ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH

DOI: 10.5336/biostatic.2018-61547

Is It Necessary to Apply the Outlier Detection for Protein-Protein Interaction Data?

Protein-Protein Etkileşim Verileri İçin Aykırı Değer Tespiti Uygulanması Gerekir mi?

Ezgi AYYILDIZ,[a]
Vilda PURUTÇUOĞLU[a]

[a] Department of Statistics,
Middle East Technical University,
Ankara, TURKEY

Received: 24.05.2018
Received in revised form: 07.08.2018
Accepted: 27.08.2018
Available Online: 07.12.2018

Correspondence:
Ezgi AYYILDIZ
Middle East Technical University,
Department of Statistics, Ankara,
TURKEY/TÜRKİYE
ayezgi@metu.edu.tr

This study was presented as oral presentation at 4th International Researchers, Statisticians and Young Statisticians Congress (IRSYSC 2018) 28 -30 April 2018, İzmir, Turkey.

ABSTRACT Objective: Outlier detection is a crucial problem in many fields. Although there are too many outlier detection methods in the literature, only a few methods suitable for dependent, sparse and high dimensional data structure. In this study, we perform various univariate and multivariate outlier detection methods as a pre-processing step before modeling the protein-protein interaction networks in order to investigate whether the outlier detection can improve the accuracy of the model. **Material and Methods:** Within the univariate approaches, we implement the z-score and Box-plot methods which are the most well-known outlier detection approaches. Besides them, we also apply the multivariate outlier detection methods, called PCOut and Sign which are based on the robust principal component analysis and the BACON method which is a distance-based approach. These methods are applicable for the data type such that the number of variables are bigger than the observation. In the analysis, we use several synthetic and real benchmark biological datasets. Then, we infer the networks with 3 network models, namely, GGM, MARS and CMARS and finally, we check the validity of models via F-measure and the accuracy measures. **Conclusion:** From the results, it has been seen that the use of outlier detection methods before the modeling cannot contribute to the performance of the models in our datasets. **Results:** Based on the results obtained from different datasets, we suggest that the estimations of protein-protein interaction networks can be made with GGM, MARS and CMARS methods without the need for an outlier detection process.

Keywords: Outlier detection; protein-protein interaction networks; multivariate adaptive regression spline; Gaussian graphical model; conic multivariate adaptive regression spline

ÖZET Amaç: Aykırı değer saptama, birçok alan için çok önemli bir sorundur. Literatürde çok fazla aykırı değer tespit yöntemi olmasına rağmen, bağımlı, seyrek ve yüksek boyutlu veri yapısı için sadece birkaç yöntem uygundur. Bu çalışmada, protein-protein etkileşim ağlarını modellemeden önce bir ön işlem adımı olarak çeşitli tek değişkenli ve çok değişkenli aykırı değer saptama yöntemlerinin modelin doğruluğunu arttırıp arttırmadığını araştırmaktayız. **Gereç ve Yöntem:** Tek değişkenli yaklaşımlarda, iyi bilinen aykırı değer tespit yaklaşımları olan z-skor ve kutu grafiği yöntemlerini uygulamaktayız. Bunların yanı sıra, dayanıklı temel bileşen analizine dayanan PCOut ve Sign yöntemlerini ve ayrıca uzaklık bazlı bir yaklaşım olan BACON yöntemine kullanmaktayız. Bu yöntemler, gözlemlerden daha fazla sayıda değişkene sahip veri tipi için uygulanabilmektedir. Analizde, çeşitli sentetik ve gerçek biyolojik veri setlerini kullanmaktayız. Daha sonra 3 farklı ağ modeli olan GGM, MARS ve CMARS ile ağları tahmin etmekte ve son olarak F ölçüsü ve doğruluk değeri ile modellerin doğruluğunu kontrol etmekteyiz. **Bulgular:** Sonuçlardan, uyguladığımız modelleme yöntemlerinden önce listelenen aykırı değer saptama yöntemlerinin kullanılmasının, bizim veri setlerimizdeki modellerin doğruluğuna katkı sağlamadığı görülmüştür. **Sonuç:** Kullandığımız farklı veri setlerinden elde ettiğimiz sonuçlara göre, protein-protein etkileşim ağlarının modellemesinde GGM, MARS ve CMARS yöntemlerinden önce aykırı değer incelemesine gerek duyulmadan tahminlerin yapılabileceğini önermekteyiz.

Anahtar Kelimeler: Aykırı değer saptama; protein-protein etkileşim ağları; çok değişkenli uyarlamalı regresyon eğrileri; Gaussian grafik modeli; konik çok değişkenli uyarlamalı regresyon eğrileri

An outlier is an observation that is significantly different from other observations.[1] In the literature, it is also known as abnormalities, discordants, deviants and anomalies.[2] Due to its importance in both the data analyses and modeling the observations, the outlier detection is a crucial problem in many research fields. There are basically two main purposes of the outlier identification. In some cases, the identification of outliers can be the primary goal since it provides new discoveries about the hidden aspects of the data.[3] In other cases, the identification of outliers can be used as a preprocessing step before fitting a statistical model and many methods have been proposed for this purpose. These methods can be divided into two parts, so called, univariate and multivariate methods. Dixon's Q test (Dixon, 1950), Grubbs test (Grubbs, 1950), z-scores (Schiffler, 1998) and *Box-plot* (Tukey, 1977) are the widely known univariate outlier detection methods.[4-7] The *Dixon's Q* test and the Grubbs test are very simple methods. The *Dixon's Q* test enables to find a single outlier. Whereas the Grubbs test can detect at most 2 outliers. They are applicable under normally distributed data. On the other hand, the *Box-plot* is the most well-known nonparametric and robust graphical method which is appropriate to identify outliers. However, the outlier detection with graphical tools is not suitable for more than three dimensions. It is necessary to apply a multivariate analysis which can consider the interactions among several variables to the multivariate data. The multivariate outlier detection methods can be basically divided into two parts, i) distance based approaches and ii) projection based approaches. Mahalanobis distance (MD) and Robust distance (RD) are the well-known distance measures.[8] Moreover, the *blocked adaptive computationally efficient outlier nominators* (BACON) is an another approach based on a robust multivariate method.[9] Unfortunately, they are not applicable for data structure where the size of the variables is much more than the observations $(p \gg n)$ due to the singularity problem in the estimation of the covariance matrix. In most of genomic datasets, there may be thousands of genes or proteins measured on few samples. Hence, $(p \gg n)$ situation is commonly observed in biological data. Here, the projection pursuit methods are appeared to be aware. Since they are not restricted with distributional assumptions and are applicable for several data structure, particularly, where the size of the variables are much more than the observations $(p \gg n)$, they are the most appropriate choice for high dimensional data. A well-known projection pursuit method is the *principal component analysis* (PCA). Additionally, there are several outlier detection approaches based on PCA such as robust PCA and spherical PCA methods.[10,11]

