Anomalies Detection in Gene Expression Matrices: Towards a New Approach

Nicoletta Del Buono[1][a], Flavia Esposito[1,2][b], Laura Selicato[1][c] and Maria Carmela Vegliante[2][d]

[1]*Members of INDAM-GNCS Research Group, Department of Mathematics, University of Bari Aldo Moro, via E. Orabona 4, I-70125, Bari Italy*

[2]*Hematology and Cell Therapy Unit, IRCCS - Istituto Tumori Giovanni Paolo II, Bari, Italy*

Keywords: Outlier Detection, Gene Expression Profiling, Clustering, Robust PCA.

Abstract: One of the main problems in analyzing real data is often related to the presence of anomalies. Anomalous cases may, in fact, spoil the resulting analysis as well as contain valuable information at the same time. In both cases, the ability to detect these occurrences is very important. Particularly, in biomedical field, a proper identification of outliers allows to develop novel biological hypotheses not taken into consideration when experimental biological data are considered. In this paper, we address the problem of detecting outlier samples in gene expression data. We propose an ensemble approach for anomalies detection in gene expression matrices based on the use of hierarchical clustering and Robust Principal Component Analysis, that allows to derive a novel pseudo mathematical classification of anomalies.

1 INTRODUCTION

Real datasets often contain observations which behave differently from the majority of data. When an occurrence is different from the dominant part of the data or is sufficiently unlikely under the assumed data probability model, it is considered as an anomaly or outlier.

Outliers may be caused by errors, but they may result from exceptional circumstances, or belong to another data population. On the one hand, anomalies may produce deleterious effect on the conclusions drawn from the data analysis, on the other hand, they may contain important information. Hence, the concern of detecting outliers lies on the interest of the outliers themselves or on the fact that they could contaminate the downstream statistical analysis.

In biomedical field, an outlier can be an abnormal sample that deviates significantly from the other samples in its class. Typically, this occurs when a sample of one class is accidentally assigned to another class. In a context of carcinogenic pathology, this may mean that such a patient's disease is a special case.

Hence, in biological datasets, a proper outlier identification could be of some interest: in fact, depending on the type of analysis to be performed, this would allow biologists to consider whether these data should be removed or not.

In this paper, we address the problem of detecting outliers in Gene Expression Profiling (GEP) data, that is microarray data which contain gene expression levels for a given number of samples labeled with a biological class (tumor type or experimental condition). In microarrays there are two main types of outliers referred to the case when instances are genes or samples, respectively (Shieh and Hung, 2009). The former is present when a gene has abnormal expression values in one or more samples from the same class. Whereas, the latter can be seen dually as samples that belong to a different class present in the data (often referred to as mislabeled samples) or as samples that do not belong to any class present in the data (called abnormal samples). The origin of these outliers can be ambiguous, they can result from an undiscovered biological class, poor class definitions, experimental errors, or extreme biological variability. Note that when we say that an anomalous sample does not belong to its class, we are not necessarily contesting the validity of its label. In fact, a sample may still be a tumor, but having expression levels that differ considerably from those of other tumor samples.

[a] https://orcid.org/0000-0001-5079-875X
[b] https://orcid.org/0000-0002-2791-9610
[c] https://orcid.org/0000-0001-9248-3879
[d] https://orcid.org/0000-0002-3165-1768

Buono, N., Esposito, F., Selicato, L. and Vegliante, M.
Anomalies Detection in Gene Expression Matrices: Towards a New Approach.
DOI: 10.5220/0010342301620169
In *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2021) - Volume 3: BIOINFORMATICS*, pages 162-169
ISBN: 978-989-758-490-9

Applying models to data affected by outliers can even produce incorrect inferences. In the past, the influence of outliers was rarely considered when analyzing data from standard microarrays. According to the new current, outlier detection is used as a preprocessing on data for their cleaning. However, it is substantial to emphasize that, in many cases, outliers may simply be the result of natural variability in the data.

In this work we propose a novel outlier detection approach, which combines Hierarchical Clustering and Robust Principal Component Analysis. This ensemble mechanism, which joins two techniques generally not adopted in this context, allows to derive a pseudo mathematical classification of outlier samples in GEP data. The obtained classification could be then used to propose a new decision-making model. The model is usually chosen based on how it separates the data into two or more clusters. We propose a data preprocessing mechanism that, independently of these, identifies the anomalies and integrates the anomalies detection tool in the context of microarrays.

