Published in final edited form as:

Health Innov Point Care Conf. 2014 October ; 2014: 300–303. doi:10.1109/HIC.2014.7038934.

Multivariate Voronoi Outlier Detection for Time Series

Chris E. Zwilling¹ and Michelle Yongmei Wang²

Michelle Yongmei Wang: ymw@illinois.edu

¹Department of Psychology, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA

²Departments of Statistics, Psychology and Bioengineering, and the Beckman Institute, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA

Abstract

Outlier detection is a primary step in many data mining and analysis applications, including healthcare and medical research. This paper presents a general method to identify outliers in multivariate time series based on a Voronoi diagram, which we call Multivariate Voronoi Outlier Detection (MVOD). The approach copes with outliers in a multivariate framework, via designing and extracting effective attributes or features from the data that can take parametric or nonparametric forms. Voronoi diagrams allow for automatic configuration of the neighborhood relationship of the data points, which facilitates the differentiation of outliers and non-outliers. Experimental evaluation demonstrates that our MVOD is an accurate, sensitive, and robust method for detecting outliers in multivariate time series data.

I. Introduction

Outlier detection in healthcare delivery and management plays an important role, though it is probably underappreciated and not well understood by many health practitioners. An outlier is a data point inconsistent with the behavior exhibited by most of the other data values [1] and, from the healthcare point of view, probably reflects the need for heightened vigilance, if not full-fledged intervention. For instance, an abnormally high glucose reading for a diabetic patient is an outlier which probably requires action or an elevated or depressed reading from a bioelectronic neural sensor might trigger the release of an electronic pulse into the human nervous system.

While outliers exist in all types of data, and detecting outliers and robust approaches remain a critical and diversified area of research [2, 3, 4], the focus of this paper is the detection of outliers in time series data. Univariate time series observations only measure a single variable independently and examples include measuring the respiratory rate every few seconds or measuring body weight each year. Measuring two or more variables simultaneously—such as systolic and diastolic blood pressure—results in multivariate time series data. Advanced brain imaging technologies like EEG or fMRI could have tens, hundreds of even thousands of variables measured simultaneously.

In this paper, we propose a general outlier detection method for multivariate time series data based on the mathematical principles of Voronoi diagrams. It is general because different attributes or features can be extracted from the data for Voronoi diagram construction. These attributes or features can be designed based on the nature of the data and the outliers, and can take parametric or nonparametric forms. This has the potential to increase the accuracy and precision of outlier detection for specific application problems.

The organization of this paper is as follows. First we review existing related outlier detection methods, focusing especially on the Multivariate Least Trimmed Squares (MLTS), with the Minimum Covariance Determinant (MCD) as the estimator. We also provide a brief introduction to Voronoi diagrams. Next, in Section III, we present our Multivariate Voronoi Outlier Detection (MVID) method for the time series data based on Voronoi diagrams, integrating the merits of MLTS. Performance evaluation of our approach is demonstrated in Section IV. Conclusions and discussions are given in Section V.

II. Background and Related Work

Time series data can contain one or more types of outliers, including additive or innovative outliers, level shifts and temporary variance changes [5]. This paper focuses on additive outliers to demonstrate the idea of our method. Additive outliers are the result of adding a value of some magnitude to some of the data points.

A. Univariate Time Series Outlier Detection

Most traditional outlier detection methods for time series data only consider univariate cases. Even in bivariate or multivariate cases, a common practice is to use these same univariate techniques on each component series [5]. This potentially creates problems because an outlier for one observed time point of one variable might affect the same observed time point for one or more other variables. In addition, ignoring the multivariate structure can create difficulties in detecting multivariate outliers. Multivariate approaches, which use the information across all component time series, have a much better chance of identifying these anomalies.

