Chapter 11

Multivariate Computing and Robust Estimating for Outlier and Novelty in Data and Imaging Sciences

Michelle Yongmei Wang and Chris E. Zwilling

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/59750

1. Introduction

Data science is an evolutionary step in interdisciplinary fields incorporating computer science, statistics, engineering, and mathematics. At its core, data science involves using automated and robust approaches to analyze massive amounts of data and to extract informative knowledge from them. Data science transforms traditional ways of analyzing problems and creates powerful new solutions. Diverse computational and analytical techniques contribute to data science. In this chapter, we review and also propose one type of data mining and pattern recognition strategy that has been under development in multiple disciplines (e.g. statistics and machine learning) with important applications ---- outlier or novelty detection [1-4].

In biomedical engineering, data science can make healthcare and medical imaging science not only more efficient but also more effective for better outcomes and earlier detection. Outlier and novelty detection in these domains plays an essential role, though it may be underappreciated and not well understood by many biomedical practitioners. From the healthcare point of view, an outlier probably reflects the need for heightened vigilance, if not full-fledged intervention. For example, an abnormally high glucose reading for a diabetic patient is an outlier which may require action. In high-dimensional medical imaging, developing automated and robust outlier detection methods is a critical preprocessing step for any subsequent statistical analysis or medical research.

An exact definition of an outlier or novelty typically depends on hidden assumptions regarding the data structure and the associated detection method, though some definitions are general enough to cope with varieties of data and methods. For example, outliers can be considered as patterns in data that do not conform to a well-defined notion of "normal" behavior, or as observations in a data set which appear to be inconsistent with the remainder of that set of data. Figure 1 shows outliers in a 2-dimensional dataset. Since most of the

observations fall into clusters N1 and N2, they are two "normal" regions; while points in region O1 as well as points o2 and o3 are outliers (in red), due to their sufficiently far distance from the "normal" regions. Identifying observations inconsistent with the "normal" data, or detecting previously unobserved emergent or novel patterns is commonly referred to as *outlier detection* or *novelty detection* [5, 6]. The distinction between novel patterns and outliers is that the novel patterns are often incorporated into the "normal" model after being detected whereas outliers are typically removed or corrected. This chapter aims to consider both detection schemes and sometimes treat them interchangeably for some general purposes.

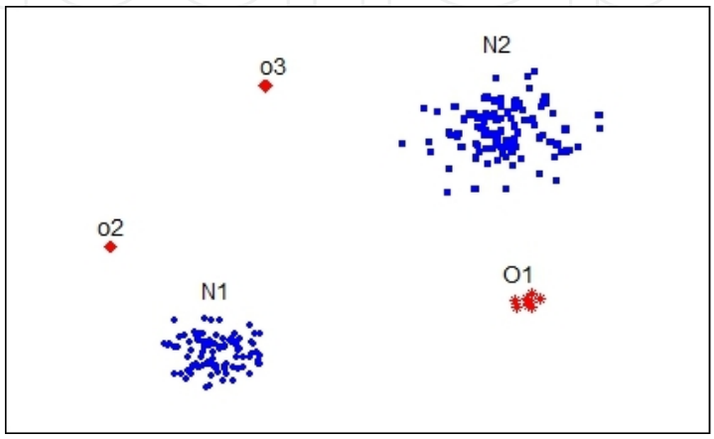

Figure 1. An example of outliers in a 2-dimensional dataset.

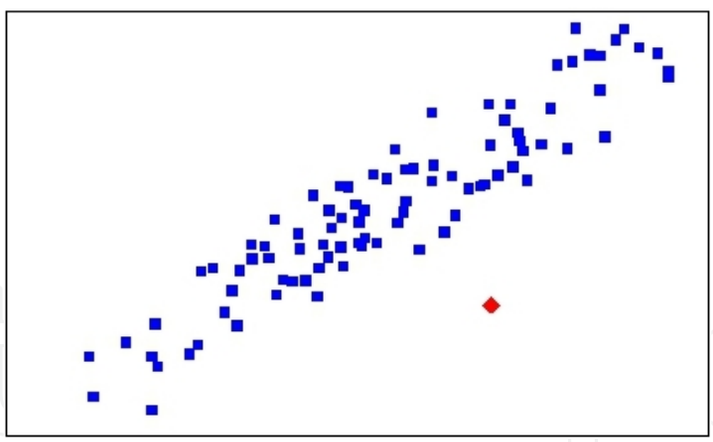

Figure 2. A 2-dimensional dataset with a multivariate outlier (lower right diamond in red).

Outlier and novelty detection methods can be divided into *univariate and multivariate* approaches [7-10]. The early univariate methods typically assume a known data distribution, for example, independently and identically distributed (i.i.d). In addition, many tests for detecting univariate outliers further assume the distribution parameters and the outlier types are known. However, these assumptions may be violated in real applications. Moreover, in many situations multivariate outliers can not be identified when each variable is examined independently. Multivariate analysis is usually required in these cases for precise outlier detection, which

allows for interactions over different variables to be taken into account within the class of data. Figure 2 illustrates 2-dimensional data points, with the lower right observation (in red) a clear multivariate outlier but not a univariate one. When analyzing each measure separately with respect to the spread of values along the two dimensions, they are close to the center of the univariate distributions. Therefore, the relationships between the two variables shall be considered when testing for outliers, leading to multivariate methods that are the focus of this chapter.