The paper is organized as following. Section 2 briefly overviews the two algorithms adopted in our proposal together with some main methodologies frequently used in gene expression field. Section 3 illustrates the experimental results obtained applying the proposed method on three different datasets (one artificial and two real medical datasets). Comparisons with techniques usually used to detect anomalies are also presented and the advantages of the proposed approach are discussed. Finally, conclusions and directions of future research are sketched in Section 4.

2 METHODS FOR OUTLIERS DETECTION

The approach we propose is based on two important techniques, which are already independently used for anomaly detection.

Clustering can be considered the most important unsupervised learning problem to find a structure in a collection of unlabeled data. A cluster is therefore a group of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters. The outliers are therefore those samples belonging to a separate micro cluster, because they are distant from most of the other data. They are usually identified by increasing the number of clusters. In particular, Hierarchical clustering allowing to select a distance measure is chosen. In gene expression data analysis, when clusters of observations with the same overall profiles need to be achieved, correlation-based distance (used as a dissimilarity measure) has to be considered the appropriate choice.

On the other hand, distance is not the only parameter to be set in clustering algorithms, also the method defining how to separate two different clusters is a task to be managed. In our experiments we used Pearson correlation distance and Average method, according to the empirical criterion assessing their stability described in Section 3.

The second technique involved in the proposed approach is Robust Principal Component Analysis (ROBPCA) method (Hubert et al., 2005), which combines the strengths of Projection-Pursuit techniques (PP) (Croux et al., 2007) and robust covariance estimation. The former is used for reducting the initial dimensionality, whereas the second, in particular the Minimum Covariance Determinant (MCD) estimator, is applied to the obtained smaller data space.

Consider an $n \times p$ data matrix $\mathbf{X} = \mathbf{X}_{n,p}$, where n indicates the number of the observations and p the original number of variables, the ROBPCA method proceeds in three main steps:

1. the data are pre-processed such that the transformed data are lying in a subspace whose dimension is at most $n-1$.
2. a preliminary covariance matrix S_0 is constructed and used for selecting the number of components k that will be retained in the sequel, yielding a k-dimensional subspace well fitted to the data.
3. data points are projected on this subspace where their location and scatter matrix are robustly estimated, from which its k nonzero eigenvalues $\ell_1, \dots \ell_k$ are computed. The corresponding eigenvectors are the k robust principal components.

Let $P_{p,k}$ be the $p \times k$ eigenvector matrix (orthogonal columns), the location estimate is denoted by the p-variate column vector $\hat{\mathrm{v}}$ and called the robust center. The scores are the entries of the $n \times k$ matrix

$$T_{n,k} = (X_{n,p} - 1_n \hat{\mathrm{v}}^\top) \cdot P_{p,k} \tag{1}$$

The k robust principal components generate a $p \times p$ robust scatter matrix S of rank k given by

$$S = P_{p,k} L_{k,k} P_{p,k}^\top \tag{2}$$

where $L_{k,k}$ is the diagonal matrix with the eigenvalues $\ell_1, \dots \ell_k$.

Similarly to classical PCA, the ROBPCA method is location and orthogonal equivariant, these properties is not trivial for other robust PCA estimators. It should be noted that dimensionality reduction approaches are widely used in the context of microarray data analysis (Esposito et al., 2020) but, for the best of our knowledge, this is the first time that ROBPCA is applied for outlier detection in microarray data.

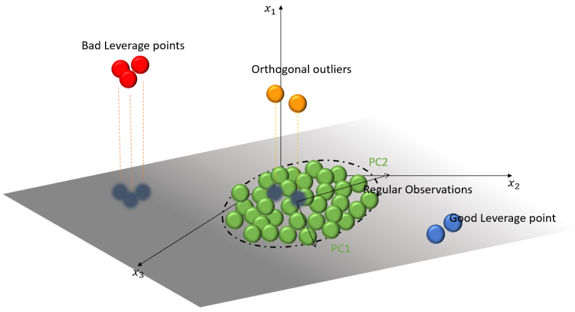

Figure 1: Outliers Classification, with $p = 3$ and $k = 2$.

