Practical Data Analysis: Designing a Real-World Outlier Detector

Leland Wilkinson
H2O.ai
Department of Computer Science, University of Illinois at Chicago.
email: leland@h2o.ai

Erin LeDell
H2O.ai
email:erin@h2o.ai

August 4, 2016

Abstract

Theory and practice do not always coincide in the world of real data analysis. This paper presents a new algorithm, called `hdoutliers`, for detecting multidimensional outliers. It is designed specifically to a) deal with a mixture of categorical and continuous variables, b) deal with the curse of dimensionality (many columns of data), c) deal with many rows of data, d) deal with outliers that mask other outliers, and e) deal consistently with unidimensional and multidimensional problems. Unlike ad hoc methods found in many machine learning papers, `hdoutliers` is based on a distributional model that allows outliers to be tagged with a probability. And unlike many methods found in the statistical literature, it presents opportunities for extending the problem to messy datasets.

1 Introduction

It is often said that data science is 90 percent data preparation and 10 percent data analysis. We can go on to say, in another wild guesstimate, that data analysis

is 90 percent discovery and 10 percent modeling. Even when we have mapped our data to a rectangular matrix of real numbers, we cannot apply statistical or machine learning models without giving a lot of thought to violations of assumptions, missing values, outliers, and other threats to the validity of our inferences. Statisticians have understood this for many years.

Practical data analysis involves more than investigating model assumptions, however. In many cases, there is a discrepancy between a theoretical statistical or machine learning model and what is needed to implement that model on real data. In practical data analysis we often encounter wide, deep, heterogeneous, sparse, or massive data that do not fit a standard model. In these cases, we need to encapsulate an analytic method in an extended algorithm designed to deal with these peculiar data structures. This effort is not part of what we normally call data preparation. Instead, it involves the analytic process itself. Unless we properly address this problem, we may not be able to find a valid answer to our analytic question.

We will use a common analytic question and an associated algorithm to illustrate our point: how do we find anomalies, specifically outliers, in a set of data? Despite more than a century of statisticians offering answers to this question, none of the traditional statistical or machine learning outlier detection methods are of much use when applied to most real data. They are all predicated on restrictive assumptions that do not apply to most practical situations.

The difficulties we face in devising a practical method are several. In working with real data, we often encounter:

- too many columns of data (the curse of dimensionality)
- too many rows of data (scalability)
- sparse data
- a mixture of categorical and continuous variables
- missing or undefined values
- outliers that mask other outliers

Before showing how we may deal with some of these problems, we will summarize related work on outlier detection.

2 Related Work

There are several excellent books and surveys on outliers (Barnett and Lewis, 1994; Hawkins, 1980; Rousseeuw and Leroy, 1987; Thode, 2002; Hadi and Simonoff, 1993; Iglewicz and Hoaglin, 1993; Anscombe, 1960; Aggarwal, 2013; Chandola et al., 2009; Hodge, 2011). We summarize the basic approaches in this section.

2.1 Univariate Outliers

The detection of outliers in the observed distribution of a single variable spans the entire history of outlier detection. It spans this history not only because it is perhaps the simplest formulation of the problem, but it also has numerous practical applicatons.

2.1.1 The Distance from the Center Rule

The word *outlier* implies lying at an extreme end of a set of ordered values – far away from the center of those values. The modern history of outlier detection emerged with methods that depend on a measure of centrality and a measure of distance from that measure of centrality. As early as the 1860's, Chauvenet (Barnett and Lewis, 1994) judged an observation to be an outlier if it lies outside the lower or upper $1/(4n)$ points of the Normal distribution. Barnett and Lewis (1994) document many other early rules that depend on the Normal distribution but fail to distinguish between population and sample variance.

Grubbs (1950), in contrast, based his rule on the sample moments of the Normal:

$$G = \frac{\max\limits_{1 \le i \le n} |x_i - \bar{x}|}{s}$$

where $\bar{x}$ and s are the sample mean and standard deviation, respectively.

Grubbs referenced G against the t distribution in order to spot an upper or lower outlier:

$$G > \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{\alpha/(2n),n-2}}{n-2+t^2_{\alpha/(2n),n-2}}}$$

If one knows that the values on a variable are sampled randomly from a Normal distribution and if the estimates of location and scale are unbiased and if one wishes to detect only the largest absolute outlier, it is a valid test.

Unfortunately, the usual sample estimates of the mean and standard deviation are not robust against outliers. So we have a circularity problem. We assume a null distribution (say, the Normal), estimate its parameters, and then use those estimates to test whether a point could have plausibly come from that distribution. But if our alternative hypothesis is that it doesn't (the usual case), then the outlier should not be included in the estimation. Barnett and Lewis (1994) discuss this problem in more detail, where they distinguish *inclusive* and *exclusive* methods. They, as well as Rousseeuw and Leroy (1987), also discuss robust estimation methods for overcoming this circularity problem.

Barnett and Lewis discuss other detection methods for non-Normal distributions. The same principals apply in these cases, namely, that the sample is random, the population distributions are known and that the parameter estimates are unbiased.

2.1.2 The Box Plot Rule

A box-plot graphically depicts a batch of data using a few summary statistics called *letter values* (Tukey, 1977; Frigge et al., 1989). The letter values in Tukey's original definition are the median and the *hinges* (medians of the upper and lower halves of the data). The hinge values correspond closely, but not necessarily, to the lower quartile (Q1) and the upper quartile (Q3). Tukey called the difference between the hinges the *Hspread*, which corresponds closely to the quantity Q3–Q1, or the Inter Quartile Range (IQR). In Tukey's version of the box-plot (see the upper panel of Figure 1), a box is drawn to span the *Hspread*. The median is marked inside the box. Whiskers extend from the edges of the box to the farthest upper and lower data points (*Adjacent values*) inside the so-called *inner fences*. The upper inner fence is the

$$upperhinge + 1.5 \times Hspread$$

and the lower inner fence is the

$$lowerhinge - 1.5 \times Hspread$$

Any data point beyond the *Adjacent values* is plotted as an outlying point. [1]

[1] Few statistics packages produce box plots according to Tukey's definition (Frigge et al., 1989).

