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 method focuses on detecting outliers within large and very large datasets using a computationally efficient procedure. Tukey’s biweight function is applied on the dataset for obtaining robust location and scale estimates of the data by filtering out the effects of extreme values. Robust Mahalanobis distances for all data points are calculated using these location and scale estimates. Next density estimation by Parzen window is utilized for computing the probability density curve of the robust Mahalanobis distances. Outliers are identified to be those points whose robust Mahalanobis distances have very low probability density.

Keywords: Outlier detection, robust estimation, parzen windows, Mahalanobis distances, large datasets, Tukey’s biweight.

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\(^1\) categorizes sub-ordering principles in four types: marginal, reduced (or aggregate), partial and conditional. For multivariate outlier study, reduced sub-ordering is more commonly 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 lose 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 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 $x_i$ by means of a distance measure, $R(x_i; \mu_0, \Sigma) = (x - x_0)' \Gamma^{-1} (x - x_0)$. $x_0$ reflects the location of a data set or underlying distribution and $\Gamma^{-1}$ applies a differential weighting to the components of the multivariate observation inversely related to their scatter. For e.g. $x_0$ might be the zero vector, or the true mean $\mu$ or the sample mean $\bar{x}$, and $\Gamma$ might be the variance-covariance matrix $V$ or its sample equivalent $S$, depending on the state of our knowledge about $\mu$ 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 $MD(x) = (x - \mu)' \Gamma^{-1} (x - \mu)$, where $\mu$ is the location estimate of the data set, and $\Gamma$ 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 $\mu$, since we are unlikely to know the location parameter of the distribution.) Maronna (1976) specifically examines robust M-estimators for $\mu$ 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 $\Psi$ and estimation of the shape matrix for some relevant elliptical distribution with density of the form $f(x) = |V| h((x-\mu)')V^{-1}(x-\mu)$ 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 $\mu$ and $V$ are obtained as iteratively derived simultaneous solutions to the equations:

$$\mu = \sum_{i=1}^{n} w_i x_i / \sum_{i=1}^{n} w_i \quad \text{... 1.1}$$

$$V = \sum_{i=1}^{n} w_i (x_i - \mu)'(x_i - \mu) / (\sum_{i=1}^{n} w_i - 1) \quad \text{... 1.2}$$

$w_i = w(R_i)/R_i$ and $R_i$ is the sample value of the reduced measure relevant to the normal distribution: that is $R_i = (x_i - \mu)'V^{-1}(x_i - \mu)$.

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 MCD 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) [1] presents a comprehensive review of the robust estimation of $\mu$ and $V$ from robust estimators of the individual elements of $\mu$ and $V$ as well as for directly ‘multivariate’ estimators of $\mu$ and $V$. Mosteller and Tukey (1977) propose a robust estimator of $V$ based on robust regression estimation: regressing $x_j$ on $x_1, x_2, \ldots, x_{j-1}$ $j = 2, 3, \ldots, p$ (where $x_j$ is the $j$th component of the typical observation vector $x$) [1].

We next present some preliminaries related to the proposed approach where we have used Tukey’s biweight function iteratively on the individual features or components of the data for obtaining a robust estimate of location.

2. Preliminaries

The new method focuses on the development of an outlier detection method suitable for large data sets. Since the basic assumption is that the underlying distribution of the data remains unknown, we chose to focus on distance-based methods guided by the reduced sub-ordering principle as mentioned in the previous section.

Obtaining a robust estimate of the covariance matrix is the first step towards successful distance-based methods. 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.

A method was required to assign weights to each of the observations such 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\(^3\) minimize functions of the deviations of the observations that are more general than the sum of squared deviations or the sum of absolute deviations. 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. A suitably chosen M-estimator will have good robustness of efficiency in large samples.

The bi-weigh\(^3\), or bisquare, is one such estimator of location, where the M-estimate \(T_n(x_1, x_2, \ldots, x_n)\) is the value of \(t\) that satisfies \(\sum_{i=1}^n \Psi(u_i) = 0\), where

\[
\Psi(u) = \begin{cases} 
  u(1 - u^2) & |u| \leq 1 \\
  0 & |u| > 1 
\end{cases}
\]

and,

\[
u = (x_i - t)/(cS_n),
\]

\(\Psi\) is the derivative of the objective function \(p(x, t)\), which is minimized for estimation of \(T_n\). \(S\) is the estimate of scale and \(c\) is a tuning constant. 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.

