Multidimensional Outlier Detection

BARUG
H2O.ai
13 September 2016

Leland Wilkinson
Chief Scientist
H2O
leland@h2O.ai

Outliers

An outlier is “an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data.”

Barnett & Lewis (1978)

“An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism.”

Hawkins (1980)

Outliers

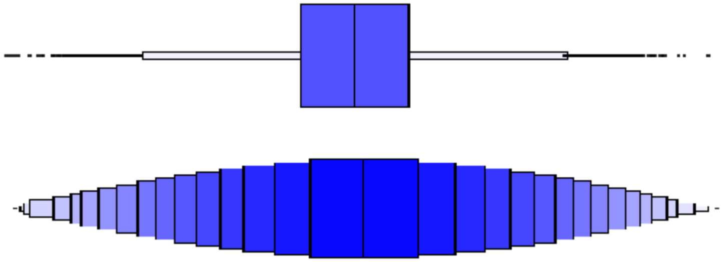
Univariate Outliers

Distance from the center rule

- Sigma rules (e.g., Six Sigma)
- Grubbs test (based on t distribution)
- Upper/lower tails of other distributions

Box (schematic) plot rule

- $upperhinge + 1.5 \times Hspread$
- $lowerhinge - 1.5 \times Hspread$
- Doesn't include *N*!

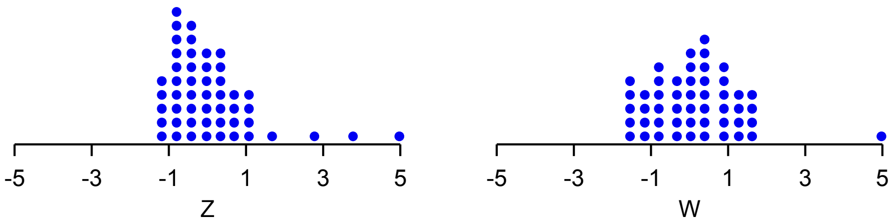


Outliers

Univariate Outliers

Gap tests

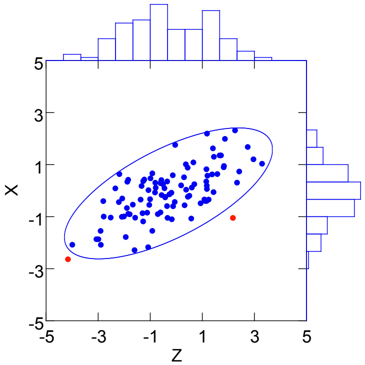
- Tukey gapping
- Dixon
- Burridge & Taylor