There are different methods to estimate the structure of networks. The Gaussian graphical model (GGM) is the well-known parametric method to obtain undirected networks.[12] However, there are some drawbacks of GGM such that it has a normality assumption and the inference of the model parameters becomes computationally demanding when the dimension of the system gets larger. On the other hand, multivariate adaptive regression splines (MARS) is a nonparametric alternative of GGM to construct the biological networks.[13] This method uses the piecewise linear splines to obtain the model and it can capture both linear and nonlinear relations between system's elements.[14] So it is suitable for the high dimensional and correlated data structures. Additionally, conic multivariate adaptive regression splines (CMARS), a modified version of MARS, is another strong alternative of GGM and also MARS.[15] The penalized residual sum of square (PRSS) is used for the parameter estimation in CMARS.[16] PRSS can control both the accuracy and the complexity of the model. From the previous studies, it is observed that the CMARS and MARS methods are the plausible alternatives of GGM in the construction of complex biological networks.[13,15] Especially, CMARS produces more accurate results than GGM and MARS.

In this study, we aim to investigate whether using an outlier detection method before modeling can improve the accuracy of fitted models and which method is the most appropriate for the protein-protein interaction data. For this purpose, we compare the accuracy of results from three different modeling approaches under several outlier detection methods. In our application, we use different synthetic and real

benchmark biological datasets. Based on our results obtained from different datasets, we observe that the removal of outliers before modeling by GGM, MARS and CMARS methods cannot conduce to improve the accuracy of the models. Furthermore, we detect that there is no unique best method for the outlier detection approaches in protein-protein interaction data.

METHODS

GAUSSIAN GRAPHICAL MODELS

The Gaussian graphical model (GGM) is a widely used undirected graphical model to construct the biological networks. GGM makes the assumption that the systems' elements, i.e., genes or proteins, have a multivariate normal distribution.[12] This model uses the conditional independency between nodes to define the structure of the graph. In other words, if two nodes are conditionally independent, no edge is drawn between these nodes in the graph. The conditionally independency is also presented with the zero entry in the precision matrix, which is the inverse of the variance-covariance matrix.

In GGM, a regression model is built for each node against all remaining nodes as described by the following form.

$$Y_{(p)} = \beta Y_{(-p)} + \varepsilon, \quad (1)$$

where $Y_{(p)}$ and $Y_{(-p)}$ denote the state of the p^{th} node and the states of the all nodes except the p^{th} node, respectively. β represents the regression coefficients and ε is the independent and normally distributed error terms. Under the normality, $Y_{(p)}$ and $Y_{(j)}$ $(j = 1,2, \dots, p)$ are conditionally independent when β_j equals to zero.

MULTIVARIATE ADAPTIVE REGRESSION SPLINES

The multivariate adaptive regression splines (MARS) model is a nonparametric approach which is suitable for high dimensional and correlated data.[14] In this model, there is no assumption between dependent and independent variables. MARS is an adaptive procedure which presents the nonlinear relation between predictors and response by means of piecewise linear functions. This model includes basis functions (BFs) rather than original variables. The MARS model can be expressed as

$$f(y) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(y) \quad (2)$$

in which β_0 represents the intercept term and β_m denotes the regression coefficients. Additionally, $h_m(y)$ presents a function that includes a set of piecewise linear splines (basis functions) and M indicates the number of basis functions in the model. The regression coefficients are estimated by minimizing the residual sum of squares.

MARS consists of two stages, namely, the forward stage and the backward stage. In the forward stage, initially an intercept term is included. Then, the basis functions are iteratively added to the model. At the end of the forward stage, a large model which overfits the data is obtained. To solve this problem, in the backward stage, the terms whose elimination cause the smallest increase in the residual squared error, are removed. Finally, the best model with the λ number of terms is estimated. In order to choose the optimal λ, the generalized cross-validation (GCV) value is used.[14,17] GCV value is defined as

$$GCV(\lambda) = \frac{\sum_{i=1}^{N}(y_i - \hat{f}_\lambda(y_i))^2}{(1-(r+cK)/N)^2}, \quad (3)$$

where r denotes the number of linearly independent basis functions, K shows the number of knot points used in the forward stage and N stands for the number of observations. Furthermore, c is the cost for the optimization of the basis function and it is generally set to 3. However, if the model is restricted to obtain an additive model, it is taken as 2. Additionally, the numerator of GCV equals to the residual sum of squares. Finally, $\hat{f}_{\lambda}(y)$ represents the estimated model with the λ number of terms.[17]

CONIC MULTIVARIATE ADAPTIVE REGRESSION SPLINES

The conic multivariate adaptive regression splines (CMARS) model is a modified version of MARS in such a way that CMARS proposes to apply the penalized residual sum of square (PRSS) and eliminates the backward stage of MARS.[16] Hereby, after obtaining a complex model with the forward stage of MARS, PRSS which is denoted by the following equation, is calculated for the parameter estimation.

$$PRSS = \sum_{i=1}^{N}(y_i - f(\bar{x}_i))^2 + \sum_{m=1}^{M_{max}} \lambda_m \sum_{|\alpha|=1}^{2} \sum_{\substack{r<s \\ r,s \in V(m)}} \int_{Q^m} \theta_m^2 [D_{r,s}^{\alpha} h_m(t^m)]^2 dt^m \quad (4)$$

in which λ_m is the penalty parameters, $V(m)$ is the variable set associated with the m^{th} basis function h_m, t^m represents the vector of variables which contributes to the m^{th} basis function, and Q^m denotes a sufficiently large feasible regions of reflected pairs where the input data are located.