Tukey designed the box plot (he called it a *schematic plot*) to be drawn by hand on a small batch of numbers. The whiskers were designed not to enable outlier detection, but to locate the display on the interval that supports the bulk of the values. Consequently, he chose the *Hspread* to correspond roughly to three standard deviations on normally distributed data. This choice led to two consequences: 1) it doesn't apply to skewed distributions, which constitute the instance many advocates think is the best reason for using a box plot in the first place, and 2) it doesn't include sample size in its derivation, which means that the box plot will falsely flag outliers on larger samples. As Dawson (2011) shows, "regardless of size, at least 30% of samples drawn from a normally-distributed population will have one or more data flagged as outliers." The top panel of Figure 1 illustrates this problem for a sample of 100,000 Normally distributed numbers. Thousands of points are denoted as outliers in the display.

Hoaglin et al. (1986) and Hoaglin and Iglewicz (1987) modified Tukey's definition to include n as a parameter, thus making the boxplot useful at least for symmetric distributions. To deal with the skewness problem, Hubert and Vandervieren (2008) and others have suggested modifying the fences rule by using a robust estimate of skewness. By contrast, Tukey's approach for this problem involved transforming the data through his *ladder of powers* Tukey (1977) before drawing the box plot.

The Letter-Value-Box-Plot (Hofmann et al., 2011) was designed to deal with both problems. The authors compute additional letter values (splitting the splits) until a statistical measure of fit is satisfied. Each letter-value region is represented by a rectangle. The lower panel of Figure 1 shows the result. On the same 100,000 Normal variables, only two points are identified as outliers.

2.1.3 The Gaps Rule

The gaps rule looks for outliers that are separated from other points instead of from a measure of central location. A simple example illustrates. Suppose we give a test to 100 students and find the mean score is 50 and the standard deviation is 5. Among these students, we find one perfect score of 100. The next lower score is 65. We might be inclined to suspect the student with a score of 100 is a genius or a cheat. On the other hand, if the perfect score is at the top of a chain of scores spaced not more than 5 points apart (assuming the same mean and standard

Surprisingly, the boxplot function in the core *R* package does not, despite its origin in Tukey's group at Bell Laboratories.

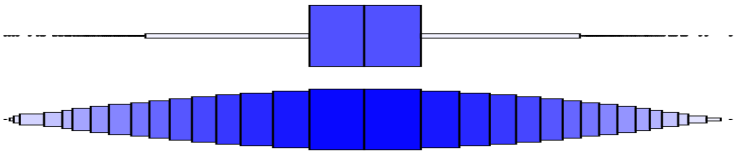

Figure 1: Outliers revealed in a box plot (Tukey, 1977) and letter values box plot (Hofmann et al., 2011). These plots are based on 100,000 values sampled from a Gaussian (Standard Normal) distribution. By definition, the data contain no probable outliers, yet the ordinary box plot shows thousands of outliers. This example illustrates why ordinary box plots cannot be trusted to discover probable outliers.

deviation overall), we might be less suspicious. Classical outlier tests would not discriminate among these possibilities.

Dixon (1951) developed an outlier test based on the gap between the largest point and the second largest point, standardized by the range of scores. His test was originally based on a normal distribution, but in subsequent publications, he developed nonparametric variations. Dixon tabulated percentage points for a range of Q statistics.

$$Q = \frac{x_n - x_{n-1}}{x_n - x_1}$$

Tukey (1977) considered the more general question of identifying gaps anywhere in a batch of scores. Wainer and Schacht (1978) adapted Tukey's gapping idea for a version of the test that weighted extreme values more than middle ones. They derived an approximate z score that could be used to test the significance of gaps.

Burridge and Taylor (2006) developed an outlier test based on the extreme-value distribution of gaps between points sampled from the Exponential family of distributions. This family is quite large (including the Normal, Exponential, Gamma, Beta, Bernoulli, Poisson, and many others).

2.2 Multivariate Outliers

2.2.1 Mahalanobis Distance

The squared Mahalanobis distance (D^2) of a multidimensional point $\mathbf{x}$ from the centroid of a multivariate Normal distribution described by covariance matrix Σ

and centroid μ is

$$D^2 = (\mathbf{x} - \mu)' \Sigma^{-1} (\mathbf{x} - \mu)$$

Figure 2 shows how this works in two dimensions. The left panel shows a bivariate normal distribution with level curves inscribed at different densities. The right panel shows the same level curves as horizontal slices through this mountain. Each is an ellipse. Distance to the centroid of the ellipses is measured differently for different directions. The weights are determined by the covariance matrix Σ. If Σ is an identity matrix, then D^2 is equivalent to squared Euclidean distance.

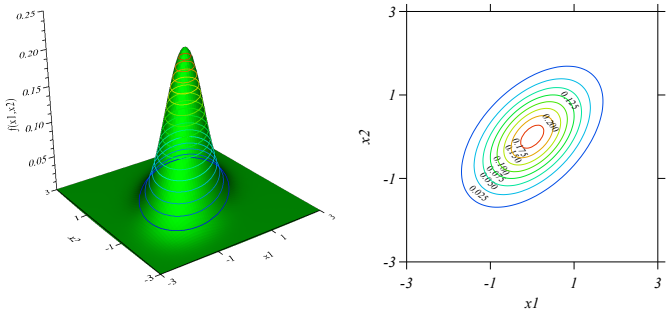

Figure 2: Mahalanobis Distance. The left panel shows a bivariate Normal distribution. The right shows level curves for that distribution. Each curve corresponds to a value of D^2.

The squared distance in the above formula is a chi-square variate. This means that, if the assumption of Normality is met, D^2 can be tested against a chi-square distribution with p degrees of freedom. As with univariate outlier tests based on a Normality assumption, this test is valid if the assumption of multivariate Normality is met. Unfortunately, this is seldom true for real data and, furthermore, estimates of the covariance matrix and centroid are far from robust. Consequently, this outlier test has limited applicability.

