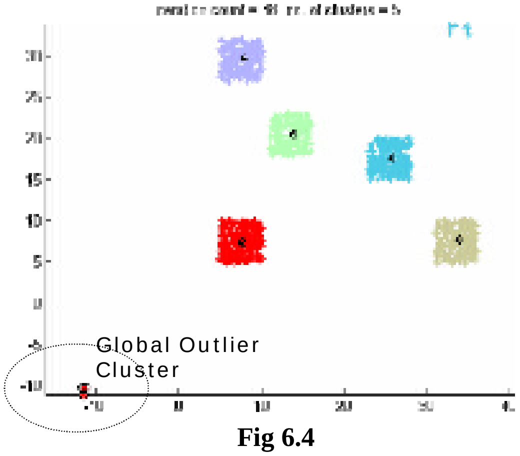


Fig 6.4

In this case, both algorithms detected the same outliers but in general, the above statistical algorithm is computationally far more efficient since it requires very few iterations (2 to 4) during the fitting of the bi-square weights and that is the only iteration this algorithm ever uses. All clustering algorithms using the partitioning method use large number of iterations for the convergence of the solution that leads to the minimization of the respective objective functions and thus sacrificing efficiency. Moreover, a perfect or even "good" clustering is very difficult to achieve even for moderately sized data sets with overlapping and highly skewed data so that the method identifying outliers based on the distance from cluster centers becomes very error prone.

We considered another data set that poses very difficult challenges for all clustering techniques using partitioning by minimizing squared sum of errors. 1500 samples with two features were generated from a uniform distribution with origin (0, 0) and radius following a uniform distribution $I_{[0,3]}$. 15 outliers were added to the data set as is depicted in Fig 6.8.

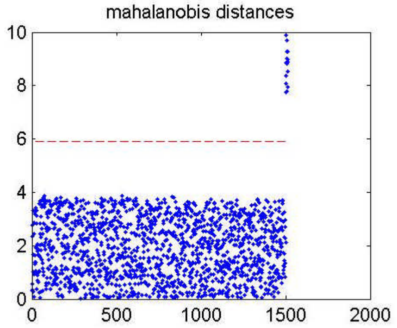


Fig 6.5

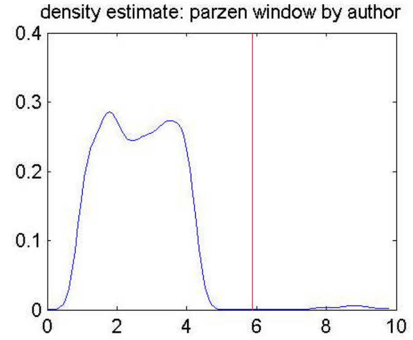


Fig 6.6

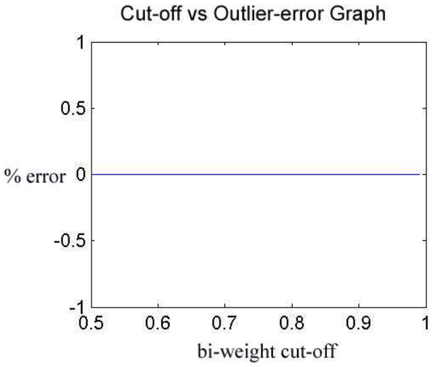


Fig 6.7

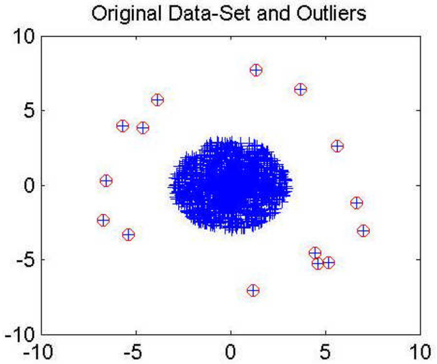


Fig 6.8

On this data set, the accuracy of this algorithm becomes markedly pronounced as we see in the figure below. For all bi-weight cut-off values in the range 0.5(0.01)0.99 the outliers were correctly identified. As far as outlier identification is concerned using statistical procedures this method far overwhelms any previous rigorous statistical outlier detection tests mostly in its inherent assumption of unknown primary distribution of the data set.

7 References

[1] Outliers in Statistical Data, 3rd Edition - Vic Barnett, Toby Lewis, ISBN: 0-471-93094-6 Hardcover 604 pages March 1994.

[2] Hardin, J., Rocke, David M. (1999), The Distribution of Robust Distances.

[3] Hoaglin, David C., Mosteller, Fredrick., Tukey, John W., Understanding Robust and Exploratory Data Analysis, ISBN: 0-471-38491-7, Wiley Interscience.

[4] Kafadar, Karen., John Tukey and Robustness Statistical Science (2003), Vol.18, issue No3, Pg 319–331.

[5] Silverman, B. W., Density Estimation for Statistics and Data Analysis.

[6] Silverman, B. W., Choosing the Window Width when Estimating a Density, Biometrika, Vol. 65, No. 1 (Apr. 1978), 1-11.

[7] Pattern Classification, 2nd Edition – Richard O. Duda, Peter E. Hart, David G. Stork ISBN: 9814-12-602-0, John Wiley & Sons, Inc.