The optimization problem in PRSS is a trade-off between the accuracy and the complexity of the model. The form of PRSS given in equation (4) can be rearranged and converted as a Tikhonov regularization problem by the following expression.[18]

$$PRSS \approx \| y - h(\bar{d})\theta \|_2^2 + \lambda \| L\theta \|_2^2, \quad (5)$$

where L is a diagonal $(M_{max}+1) \times (M_{max}+1)$ dimensional matrix and θ denotes a $((M_{max}+1) \times 1)$ -dimensional vector of parameters, and finally $\|\cdot\|_2^2$ stands for the L_2 –norm of the given term. To solve this regularization problem, the conic quadratic programming can be applied.

In this study, to construct the networks by using MARS and CMARS methods, we use the same logic as given in equation (1). Hereby, to define links of a specified protein, we regress each node in the graph against all the remaining nodes and estimate the regression coefficients to detect the relation between corresponding nodes. We apply this process iteratively. That means, in each step different protein can be the response variable and the remaining are the explanatory variables. Additionally, the significant regression coefficients of the estimated models indicate that there is a relation between the response variable and the corresponding explanatory variables. Then, this relation can be represented by a link in the undirected network.

OUTLIER DETECTION METHODS

In this section, we present the brief mathematical details of the well-known outlier detection approaches under univariate and multivariate methods.

Univariate Methods

In univariate data, graphical methods, such as scatter plot and Box-plot, and statistical tests, such as Dixon's Q test (Dixon, 1950), Grubbs test (Grubbs, 1950) and z-scores (Schiffler, 1998) can be performed to detect outlier observations.[4-6] Accordingly, by using graphs, the observations which show different patterns visually from the rest of data, can be considered as outliers.

In this branch, the *Box-plot* (Tukey, 1977) is the most well-known graphical method which is appropriate to identify outliers for univariate data.[7] It is a nonparametric and robust approach, since it is based on quartiles. In this method, initially the box is constructed with the 1st (Q_1), 3rd (Q_3) quartiles and a line, which represents the median. Then, the interquartile range (IQR) is calculated within the range of 1st and 3rd quartiles indicating the length of the box. In the decision of outliers Tukey (1977) uses measurements which are smaller than $(Q_1 - 1.5 \times \text{IQR})$ or greater than $(Q_3 + 1.5 \times \text{IQR})$. Even though the graphical approaches generally, are the most preferable methods to investigate outliers due to their friendly usage and simplicity, they cannot be applicable for more than 2-dimensional datasets.

On the other hand, among statistical tests, the *Dixon's Q* approach is a very simple test which enables to find a single outlier. Hence, this test is suitable for data containing few number of observations and it is solely applicable under normally distributed data. Furthermore, the test cannot be implemented more than once in a dataset since it can identify only one outlier in a given set.

The *Grubbs'* test is an another outlier detection method for the univariate data. Similar to the Dixon's Q test, it applies the normality assumption and finds the value that is the furthest from the sample mean as an outlier. Therefore, the tested data are taken as the minimum and maximum values. If the test is applied as a one-sided test, a single outlier can be identified by choosing either the minimum or maximum value of the dataset. On the other side, the two-sided Grubbs' test is conducted both the minimum and maximum values are assigned as outliers.

The z-*scores,* also called the standardized value, is the most common and general method among alternatives and represents the distance between the observations as well as their mean in units of the standard deviation by computing the following expression for each observation x_i.

$$z_{ij} = \frac{x_{ij} - \bar{x}_i}{s} \quad i = 1, \ldots, n \text{ and } j = 1, \ldots, p \tag{6}$$

in which $\bar{x}$ is the sample mean and s denotes the standard deviation of sample. In the testing procedure, the z-scores greater than 3 or less than -3 are considered to be an outlier. Whereas, the major drawback of this method is that it is based on the mean and the standard deviation which are not robust measures against outliers.

On the other hand, there is also an extended version of z-score, called as the *modified z-score* method, proposed by Iglewicz and Hoaglin (1993).[19] This approach uses the median and the median absolute deviation (MAD) instead of the mean and the standard deviation. The median and MAD are robust measures of the central tendency and the dispersion, respectively. Moreover, if the absolute value of this score is greater than 3.5, the corresponding observation is assigned as an outlier.

Multivariate Methods

As mentioned earlier, the outlier detection with graphical tools is not suitable for more than three dimensions. Furthermore, under multivariate dimensional data, analyzing each dimension separately is not an appropriate way since some observations may be outliers merely in the multivariate space. Therefore, it is necessary to apply a multivariate analysis which can consider the interactions among several variables. Moreover, the datasets with multiple outliers can suffer from masking and swamping effects. In the masking effect, one outlier masks a second outlier in such a way that an observation can be an outlier by itself, however cannot be observed in the presence of the first outlier. Thus, other outliers come out if and only if the first outlier is deleted. On the other hand, in the swamping effect, one outlier swamps a second observation in such a way that when the first outlier is deleted, the second observation becomes a

non-outlying observation.[20] Because of these effects, the identification of outliers in the multivariate data becomes a challenging problem.