Another related topic is robust statistics for estimation that can handle outliers or at least is less sensitive to the influence of outliers. Robust statistics perform well with data drawn from a wide range of probability distributions, especially for distributions that are not normally distributed. Robust statistical methods have been developed for many common problems, such as estimating data properties including location and scatter or estimating model parameters as in regression analysis [10, 11]. One motivation is to produce statistical methods that are not unduly affected by outliers. Another motivation is to provide methods with good performance when there are small departures from a parametric distribution. A typical procedure or example of the former case is for multivariate estimation of location and covariance as well as for multivariate outlier detection. In this case, as a first step, the approaches often try to search for a minimum number of observations with a certain degree of confidence being outlier-free. Based on this starting subset, location and covariance can be estimated robustly. In a second step, outliers can be identified through computing the observations' distances with respect to these initial estimates.

In this chapter, we review and also propose statistical and machine learning approaches for outlier and novelty detection, as well as robust methods that can handle outliers in data and imaging sciences. In particular, robust statistical techniques based on the Minimum Covariance Determinant (MCD) are introduced in Section 2, which include a classical and fast computation scheme of MCD and a few robust regression strategies. We present our newly developed multivariate Voronoi outlier detection (MVOD) method for time series data and some preliminary results in Section 3. This approach copes with outliers in a multivariate framework via designing and extracting effective attributes or features from the data; Voronoi diagrams allow for automatic configuration of the neighborhood relationship of the data points, facilitating the differentiation of outliers and non-outliers. Section 4 reviews varieties of machine learning methods for novelty detection, with a focus on probabilistic approaches. In Section 5, we present some existing and new technologies related to outliers and novelty in the area of imaging sciences. Section 6 provides concluding remarks of the chapter.

2. Robust statistical methods using Minimum Covariance Determinant (MCD)

The Minimum Covariance Determinant (MCD) estimator is a highly robust estimator of multivariate location and scatter. Since estimating the covariance matrix is the cornerstone of

many multivariate statistical methods, the MCD has also been used to develop robust and computationally efficient multivariate techniques.

2.1. MCD and its fast computing algorithm

Given a dataset consisting of p variables and n observations, i.e. a $n \times p$ data matrix, we can represent this multivariate data as $X = (x_1, \ldots, x_n)'$, where x_i, for $i = 1, \ldots, n$, is the ith observation and a p-dimensional vector. A classical distance measure, Mahalanobis distance (MD), is given in Equation (1); it only depends on the sample mean ($\hat{\mu}_{MD}$) and sample covariance matrix ($\hat{\Sigma}_{MD}$), both of which are computed from the entire set of data.

$$MD(x) = \sqrt{(x - \hat{\mu}_{MD})' \hat{\Sigma}_{MD}^{-1} (x - \hat{\mu}_{MD})} \tag{1}$$

A point with a larger Mahalanobis distance will lie further away from the center of the data cloud than a point with a smaller Mahalanobis distance. A robust distance (RD) measure is achieved if we substitute the MCD estimate of mean ($\hat{\mu}_{MCD}$) and covariance ($\hat{\Sigma}_{MCD}$) into Equation (1), which yields Equation (2).

$$RD(x) = \sqrt{(x - \hat{\mu}_{MCD})' \hat{\Sigma}_{MCD}^{-1} (x - \hat{\mu}_{MCD})} \tag{2}$$

The classical estimates can be sensitive to outliers, while the MCD estimate is robust [8, 12, 13]. The MCD relies on a subset of the total observations. Choosing this subset makes the algorithm robust because it is less sensitive to the influence of outlying points. Figure 3 illustrates the difference between these two estimates; it is a scatterplot of the distances for an example dataset with 70 observations and 2 variables (i.e. $n = 70$, $p = 2$). The two ellipses are two outlier thresholds, determined by the 0.975 chi-square quantile with 2 degrees of freedom when the classical and robust estimates are used, respectively. The dashed blue ellipse marks off the 97.5% outlier threshold for the classical Mahalanobis distance, suggesting that two observations lying beyond the ellipse are outliers. The 97.5% outlier threshold for the robust distance measure is marked off by the solid red ellipse, suggesting ten points are outliers.

The MCD has a user-determined parameter, h, which specifies the size of the subset of data to base the estimate upon. It is constrained by $[(n + p + 1)/2] \leq h \leq n$. The h observations are chosen such that the determinant of the sample covariance matrix is minimal (but not minimized in the formal sense, because it relies on a sampling algorithm instead of a loss function). The MCD is optimally designed for elliptically symmetric unimodal distributions, such as the commonly encountered multivariate normal distribution. The MCD is most robust when $h = [(n + p + 1)/2]$. But this causes low efficiency [14] (at least for normal probability distributions), which can be increased (while retaining high robustness) by applying reweighted estimators [15, 16]. Robust statistical estimators are commonly evaluated both on their breakdown value and influence functions. The MCD is a high breakdown estimator and its influence function appears bounded, which is desirable. An alternative strategy that employs

http://dx.doi.org/10.5772/59750

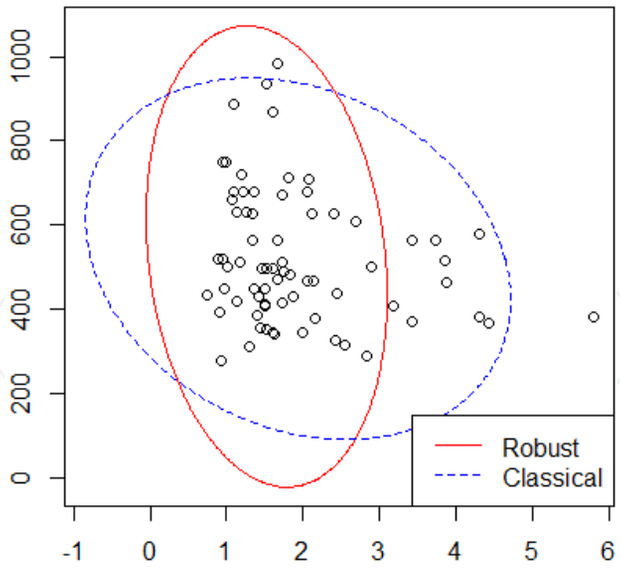

Figure 3. Outlier thresholds, as represented by ellipses, based on a classical and a robust scheme.