Rousseeuw and Zomeren (1990) introduce a robust Mahalanobis Distance estimator that can be used to overcome some of these problems. Stahel (1981) and Donoho (1982) have developed a similar estimator with a high breakdown point. In both cases, the computations are intensive and require additional designs to allow them to scale to big data.

2.2.2 Multivariate Gap Tests

Multivariate data do not have a simple ordering for computing gaps between adjacent points. There have been several attempts at getting around this problem. Rohlf (1975) proposed using the edge lengths of the geometric minimum spanning tree (MST) as a single distribution measure. Assuming these edges follow a gamma distribution, one could construct a gamma probability plot on them or examine the upper tail for judgments on outliers. There are problems with this method, however, when variates are correlated (Caroni and Prescott, 1995). Similar methods based on the MST have been proposed (Lin et al., 2008; Peter and Victor, 2010), but they suffer from the same problem.

2.2.3 Clustering

A popular multivariate outlier detection method has been to cluster the data and then look for any points that are far from their nearest cluster centroids (Zahn, 1971; Jiang et al., 2001; Pamula et al., 2011; Jobe and Pokojovy, 2015). This method works reasonably well for moderate-size datasets with a few singleton outliers. Most clustering algorithms do not scale well to larger datasets, however.

A related approach, called Local Outlier Factor (LOF) (Breunig et al., 2000), is derived from density-based clustering. Like DBSCAN clustering (Ester et al., 1996), it is highly sensitive to the choice of input parameter values and doesn't scale well.

Most importantly, clustering outlier methods are not based on a probability model (see Fraley and Raftery (2002) for an exception) so they are susceptible to false negatives and false positives. We will show one remedy in Section 3.3.2.

3 A New Multivariate Outlier Algorithm

A new algorithm `hdoutliers` is designed to meet several criteria at once:

- It allows us to identify outliers in a mixture of categorical and continuous variables.
- It deals with the curse of dimensionality by exploiting random projections for large p (number of dimensions).
- It deals with large n (number of points) by exploiting a one-pass algorithm to compress the data.

- It deals with the problem of *masking* Barnett and Lewis (1994), in which clusters of outlying points can elude detection by traditional methods.
- It works for both single-dimensional and multi-dimensional data.

3.1 The Algorithm

1. If there are any categorical variables in the dataset, convert each categorical variable to a continuous variable by using Correspondence Analysis Greenacre (1984).
2. If there are more than 10,000 columns, use random projections to reduce the number of columns to $p = 4\log n/(\varepsilon^2/2 - \varepsilon^3/3)$, where ε is the error bound on squared distances.
3. Normalize the columns of the resulting n by p matrix X.
4. Let $row(i)$ be the ith row of X.
5. Let $radius = .1/(\log n)^{1/p}$.
6. Initialize *exemplars*, a list of exemplars with initial entry $[row(1)]$.
7. Initialize *members*, a list of lists with initial entry $[1]$; each exemplar will eventually have its own list of affiliated member indices.
8. Now do one pass through X:

forall the $row(i)$, $i = 1,\ldots,n$ **do**
  d = distance to closest exemplar, found in $exemplars(j)$
  if $d < radius$ **then**
    add i to $members(j)list$
  else
    add $row(i)$ to *exemplars*
    add new list to *members*, initialized with $[i]$
  end
end

9. Now compute nearest-neighbor distances between all pairs of exemplars in the *exemplars* list.

10. Fit an Exponential distribution to the upper tail of the nearest-neighbor distances and compute the upper $1-\alpha$ point of the fitted cumulative distribution function (CDF).

11. For any exemplar that is significantly far from all the other exemplars based on this cutpoint, flag all entries of *members* corresponding to *exemplar* as outliers.

3.2 Comments on the Algorithm

1. Correspondence Analysis (CA) begins by representing a categorical variable with a set of dummy codes, one code (1 or 0) for each category. These codes comprise a matrix of 1's and 0's with as many columns as there are categories on that variable. We then compute a principal components decomposition of the covariance matrix of the dummy codes. This analysis is done separately for each of k categorical variables in a dataset. CA scores on the rows are computed on each categorical variable by multiplying the dummy codes on that row's variable times the eigenvectors of the decomposition for that variable. [2]

2. The Johnson and Lindenstrauss (1984) lemma states that if a metric on X results from an embedding of X into a Euclidean space, then X can be embedded in R^p with distortion less than $1+\varepsilon$, where $p \sim O(\varepsilon^2 \log n)$. Remarkably, this embedding is achieved by projecting onto a p-dimensional subspace using random Gaussian coefficients. Because our algorithm depends only on a similarity transformation of Euclidean distances, we can logarithmically reduce the complexity of the problem through random projections and avoid the curse of dimensionality. The number of projected columns based on the formula in this step was based on $\varepsilon = .2$ for the analyses in this paper. The value 10,000 is the lower limit for the formula's effectiveness in reducing the number of dimensions when $\varepsilon = .2$.

3. X is now bounded by the unit (hyper) cube.

4. A *row* represents a p-dimensional vector in a finite vector space.

5. The value of *radius* is designed to be well below the expected value of the distances between $n(n-1)/2$ pairs of points distributed randomly in a p dimensional space.

6. The *exemplars* list contains a list of row values representing clusters of points.

7. The *members* list of lists contains one list of indices for each exemplar that point to rows represented by that exemplar.

8. The Leader algorithm (Hartigan, 1975) in this step creates clusters in one pass through the data. It is equivalent to centering balls in p dimensional space on points considered to be exemplars. Unlike k-means clustering, the Leader algorithm centers clusters on actual data points rather than on centroids and it involves only one pass through the data. In rare instances, the resulting clusters can be dependent on the order of the data, but not enough to affect the identification of outliers because of the large number of clusters produced. We are characterizing a high-dimensional density by covering it with many small balls.

9. The number of clusters resulting from *radius* applied even to large numbers of data points is small enough to allow the simple brute-force algorithm for finding nearest neighbors.

10. We use a modification of the Burridge and Taylor (2006) algorithm due to Schwarz (2008). For all examples in this paper, α (the critical value) was set to .05.