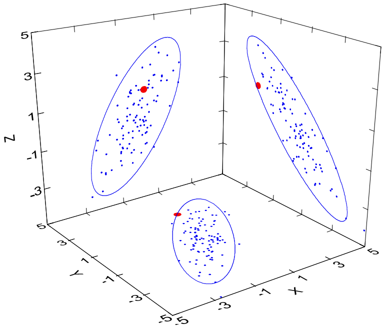


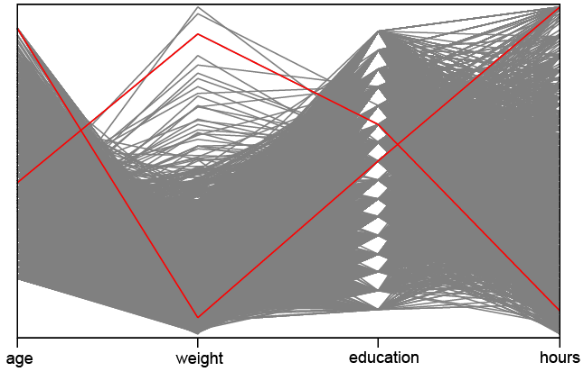
Outliers

Multivariate Outliers

Can’t use visualization to find multivariate outliers







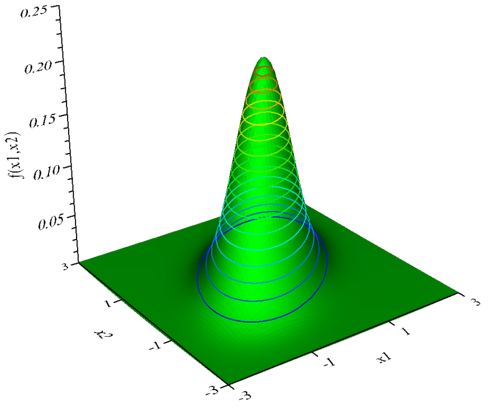
Outliers

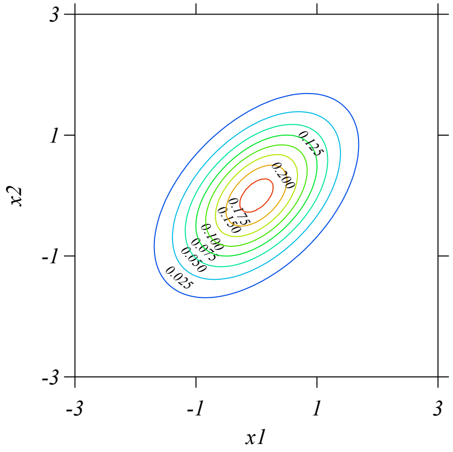
Multivariate Outliers

Distance from centroid

Mahalanobis Distance

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





Robust Mahalanobis Distance

Outliers

Multivariate Outliers

Gap tests

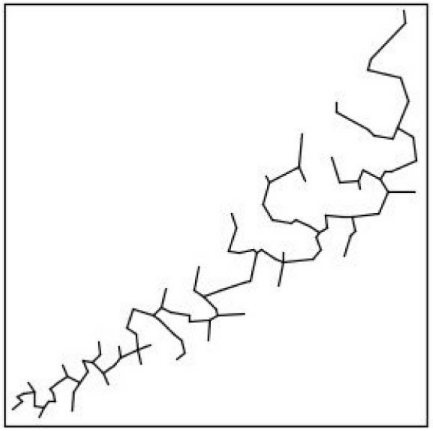
Geometric Minimum Spanning Tree (MST)

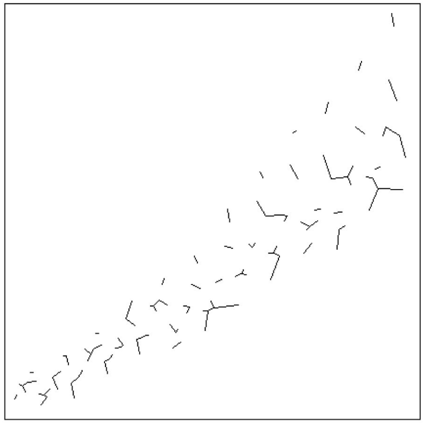
Nearest neighbor (NN)

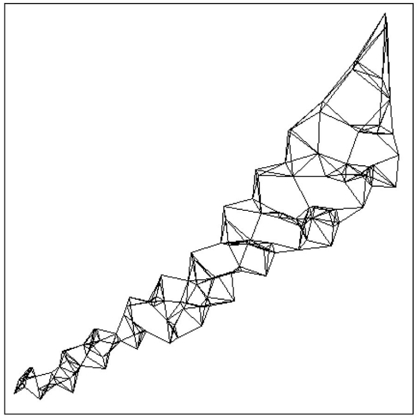
K-Nearest neighbor (KNN)

Local Outlier Factor (LOF)

Probabilistic Local Outlier Factor (PLOF)







Outliers

Multivariate Outliers

HDoutliers

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*, in which clusters of outlying points can elude detection by traditional methods.

It works for both single-dimensional and multi-dimensional data.

Outliers

HDoutliers

1. If there are any categorical variables in the dataset, convert each categorical variable to a continuous variable by using Correspondence Analysis

Outliers

HDoutliers

2. If there are more than 10,000 columns, use random projections to reduce the number of columns to

$$p = 4 \log n / (\epsilon^2/2 - \epsilon^3/3)$$

where ϵ is the error bound on squared distances.