Delaunay triangulation to identify a robust outlier-free subsample in an adaptive way was presented in [17].

Computing the exact MCD is possible but computationally difficult, as it requires the evaluation of all $\binom{n}{h}$ subsets of size h. Even though the MCD is a powerful robust estimator, it has only become widely used since the development of the so-called Fast-MCD algorithm [18] which we summarize below. Assume we have a dataset $X = (x_1, \ldots, x_n)'$ and let $H_1 \subset \{1 \ldots n\}$ represent a h-subset of length constrained by $[(n + p + 1)/2] \leq h \leq n$. Denote this first h-subset as H_1 and it is randomly chosen from the entire dataset. Compute the mean $\hat{\mu}_{MCD,1}$ and covariance matrix $\hat{\Sigma}_{MCD,1}$ of H_1, as well as the determinant of $\hat{\Sigma}_{MCD,1}$, denoted as $det(\hat{\Sigma}_{MCD,1})$. Then compute the distance of all n observations in the entire dataset (and not just the h comprising the initial subset) using Equation (2). Next, these distances are ordered from smallest to largest. Retain an equivalent number of observations from this ordering as chosen in the initial h-subset; but instead of being chosen arbitrarily as in the initial subset, these are chosen such that they have the smallest distances as defined by the order statistics. Call this subset of observations H_2, and compute $\hat{\mu}_{MCD,2}$, $\hat{\Sigma}_{MCD,2}$ and $det(\hat{\Sigma}_{MCD,2})$. Now Equation (3) must be true:

$$det\ (\hat{\Sigma}_{MCD,2}) \leq det(\hat{\Sigma}_{MCD,1}) \tag{3}$$

Going from H_1 to H_2 is called a C-step for "Concentration step", because the algorithm concentrates on the h observations with the smallest distances and $det(\hat{\Sigma}_{MCD,2})$ is more concentrated (or equivalently, has a smaller determinant). This C-step is repeated numerous or sufficient times, with each iteration using a different initial h-subset. The 10 subsets that yield the smallest determinants overall are retained and further concentrated until convergence is met.

2.2. Robust multivariate regression and Multivariate Least-Trimmed Squares (MLTS) estimator

Section 2.1 introduced the robust MCD estimator and showed how the MCD can be computed efficiently. In this section, we review different frameworks for applying the MCD estimator to multivariate regression. These methods offer robust alternatives to standard multiple regression analysis.

We first look at robust multivariate regression in [19]. Suppose we have a full dataset of predictors and responses containing no outliers; computing the regression parameter estimates from the full dataset using a least squares procedure will yield accurate results. With outliers present in the dataset, the MCD is used to search for a subset of size h whose covariance matrix has the smallest determinant with h constrained by $[(n + p + q + 1)/2] \leq h \leq n$, where p and q are respectively the numbers of variables for the predictor and response matrices, and n is the number of observations. Then, using this subset h, the sample mean and covariance estimates are calculated, which would allow for accurate estimation of the regression coefficients and covariance matrix of the errors in the presence of outliers.

Different from the above robust multivariate regression, the multivariate least trimmed squares (MLTS) estimator in [20] first fits a regression model to the subset of data and then calculates the covariance matrix of the residuals. The estimator is defined by minimizing a trimmed sum of squared Mahalanobis distances, and can be computed by a fast algorithm. Let us consider the classical multivariate regression framework. Assume we have a sample of data defined as $Z_n = \{(x_i, y_i); i = 1, \ldots, n\}$ and let $X = (x_1, \ldots, x_n)'$ denote the design (or predictor) matrix and $Y = (y_1, \ldots, y_n)'$ denote the response matrix. The regression model is:

$$Y = X\beta + \varepsilon \tag{4}$$

The classical least squares estimator for the regression parameter is given by:

$$\hat{\beta}_{LS} = (X'X)^{-1}X'Y \tag{5}$$

and the classical estimator of the scatter matrix is:

$$\hat{\Sigma}_{LS} = \frac{1}{n - p}(Y - X\hat{\beta}_{LS})'(Y - X\hat{\beta}_{LS}) \tag{6}$$

These classical least squares estimators are sensitive to outliers. A robust alternative to these estimators based on the residuals is achieved as below. For any $\beta \in \mathbb{R}^{p+q}$, let $r_i(\beta) = y_i - \beta' x_i$ denote the residuals from the fitted regression model. Furthermore let $\mathcal{H} = \{H \subset \{1, \ldots, n\} \mid \#H = h\}$ be the collection of all subsets of size h. For any $H \in \mathcal{H}$ denote $\hat{\beta}_{LS}(H)$ the least-squares fit solely on the observations $\{(x_j, y_j); j \in H\}$. In addition, for all $H \in \mathcal{H}$ and $\beta \in \mathbb{R}^{p+q}$ denote the covariance matrix of the residuals with respect to the fit β, belonging to subset H as:

$$\mathrm{cov}(H,\beta)=\frac{1}{h}\sum_{j\in H}\left(r_j(\beta)-\bar{r}_H(\beta)\right)\left(r_j(\beta)-\bar{r}_H(\beta)\right)' \tag{7}$$

where $\bar{r}_H(\beta)=\frac{1}{h}\sum_{j\in H} r_j(\beta)$. If we let $\hat{\Sigma}_{LS}(H)=\mathrm{cov}(H,\hat{\beta}_{LS}(H))$ for any $H\in\mathcal{H}$, the MLTS estimator is defined as:

$$\hat{\beta}_{MLTS}(Z_n)=\hat{\beta}_{LS}(\hat{H}) \tag{8}$$

where $\hat{H}\in\underset{H\in\mathcal{H}}{\mathrm{argmin}}\ det\ \hat{\Sigma}_{LS}(H)$. The covariance of the errors can be estimated by

$$\hat{\Sigma}_{MLTS}(Z_n)=c_\alpha\hat{\Sigma}_{LS}(\hat{H}) \tag{9}$$

where c_α is a consistency factor. The observations corresponding to the residuals with the smallest determinant of the covariance matrix can then be used to give robust results for the regression parameters.

Using the MLTS as a means to estimate the parameters of the Vector Autoregressive (VAR) Model was presented in [21]. The VAR model is popular for modeling multiple time series. Estimation of its parameters based a typical least squares method is unreliable when outliers are present in the data. Development of robust procedures for multiple time series analysis is more crucial than for univariate time series analysis due to the data correlation structure. Experimental results in [21] show that applying the reweighted MLTS procedure to the VAR model leads to robust multivariate regression estimators with improved performance.

3. Multivariate Voronoi Outlier Detection (MVOD) for time series

In order to better analyze multivariate time series data, we have recently proposed a general outlier detection method 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. This has the potential to increase the accuracy and precision of outlier detection for specific application problems.

3.1. Background on Voronoi diagram

Our new method requires a Voronoi diagram, which is composed of Voronoi cells [22]. A Voronoi diagram is a way of dividing space into regions. Assume we have a set S of n points, $p_1, \ldots, p_n$ in the Euclidean 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. This is expressed formally in Equation (10):

$$V(p_i)=\{q \mid \text{dist}(p_i, q)\leq \text{dist}(p_j, q),\ \forall\ j\neq i\} \tag{10}$$

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

Figure 4 shows part of a Voronoi diagram, assuming Euclidean distance between the points. If one used a different distance metric, the Voronoi diagram would be configured differently. The plane is decomposed into n convex polygonal regions, one for each p_i. Vertices (or nodes) are called *Voronoi vertices* and are equidistant to three or more sites. *Voronoi edges* are the segments defined as the boundaries between two Voronoi cells and contain all the points in the plane equidistant to the two nearest sites. The boundaries of a Voronoi cell $V(p_i)$ cannot exceed n - 1 edges. Three important 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.*

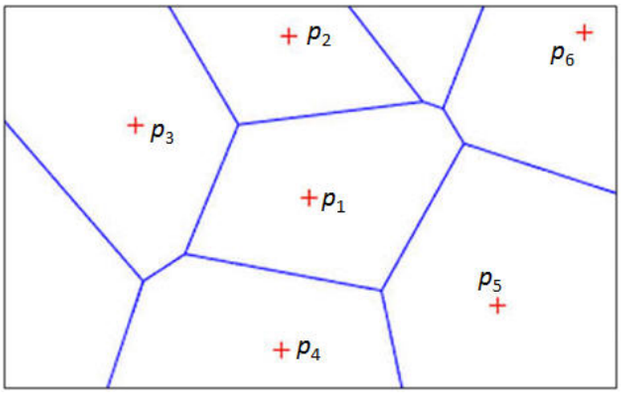

Figure 4. A subset of Voronoi cells from a Voronoi diagram.

3.2. Our proposed MVOD method

The Voronoi Outlier Index (VOInd) used in our Multivariate Voronoi Outlier Detection (MVOD) method is based upon the Voronoi notion of 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 4 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

http://dx.doi.org/10.5772/59750

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. Here, we present one such case of the MVOD framework for feature extraction; but many others are also possible, including nonparametric forms. Figure 5 overviews the process and the rest of this subsection explains the steps in more detail.

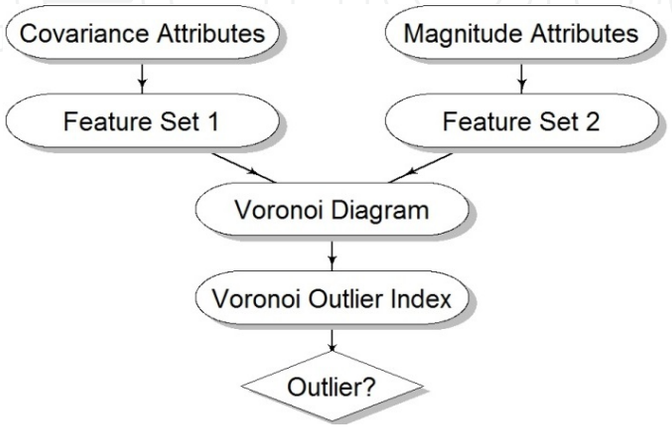

Figure 5. Flow chart of our MVOD procedure for outlier detection.

Feature Set 1 (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 and then use Multivariate Least Trimmed Squares (MLTS) introduced in Section 2.2, which computes a determinant of the covariance matrix *without that point*, to yield a single feature value. Removing one observation at a time is not part of the original MLTS method; we introduced this modification into the procedure to show 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 Set 2 (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 [23]. 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. Here, we estimate the parameters for a Multivariate Vector Auto-Regressive (MVAR) model because our simulated data are generated from this type of 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_i denote the ith observation of the time series of length n, and r_i denote the residual after fitting the MVAR model for each observation. Then, for all $i=1, \ldots, n$ we compute this feature value as in Equation (11):

$$\sum_{1}^{n} abs(y_i) \times \sum_{1}^{n} r_i^2 \tag{11}$$

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.