11. Flagging all members of an outlying cluster means that this algorithm can identify outlying sets of points as well as outlying singletons.

3.3 Validation

We validate `hdoutliers` by examining its performance with regard to 1) false positives and 2) false negatives. If the claims for the algorithm are true, then we should expect it 1) to find outliers in random data not more than 100α percent

[2]Computing the decomposition separately for each categorical variable is equivalent to doing an MCA separately for each variable instead of pooling all the categorical variable dummy codes into one matrix.

Table 1: Empirical level of `hdoutliers` test based on null model with Gaussian variables and critical value $\alpha = .05$.

	p=1	**p=5**	**p=10**	**p=100**
n=100	.011	.040	.018	.012
n=500	.015	.035	.027	.020
n=1000	.017	.045	.027	.024

of the time and 2) not to miss outliers when they are truly due to mixtures of distributions or anomalous instances.

3.3.1 False Positives

- Table 1 contains results of a simulation using random distributions. The entries are based on 1,000 runs of `hdoutliers` on normally distributed variables with α (the critical value) set to .05. The entries show that `hdoutliers` is generally conservative.
- The results were similar for random Uniform variables.

3.3.2 False Negatives

- Figure 3 shows that `hdoutliers` identifies the inlier in the center of both one-dimensional and two-dimensional configurations.
- Table 2 shows that `hdoutliers` identifies the outlier in a table defined by two categorical variables. The data consist of two columns of strings, one for {A,B,C,W} and one for {A,B,C,X}. There is only one row with the tuple $\langle W, X \rangle$. The `hdoutliers` also handles mixtures of categorical and continuous variables.

Table 2: Crosstab with an outlier (red entry)

	A	**B**	**C**	**X**
A	100	0	0	0
B	0	100	0	0
C	0	0	100	0
W	0	0	0	**1**

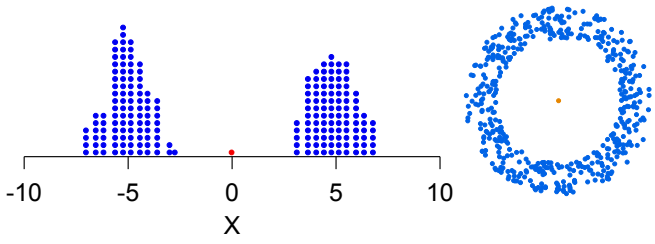

Figure 3: Inlier datasets; `hdoutliers` correctly identifies the inliers.

4 Visualization

This section illustrates the adaptability of `hdoutliers` to a variety of applications. First, however, we document the shortcomings of popular visualization methods for identifying multivariate outliers.

4.1 Low-dimensional visualizations cannot be used to discover multivariate outliers

There have been many outlier identification proposals based on looking at axis-parallel views or low-dimensional projections (usually 2D) that are presumed to reveal high-dimensional outliers (e.g., Kandogan (2012); Hinneburg et al. (1999); Kandogan (2001); Kriegel et al. (2009)). This approach is infeasible. Figure 4 shows why. The data are samples from a multivariate Normal distribution. The left panel plot illustrates the problem for two dimensions. The figure incorporates a 95 percent joint confidence ellipse based on the sample distribution of points. Two points are outside this ellipse. The red point on the left is at the extreme of both marginal histograms. But the one on the right is well inside both histograms. Examining the separate histograms would fail to identify that point.

The right panel plot shows the situation for three dimensions. The three marginal 2D plots are shown as projections onto the facets of the 3D cube. Each confidence ellipse is based on the pairwise plots. The red outlying point in the joint distribution is inside all three marginal ellipses. The 2D scatterplots fail to reveal the 3D outlier. The situation gets even worse in higher dimensions.

Some authors have proposed methods for finding low-dimensional views based on projection pursuit or ad hoc projections (e.g., Breaban and Luchian (2013); Ruiz-Gazen et al. (2010)). This approach is relatively ineffective for visualizing

outliers in higher-dimensional datasets because many projections are required to discriminate outliers. Furthermore, most outlier-seeking projection methods are impractical on large datasets.

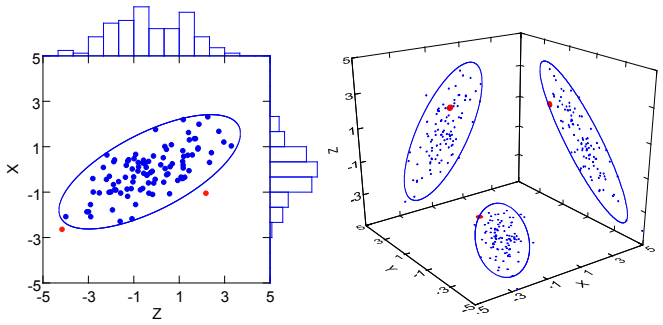

Figure 4: 2D (left) and 3D (right) joint outliers. The figures show why lower-dimensional projections cannot be used to discern outliers.

Parallel coordinates have been advocated for the purpose of outlier detection (Novotny and Hauser, 2006). Figure 5 shows why this is infeasible. The figure contains a parallel coordinates display on four variables from the Adult dataset in the UCI dataset repository (Kohavi and Becker, 1996). The `hdoutliers` algorithm discovered two outliers out of 32,561 cases. The profiles appear to run through the middle of the densities even though they are multivariate outliers. Although parallel coordinates are generally useless for discovering outliers, they can be useful for inspecting outlier profiles detected by a statistical algorithm.

4.2 Using statistical algorithms to highlight outliers in visualizations

While visualizations cannot be used to detect multidimensional outliers, they are invaluable for inspecting and understanding outliers detected by statistical methods. This section covers a variety of visualizations that lend themselves to outlier description.