The Johnson-Lindenstrauss 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 $1 + \epsilon$, where $p \sim O(\epsilon^2 \log n)$

The value 10,000 is the lower limit for the formula's effectiveness in reducing the number of dimensions when $\epsilon = .2$.

Outliers

HDoutliers

3. Normalize the columns of the resulting *n* by *p* matrix X.

The columns are now scaled on [0, 1]

Outliers

HDoutliers

4. Let $row(i)$ be the ith row of X.

5. Let $radius = .1/(\log n)^{1/p}$.

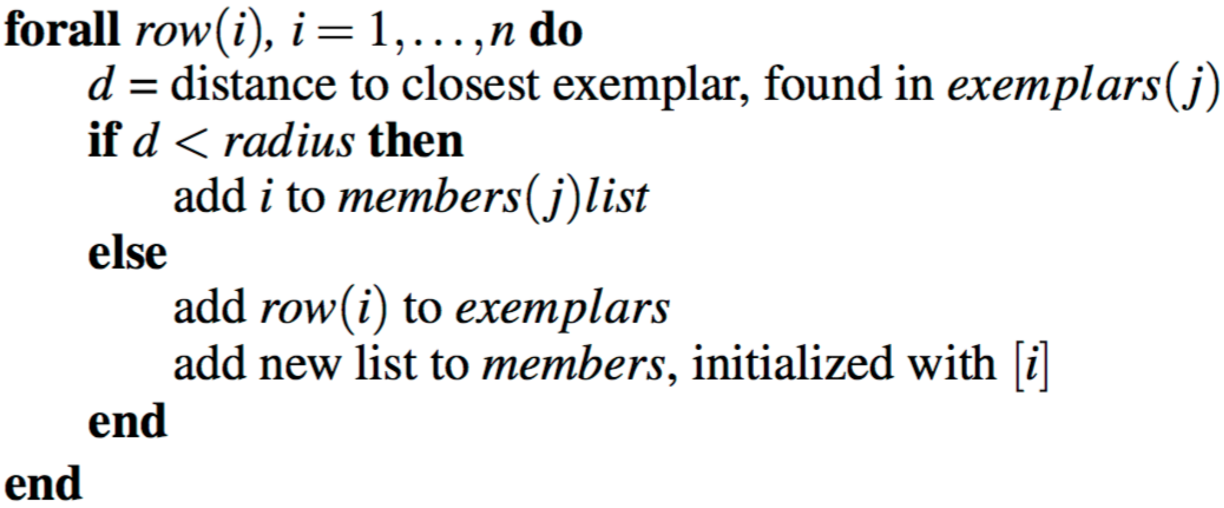
6. Initialize $exemplars$, a list of exemplars with initial entry $[row(1)]$.

7. Initialize $members$, a list of lists with initial entry [1]; each $exemplar$ will eventually have its own list of affiliated member indices.

Outliers

HDoutliers

8. Now do one pass through X. This is Hartigan’s *leader* algorithm.



```
forall $row(i)$, $i = 1, \ldots, n$ 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
```

Outliers

HDoutliers

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).

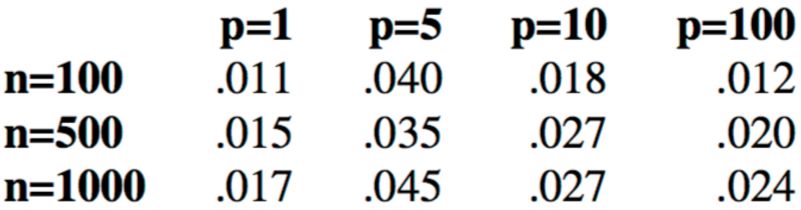
11. For any *exemplar* that is significantly far from all the other *exemplars* based on this cut point, flag all entries of *members* corresponding to *exemplar* as outliers.