Voronoi Outlier Index (VOInd): Given the two feature sets (from the above procedures) that can be used as the x-coordinates and y-coordinates for the data, we construct the Voronoi diagram based on Section 3.1 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 nearest neighbors. This is divided by the denominator term which is the number of nearest neighbors, yielding an average density, as expressed in Equation (12):

$$\text{VOInd}(p_i) = \sum_{O \in V_{NN}(p_i)} \text{dist}(p_i, o) / \left| V_{NN}(p_i) \right| \tag{12}$$

Note that a Voronoi outlier factor is used in [24] 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.

3.3. Experimental evaluation and results

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 [25]. 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

http://dx.doi.org/10.5772/59750

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.

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.

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

Table 1. Definition of True (TPR) and False (FPR) Positive Rate.

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.

			True Positive Rate *Number of Outliers*			**False Positive Rate** *Number of Outliers*		
			5	*10*	*15*	*5*	*10*	*15*
Magnitude	*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

Table 2. True and False Positive Rates for MVOD and MLTS with 5, 10, or 15 outliers of magnitudes 1, 2, 3, 4, or 5.

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.

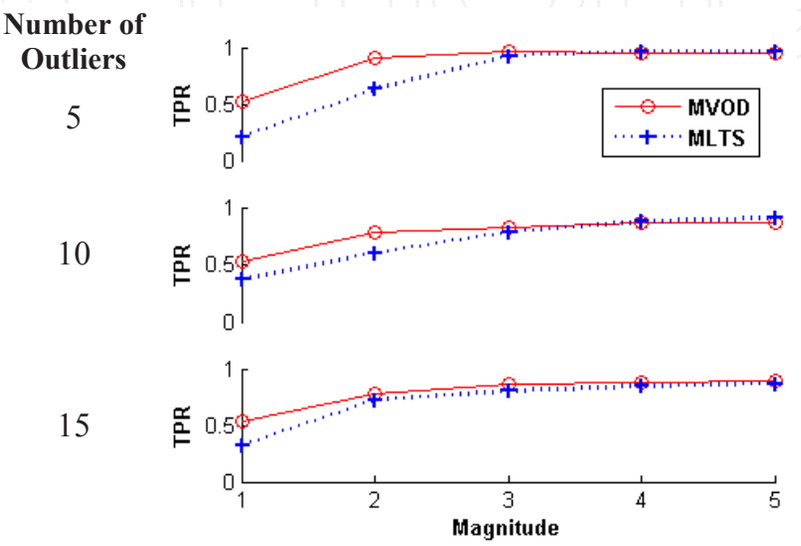

Figure 6. True Positive Rate (TPR, y-axis) for MVOD and MLTS for 5 outliers (top panel), 10 outliers (middle panel), and 15 outliers (bottom panel) with outlier magnitudes of 1, 2, 3, 4 or 5 (x-axis).

Results: Table 2 and Figure 6 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, which are often difficult for an outlier detection algorithm. 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.

4. Machine learning methods for novelty detection

Novelty detection can be considered as the task of classifying test data that differ in some respect from the data that are available during training. This may be approached within the

framework of "one-class classification" [3], in which a model is built to describe "normal" training data. Novelty detection methods can be categorized into several areas such as probabilistic, distance-based, reconstruction-based, domain-based, and information-theoretic techniques. In this section, we mainly introduce the first category of probabilistic approaches, and briefly summarize the others.

4.1. Probabilistic approaches

Probabilistic approaches to novelty detection are based on estimation of the generative probability density function of the data, which may then be used to define thresholds for the boundaries of "normality" in the data space and test whether or not a test sample is from the same distribution. Statistical hypothesis tests are the simplest statistical techniques for novelty detection [5]. Among the different statistical tests for novelty detection, here we concentrate on more advanced statistical modeling methods involving complex, multivariate data distributions. Techniques for estimating the underlying data density from multivariate training data broadly fall into parametric and nonparametric methods. The former imposes a restrictive model on the data, leading to a large bias when the model does not fit the data; the later builds up a very flexible model with fewer assumptions but requires a large sample size for a reliable fit of all free parameters when the model size becomes large.

In parametric approaches, the widely used distribution form for continuous variables is Gaussian. The involved parameters are estimated from the training data via *maximum likelihood estimates* (MLE), for which a closed-form analytical solution is available for a Gaussian distribution. More complex data distribution forms can be modeled through mixture models (e.g. Gaussian Mixture Models, or GMMs for short), or other mixtures of different types of distributions (e.g. the gamma, the Poisson, the Student's t, and the Weibull distributions) [26, 27]. When the form of the data distribution is not available, Gaussian distribution is usually taken due to its convenient analytical properties. The parameters of the GMMs can be estimated with maximum likelihood methods (using optimization algorithms including conjugate gradients or expectation-maximization, EM) or with Bayesian methods (e.g. variational Bayes) [26]. Besides the requirement of large numbers of training examples in estimating model parameters, another limitation of parametric methods is that the chosen data distribution form and the model generating the data may not match well. Despite the limitations, GMMs have been a popular scheme for novelty detection. The other strategy for novelty detection is to utilize time-series approaches, for example, the stochastic process of *Autoregressive Integrated Moving Average* (ARIMA), which can be used to predict the next data point and determine whether or not it is artefactual [28]. State-space models are also typically used for novelty detection in time-series data, assuming there is some underlying hidden state that generates the observations and this hidden state evolving through time [29]. The Hidden Markov Model (HMM) and the Kalman filter are two common state-space models for novelty detection.

Non-parametric methods do not assume a fixed structure of a model; the model grows in size as necessary to fit the data and accommodate the data complexity. A common non-parametric approach for probabilistic density estimation is the kernel density estimator [26], which

estimates the probability density function with a large number of kernels over the data space. The kernel density estimator places a kernel (e.g. Gaussian) on each data point and then sums the contributions from a localized neighborhood of the kernel. This is the so-called *Parzen windows* estimator [30], which has been used for novelty detection in a number of applications including mammographic image analysis [31]. One-class classification based on Gaussian Processes (GPs) has been developed and used recently [32]. This technique also takes a point-wise approach to novelty detection, which divides the data space into regions with high and low supports depending on whether those regions are close to those occupied by "normal" training data, or not. A related way of detecting novelty is based on the well-established area of *changepoint* detection [33], with the goal of determining whether the generative distribution of a sequence of observations has remained stable or has undergone some abrupt change. Here in addition to detecting whether a change has occurred or not, another aim is to estimate the time that the change has occurred. When applied in a batch or online setting, the idea of changepoint detection to the retrospective problem is identifying a test statistic suitable for testing the hypothesis that a change has occurred versus the one that no change has occurred. A likelihood ratio statistic, as well as others [33, 34], would be appropriate.

4.2. Other categories

Distance-based approaches, such as clustering or nearest-neighbor methods [35-37], are another types of techniques that can be used for classification or for estimating the probability density function of data. The underlying assumption is that "normal" data are tightly clustered, while novel data occur far from their nearest neighbors. These methods use well-defined distance metrics to compute the distance (e.g. similarity measure) between two data points.

Reconstruction-based methods involve training a regression model with the training data [3, 38, 39]. The distance between the test vector and the output of the system (i.e. the reconstruction error) can be related to the novelty score, which would be high when "abnormal" data occurs. For instance, neural networks can be used in this way and show many of the same advantages for novelty detection as they do for typical classification applications. Another type of reconstruction-based novelty detection is subspace-based techniques. They assume that data can be projected or embedded into a lower dimensional subspace, which makes better discrimination of "normal" and "abnormal" data easier.

Domain-based methods often aim to describe a domain that contains "normal" data through a boundary around the "normal" class following the distribution of the data without explicitly providing a distribution [40, 41]. These techniques are usually insensitive to the specific sampling and density of the interested class. The location to the boundary is the criterion for determining the class membership of unknown data. Novelty detection support vector machines (SVMs) are the "one-class SVMs", which set the location of the novelty boundary only based on the data lying closest to it in the transformed feature space. That is, the novelty boundary is determined without considering the data that are not support vectors.

Information-theoretic methods calculate the information content of a dataset with measures such as entropy, relative entropy, and Kolmogorov complexity, etc. [42, 43]. The key idea is that novel data alter the information content in a dataset significantly. A common procedure

http://dx.doi.org/10.5772/59750

is: metrics are computed using the entire dataset and then the subset of points whose elimination from the dataset causes the largest difference in the metric are identified. The data contained in this subset is then assumed to be novel data.

5. Robust estimator and outlier detection in high-dimensional medical imaging

The statistical analysis of medical images is challenging, not only because of the high-dimensionality and low signal-to-noise ratio of the data, but also due to varieties of errors in the image acquisition processes, such as scanner instabilities, acquisition artifacts, and issues associated with the experimental protocol [44]. Furthermore, populations under study typically present high variability [45, 46], and therefore the corresponding imaging data may have uncommon though technically correct observations. Such outliers deviating from normality could be numerous. With emergence of large medical imaging databases, developing automated outlier detection methods turns out to be a critical preprocessing step for any subsequent statistical analysis or group study. In addition, medical imaging data are usually strongly correlated [47]; outlier detection approaches based on multivariate models are thus crucial and desirable. Procedures using the classical MCD estimator are not well-suited for such high-dimensional data.

In [48], several extensions to the classical outlier detection framework are proposed to handle high-dimensional imaging data. Specifically, the MCD robust estimator were modified so that it can be used for detecting outliers when the number of observations is small compared to the number of available features. This is achieved through introducing regularization in the definition and estimation of the MCD. Three regularization procedures were presented and compared: l_1 regularization (*RMCD - l_1*); l_2 regularization or ridge regularization (*RMCD - l_2*); and random projections (*RMCD - RP*). The idea of *RMCD - RP* is to run the MCD estimator on datasets of reduced dimensionality, and this dimensionality reduction is done by projecting to a randomly selected subspace. In addition, the parametric approach of the regularized MCD estimators is compared to a non-parametric procedure, the One-Class SVM algorithm (see Section 4). Experimental results on both simulated and real data show that l_2 regularization performs generally well in simulations, but random projections outperform it in practice on non-Gaussian, and more importantly, on real neuroimaging data. One-Class SVM works well on unimodal datasets, and it has a strong potential if their parameters can be set correctly.

Outlier detection methods described above can serve as a statistical control on subject inclusion in neuroimaging. However, sometimes it is controversial regarding whether or not outliers should be discarded, and, if so, what tolerance to use. An alternative strategy is to utilize outlier-resistant techniques for statistical inference, which would also compensate for inexact hypotheses including data normality and homogeneous dataset. Robust techniques are especially useful when a large number of regressions are tested and assumptions cannot be evaluated for each individual regression, as with neuroimaging data.

Both individual subject and group analyses are required in neuroimaging. At a typical single subject level, a multiple regression model is used for the time series data at each voxel [49,

50], and outliers (or other assumption violations) in the time series would impact the model fitness. Robust regression can minimize the influence of these outliers. At the group level, after spatial normalization, a common strategy is to first save the regression parameters for each subject at each voxel and then perform a test on the parameter values. Robust regression used at this level can minimize the influence of outlying subjects. Wager et al. [51] used simulations to evaluate several robust techniques against ordinary least squares regression, and apply robust regression to second-level group analyses in three real fMRI datasets. Experimental results demonstrate that robust Iteratively Reweighted Least Squares (IRLS) at the second level is computationally efficient; it increases statistical power and decreases false positive rates when outliers are present. Without the presence of outliers, IRLS controls false positive rates at an appropriate level. In summary, IRLS shows significant advantages in group data analysis and in the hemodynamic response shape estimation for fMRI time series data.

6. Conclusions

Outlier and novelty detection is a primary step in many data mining and analysis applications, including healthcare and medical research. In this chapter, statistical and machine learning methods for outlier and novelty detection, and robust approaches for handling outliers in data and imaging sciences were introduced and reviewed. Particularly, we also presented our new method for outlier detection in time series data based on the Voronoi diagram (i.e. MVOD). There are several key advantages of our method. First, it copes with outliers in a multivariate framework by accounting for 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. Lastly, Voronoi diagrams capture the geometric relationship embedded in the data points. Initial experimental results show that our MVOD method can lead to accurate, sensitive, and robust identification of outliers in multivariate time series.

It is often difficult to reach a precise definition of outlier or novelty, and suggesting an optimal approach for outlier or novelty detection is even more challenging. The variety of practical and theoretical considerations arising in real-world datasets lead to the variety of techniques utilized [52]. Therefore, there is no single universally applicable detection method due to the large variety of considerations, which could include the application domain, the type of data such as dimension, and the availability of training data, etc. Based on the application and the nature of the associated data, developing suitable computational methods that can robustly and efficiently extract useful quantitative information from big data is still a current challenge and gaining increasing interest in data and imaging sciences.

Acknowledgements

This work is supported in part by a grant from the National Institute of Health, K25AG033725.

http://dx.doi.org/10.5772/59750

Author details

Michelle Yongmei Wang[1,2,3,4*] and Chris E. Zwilling[2]

*Address all correspondence to: ymw@illinois.edu

1 Department of Statistics, University of Illinois at Urbana-Champaign, USA

2 Department of Psychology, University of Illinois at Urbana-Champaign, USA

3 Department of Bioengineering, University of Illinois at Urbana-Champaign, USA

4 Beckman Institute, University of Illinois at Urbana-Champaign, USA

References

[1] Aggarwal CC. Outlier Analysis. New York: Springer Science + Business Media; 2013.

[2] Markou M, Singh S. Novelty detection: a review---part 1: statistical approaches. Signal Processing 2003a; 83(12): 2481-2497.

[3] Markou M, Singh S. Novelty detection: a review---part 2: neural network based approaches. Signal Processing 2003b; 83(12): 2499-2521.

[4] Zwilling CE, Wang MY. Multivariate Voronoi outlier detection for time series. In: Proc. IEEE Healthcare Innovation Point-Of-Care Technologies Conference 2014; in press.

[5] Barnett V, Lewis T. Outliers in Statistical Data. John Wiley and Sons; 1994.

[6] Tarassenko L, Clifton DA, Bannister PR, King S, King D. Novelty Detection. In: Boller C, Chang F-K, Fujino Y (eds.) Encyclopedia of Structural Health Monitoring. 2009. Chapter 35.

[7] Davies L, Gather U. The identification of multiple outliers. Journal of American Statistical Association 1993; 88(423): 782-792.

[8] Rousseeuw P. Multivariate estimation with high breakdown point. In: W. Grossmann et al. (eds.) Mathematical Statistics and Applications. Budapest: Akademiai-Kiado; 1985. Vol. B, p283-297.

[9] Ben-Gal I. Outlier detection. In: Maimon O, Rockach L (eds.) Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers. Kluwer Academic Publishers; 2005. Chapter 1.

[10] Becker C, Fried R, Kuhnt S, editors. Robustness and Complex Data Structures. Berlin Heidelberg: Springer-Verlag; 2013.

[11] Huber PJ, Ronchetti EM. Robust Statistics. John Wiley & Sons, Inc.; 2009.

[12] Rousseeuw PJ. Least median of squares regression. Journal of the American Statistical Association 1984; 79: 871-880.

[13] Hubert M, Debruyne M. Minimum covariance determinant. Wiley interdisciplinary reviews: Computational statistics 2010; 2: 36-43.

[14] Croux C, Haesbroeck G. Influence function and efficiency of the Minimum Covariance Determinant. Journal of Multivariate Analysis 1999; 71: 161-190.

[15] Lopuhaa HP, Rousseeuw PJ. Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. Annals of Statistics 1991; 19: 229-248.

[16] Lopuhaa HP. Asymptotics of reweighted estimators of multivariate location and scatter. Annals of Statistics 1999; 27: 1638-1665.

[17] Liebscher S, Kirschstein, Becker C. RDELA –--A Delaunay-triangulation-based, location and covariance estimator with high breakdown point. Statistics and Computing 2013; 23: 677-688.

[18] Rousseeuw PJ, Driessen KV. A fast algorithm for the minimum covariance determinant estimator. Technometrics 1999; 41(3): 212-223.

[19] Rousseeuw PT, Aelst SV, Driessen KV, Agullo J. Robust multivariate regression. Technometrics 2004; 46(3): 293-305.

[20] Agullo J, Croux C, Aelst SV. The multivariate least-trimmed squares estimator. Journal of Multivariate Analysis 2008; 99: 311-338.