4.2.1 Time Series Outliers

Detecting time series outliers requires some pre-processing. In particular, we need to fit a time series model and then examine residuals. Fitting parametric models

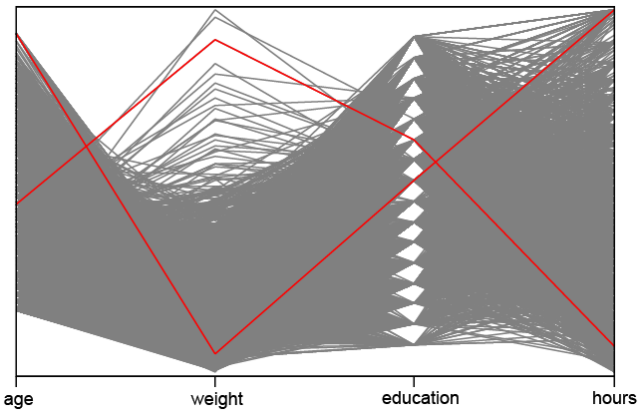

Figure 5: Parallel coordinates plot of five variables from the Adult dataset in the UCI data repository. The red profiles are multivariate outliers.

like ARIMA (Box and Jenkins, 1976) can be useful for this purpose, but appropriate model identification can be complicated. A simpler approach is to fit a nonparametric smoother. The example in Figure 6 was fit by a kernel smoother with a biweight function on the running mean. The data are measurements of snowfall at a Greenland weather station, used in Wilkinson (1999b). The outliers (red dots) are presumably due to malfunctions in the recording equipment.

Computing outlying series for multiple time series is straightforward with the `hdoutliers` algorithm. We simply treat each series as a row in the data matrix. For n series on p time points, we have a p-dimensional outlier problem. Figure 7 shows series for 20 years of the Bureau of Labor Statistics Unemployment data. The red series clearly indicate the consequences of the Great Recession. This example illustrates why a probability-based outlier method is so important. We could rank the series by their average levels of unemployment or use one of the other ad-hoc multidimensional outlier detectors, but we would have no way of knowing how many at the top are significant outliers.

4.2.2 Ipsative Outliers

An *ipsative* outlier is a case that is an outlier with respect to itself. That is, we standardize values within each case (row) and then look for outliers in each standardized profile. Any profile with an outlier identified by `hdoutliers` is considered noteworthy; in other words, we can characterize a person simply by referring

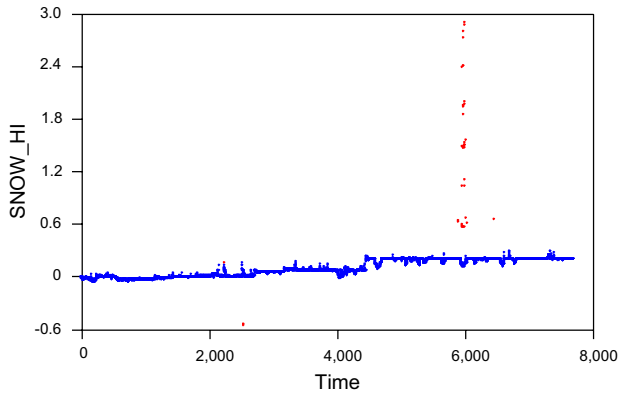

Figure 6: Outlying measurements of snow cover at a Greenland weather station.

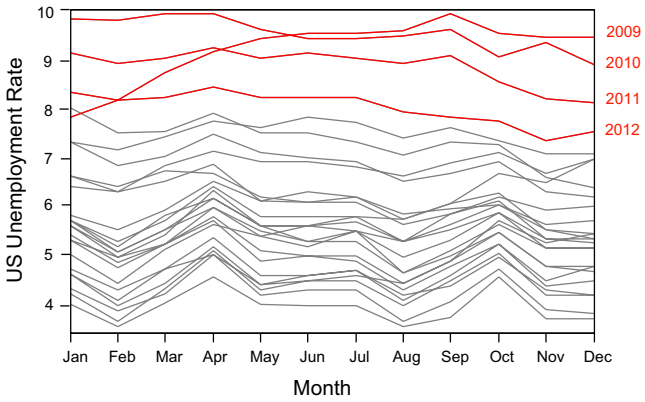

Figure 7: US Unemployment series outliers. The shock and ensuing recovery from the Great Recession is clearly indicated in the outliers.

to his outliers. It is easiest to understand this concept by examining a graphic. Figure 8 shows an outlying profile for a baseball player who is hit by pitches more frequently than we would expect from looking at his other characteristics. This player may not be hit by pitches significantly more than other players, however. We are instead interested in a player with a highly unusual profile that can be described simply by his outlier(s). In every other respect, the player is not necessarily noteworthy. This method should not be used, of course, unless there are enough features to merit computing the statistical outlier model on a case.

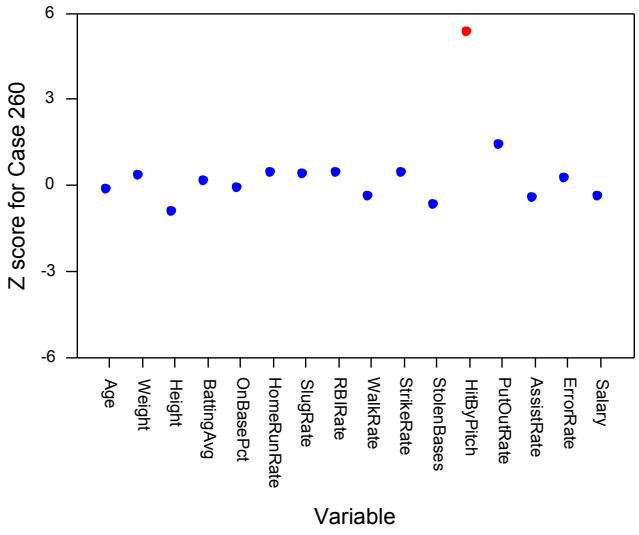

Figure 8: One baseball player's profile showing an outlier (hit by pitch) that deviates significantly from his other features.

4.2.3 Text Outliers

An important application for multivariate outlier detection involves document analysis. Given a collection of documents (Twitter messages, Wikipedia pages, emails, news pages, etc.), one might want to discover any document that is an outlier with respect to the others. The simplest approach to this problem is to use a bag-of-words model. We collect all the words in the documents, stem them to resolve variants, remove stopwords and punctuation, and then apply the tf-idf measure (Salton and Buckley, 1988) on the words within each document. The resulting vectors for each document are then submitted to `hdoutliers`.

Figure 9 shows the results for an analysis of 21 novels from the Guttenberg Web site (Hart, 2016). This problem requires the use of random projections. Before projection, there are 21,021 columns (tf-idf measures) in the dataset. After projection there are 653. Not surprisingly, *Ulysses* stands out as an outlier. Distinctively, it contains numerous neologisms.

Tristram Shandy was identified by `hdoutliers` as the second largest, but not significant, outlier. It too contains numerous neologisms. These two novels lie outside most of the points in Figure 9. Not all multivariate outliers will fall on the periphery of 2D projections, however, as we showed in Section 4.2.

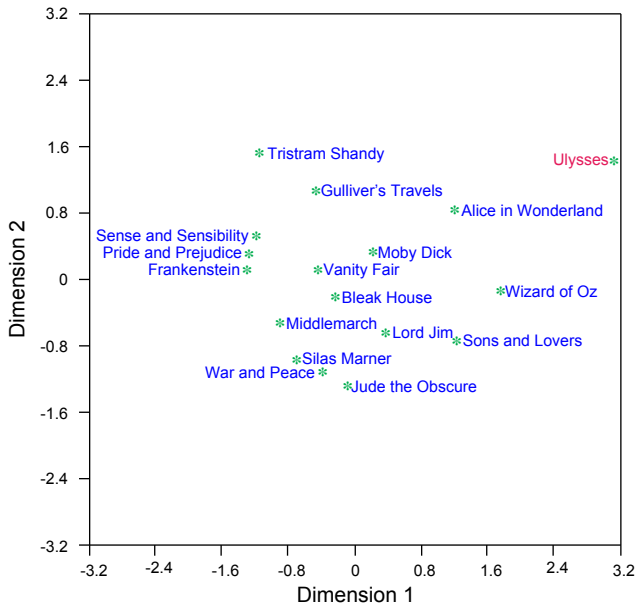

Figure 9: Document outliers. Nonmetric multidimensional scaling on matrix of Spearman correlations computed on tf-idf scores. The stress for this solution is .163 and one document (*Ulysses*) is flagged as an outlier by `hdoutliers`.

4.3 Graph Outliers

There are several possibilities related to finding outliers in graphs. One popular application is the discovery of outliers among nodes of a network graph. The best way to exploit `hdoutliers` in this context is to featurize the nodes. Common candidates are Prominence, Transitivity (Watts-Strogatz Clustering Coefficient), Closeness Centrality, Betweenness Centrality, Node Degree, Average Degree of Neighbors, and Page Rank (Newman et al., 2006). Figure 10 shows an example for the Les Miserables dataset (Knuth, 1993). The nodes were featurized for Betweenness Centrality in order to discover any extraordinarily influential characters. Not surprisingly, Valjean is connected to significantly more characters than anyone else in the book.

An alternative application involves discovering outlying graphs in a collection of graphs. For this problem, we need to find a way to characterize a graph and to derive a distance measure that can be fed to `hdoutliers`. This application depends on assuming the collection of graphs is derived from a common

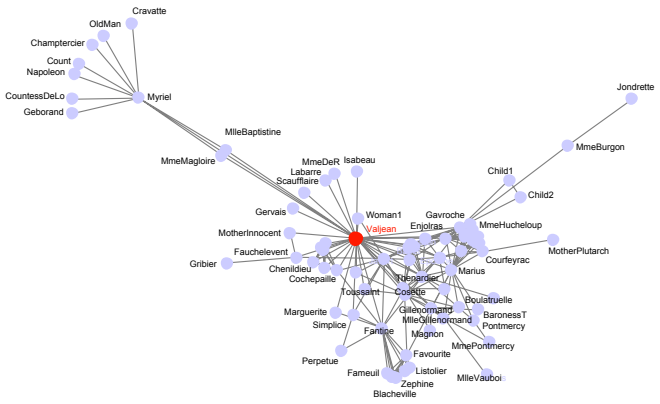

Figure 10: Les Miserables characters network graph. Valjean is identified as outlying on Betweenness Centrality.

population model and that any outliers involve a contamination from some alternative model. We need a measure of the distance between two graphs to do this. Unfortunately, graph matching and related graph edit distance calculations have impractical complexities. Approximate distances are easier to calculate, however (Umeyama, 1988). The approach we take is as follows:

First, we compute the adjacency matrix for each graph. We then convert the adjacencies above the diagonal to a single binary string. When doing that, however, we have to reorder the adjacency matrix to a canonical form; otherwise, arbitrary input orderings could affect distance calculations on the string. A simple way to do this is to compute the eigendecomposition of the related Laplacian matrix and permute the adjacencies according to the ordering of the values of the eigenvector corresponding to the smallest nonzero eigenvalue. After permuting and encoding the adjacency matrices into strings, we compute the Levenshtein distances (Levenshtein, 1966) between pairs of strings. Finally, we assemble the nearest-neighbor distances from the resulting distance matrix and subject them to the `hdoutliers` algorithm.

Figure 11 shows an example of this approach using the Karate Club graph (Zachary, 1977). We generated 15 random minimum spanning tree graphs having the same number of nodes as the Karate Club graph. Then we applied the above procedure to identify outliers. The Karate Club graph was strongly flagged as an outlier by the algorithm.

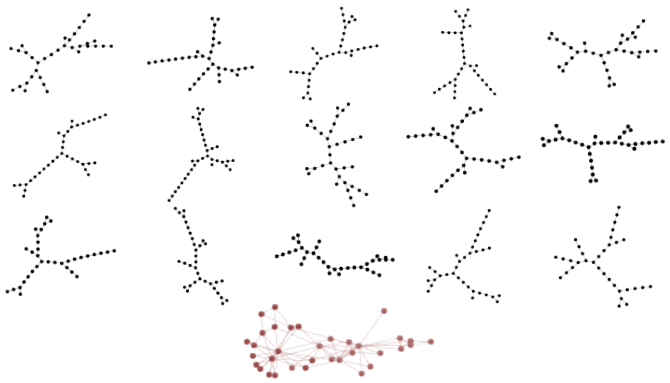

Figure 11: Karate Club graph (red) is an outlier with respect to comparably scaled random minimum spanning tree graphs.

4.3.1 Scagnostics Outliers

Scagnostics (Wilkinson et al., 2005) can be used to identify outlying scatterplots. Because the calculations are relatively efficient, these measures can be computed on many thousands of plots in practical time. This outlier application is multivariate, because there are nine scagnostics for each scatterplot, so a multivariate detection algorithm like `hdoutliers` is required.

Figure 12 shows two outlying scatterplots identified by `hdoutliers` when applied to a dataset of baseball player characteristics featured in Wilkinson et al. (2006). While the left plot in the figure is clearly unusual, the surprising result is to see an evidently bivariate Normal scatterplot of Weight against Height in the right plot. Although the dataset includes many physical and performance features of real baseball players, the type of Normal bivariate distribution found in many introductory statistics books is an outlier among the 120 scatterplots considered in this example. This result should motivate authors writing tutorials on data analysis to include examples beyond Normal distributions.

4.3.2 Geographic Outliers

We can compute spatial outliers using the `hdoutliers` algorithm. More frequently, however, maps are a convenient way to display the results of outlier detection on other variables. Figure 13 shows an example of outlier detection on marriage and divorce rates by US state. Nevada is clearly an outlier.

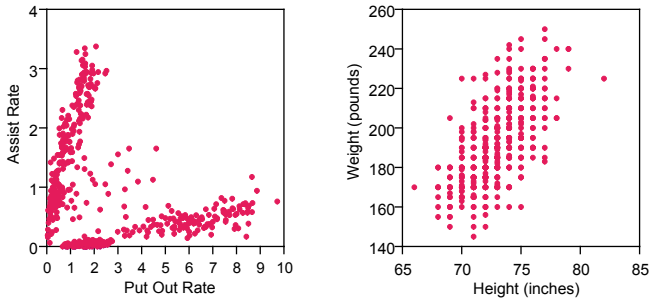

Figure 12: Scatterplot outliers based on Scagnostics computed on 120 scatterplots of baseball player features.

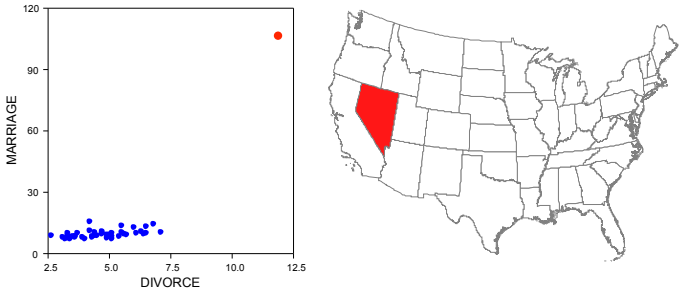

Figure 13: Marriage and Divorce rates in the US. There is one state that is an outlier.

5 Conclusion

There is a huge assortment of papers on outlier detection in the statistics and machine learning communities; only a fraction is cited here. While many of the machine learning approaches are ingenious, few rest on a statistical foundation that takes risk into account. If we label something as an outlier, we had better be able to quantify or control our risk. Methods that do not do this, that simply rank discrepancies or flag observations above an arbitrary threshold (like most machine learning outlier algorithms), can lead to inconsistent results.

When we do base an algorithm on classical or Bayesian statistical models, however, we have to consider the peculiarities of the many different real datasets that we hope to cover. The `hdoutliers` algorithm illustrates some of these considerations. Statisticians can learn from computer scientists how to adapt to unusual or massive data configurations. And computer scientists can learn from

statisticians how to incorporate probability into their analytic algorithms.

References

Aggarwal, C. (2013). *Outlier Analysis*. Springer Verlag.

Anscombe, F. (1960). Rejection of outliers. *Technometrics*, 2:123–147.

Atkinson, A. (1985). *Plots, Transformations and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*. Oxford University Press.

Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data*. John Wiley & Sons.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley & Sons.

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control (rev. ed.)*. Holden-Day, Oakland, CA.

Breaban, M. and Luchian, H. (2013). Outlier detection with nonlinear projection pursuit. *International Journal of Computers Communications & Control*, 8(1):30–36.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000). LOF: Identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, SIGMOD '00, pages 93–104, New York, NY, USA. ACM.

Burridge, P. and Taylor, A. (2006). Additive outlier detection via extreme-value theory. *Journal of Time Series Analysis*, 27:685–701.

Caroni, C. and Prescott, P. (1995). On Rohlf's method for the detection of outliers in multivariate data. *Journal of Multivariate Analysis*, 52:295–307.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM Comput. Surveys*, 41:15:1–15:58.

Cleveland, W. S. (1985). *The Elements of Graphing Data*. Hobart Press, Summit, NJ.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman & Hall, London.

Dawson, R. (2011). How significant is a boxplot outlier? *Journal of Statistics Education*, 19.

Dixon, W. (1951). Ratios involving extreme values. *Annals of Mathematical Statistics*, 22:68–78.

Donoho, D. (1982). Breakdown properties of multivariate location estimators. Technical report, Harvard University.

Donoho, D. and Huber, P. (1983). The notion of breakdown point. In Bickel, P., Doksum, K., and Hodges, J., editors, *A Festschrift for Erich L. Lehman*, pages 157–184. Wadsworth, Belmont, CA.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Simoudis, E., Han, J., and Fayyad, U. M., editors, *Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231. AAAI Press.

for Artificial Intelligence, G. R. C. (2016). dataset: dfki-artificial-3000-unsupervised-ad.csv. `http://madm.dfki.de/downloads`. Accessed: 2016-02-08.

Fraley, C. and Raftery, A. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97:611–631.

Frigge, M., Hoaglin, D., and Iglewicz, B. (1989). Some implementations of the boxplot. *The American Statistician*, 43:50–54.

Gnanadesikan, R. (1977). *Methods for statistical data analysis of multivariate observations*. John Wiley & Sons, New York.