Outliers

HDoutliers

False positives

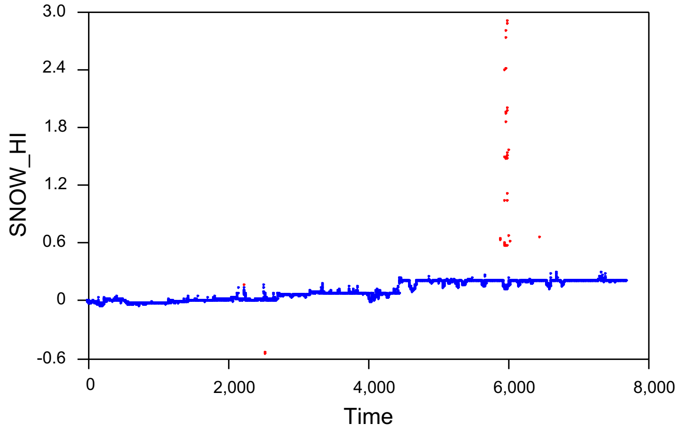
Empirical level of HDoutliers test based on null model with Gaussian variables and critical value α = .05.



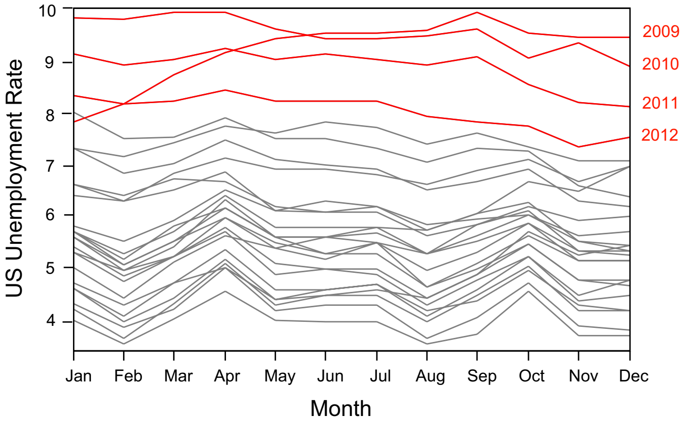
	p=1	**p=5**	**p=10**	**p=100**
n=100	.011	.040	.018	.012
n=500	.015	.035	.027	.020
n=1000	.017	.045	.027	.024

Outliers

Time Series



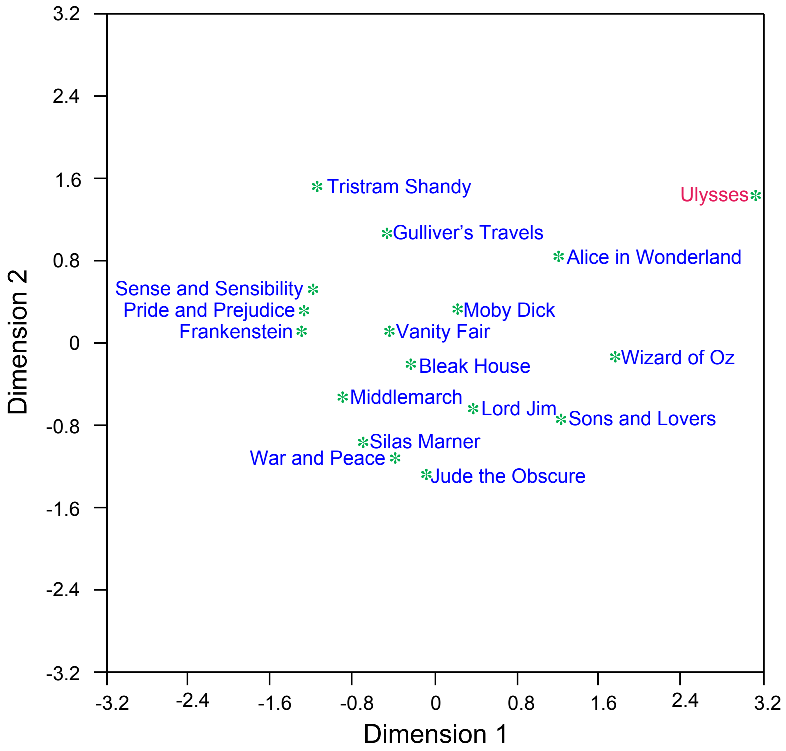
Kernel smoother with a biweight function on the running mean. The data are measurements of snowfall at a Greenland weather station.