[21] Croux C, Joossens K. Robust estimation of the vector autoregressive model by a least trimmed squares procedure. In: Proceedings in Computational Statistics 2008; p489-501.

[22] Preparata FP, Shamos MI. Computational Geometry-An Introduction. Springer; 1985.

[23] Pearson, RK. Exploring Data in Engineering, the Sciences and Medicine. Oxford University Press; 2011.

[24] Qu J. Outlier detection based on Voronoi diagram. In: Proceedings of the ADMA International Conference on Advanced Data Mining and Applications 2008; p516-523.

[25] 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.

[26] Bishop CM. Pattern Recognition and Machine Learning. Springer, New York; 2006.

[27] Carvalho A, Tanner M. Modelling nonlinear count time series with local mixtures of poisson autoregressions. Comput. Stat. Data Anal. 2007; 51(11): 5266–5294.

[28] Hoare S, Asbridge D, Beatty P. On-line novelty detection for artefact identification in automatic anaesthesia record keeping. Med. Eng. Phys. 2002; 24(10): 673–681.

[29] Quinn J, Williams C. Known unknowns: novelty detection in condition monitoring. In: Marti J et al. (eds.) Pattern Recognition and Image Analysis, LNCS 4477. 2007. p1–6.

[30] Parzen E. On estimation of a probability density function and mode. Ann. Math. Stat. 1962; 33(3): 1065–1076.

[31] Tarassenko L, Hayton P, Cerneaz N, Brady M. Novelty detection for the identification of masses in mammograms. In: Proceedings of the 4th International Conference on Artificial Neural Networks, IET. 1995. p442–447.

[32] Kemmler M, Rodner E, Denzler J, One-class classification with Gaussian processes. In: Asian Conference on Computer Vision (ACCV), vol. 6493. 2011. p489–500.

[33] Basseville M, Nikiforov IV. Detection of Abrupt Changes: Theory and Application. Prentice Hall, Englewood Cliffs; 1993.

[34] Reeves J, Chen J, Wang XL, Lund R, Lu QQ. A review and comparison of change-point detection techniques for climate data. J. Appl. Meteorol. Climatol. 2007; 46(6): 900–915.

[35] Pires A, Santos-Pereira, C. Using clustering and robust estimators to detect outliers in multivariate data. In: Proceedings of the International Conference on Robust Statistics. 2005.

[36] Yong S, Deng J, Purvis M, Wildlife video key-frame extraction based on novelty detection in semantic context. Multimed. Tools Appl. 2013; 62(2): 359–376.

[37] Hautamaki V, Karkkainen I, Franti P. Outlier detection using k-nearest neighbor graph. In: Proceedings of the 17th International Conference on Pattern Recognition, vol. 3. 2004. p430–433.

[38] Kit D, Sullivan B, Ballard D. Novelty detection using growing neural gas for visuospatial memory. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. 2011, p1194–1200.

[39] Xiao Y, Wang H, Xu W, Zhou J. L1 norm based KPCA for novelty detection. Pattern Recognit. 2013; 46(1): 389–396.

[40] Schölkopf B, Williamson R, Smola A, Shawe-Taylor J, Platt J. Support vector method for novelty detection. Adv. Neural Inf. Process. Syst. 2000; 12(3): 582–588.

[41] Le T, Tran D, Ma W, Sharma D. Multiple distribution data description learning algorithm for novelty detection. Adv. Knowl. Discov. Data Min. 6635. 2011. p246–257.

[42] He Z, Deng S, Xu X, Huang J. A fast greedy algorithm for outlier mining. Adv. Knowl. Discov. Data Min. 3918. 2006. p567–576.

[43] Filippone M, Sanguinetti G. Information theoretic novelty detection. Pattern Recognition 2010; 43(3): 805–814.

[44] Wang MY, Zhou C, Xia J. Statistical analysis for recovery of structure and function from brain images. In: Komorowska MA, Olsztynska-Janus S (eds.) Biomedical Engineering, Trends, Researches and Technologies. 2011. p169-190.

[45] Chen G, Fedorenko E, Kanwisher NG, Golland P. Deformation-invariant sparse coding for modeling spatial variability of functional patterns in the brain. In: Proc. Neural Information Processing Systems Workshop on Machine Learning and Interpretation in Neuroimaging, LNAI 7263. 2012. p68-75.

[46] Staib LH, Wang YM. Methods for nonrigid image registration. In: Bayro-Corrochano E (ed.) Handbook of Goemetric Computing: Applications in Pattern Recognition, Computer Vision, Neuralcomputing, and Robotics. Springer-Verlag; 2005. p571-602.

[47] Wang MY, Xia J. Unified framework for robust estimation of brain networks from fMRI using temporal and spatial correlation analyses. IEEE Trans. on Medical Imaging 2009; 28(8): 1296-1307.

[48] Fritsch V, Varoquaux G, Thyreau B, Poline J-B, Thirion B. Detecting outliers in high-dimensional neuroimaging datasets with robust covariance estimators. Medical Image Analysis 2012; 16(7): 1359-1370.

[49] Worsley KJ, Friston KJ. Analysis of fMRI time-series revisited — Again. NeuroImage 1995; 2(3): 173 – 181.

[50] Worsley KJ, Poline JB, Friston KJ, Evans AC. Characterizing the response of PET and fMRI data using multivariate linear models. NeuroImage 1997; 6(4): 305–319.

[51] Wager TD, Keller MC, Lacey SC, Jonides J. Increased sensitivity in neuroimaging analyses using robust regression. NeuroImage 2005; 26: 99-113.

[52] Singh K, Upadhyaya S. Outlier detection: applications and techniques. International Journal of Computer Science Issues 2012; 9(1): 307-323.