B. Least Trimmed Squares and Minimum Covariance Determinant

The Multivariate Least Trimmed Squares (MLTS) [6] is a robust approach for estimating the vector autoregressive model while handling outliers in the data. It relies on a popular statistical procedure, called the Minimum Covariance Determinant (MCD) that performs fast and efficient statistical outlier detection [7, 8]. MCD finds h observations (out of n) whose covariance matrix has the lowest determinant.

Assume we have a data set with p variables with $i = 1 \dots n$ observations. Take a subset of these n observations, h , where h is constrained by $[(n+p+1)/2] \leq h \leq n$. The MCD algorithm selects a subset of randomly selected observations of size h , computes the mean T_I , the variance/covariance matrix S_I and then determines the distance d for each data point x_i in n :

$$d(i) = \sqrt{(\mathbf{x}_i - \mathbf{T}_1)^t \mathbf{S}_1^{-1} (\mathbf{x}_i - \mathbf{T}_1)}. \quad (1)$$

The obtained distances from (1) are next sorted from smallest to largest, and the h smallest are retained as h_2 . From h_2 a new mean $\mathbf{T}_2$ and variance/covariance matrix $\mathbf{S}_2$ are computed. The relationship expressed in (2) now holds between the determinant ($\det$) of the two variance/covariance matrices $\mathbf{S}_1$ and $\mathbf{S}_2$:

$$\det(\mathbf{S}_2) \leq \det(\mathbf{S}_1). \quad (2)$$

These steps are repeated 500 times (when n is less than 500), where a different subset h is chosen for each iteration. The subset yielding the smallest overall determinant is then used for further statistical analysis.

The MCD framework has been extended to a regression framework for time series data [9]. The difference between the MCD and MLTS is that the latter obtains the residuals through a least squares regression. Then, instead of determining the h from the raw data as in the MCD, the MLTS selects the h observations with the smallest determinant of the covariance matrix of the residuals. So in the MLTS the joint variability (of the predictor and response variables) is modeled by using residuals. Also, by design, the MLTS method returns a vector of 1's and 0's, where a 0 indicates whether an observation is an outlier. The MLTS method requires the user to set a parameter, alpha, which reflects the user's belief about the number of outliers in the dataset.

C. Voronoi Diagram

Our new method requires a Voronoi diagram, which is composed of Voronoi cells [10]. Assume we have a set S of n points, $p_1 \dots p_n$ in a plane. Let $V(p_i)$ denote a Voronoi cell, which is a subdivision of the plane where the set of points q are closer or as close to p_i than to any other point in S . Formally:

$$V(p_i) = \{q \mid \text{dist}(p_i, q) \leq \text{dist}(p_j, q), \forall j \neq i\}, \quad (3)$$

where dist is the Euclidean distance function. The set of all Voronoi cells for all n points comprises a Voronoi diagram.

Figure 1 shows part of a Voronoi diagram. The plane is decomposed into n convex polygonal regions (with $n = 6$ here), one for each p_i . Vertices are called *Voronoi vertices* and *Voronoi edges* are defined as the boundaries between two Voronoi cells. The boundaries of a Voronoi cell $V(p_i)$ cannot exceed $n-1$ edges. Three theorems apply to Voronoi diagrams:

Theorem 1: Every nearest neighbor of p_i defines an edge of the Voronoi polygon $V(p_i)$.

Theorem 2: Every edge of the Voronoi polygon $V(p_i)$ defines a nearest neighbor of p_i .

Theorem 3: For $n \geq 3$, a Voronoi diagram on n points has at most $2n-5$ vertices and $3n-6$ edges.

III. Research Methodology

The Voronoi Outlier Index (VOInd) used in our Multivariate Voronoi Outlier Detection (MVOD) method is based upon Voronoi nearest neighbors. For a point p_i of set S , the nearest neighbors of p_i defined by the Voronoi polygon $V(p_i)$ are the Voronoi nearest neighbor of p_i denoted as $V_{NN}(p_i)$. In Figure 1 the nearest Voronoi neighbors to point p_1 are p_2, p_3, p_4, p_5 and p_6 . For each point in the data set, our method uses the nearest neighbors to compute an index (i.e. VOInd) of how likely that point is an outlier. It is multivariate because it aggregates information across all individual time series, thus retaining features which might be common to the entire interlocking set of variables.

Our method is based upon the geometric principles of Voronoi diagrams for defining the neighborhood relationship of the data points and this facilitates the assignment of group or data membership (i.e. outliers and non-outliers). Construction of a two dimensional Voronoi diagram requires two coordinates for each data point. Based on the nature of the data and the nature of the outliers to be identified, we can embed their attributes into the coordinates via extracting different valid features from the data. In this paper, we present one such case of the MVOD framework for feature extraction; but many others are also possible, including nonparametric forms. Figure 2 overviews the process and the rest of this section explains the steps in more detail.

Feature value for the x-coordinate

In order to determine how a single observation (at the same time point, across all time series under consideration) affects the covariance matrix (which is a measure of inter-relationship among the individual time series), we remove a given point from the data set, compute the covariance matrix *without that point* and then take the determinant of this covariance matrix to yield a single feature value which shows the effect of *removing* that observation from all the time series. This determinant is known as the generalized variance and can be interpreted as a volume. If we have outlying observations, the volume will be larger. But if we remove those outlying observations, the volume will be smaller.

Feature value for the y-coordinate

This is a two-step process. In Step 1, we take the absolute value of all time series points for all variables, which provides some information about the total magnitude each time series contributes. However, sometimes, magnitude alone is not sufficient for outlier detection as some data may have less extreme values than those data with the largest magnitudes but are actually outliers for the data [1]. One way to address this issue is by using the residual to calculate the feature value for the y-coordinate. So Step 2 consists of fitting an appropriate model to the multivariate time series data and then computing the residuals. In this paper, we estimate the parameters for a Multivariate Vector Auto-Regressive (MVAR) model because our simulated data are generated from this model. Once these residuals are obtained for each time series, they are squared and then summed across all time series. Finally, the feature value for the y-coordinate is determined by multiplying together the results of Step 1 and Step 2. Specifically, let y denote a vector of time series data sets, y_i denote the i th time series

in $\mathbf{y}$ (which all have length n), and r denote the residual after fitting the MVAR model. Then, we compute this feature value as:

$$\sum_1^n \text{abs}(\mathbf{y}) \times \sum_1^n r^2. \quad (4)$$

Although a regression model is used here in Step 2 to extract the feature value, in fact, our method does not require this model. With either Step 1 or Step 2 alone, we will have a corresponding nonparametric or parametric basis, both of which could be suitable for different applications or datasets. While we did not consider a non-parametric example in this paper, one could, for instance, use the absolute value of the time series as one coordinate and the difference in magnitude between adjacent observations as the second coordinate, thus yielding a non-parametric basis for outlier detection.

Voronoi Outlier Index (VOInd)

Given the x-coordinates and y-coordinates for the data using the above procedures, we construct the Voronoi diagram based on Section II.C and compute a Voronoi Outlier Index (VOInd) for each time series point. The VOInd for point p_i has as its numerator the sum of the Euclidean distance ($dist$) between each point and all its neighbors. This is divided by the denominator term which is the number of neighbors, yielding an average density. That is:

$$VOInd(p_i) = \frac{\sum_{o \in V_{NN}(p_i)} dist(p_i, o)}{|V_{NN}(p_i)|}. \quad (5)$$

Note that Qu [11] used a Voronoi outlier factor as the index which, however, was completely univariate in nature, since the x- and y-coordinates were based on a univariate time series. One of our primary motivations for this study is to create a novel and general MVOD method, which can detect outliers in time series data in a multivariate framework with multiple, interlocking sets of variables.

IV. Experimental Evaluation

A. Simulation Setup and Data Generation

For each analysis, 5 multivariate autoregressive time series, each containing 100 observations, were simulated 25 times using published Matlab code [12]. The time series were generated using a Gaussian process with mean 0 and standard deviation 1. The variance/covariance matrix contained all 1's on the diagonal and all 0's on the off-diagonals. A total of 12 different unique multivariate time series were constructed, each with differing numbers of outliers and strength/magnitude of the outliers. 5, 10 or 15 outliers were introduced into a time series and the magnitude of those outliers was 1, 2, 3, 4 or 5. All combinations of number of outliers and outlier magnitude were constructed; but they were never mixed. For instance, if we introduced 5 outliers of magnitude 3 into a simulated time series, only 5 outliers of magnitude 3 were used for all 25 simulated time series in that set. The observation to which the outliers were introduced into the time series was always determined randomly. Once the observations had been selected for outlier introduction, the

same number of outliers for the given magnitude was added or subtracted (if the original observation was negative) to *each* of the five components of the multivariate time series.

B. Validation Criteria and Procedure

We validated and compared the performance of our new Voronoi Outlier Detection (MVOD) method with the MLTS, using True and False Positive Rates (TPR and FPR) as defined in Table 1.

The alpha parameter in the MLTS method determines both the size of the subset to use as well as a critical value in a chi-square distribution. If an observation is greater than this threshold in the chi-square distribution, then the MLTS method flags the observation as an outlier. However, it is critical to note that a one-to-one correspondence does not exist between the alpha value chosen, and the number of outliers flagged. For instance, one could set alpha at 0.10 but only have 2 out of 100 observations flagged as outliers. Partly for this reason we considered a range of alpha values and then averaged across this range to fairly compare with the MVOD method. For all simulated time series, we considered alpha between 0.01 and 0.20.

In the results presented next, we obtained the TPR and FPR for the two methods in the following way. For a given number of outliers with a specific outlier magnitude, we averaged a total of five cases. The five cases averaged *always* included the threshold (MVOD) or alpha value (MLTS) corresponding with the number of outliers, but also contained the preceding four cases as well. For instance, in the 10 outlier case, we took the results for threshold = 10 (MVOD), as well as thresholds of 9, 8, 7 and 6. In the corresponding MLTS case, we would have taken alpha = 0.10, 0.09, 0.08, 0.07 and 0.06. The TPR and FPR for each of these five cases for each method were averaged to obtain the values shown in Table 2.

C. Results

Table 2 and Figure 3 show that in terms of the TPR, our method outperforms the MLTS when the outlier strength is low (i.e. magnitudes of 1 and 2) and has slightly better performance than the MLTS for medium outlier strength (i.e. magnitude of 3), while the two approaches are comparable when the outlier strength is high (i.e. magnitude of 4 and 5). As evident from Table 2, for the FPR, the two methods have similar behavior, with negligible difference on the order of 10^{-2} . Additionally, the number of outliers (5, 10 or 15) does not have an obvious effect on either of the methods. In summary, the experiments demonstrate that our MVOD method can work effectively and accurately in detecting the outliers from the multivariate time series data. Compared to MLTS, the MVOD is more sensitive in detecting the small magnitude outliers. Furthermore, both our MVOD and the MLTS work reasonably well for a wide range of contamination levels. That is, both methods are quite robust to the number of outliers in the dataset.

V. Conclusions

We have presented a new and general method for outlier detection in time series data based on a Voronoi diagram. There are several key advantages of the approach. First, it copes with

outliers in a multivariate framework by accounting for the multivariate structure in the data. Second, it is flexible in extracting valid features for differentiating outliers from non-outliers, in the sense that we have the option of using or not using a parametric model (such as a regression model, as in this paper). Lastly, Voronoi diagrams capture the geometric relationship embedded in the data points. Experimental results show that our MVOD method can lead to accurate, sensitive, and robust identification of outliers in the multivariate time series.

A future direction is to test the performance of the method for other types of outliers, besides the additive outliers that have been investigated in this paper. We also plan to further validate the approach by comparing with a univariate version of the method.

Acknowledgments

Research supported by NIH grant K25AG033725.

References

1. Pearson, RK. Exploring Data in Engineering, the Sciences and Medicine. Oxford University Press; 2011.
2. Rousseeuw PJ, Hubert M. Robust statistics for outlier detection. Wiley interdisciplinary reviews: Data Mining and Knowledge Discovery. 2011; 1:73–79.
3. Qi GJ, Aggarwal CC, Huang TS. On Clustering heterogeneous social media objects with outlier links. Proceedings of the fifth ACM International Conference on Web Search and Data Mining. 2012:553–562.
4. Muler N, Yohai VJ. Robust estimation for vector autoregressive models. Computational Statistics and Data Analysis. 2013; 65:68–79.
5. Tsay RS, Pena D, Pankratz A. Outliers in multivariate time series. Biometrika. 2000; 87:789–804.
6. Rousseeuw PT, Aelst SV, Driessen KV, Agullo J. Robust multivariate regression. Technometrics. 2004; 46(3):293–305.
7. Rousseeuw PJ, Driessen KV. A fast algorithm for the minimum covariance determinant estimator. Technometrics. 1999; 41(3):212–223.
8. Hubert M, Debruyne M. Minimum covariance determinant. Wiley interdisciplinary reviews: Computational statistics. 2010; 2:36–43.
9. Croux C, Joossens K. Robust estimation of the vector autoregressive model by a least trimmed squares procedure. Proceedings in Computational Statistics. 2008:489–501.
10. Preparata, FP.; Shamos, MI. Computational Geometry-An Introduction. Springer; Heidelberg: 1985.
11. Qu J. Outlier detection based on Voronoi diagram. Proceedings of the ADMA International Conference on Advanced Data Mining and Applications. 2008:516–523.
12. Neumaier A, Schneider T. Algorithm 808: ARfit - A Matlab package for the estimation of parameters and eigenmodes of multivariate autoregressive models. ACM Transactions Mathematical Software. 2001; 27:58–65.

Figure 1. A subset of Voronoi cells from a Voronoi diagram

Figure 2. Flow chart of our MVOD procedure for outlier detection

Figure 3. True Positive Rate (TPR, y-axis) for MVOD and MLTS for 5 Outliers (top panel), 10 Outliers (middle panel), and 15 Outliers (bottom panel) with outlier magnitudes of 1, 2, 3, 4 or 5 (x-axis).

Table 1

Definition of True (TPR) and False (FPR) Positive Rate.

		Outlier in data? (Gold Standard)		Definition of TPR and FPR
		<i>Yes</i>	<i>No</i>	<i>True Positive Rate (TPR)</i> = $TP/(TP + FN)$
Detected Outlier?	<i>Yes</i>	TP	FP	<i>False Positive Rate (FPR)</i> = $FP/(FP + TN)$
	<i>No</i>	FN	TN	

True and False Positive Rates for MVOID and MLTS With 5, 10, or 15 Outliers of Magnitude 1, 2, 3, 4, or 5.

Table 2

		True Positive Rate		False Positive Rate				
		<i>Number of Outliers</i>						
		5	10	15	5	10	15	
<i>Magnitude</i>	1	MVOD	0.52	0.52	0.54	0.037	0.065	0.094
		MLTS	0.21	0.37	0.32	0.012	0.028	0.047
	2	MVOD	0.91	0.79	0.78	0.025	0.041	0.056
		MLTS	0.63	0.61	0.73	0.002	0.012	0.011
	3	MVOD	0.96	0.83	0.86	0.023	0.037	0.045
		MLTS	0.93	0.78	0.80	0.004	0.006	0.005
	4	MVOD	0.96	0.86	0.88	0.023	0.034	0.042
		MLTS	0.97	0.87	0.85	0.002	0.002	0.003
	5	MVOD	0.96	0.86	0.90	0.023	0.034	0.039
		MLTS	0.96	0.90	0.87	0.002	0.002	0.002