The multivariate outlier detection methods can be basically divided into two parts, i) distance based approaches and ii) projection based approaches. The main idea of the distance based methods is to detect outliers by computing a measure of how far each point is from the center of the data. A well-known distance measure in this branch is the *Mahalanobis* distance (MD). For an observation x, MD is defined as the following way.

$$MD(x) = \sqrt{(x - \bar{x})^t S^{-1} (x - \bar{x})}, \tag{7}$$

where $\bar{x}$ is the sample mean and S refers to the sample covariance matrix. In this equation, $(\cdot)^t$ and $(\cdot)^{-1}$denote the transpose and the inverse of the given vector or matrix, respectively. Once MDs are computed, the data point which has the biggest MD is labeled as the outlier. Moreover, the MD values under the multivariate normal data are approximately Chi-square distributed. On the other side, a drawback of this measure is that it includes a sample mean and a covariance which are not robust, resulting in affecting from outliers. Thereby, in order to obtain reliable results, MD needs to be estimated with robust measures. One of the most widely used robust estimators for the multivariate location and the scatter is the *minimum covariance determinant* (MCD).[21] If the MCD estimators are used in place of MD, the robust distance (RD) can be computed as.[22]

$$RD(x) = \sqrt{(x - \hat{\mu}_{MCD})^t \hat{\Sigma}_{MCD}{}^{-1} (x - \hat{\mu}_{MCD})} \tag{8}$$

in which $\hat{\mu}_{MCD}$ and $\hat{\Sigma}_{MCD}$ denote the MCD estimates of location and covariance parameters, respectively.

The *blocked adaptive computationally efficient outlier nominators* (BACON) is an another approach based on a robust multivariate method.[9] BACON is an iterative approach which uses the methods of Hadi (1992,1994).[8,23] But it is computationally more efficient. Therefore, it can be applied for the datasets which include thousands of observations. The Hadi's method initially defines a clean basic subset of the data and this subset is presumed to be outlier- free. Then, the subset iteratively grows by adding one observation in each step with a forward search. This computation can be computationally very demanding for large datasets. On the other hand, in the BACON method, a potentially large number of points, called the block of points, are added to the basic subset in each iteration, rather than adding one by one. As a result, BACON ends up with very few steps without taking into account the sample size.[9]

On the other hand, the main idea of the projection pursuit method is to obtain lower dimensional projections which include useful information for discovering the structure of data. Because these methods convert the data via a suitable projection to easily visualize the outliers. Moreover, they are not restricted with distributional assumptions and are applicable for several data structure, particularly, where the size of the variables is much more than the observations $(p \gg n)$.

A well-known projection pursuit method is the *principal component analysis* (PCA). This dimension reduction approach uses a few number of principal components (PC) to represent the data. Here, the direction of orthogonal components is defined by maximizing the variance and a meaningful information can be obtained with a small number of components. In this case the remaining majority of components are accepted as noise and are not counted within the total variance.[10] If the PCA method is applied to detect

outliers, it is necessary to use the robust estimator of the covariance matrix since the variance is highly sensitive to outliers.[3] Filzmoser et al. (2008) propose a *robust PCA* (PCOut) method for this purpose.[10] This method consists of two main parts. The first one is to detect the location outliers (e.g., mean-shift outliers) and the other part is to identify the scatter outliers (e.g., variance inflation outliers). In this computation, initially, the PCOut method robustly scales each variable by using the coordinate wise median and the median absolute deviation (MAD). Then, a semi-robust PCA is obtained. These newly obtained small numbers of PCs explain at least 99% of the total variance. By this way, this method can be easily applied to the data which have larger number of variables than observations ($p \gg n$) as it solves the singularity problem of the covariance matrix. In PCOut, the absolute value of a robust kurtosis measure is calculated for each component to detect the location outliers. On the other hand, the scatter outliers are detected via the semi robust PCs.

Another projection method which is also based on the robust principle components is suggested by Locantore et al. (1999).[11] In this method, first of all, all the observations are projected onto the boundary of a sphere (or an ellipsoid) with a unit radius. Then, the spatial median and the spatial sign covariance matrix are estimated to obtain robust estimators. Finally, the robust Mahalonobis distance can be computed for each observation. Similar to the other methods, the observations with large distances are the possible outliers. Locantore et al. (1999) call this method as the *spherical PCA*, whereas, Filzmoser et al. (2008) refer the same method as the *Sign* approach in their R package, called mvoutlier.[10,11]

In this study, we use several univariate and multivariate outlier detection methods to detect outliers in different dimensional biological data. Within the univariate methods, we select the z-score and Box-plot (IQR) approaches as they are the most common approaches in this field and other methods have certain limitations in the application. For example, the Dixon's test is suitable only for the data whose sample size is in the range 3 -30. Furthermore, the Dixon's test and the Grubbs' test can detect at most 2 outliers in a dataset. On the other hand, within multivariate methods, we apply PCOut, Spherical PCA, i.e., Sign, and BACON methods. The first two methods are the projection pursuit approaches and are more suitable for the data structure where the size of the variables are much more than the observations ($p \gg n$). This situation is commonly observed in biological data. Because in most of genomic datasets, there may be thousands of genes or proteins measured on few samples. Majority of the outlier detection methods, specifically, distance based methods, are not applicable under the $p \geq n$ situation due to the singularity problem in inference of the covariance matrix. In the following section, the application of the methods used in this study are explained in details.