Each series is a row in the data matrix. For *n* series on *p* time points, we have a *p*-dimensional outlier problem. This figure shows series over 20 years of the Bureau of Labor Statistics Unemployment data.

Outliers

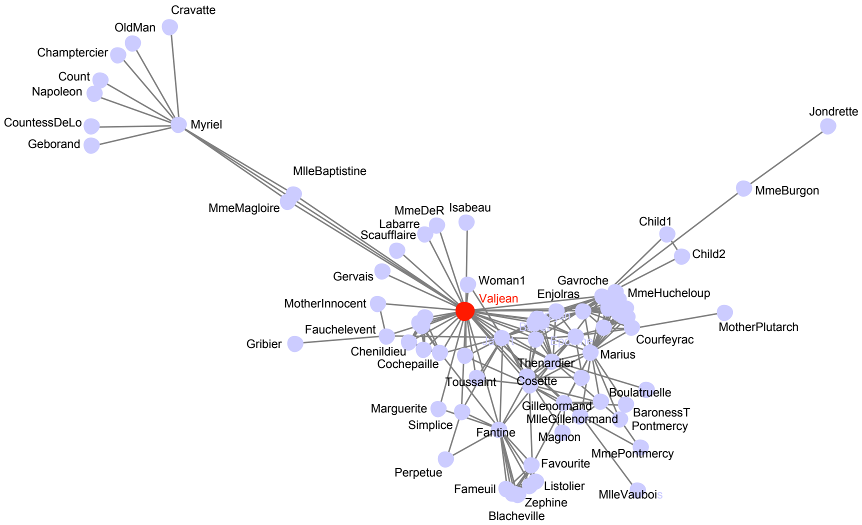
Text analysis of documents



Bag-of-words model. Collect all the words in the documents, stem them to resolve variants, remove stopwords and punctuation, and then apply the tf-idf measure. These data required the use of random projections. Before projection, there were 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.

Outliers

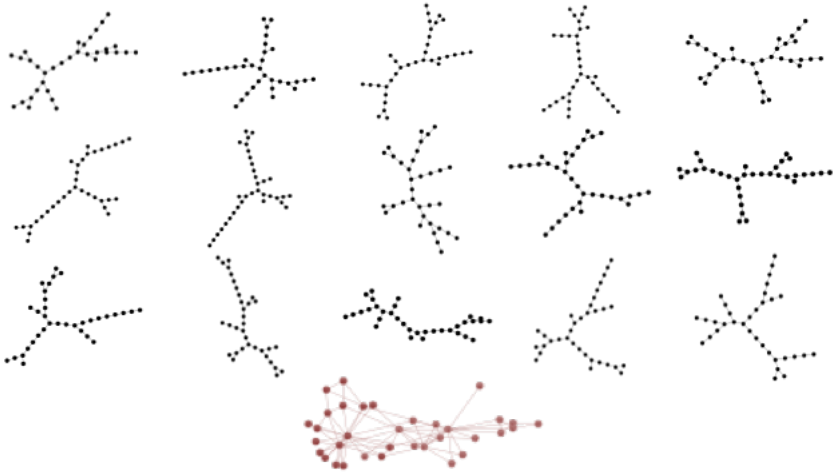
Graphs



This figure shows a graph layout for the Les Miserables dataset. 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.

Outliers

Graphs



1 Compute the adjacency matrix for each graph.
2 Compute the eigendecomposition of the Laplacian matrix.
3 Reorder the adjacencies using this decomposition.
4 Convert the adjacencies above the diagonal to a string.
5 Compute the Levenshtein distances between pairs of strings.
6 Find the nearest-neighbor distances.
7 Subject them to the HDoutliers algorithm.

This figure shows an example using the Karate Club graph. 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.

Outliers

References

Barnett, V. and Lewis, T. (1978). *Outliers in Statistical Data*. John Wiley & Sons.

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.

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

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

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

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

Outliers

Thanks!

Chris Fraley implemented the R package from my Java code.

fraley@u.washington.edu@

leland@h2o.ai

https://www.cs.uic.edu/~wilkinson/

Email me or Chris for tech support: