# Rare Event Discovery And Event Change Point In Biological Data Stream

T. Jagadeeswari [1] M.Tech(CSE) MISTE, B. Mahalakshmi [2]M.Tech(CSE)MISTE,
N. Anusha [3]M.Tech(CSE)
Department of Computer Science and Engineering[1,2,3]
Vardhaman College of Engineering, JNTU Hyderabad

## Abstract

*Outlier detection is currently an important and active research problem in many fields and is involved in numerous applications. This paper applies minimum volume ellipsoid (MVE) with principle component analysis (PCA) extension, a powerful algorithm for detecting multivariate outliers. If the data points exceed the cut-off value, the mahalanobis distance is used for the outliers. The paper also compares the performance of the suggested frame work with statistical methods to demonstrate its validity through simulation and experimental applications for incident detection in the field of bioinformatics. The observable results are obtained from biological samples such as glucose, acids and protein synthesis and are shown in the graphs.*

*Keywords: outlier detection, PCA, MVE, Mahalanobis distance.*

## 1. Introduction

An outlier is an observation (or measurement) that is different with respect to the other values contained in a given data set. Such irregularities or rare events can indicate an error in the data, or abnormal behaviour of the underlying system. There are different definitions of “outlier” the most commonly referred ones are:

- “An outlier is an observation that deviates so much from other observations as to arouse suspicions that is was generated by a different mechanism “ (Hawkins, 1980).
- “An outlier is an observation (or subset of observations) which appear to be inconsistent with the remainder of the dataset” (Barnet & Lewis, 1994).
- “An outlier is an observation that lies outside the overall pattern of a distribution” (Moore and McCabe, 1999).

Many data mining algorithms try to minimise the influence of outlier in data sets. They are extensively used in a wide variety of applications such as fraud detection in credit card transactions, intrusion detection in cyber security, identifying novel molecular structures in the field of bioinformatics as part of pharmaceutical research, loan application processing of problematic customers, etc. Their importance in data is due to the fact that they can translate into actionable information in a wide variety of applications.

### 1.1. Defining Outliers

Outliers are patterns in data that do not conform to a well defined notion of normal behaviour. Figure 1 illustrates outliers in a simple 2-dimensional data set.

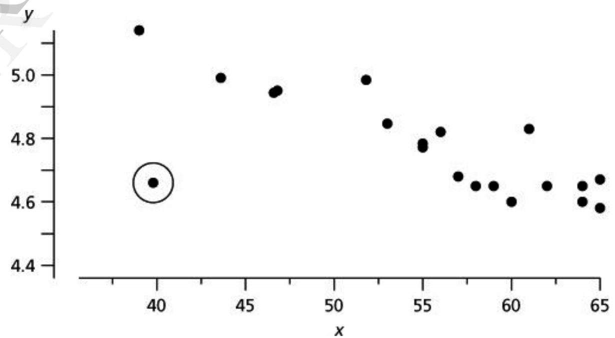


**Figure 1.**Outliers in 2 dimensional datasets

## 2. Problem identification

Bioinformatics is the application of information technology in field of molecular biology. Bioinformatics entails the creation and advancement of databases, algorithms, computational and statistical techniques, and theory to solve formal and practical problems arising from the management and analysis of biological data. The primary goal of bioinformatics is to increase our understanding of biological processes. The focus is on developing and applying computationally intensive techniques (e.g., data mining and machine learning algorithms) to achieve this goal. In order to identify the irregularities in bio informatics one of technique used is Principal component analysis.

Principal Component Analysis (PCA), developed by Karl Pearson in 1901, is a simple, non parametric method of extracting relevant information from confusing data. The aim of this method is to reduce the dimensionality of multivariate data and is a linear transformation that transforms the data to a new coordinate system. This method calculates the covariance matrix, because covariance is always measured between two dimensions. It measures how much the dimensions vary from the mean with respect to one another. If we calculate the covariance between one dimension and itself, we will get the variance of that dimension. The covariance matrix describes all relationships between pairs of measurements in the considered data set.

The basic formula for the covariance is

$$\mathrm{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)}$$

Where X and Y are two separate dimensions of data. By getting the covariance matrix, the Eigen vector and Eigen values are calculated. Before outlier detection the observations on the eigenvectors are scored with positive Eigen values for co-variance matrix.