APPLICATION

DESCRIPTION OF DATASETS

To evaluate the performance of outlier detection methods under biological data, we use different synthetic and real benchmark datasets. In this application, the synthetic data are taken from benchmark synthetic data tools such as SynTRen (Synthetic transcriptional regulatory networks) and GeneNetWeaver (GNW).[24,25] The synthetic gene expression datasets from these tools approximate the experimental data and enable us to validate the methods under different network constructions as they present the true network model too. The names of all datasets used in this study and their numbers of genes as well as samples are listed (Table 1). The number of outliers and their percentages which are detected and removed in each dataset are also shown (Table 1). Additionally, the biological details of all datasets and the results of the analysis are presented in the following part.

TABLE 1: List of the datasets and the number of outliers detected under PCOut, Sign, Z-Score and Box-Plot methods. Percentages of outliers are shown in parentheses. n represents the number of samples per protein and p denotes the number of proteins.

ID	Dataset	n	p	PCOut	Sign	Z-Score	Box-Plot
1	Insilico-size 10-1	10	10	3 (30)	0 (0)	0 (0)	1 (10)
2	Insilico-size 10-3	10	10	2 (20)	1 (10)	0 (0)	2 (20)
3	Insilico-size 10-5	10	10	2 (20)	2 (20)	0 (0)	5 (50)
4	Insilico-size 100-1	100	100	25 (25)	24 (24)	33 (33)	81 (81)
5	Insilico-size 100-4	100	100	13 (13)	18 (18)	40 (40)	85 (85)
6	Gene expression (Toy)	64	64	16 (25)	18 (28)	16 (25)	41 (64)
7	NCI-60 cell lines (p53)	33	20	13 (39)	13 (39)	0 (0)	1 (3)
8	Human gene expression	60	100	18 (30)	16 (27)	15 (25)	51 (85)
9	Egeod-9891	285	11	43 (15)	26 (9)	19 (7)	45 (16)
10	Cell signal data	11672	11	4195 (36)	3792 (32)	1112 (10)	4728 (40)

Gene Expression Datasets from DREAM 4

In our analyses, the gene expression datasets from DREAM 4 (2009) in-silico size 10 and size 100 are used.[26] DREAM, International Dialogue for Reverse Engineering Assessments and Methods competition, publishes challenges in network inference and the DREAM 4 in-silico size 10-challenge consists of five networks involving 10 genes. Moreover, the in-silico size 100-challenge includes five networks involving 100 genes. For each network, multifactorial data are available and contain steady-state levels of variations in the network based on gene expression characteristics of two well-studied systems, namely, Escherichia coli (E.coli) and Saccharomyces cerevisiae (S.cerevisiae). These datasets are accessible from the R package DREAM4. Hence, in our analyses, we use three datasets from in-silico size 10 and two datasets from in-silico size 100.

Gene Expression Data (toy)

We compare the performance of three methods with a simulated toy example of gene expression data generated by the GNW generator by using known biological interaction networks of E.coli. The toy data contain 64 samples and 64 genes, and are available in the R package grndata.[27]

PGC-1A Pathway

As a real biological dataset, first of all, we apply a part of the NCI-60 cell lines (p53) data which include "Biocarta-PGC1A-Pathway" described in the study of Rahmatallah et al. (2014).[28] The p53 dataset consists of the gene expression profiling of 33 p53 mutated (MUT) cancer cell lines and is taken from the R package GANPAdata.[29] Here, we deal with 20 genes which constitute the PGC1A pathway. The peroxisome proliferator-activated receptor gamma coactivator-1 alpha (PGC-1A) is a tissue-specific coactivator that coordinates transcriptional programs important for energy metabolism and energy homeostasis. Thereby, inappropriate increases in the PGC-1A activity are linked to a number of pathological conditions including heart failures and diabetes. On the other side, PGC-1A is a coactivator for many factors including, CBP, Scr-1, PPARa, GR (glucocorticoid receptor), THR (thyroid hormone receptor), several orphan receptors and MEF2.

Human Gene Expression Data

For the second real dataset, we implement large-scale human gene expression data originally described in the works of Bhadra and Mallick (2013), Chen and Chen, (2008) and Stranger et al. (2007).[30-32] This dataset

includes the gene expression of B-lymphocyte cells from the Utah residents with Northern and Western European ancestry sample. The genes of 60 unrelated individuals are probed for 100 different transcripts. Here, the focus is on the 3125 Single Nucleotide Polymorphisms (SNPs) that are found in the 5' UTR (untranslated region) of mRNA (messenger RNA) with a minor allele frequency greater than 0.1. The UTR has an important role in the regulation of gene expressions. From the 55 biologically validate links, 45 have the names of the transcription factor and target genes in the network of gene expression data.[30] Therefore, for the inference of all models we use these 45 links for the calculation of the accuracy measures.

Human Ovarian Tumour Samples (E-GEOD-9891)

As a third real biological dataset, we apply the transcription profiling of 285 humans ovarian tumour samples named as E-GEOD-9891. The data are cohorts of 285 patients with epithelial ovarian, primary peritoneal, or fallopian tube cancer, diagnosed between 1992 and 2006. The samples are collected through Australian Ovarian Cancer Study with a sample size (n=206), Royal Brisbane Hospital (n=22), Westmead Hospital (n=54) and Netherlands Cancer Institute (n=3).[33] In this dataset, there are 11 target genes which are chosen to identify the novel molecular subtypes of the ovarian cancer by the gene expression profiling with a linkage to clinical and pathological features.[33]

Cell Signaling Pathway

Finally, we use cell signaling data from Sachs et al. (2005).[34] This dataset contains the flow cytometry results of 11 phosphorylated proteins and phospholipids measured on 11672 red blood cells. These components belong to the cellular protein-signaling network of human immune system cells. Therefore, the aim of the construction of this network is to understand the native-state tissue signaling biology, complex drug actions and the disfunctional signals in diseased cells.[34]

RESULTS

In this application, we evaluate the performance of outlier detection methods under several biological datasets. Our main aim is to investigate whether an outlier detection approach is necessary as a pre-processing step before modeling. For this purpose, as mentioned earlier, we perform different outlier detection methods designed for univariate and multivariate cases. We apply z-score and Box-plot methods within the univariate approaches. Additionally, we use PCOut, Sign method and BACON as multivariate methods. First of all, we detect the outliers in each dataset by using all these methods and remove the outliers. Then, we check the accuracy of estimated systems with these outlier-free datasets under 3 network models, namely, GGM, MARS and CMARS.