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, \(T_{k-1}\):

\[
T_k = \frac{\Sigma_{i=1}^n w_i x_i}{\Sigma_{i=1}^n w_i}, \quad w_i = w(u_i), \quad u_i = (x_i - T_{k-1})/(cS_{k-1}),
\]

\[
w(u) = (1 - u^2)^2, \quad |u| \leq 1
\]

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). In the next section, we introduce our proposed outlier detection algorithm, which is based on the application of Tukey’s bi-weight function.

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 \(x_{bw}\) be the location estimate, which we initially take to be the median, \(S\) a scale estimate based on the median absolute deviation from \(x_{bw}\),

\[
S = \text{median}(|x_i - x_{bw}|),
\]

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

\[
w_i = \begin{cases} 
  (1 - ((x_i - x_{bw})/cS)^2)^2, & \text{when } ((x_i - x_{bw})/cS)^2 < 1 \\
  0, & \text{otherwise.}
\end{cases}
\]

\[
x_{bw} = \Sigma^p w_i x_i / \Sigma^p w_i, \quad i = 1, 2, \ldots, n
\]

and \(S = \text{median}(|x_i - x_{bw}|)\).

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 \(c = 6\).\(^3\)

We used the weights of the observations to calculate the weighted mean and the weighted covariance matrix as,

\[
m = \Sigma_{i=1}^n w_i x_i / \Sigma_{i=1}^n w_i, \quad C = (\Sigma_{i=1}^n w_i x_i x_j^T / \Sigma_{i=1}^n w_i) - m'm,
\]

where, the subscript indicates the observation number (1, \ldots, \(n\)), \(m\) is an \([1 \times p]\) matrix and \(C\) is a \([p \times p]\) matrix. For each sample \(x_{ij}\), 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 \( x_i \).

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. It is to be noted that points near the tail of the primary distribution received reduced weights, yet their full weights were necessary for an accurate estimate.

A mapping from a continuous \( w(u) \) to a 0-1 step function was needed that would give a 0 weight to the outliers so that their influence while calculating the location and scale estimates was eliminated. Empirical studies reveal that a cut off value between 0.75 and 0.90 yields good outlier identification accuracy.

At this point, it is ambiguous as to what the distribution of the robust Mahalanobis distances is that determines whether a data point is an outlier. If the data are normally distributed, the distribution of the robust Mahalanobis distances is non-parametrically, using Parzen windows. A central peak describes the primary distribution of the data, but we look for the rejection boundary where the density curve flattens and rises again to form small secondary peaks after the central peak.

To calculate the sample density, we make use of the kernel density estimate of Silverman (1986)[4]. For each of the \( n \) robust Mahalanobis distances \( RMD_i \), the sample density at a point signifying distance value \( d \) \( f_n(d) \)) is given by

\[
f_n(d) = \frac{1}{nh} \sum_{i=1}^{n} K((d - RMD_i)/h) \quad \ldots \ldots 2.12
\]

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 \( d \), we form a sequence of regions \( R_1, R_2, \ldots, R_n \), containing \( d \) such that, the first region is to be used with one sample point, the second with two, and so on. We assume \( R_n \) is a p-dimensional hypercube and \( h_n \), the length of an edge of the hypercube. If \( V_n \) is the volume of \( R_n \), then \( V_n = h_n^n \). If \( k_n \) be the number of samples falling in \( R_n \), and \( f_n(d) \) be the n\textsuperscript{th} estimate of \( f(d) \), then the probability mass of a sample falling in the window is \( p_n = k_n/n \), and the density estimated is \( f_n(d) = (k_n/n)/V_n \) (mass/volume).

Our goal is to design a sequence of windows \( V_1, V_2, \ldots, V_n \) at point \( d \), so that, as \( n \to \infty \), \( f_n(d) \to f(d) \), where \( f(d) \) is the true density at \( d \). It is to be noted that certain conditions for the window design must be satisfied viz.:

- \( \lim_{n \to \infty} V_n = 0 \), i.e., Increasing spatial resolution.
- \( \lim_{n \to \infty} k_n = \infty \), Assuming \( p(x) \neq 0 \) and large samples at each point.
- \( \lim_{n \to \infty} k_n/n = 0 \), i.e., \( k_n \) grows in an order smaller than \( n \).

Parzen windows as explained in Duda and Hart[6] consider a volume \( V_n \) as a function of \( n \), e.g. \( V_n = 1/n \), where the focus is to shrink the size of \( V_n \) gradually. It also considers a window function \( \phi \), such that, \( \phi(u) \geq 0 \) with \( \int_{\Omega} \phi(u) du = 1 \), where \( \Omega \) is the domain of \( u \).