For a square matrix *A* of order n, the number **λ** is an Eigen value if and only if there exists a non-zero vector *C* such that

$$AC=\lambda C$$

Using the matrix multiplication properties, we obtain

$$(A - \lambda I_n)C=0$$

Where $I_n$ is the unit vector

Covariance measures the degree to which two variables change or vary together (i.e. co-vary). On the one hand, the covariance of two variables is positive if they vary together in the same direction relative to their expected values (i.e. if one variable moves above its expected value, then the other variable also moves above its expected value). On the other hand, if one variable tends to be above its expected value when the other is below its expected value, then the covariance between the two variables is negative. If there is no linear dependency between the two variables, then the covariance is 0.

Correlation is a measure of the relation between two or more variables. The correlation coefficient ρX, Y between two random variables X and Y values

$$\rho xy = \frac{cov(X,Y)}{\sigma_x \, \sigma_y}$$

In higher-dimensional spaces, the problem is even more difficult. Traditional multivariate outlier-detection methods are based on the calculation of the generalized squared (Mahalanobis) distances for each data point. Mahalanobis distances are in essence weighted Euclidean distances, the distance of each point from the centre of the distribution is weighted by the inverse of the sample variance-covariance matrix. Unfortunately, outliers greatly inflate the covariance matrix and can therefore effectively mask their own existence.

To counter this masking problem, Rousseeuw (1985) introduced the robust Minimum Volume Ellipsoid (MVE) method for detection of outliers in multidimensional data. By the term change point, we mean a time point at which the data properties suddenly change.

## 3. Implementation and design methodology

The suggested PCA-MVE method is for finding Outliers in Bioinformatics datasets. The sample data is generated through biological tools and taken in form of numeric values then PCA method is applied. In PCA, the Covariance and Correlation matrix for the data is found. Then by applying MVE algorithm, whether it is an Outlier or a Cluster is known. By giving the weight as 0 for Outlier and 1 for Cluster, the Outliers are found and then projected in the form of a Graph.

### 3.1. Procedure

Input: Take biological samples such as acids, glucose and Protein synthesis.
Step 1: Use bioinformatics tool to generate or create numeric data set.
Step 2: Apply PCA for sample dataset.
Step 3: Calculate covariance matrix and then find Eigen vector, Eigen values and correlation matrix for thedata.
Step 4: Then apply MVE to know whether it is an outlier or cluster by giving the weights as 0 for outlier and 1 for cluster.
Step 5: Apply mahalanobis distance, if data point exceeds the cut-off value.
Step 6: Generate reports to visualize the outliers in the data set by using statistical charts.

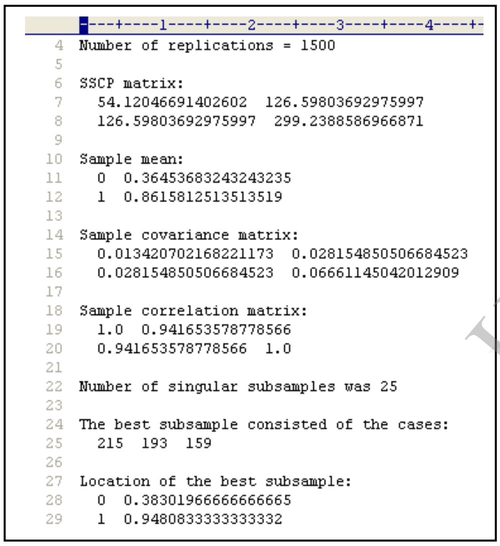
## 4. Experimental results

Taking biological samples such acids, glucose and protein synthesis and generate above procedure
The sample input data is as follows:

**Table 1**. The sample data

| X | Y |
|---|---|
| 0.40078200 | 0.952314 |
| 0.38408900 | 0.889371 |
| 0.44236600 | 0.953902 |
| 0.41746100 | 0.947438 |
| 0.47986600 | 0.935342 |
| 0.36915900 | 0.959396 |
| 0.39317700 | 0.968266 |
| 0.44826500 | 0.870872 |
| 0.34656500 | 0.884370 |
| 0.42260000 | 0.940137 |
| 0.42777000 | 0.954561 |
| 0.40961500 | 0.920041 |
| 0.35421200 | 0.923201 |
| 0.45138200 | 0.990933 |

The mean, covariance, correlation values are generated as shown below



```
Number of replications = 1500

SSCP matrix:
  54.12046691402602  126.59803692975997
  126.59803692975997  299.2388586966871

Sample mean:
  0  0.36453683243243235
  1  0.8615812513513519

Sample covariance matrix:
  0.013420702168221173  0.028154850506684523
  0.028154850506684523  0.06661145042012909

Sample correlation matrix:
  1.0  0.941653578778566
  0.941653578778566  1.0

Number of singular subsamples was 25

The best subsample consisted of the cases:
  215  193  159

Location of the best subsample:
  0  0.3830196666666665
  1  0.9480833333333332
```

**Figure 2.** The mean, covariance, correlation values

## 4.1. Statistical charts for representing the outliers

The statistical graphs are useful for representing the data in a meaningful way. A good graph conveys information quickly and easily to the user. Graphs highlight the hidden features of the data.

**4.1.1. Box and whisker.** The box and whisker is one of the important graphs for outlier detection to show the spread of the data. The diagram is made up of a "box", which lies between the upper and lower quartiles. The median can also be indicated by dividing the box into two. The "whiskers" are straight lines extending from the ends of the box to the maximum and minimum values.

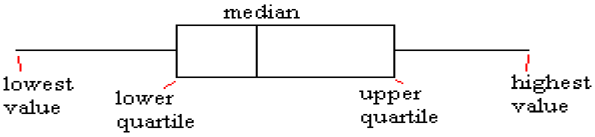


**Figure 3.** Box and whisker diagram

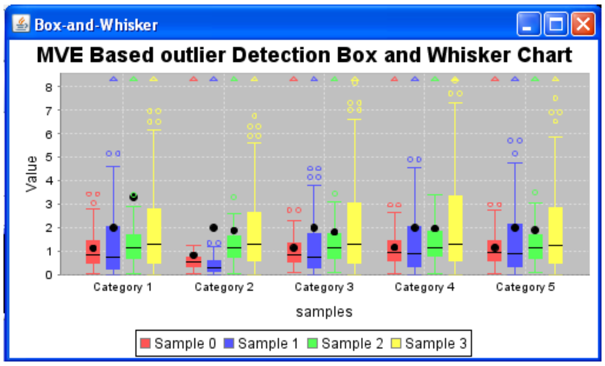


**Figure 4.** The MVE based outlier detection box and whisker chart

**4.1.2. Linear and power regression models.** Regression analysis is a statistical tool for the investigation of relationships between variables. Linear regression is the most widely used of all statistical techniques. It is the study of linear (i.e., straight-line) relationships between variables, usually under an assumption of normally distributed errors. Where as the following two models such that linear regression model and power regression model which shows the outliers deviated from the data set.

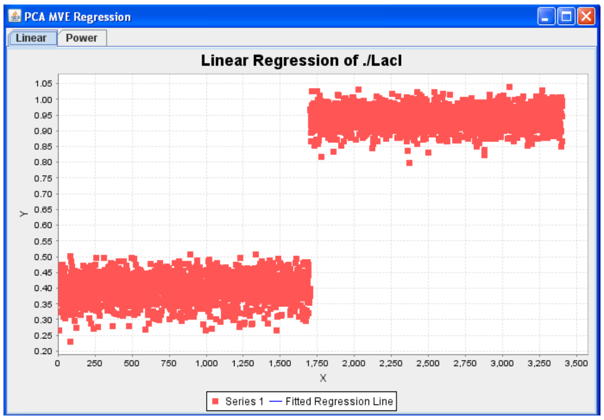


**Figure 5.** The linear regression model

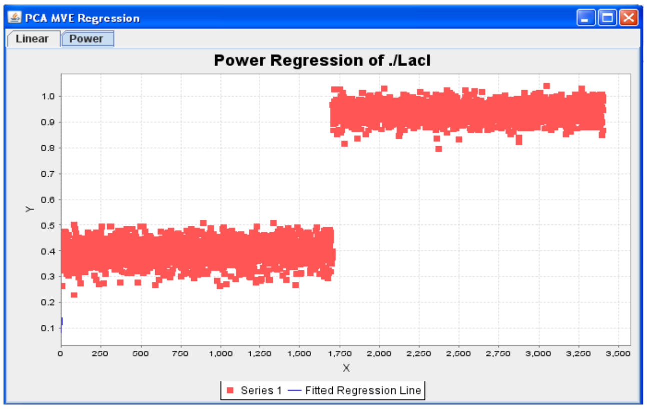


**Figure 6**. The power regression model

## 4. Conclusion

Bioinformatics is an inter-disciplinary field that involves use of Information Technology in developing and improving methods for analysing biological data and enhance our biological knowledge. This paper suggests the use of PCA-MVE, which is one of the data mining techniques, along with Mahalanobis distance to detect the outliers in two dimensional or multivariate data sets. Such detection of the outliers helps in identifying errors in biological data and thus proves to be of immense significance in the applications of bioinformatics like accurate diagnosis of diseases, Personalised medicine, Preventive medicine etc.

## 5. Bibilography

[1] Jun-ichi Takeuchi and Kenji Yamanishi (2006) Unifying Framework for Detecting Outliers and Change Points from Time Series

[2] Hodge, V.J. and Austin, J. (2004) A survey of outlier detection methodologies.

[3] Romy Shioda and Levent Tuncel (2005) Clustering via Minimum Volume Ellipsoid

[4] Hongxia Pang, Jiaowei Tang, Su-Shing Chen, and Shiheng TaLac Statistical distributions of optimal global alignment scores of random protein sequences

[5] V. Barnett and T. Lewis 1994 "Outliers in Statistical Data,"JohnWiley&Sons

[6] Wold, S., Esbensen, K., Geladi, P. (1987).Principal Components Analysis. Chemometrics and Intelligent Laboratory Systems. 2, 37-55.

[7] T.Cover and J.A. Thomas, Elements of Information Theory. Wiley-International, 1991

[8] M. Huskova, "Nonparametric Procedures for Detecting a Change in Simple Linear Regression Models," Applied Change Point Problems in Statistics, 1993.

[9] K. Yamanishi and J. Takeuchi 2001 "Discovering Outlier Filtering Rules from Unlabeled Data,"Proc. Fourth Workshop Knowledge Discovery and Data Mining, pp. 389- 394.

[10] K. Yamanishi and J. Takeuchi, "A Unifying Approach to Detecting Outliers and Change-Points from Nonstationary Data," Proc of the EighthACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, 2002.

[11] T. Fawcett and F. Provost, "Activity Monitoring: Noticing Interesting Changes in Behavior," Proc. ACM-SIGKDD Int'l conf. Knowledge Discovery and Data Mining, pp. 53-62, 1999.

[12] V. Guralnik and J. Srivastava, "Event Detection from Time Series Data," Proc. ACM-SIGKDD Int'l Conf. Knowledge Discovery and DataMining, pp. 33-42, 1999

[13] D.M. Hawkins, "Point Estimation of Parameters of Piecewise Regression Models," J Royal Statistical Soc. Series C, vol. 25, no. 1, pp. 51-57, 1976.

[14] Kenji Yamanishi and Jun-ichi Takeuchi, "A Unifying Framework for Detecting Outliers and Change Points from Non-Stationary Time Series Data" in IEEE transactions on knowledge and data engineering vol17. No.6, June 2006

[15] Barnett, V. and Lewis, T.: 1994, Outliers in statistical Data. John Wiley & Sons.,3 edition.

[16] Aggarwal, C. C. and Yu, P. S.: 2001,Outlier detection for High Dimensional Data'.In: Proceedings of the ACM SIGMOD Conference 2001

[17] V. Guralnik and J. Srivastava, "Event Detection from Time Series Data," Proc. ACM-SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 33-42, 1999

Jagadeeswari Tanukonda received the Master of Computer Applications degree from Andhra University Vizag in 2002 and the Master of Technology degree in Computer Science and Engineering from Jawaharlal Nehru Technological University, Hyderabad in 2010. She is currently working as an Assistant Professor in the Department of Computer Science and Engineering ,Vardhaman College of Engineering, Hyderabad. Her research interests include Data Mining, Database Management Systems, and Software Engineering.

B. Mahalakshmi received the Master of Computer Applications degree from IGNOU New Delhi in 2006 and the Master of Technology degree in Computer Science and Engineering from Jawaharlal Nehru Technological University, Hyderabad in 2012. She is currently working as an Assistant Professor in the Department of Computer Science and Engineering, Vardhaman College of Engineering, Hyderabad. Her research interests include Data Mining, Operating System, and Network Security.

Anusha Nalabothu received the Bachelor of Technology in Computer Science and Engineering from Nagarjuna University, Vijayawada in 2009 and the Master of Technology degree in Computer Science and Engineering from Nagarjuna University, Vijayawada in 2011. She is currently working as an Assistant Professor in the department of Computer Science and Engineering, Vardhaman College of Engineering, Hyderabad. Her research interests include Data Mining, Operating System, Computer Architecture and Organization and Network Security.