Hereby, under such network constructions, the significant regression coefficients of all the estimated models are used to indicate the relation between the response variable and the corresponding explanatory variables. Then, this relation can be represented by a link (also named as edges between proteins) in the undirected network. By this way, the estimates of coefficients are applied to construct an *adjacency matrix* which is composed of zero and one values. Here, zero implies no link, one stands for link between each pair of proteins. Therefore, in the end of the computation, the ordinary regression problem becomes a classification problem in the construction of the estimated networks and thus, in the assessment of the models, we use the binary classification performance measures. For this calculation, we need certain scores and thereby, initially compute true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Here, TP implies the number of correctly classified objects that have positive labels that means there is an edge, TN indicates the number of correctly classified objects that have negative labels which denotes no edge, FP shows the number of misclassified objects that have negative labels, and FN

presents the number of misclassified objects that have positive labels. Then, we calculate the accuracy and F-measure as shown below.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \cdot \tag{9}$$

$$\text{F} - \text{measure} = \frac{2TP}{2TP+FP+FN} \cdot \tag{10}$$

From the expressions above, the accuracy is the ratio of correctly classified genes in both labels to the total of all classified genes. That means the ratio of correctly estimated links (both positive and negative) over total links. Additionally, F-measure indicates the overlap between actual and predicted classes.

After these calculations, we compare the accuracy results of models under the full data which includes potential outliers with the results of outlier-free datasets. All the accuracy results for each dataset and each method are given (Table 2 and Table 3). As presented, removing the outlier from our datasets which are

TABLE 2: Comparison of F-measure and accuracy values of Z-score and Box-plot methods under CMARS, GGM and MARS approaches for datasets which are listed in TABLE 1. (*) represents that outliers are not detected and (-) denotes that the models are not computable.

		Full data		Z-Score		Box-Plot	
Data	Methods	F-Measure	Accuracy	F-Measure	Accuracy	F-Measure	Accuracy
1	CMARS	0.647	0.760	*	*	0.563	0.720
	GGM	0.454	0.760	*	*	0.454	0.760
	MARS	0.616	0.800	*	*	0.560	0.780
2	CMARS	0.563	0.720	*	*	0.533	0.720
	GGM	0.416	0.720	*	*	0.416	0.720
	MARS	0.538	0.760	*	*	0.384	0.680
3	CMARS	0.583	0.800	*	*	-	-
	GGM	0.476	0.780	*	*	0.476	0.780
	MARS	0.522	0.780	*	*	0.476	0.780
4	CMARS	0.296	0.943	0.268	0.942	-	-
	GGM	-	-	0.373	0.966	0.373	0.966
	MARS	0.298	0.949	0.295	0.951	0.362	0.963
5	CMARS	0.267	0.934	0.280	0.939	0.277	0.946
	GGM	-	-	0.326	0.959	0.326	0.959
	MARS	0.258	0.939	0.280	0.946	0.321	0.956
6	CMARS	0.347	0.842	0.318	0.837	-	-
	GGM	0.236	0.899	0.780	0.991	0.236	0.899
	MARS	0.342	0.895	0.326	0.891	0.283	0.898
7	CMARS	0.387	0.810	*	*	0.387	0.810
	GGM	0.532	0.675	*	*	0.500	0.620
	MARS	0.432	0.790	*	*	0.377	0.785
8	CMARS	0.805	0.978	0.733	0.991	0.636	0.987
	GGM	0.178	0.831	0.168	0.817	0.224	0.911
	MARS	0.612	0.983	0.594	0.983	0.638	0.988
9	CMARS	0.825	0.703	0.837	0.719	0.790	0.653
	GGM	0.167	0.091	0.167	0.091	0.167	0.091
	MARS	0.469	0.306	0.487	0.322	0.506	0.339
10	CMARS	0.715	0.835	0.594	0.785	0.594	0.785
	GGM	0.448	0.736	0.594	0.785	0.625	0.752
	MARS	0.690	0.785	0.649	0.686	0.559	0.570

TABLE 3: Comparison of F-measure and accuracy values of PCOut, Sign and BACON methods under CMARS, GGM and MARS approaches for datasets which are listed in TABLE 1. (-) denotes the models that are not computable.