Greenacre, M. (1984). *Theory and Applications of Correspondence Analysis*. Academic Press.

Grubbs, F. (1950). Sample criteria for testing outlying observations. *The Annals of Mathematical Statistics*, 21:27–58.

Hadi, A. and Simonoff, J. (1993). Procedures for the identification of multiple outliers in linear models. *Journal of the American Statistical Association*, 88:1264–1272.

Hart, M. (2016). Project Gutenberg. `https://www.gutenberg.org`.

Hartigan, J. (1975). *Clustering Algorithms*. John Wiley & Sons, New York.

Hawkins, D. (1980). *Identification of Outliers*. Chapman & Hall/CRC.

Hinneburg, A., Keim, D., and Wawryniuk, M. (1999). HD-Eye: Visual mining of high-dimensional data. *IEEE Computer Graphics and Applications*, 19(5):22–31.

Hoaglin, D. and Iglewicz, B. (1987). Fine tuning some resistant rules for outlier labeling. *Journal of the American Statistical Association*, 82:1147–1149.

Hoaglin, D., Iglewicz, B., and Tukey, J. (1986). Performance of some resistant rules for outlier labeling. *Journal of the American Statistical Association*, 81:991–999.

Hodge, V. (2011). *Outlier and Anomaly Detection: A Survey of Outlier and Anomaly Detection Methods*. LAP LAMBERT Academic Publishing.

Hofmann, H., Kafadar, K., and Wickham, H. (2011). Letter-value plots: Boxplots for large data. Technical report, had.co.nz.

Hubert, M. and Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational Statistics & Data Analysis*, 52:5186–5201.

Iglewicz, I. and Hoaglin, D. (1993). How to detect and handle outliers. In Mykytka, E., editor, *The ASQC Basic References in Quality Control: Statistical Techniques*. ASQC.

Jiang, M., Tseng, S., and Su, C. (2001). Two-phase clustering process for outliers detection. *Pattern Recognition Letters*, 22:691–700.

Jobe, J. and Pokojovy, M. (2015). A cluster-based outlier detection scheme for multivariate data. *Journal of the American Statistical Association*, 110:1543–1551.

Johnson, W. B. and Lindenstrauss, J. (1984). Lipschitz mapping into Hilbert space. *Contemporary Mathematics*, 26:189–206.

Kandogan, E. (2001). Visualizing multi-dimensional clusters, trends, and outliers using star coordinates. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '01, pages 107–116, New York, NY, USA. ACM.

Kandogan, E. (2012). Just-in-time annotation of clusters, outliers, and trends in point-based data visualizations. In *Proceedings of the 2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, VAST '12, pages 73–82, Washington, DC, USA. IEEE Computer Society.

Knuth, D. (1993). *The Stanford GraphBase: A Platform for combinatorial computing*. Addison-Wesley, Reading, MA.

Kohavi, R. and Becker, B. (1996). Adult data set. `http://archive.ics.uci.edu/ml/datasets/Adult`.

Kriegel, H.-P., Kröger, P., Schubert, E., and Zimek, A. (2009). Outlier detection in axis-parallel subspaces of high dimensional data. In *13th Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD 2009)*, Tangkok, Thailand.

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10:707–710.

Lin, J., Ye, D., Chen, C., and Gao, M. (2008). Minimum spanning tree based spatial outlier mining and its applications. In *Proceedings of the 3rd International Conference on Rough Sets and Knowledge Technology*, pages 508–515, Berlin, Heidelberg. Springer-Verlag.

Newman, M. E. J., Barabasi, A.-L., and Watts, D. J. (2006). *The Structure and Dynamics of Networks*. Princeton University Press.

Novotny, M. and Hauser, H. (2006). Outlier-preserving focus+context visualization in parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):893–900.

Pamula, R., Deka, J., and Nandi, S. (2011). An outlier detection method based on clustering. In *International Conference on Emerging Applications of Information Technology, EAIT*, pages 253–256.

Peter, S. and Victor, S. (2010). Hybrid – approach for outlier detection using minimum spanning tree. *International Journal of Computational & Applied Mathematics*, 5:621.

Rohlf, F. (1975). Generalization of the gap test for the detection of multivariate outliers. *Biometrics*, 31:93–101.

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

Rousseeuw, P. and Zomeren, B. V. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85:633–651.

Rousseeuw, P. J. and Leroy, A. (1987). *Robust Regression & Outlier Detection*. John Wiley & Sons.

Ruiz-Gazen, A., Marie-Sainte, S. L., and Berro, A. (2010). Detecting multivariate outliers using projection pursuit with particle swarm optimization. In *COMPSTAT2010*, pages 89–98.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

Schwarz, K. (2008). *Wind Dispersion of Carbon Dioxide Leaking from Underground Sequestration, and Outlier Detection in Eddy Covariance Data using Extreme Value Theory*. PhD thesis, Physics, University of California, Berkeley.

Stahel, W. (1981). Breakdown of covariance estimators. Technical report, Fachgrupp fur Statistik.

Thode, H. (2002). *Testing For Normality*. Taylor & Francis.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley Publishing Company, Reading, MA.

Umeyama, S. (1988). An eigendecomposition approach to weighted graph matching problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10:695–703.

Wainer, H. and Schacht, S. (1978). Gapping. *Psychometrika*, 43:203–212.

Wilkinson, L. (1999a). Dot plots. *The American Statistician*, 53:276–281.

Wilkinson, L. (1999b). *The Grammar of Graphics*. Springer-Verlag, New York.

Wilkinson, L., Anand, A., and Grossman, R. (2005). Graph-theoretic scagnostics. In *Proceedings of the IEEE Information Visualization 2005*, pages 157–164. IEEE Computer Society Press.

Wilkinson, L., Anand, A., and Grossman, R. (2006). High-dimensional visual analytics: Interactive exploration guided by pairwise views of point distributions. *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1363–1372.

Zachary, W. (1977). An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33:452–473.

Zahn, C. T. (1971). Graph-theoretical methods for detecting and describing Gestalt clusters. *IEEE Transactions on Computers*, C-20:68–86.