Using Parzen windows, we can estimate the density in the following way: \( \phi((d - RMD)/h_n) = 1 \) if \( RMD \) 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

\[
k_n = \sum_{i=1}^{n} \phi((d - RMD)/h_n) \quad \ldots \ldots 2.13
\]

\[
f_n(d) = \frac{1}{nh} \sum_{i=1}^{n} \phi((d - RMD)/h_n)/V_n
\]

\[
= \frac{1}{nh} \cdot \frac{1}{\sum_{i=1}^{n} \phi((d - RMD)/h_n)} \quad \ldots \ldots 2.14
\]

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

Silverman (1978)[5] 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)[4] recommends \( h = 0.9A^{-1/5} \), where \( A = \min \{ \text{standard deviation}, \text{interquartile range}/1.34 \} \).

We choose the rejection point as that value where the slope is sufficiently close to zero after the central peak has ended. We estimate the slope by first differences; In practice, the rejection point should be taken as the first value of the robust Mahalanobis distance corresponding to which the density slope enters a confidence interval containing zero i.e., after attaining the central peak, the slope has a large negative value and increases towards zero. Let the value of Mahalanobis...
distance where the slope is sufficiently close to zero (and also where the density is less than 0.25 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.

It is to be noted that, 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.

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 Tukey’s biweight for each data point, fitting each component separately. O(npt) [t is the number of iterations required for convergence of the weights. 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(n)

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(np^2 + p^3)

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 (1982)\(^7\) FFT method, the order of computation becomes O(nlog(n)).

6. As a test of discordancy, 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 hundreds of simulation on an overlapped correlated data with 10000 sample points each having 10 features generated randomly from a mixture of Poisson distributions as follows:

The last component of the first 8000 observations was randomly generated from a Poisson distribution with \( \lambda = 12 \). The last component of the next 2000 observations was randomly generated from a Poisson distribution with \( \lambda = 42 \). Components 1-9 of the rest of the data were randomly generated from Poisson distribution with \( \lambda = 12 \). Finally, to the entire dataset was added a sample of size 10000 x 1 generated randomly from a 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 in training is calculated to be:

\[
100 \times \frac{\text{#true outliers} - \text{#observed outliers}}{\text{#true outliers}}
\]

Fig. 1 shows the plot of the robust Mahalanobis distances for this dataset, whereas fig. 2 shows the probability density curve of these distances estimated by Parzen windows. Averaging the results of all simulations of datasets generated by this process for cut-off 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. Fig. 3 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, although, a cut-off value of 0.97 yields the lowest outlier-detection error rates, the erratic behavior of the cut-off vs. error graph is a further area of research. However, 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. Outlier detection error rates were nil in some other synthetically generated data where there existed a clear separation between the outliers and inliers.

We consider this algorithm to be a success against the established statistical outlier detection methods in [1,2] where the underlying distribution of the data needs to be known. We believe that further refinements to this algorithm are necessary to accurately determine outliers. Although no outlier detection algorithm is a solution for all data set scenarios, this algorithm dominated by the time complexity required in computing the density – $O(n\log n)$ according to [7], is a sufficient improvement over combinatorial algorithms like the Minimum Covariance Determinant (MCD)\(^8\). However, in cases where $p$ is very large this method proves to be computationally expensive.

We considered another artificial data set consisting of 1500 samples with two features randomly 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 depicted in Fig 4.

On this data set, the accuracy of this algorithm becomes markedly pronounced as we see in the fig. 4(d) below. For all bi-weight cut-off values in the range $0.5(0.01)0.99$ the outliers were correctly identified.

Finally, we executed the above algorithm on another dataset, which we named the Squares dataset, with 800 sample points with 2 features generated randomly from a uniform rectangular distribution with varying parameters to obtain 6 major classes of data and few outlying data as in Fig. 5(b).

Most of the inlier classes have equal distribution of points except one which was deliberately created with the aim of illustrating the fact that a class containing few data points may not be outliers with respect to the entire data set. As usual the outliers were detected correctly. Similar other experiments have established the efficacy of the proposed method.
6. Conclusion

The proposed method using robust location and scale evaluation as well as non-parametric density estimation of the robust Mahalanobis distances is very effective in detecting outliers and is evident empirically. An established statistical test of discordancy had been replaced by a simple method to find the rejection point. Although such a method lacks mathematical sophistication, no assumption about the distribution of the data or the robust Mahalanobis distances had been made.

An interesting question is whether simultaneous outlier detection and a “good” clustering of the data, possible through this method? In the proposed method, only two possible clusters are created – 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 rejection region may represent distinct density clusters.

Several peaks of the density curve, however, do not in general represent separate clusters. This method
is probably not suitable for clustering applications without refinements.

7. References