		Full data		PCout		Sign		BACON	
Data	Methods	F-Measure	Accuracy	F-Measure	Accuracy	F-Measure	Accuracy	F-Measure	Accuracy
	CMARS	0.647	0.760	-	-	0.647	0.760	0.563	0.720
1	GGM	0.454	0.760	0.454	0.760	0.454	0.760	0.454	0.760
	MARS	0.616	0.800	0.616	0.800	0.616	0.800	0.435	0.740
	CMARS	0.563	0.720	0.533	0.720	0.500	0.680	0.514	0.660
2	GGM	0.416	0.720	0.416	0.720	0.416	0.720	0.416	0.720
	MARS	0.538	0.760	0.384	0.680	0.462	0.720	0.462	0.720
	CMARS	0.583	0.800	0.560	0.780	0.560	0.780	0.435	0.740
3	GGM	0.476	0.780	0.476	0.780	0.470	0.780	0.476	0.780
	MARS	0.522	0.780	0.584	0.800	0.545	0.800	0.500	0.760
	CMARS	0.296	0.943	0.258	0.941	0.281	0.942	0.263	0.938
4	GGM	-	-	0.373	0.966	0.373	0.966	0.373	0.966
	MARS	0.298	0.949	0.373	0.966	0.288	0.950	0.328	0.948
	CMARS	0.267	0.934	0.281	0.938	0.253	0.934	0.259	0.934
5	GGM	-	-	0.326	0.959	0.326	0.959	0.326	0.959
	MARS	0.258	0.939	0.332	0.959	0.285	0.943	0.259	0.938
	CMARS	0.347	0.842	-	-	0.295	0.832	0.205	0.807
6	GGM	0.236	0.899	0.236	0.899	0.236	0.899	0.236	0.899
	MARS	0.342	0.895	0.325	0.897	0.325	0.895	0.328	0.888
	CMARS	0.387	0.810	0.361	0.805	0.413	0.815	0.344	0.790
7	GGM	0.532	0.675	0.400	0.565	0.397	0.560	0.370	0.405
	MARS	0.432	0.790	0.367	0.775	0.438	0.795	0.351	0.815
	CMARS	0.805	0.978	0.709	0.991	0.704	0.991	-	-
8	GGM	0.178	0.831	0.199	0.852	0.207	0.859	-	-
	MARS	0.612	0.983	0.656	0.986	0.645	0.985	-	-
	CMARS	0.825	0.703	0.778	0.636	0.825	0.703	-	-
9	GGM	0.167	0.091	0.167	0.091	0.167	0.091	-	-
	MARS	0.469	0.306	0.429	0.273	0.448	0.289	-	-
	CMARS	0.715	0.835	0.594	0785	0.594	0.785	0.367	0.686
10	GGM	0.448	0.736	0.658	0.785	0.690	0.785	0.392	0.719
	MARS	0.690	0.785	0.554	0.587	0.559	0.570	0.560	0.521

detected by the z-score or Box-plot methods cannot increase the performance of our models (Table 2). Generally, the F-measure and accuracy values either remain the same or small decreases occur. Additionally, the z-score method does not identify an outlier in some datasets (Table 2). On the other hand, the results of the multivariate outlier detection methods, namely, PCOut, Sign and BACON, are presented (Table 3). From the results, we can see that for the first five datasets which are synthetic data, the PCOut method is slightly better than sign and BACON, especially, for the higher dimensional Data 4 and Data 5. Additionally, for the remaining real datasets, regression methods with the raw datasets give better results. Rarely, the Sign method gets the upper hand, but the increases in the F-measure and accuracy are very low. In other words, overall, there is no any outlier detection method which outperform the others. Finally, when we compare the regression models, we can observe that the performance of CMARS is the best and GGM is the worst as declared in the previous studies.[13,15] As a result, it is seen that such a pre-processing step cannot improve the accuracy of the regression models under our selected datasets and there are only a few methods applicable for sparse and high-dimensional systems.

DISCUSSION

Outlier detection is an important step in both the data analysis and modeling the observations in many research fields and therefore, there are many methods for this purpose in the literature. Dixon's Q test, Grubbs test, z-scores and Box-plot are the well-known fundamental outlier detection methods for univariate data. However, the datasets with multiple outliers can suffer from masking and swamping effects. Therefore, it is necessary to apply a multivariate analysis which can consider the interactions among several variables. Hence, there are different methods suitable for multivariate data which are basically classified into two parts, namely, distance based and projection based methods. The widespread distance measures are the Mahalanobis distance and robust distance. Unfortunately, these methods are not applicable under the $p \geq n$ situation due to the singularity problem in the estimation of the covariance matrix. On the other hand, PCA, robust PCA, Sign methods and BACON are widely used projection based methods. The projection based approaches can obtain lower dimensional projections and include useful information for discovering the structure of data. The main advantage of the projection based methods is that they are applicable for the data type which has larger number of variables than observations. This sort of data is frequently observed in biological fields and can cause the computational problems for many methods due to their sparse and high dimensional structures. Therefore, there are only a few methods suitable for this data type.

CONCLUSION

In this study, we have investigated whether an outlier detection approach is necessary as a pre-processing step before modeling the biological networks, especially, the protein-protein interaction data. For this purpose, we have searched several outlier detection methods designed for univariate and multivariate data. We have chosen the z-score and Box-plot as univariate methods and robust PCA (PCOut), Sign method and BACON within multivariate methods. We have performed all these outlier detection methods as a pre-processing step and removed the outliers from the data before modeling. Additionally, for modeling we have applied GGM, MARS and CMARS approaches. Then, we have checked the accuracy of the estimated models by using accuracy measures such as, F-measure and accuracy values. For this purpose, we have used several synthetic and real benchmark biological datasets which have different dimensions and different structures. From the results, it has seen that the outlier detection as a pre-processing step cannot improve the accuracy of the models apart from some cases, where the regression models give slightly better results when outliers are detected and removed with the PCOut or Sign methods. However, the increases in the accuracy values are very low. Hence, within the outlier detection methods, we have not found any method which outperforms the others in the construction of our biological networks.

As a future work, we consider to suggest an alternative outlier detection method which is applicable, particularly, for sparse and high dimensional datasets. Furthermore, we think to propose new approaches to improve the accuracy of the fitted models on the construction of the protein-protein interaction networks.

Acknowledgement

The authors thank to the editor and anonymous referees for their valuable comments which improve the quality of our paper. Furthermore, they would like to thank the BAP project at Middle East Technical University (Project no: BAP- 01-09-2017-002) for its support.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and / or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

Idea/Concept: *Ezgi Ayyıldız, Vilda Purutçuoğlu*; ***Design:*** *Ezgi Ayyıldız, Vilda Purutçuoğlu*; ***Control/Supervision:*** *Ezgi Ayyıldız, Vilda Purutçuoğlu*; ***Analysis And/Or Interpretation:*** *Ezgi Ayyıldız, Vilda Purutçuoğlu*; ***Literature Review:*** *Ezgi Ayyıldız, Vilda Purutçuoğlu*; ***Writing The Article:*** *Ezgi Ayyıldız, Vilda Purutçuoğlu*; ***Critical Review:*** *Ezgi Ayyıldız, Vilda Purutçuoğlu*

REFERENCES

1. Hawkins DM. Introduction. Identification of Outliers. 1st ed. London: Chapman & Hall; 1980. p.1-12.
2. Aggarwal CC. An introduction to outlier analysis. Outlier Analysis. 1st ed. New York: Springer-Verlag; 2013. p.1-33.
3. Hadi AS, Imon AHMR, Werner V. Detection of outliers. Wiley Interdiscip Rev Comput Stat 2009;1(1):57-70.
4. Dixon WJ. Analysis of extreme values. Ann Math Statist 1950;21(4):488-506.
5. Grubbs FE. Sample criteria for testing outlying observations. Ann Math Statist 1950;21(1):27-58.
6. Schiffler RE. Maximum Z scores and outliers. Am Stat 1998;42(1):79-80.
7. Tukey JW. Schematic summaries: Box and whisker plots. Exploratory Data Analysis. 18th ed. Reading, MA: Addison-Wesley; 1977. p.39-56.
8. Hadi AS. Identifying multiple outliers in multivariate data. J R Stat Soc Ser B 1992;54(3):761-71.
9. Billor N, Hadi AS, Velleman PF. BACON: blocked adaptive computationally efficient outlier nominators. Comput Stat Data Anal 2000;34:279-98.
10. Filzmoser P, Maronna RA, Werner M. Outlier identification in high dimensions. Comput Stat Data Anal 2008;52(3):1694-711.
11. Locantore N, Marron JS, Simpson DG, Tripoli N, Zhang JT, Cohen KL, et al. Robust principal component analysis for functional data. Test 1999;8(1):1-73.
12. Whittaker J. Graphical Gaussian models. Graphical Models in Applied Multivariate Statistics. 1st ed. New York: John Wiley & Sons; 1990. p.155-97.
13. Ayyıldız E, Ağraz M, Purutçuoğlu V. MARS as an alternative approach of Gaussian graphical model for biochemical networks. J Appl Stat 2017;44(16):2858-76.
14. Friedman JH. Multivariate adaptive regression splines. Ann Statist 1991;19(1):1-67.
15. Ayyıldız E, Purutçuoğlu V, Weber GW. Loop-based conic multivariate adaptive regression splines is a novel method for advanced construction of complex biological networks. EJOR 2018;270:852-61.
16. Yerlikaya-Özkurt F, Batmaz I, Weber GW. A review and new contribution on conic multivariate adaptive regression splines (CMARS): a powerful tool for predictive data mining. In Modeling, Dynamics, Optimization and Bioeconomics I. Vol 73. Chapter 38. Switzerland: Springer International Publishing; 2014. p.695-722.
17. Hastie T, Tibshirani R, Friedman J. MARS: Multivariate adaptice regression splines. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York: Springer Verlag; 2009. p.321-9.
18. Weber GW, Batmaz İ, Köksal G, Taylan P, Yerlikaya-Özkurt F. CMARS: A new contribution to nonparametric regression with multivariate adaptive regression splines supported by continuous optimization. Inverse Probl Sci Eng 2012;20:371-400.
19. Iglewicz B, Hoaglin DC. Outlier labeling. How to Detect and Handle Outliers. 1st ed. Milwaukee: WI ASQC Quality Press; 1993. p.10-5.
20. Ben-Gal I. Outlier detection. In: Maimon O, Rokach L, eds. Data Mining and Knowledge Discovery Handbook. 1st ed. Boston: Springer; 2005. p.131-46.
21. Rousseeuw PJ, Van Driessen K. A fast algorithm for the Minimum Covariance Determinant estimator. Technometrics 1999;41:212-23.
22. Hubert M, Debruyne M. Minimum covariance determinant. Wiley Interdiscip Rev Comput Stat 2010;2(1):36-43.
23. Hadi AS. A modification of a method for the detection of outliers in multivariate samples. J R Stat Soc Ser B 1994;56(2):393-6.
24. Van den Bulcke T, Van Leemput K, Naudts B, van Remortel P, Ma H, Verschoren A, et al. SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. BMC Bioinformatics 2006;7(1):43.
25. Schaffter T, Marbach D, Floreano D. GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. Bioinformatics 2011;27(16):2263-70.
26. Shannon P. DREAM4: Synthetic expression data for gene regulatory network inference from the 2009 DREAM4 challeng. R Package Version 1.16.0; 2013. Doi: 10.18129/B9.bioc.DREAM4.
27. Bellot P, Olsen C, Meyer PE. grndata: synthetic expression data for gene regulatory network inference. R package version 1.12.0; 2014. Doi: 10.18129/B9.bioc.grndata.

28. Rahmatallah Y, Emmert-Streib F, Glazko G. Gene sets net correlations analysis (GSNCA): a multivariate differential coexpression test for gene sets. Bioinformatics 2014;30(3):360-8.
29. Fang Z, Tian W, Ji H. A network-based gene-weighting approach for pathway analysis. Cell Res 2012;22(3):565-80.
30. Bhadra A, Mallick BK. Joint high-dimensional Bayesian variable and covariance selection with an application to eQTL analysis. Biometrics 2013;69(2):447-57.
31. Chen J, Chen Z. Extended Bayesian information criteria for model selection with large model spaces. Biometrika Trust 2008;95(3):759-71.
32. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, et al. Population genomics of human gene expression. Nat Genet 2007;39(10):1217-24.
33. Tothill RW, Tinker AV, George J, Brown R, Fox SB, Johnson DS, et al. Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. Clin Cancer Res 2008;14(16):5198-208.
34. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP. Causal protein-signaling networks derived from multiparameter single-cell data. Science 2005;308(5721):523-9.