An advantage of using PCA related techniques lies on the possibility of classify outlier according to their position in the projected subspace. An example of this is depicted in Figure 1, where four type of outliers according to the location of the observations, can be distinguished. The regular observations that form one homogeneous group near the PCA subspace generated by the principal components. The good leverage points, which lie on the same plane as the PCA subspace but away from the regular observations. The orthogonal outliers, which have a large orthogonal distance to the PCA subspace but their projection is on the PCA subspace. Finally, the bad leverage points, which have a large orthogonal distance and whose projection on the PCA subspace is remote from the regular observations.

To understand and quantify how far an observation is from the center of the ellipse, defined by regular observations (score = 0), two distances are adopted. The Score Distances SD_i,

$$SD_i = \sqrt{\sum_{j=1}^{k} \frac{t_{ij}^2}{\ell_j}},$$

is a measure of the distance between an observation belonging to the PCA k-dimensional subspace and the origin of that subspace. The Orthogonal Distance OD_i,

$$OD_i = ||x_i - \hat{\mu} - P_{p,k} t_i'||,$$

where ℓ are the eigenvalues of the dispersion matrix MCD and t_{ij} are the robust scores for each $j = 1, \ldots, k$, $\hat{\mu}$ is the robust estimate of the center, that measures the deviation (i.e. the lack of adaptation) of an observation from the PCA k-dimensional subspace.

Based on these measures, a plot, namely diagnostic plot or outlier map, can be constructed to distinguish between regular observations and the three types of outliers. An example of this plot is depicted in Figure 2, with the robust score distance SD_i and the orthogonal distance OD_i on the horizontal and vertical axis, respectively. In the diagnostic plot the

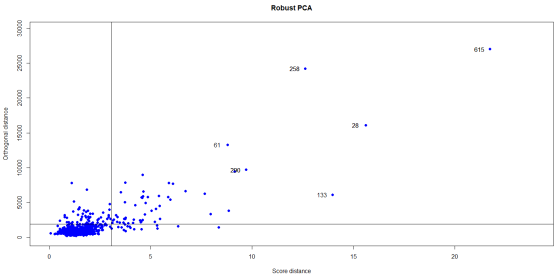

Figure 2: Outlier map obtained with ROBPCA (from the Rospca package available in R).

first quadrant at the top right contains the bad outliers, the second quadrant on the left the orthogonal outliers, the third quadrant the regular observations, finally the fourth quadrant contains the good leverage points, based on the previous classification.

To classify the observations, two cutoff lines are then drawn according to the data. The cutoff value on the horizontal axis is obtained from the 0.975 quantile of the χ-square distribution with k degrees of freedom:

$$SD > \sqrt{\chi^2_{k,.975}}.$$

The cutoff values for orthogonal distances are obtained using the Wilson-Hilferty approximation for a Chi-Squared distribution, i.e. the orthogonal distances, raised to $2/3$, are distributed approximately normally. Therefore, the cutoff values for anomalous observations are given by

$$OD > (\hat{\mu} + \hat{\sigma} z_{.975})^{\frac{3}{2}}$$

where $\hat{\mu}$ and $\hat{\sigma}$ are the MCD estimates of μ and σ of the above normal distribution. As for standard PCA approaches, also its robust variants need a criterion to choose the number of its components. An example of this is proposed in (Hubert et al., 2005) and selects k components according to $\sum_{j=1}^{k} \ell_i / \sum_{j=1}^{r} \ell_i \approx 90$ or, for instance, such that $\frac{\ell_k}{\ell_1} \geq 10^{-3}$, where ℓ_j are the eigenvalues of S_0, the robust covariance matrix of the data and r its rank. In the experimentation carried out, k was always chosen using this criterion.

In gene expression matrix analysis, anomaly detection is commonly performed using Bioconductor package *arrayQuality* (Paquet and Yang, 2010). This performs the univariate analysis based on two independent methods that assign a rank for each column. The first technique, taking one sample at time, compares its probability distribution to the one from the whole dataset with a Kolmogorov-Smirnov statistics. Instead, the other technique simply ranks each sample according the sum across all genes. At the end, the outlier samples is chosen according to the standard univariate detection approach applied on the the rank score.

3 PROPOSED APPROACH

To detect anomalous samples, we use the combined approach of Hierarchical Clustering and Robust PCA for our data typology performed sequentially. The first technique provides a preliminary view of the dataset, the choice of distance is validated by the Cophenetic Correlation Coefficient (CCC) and the adaptation to clustering by the Silhouette coefficient. The second technique characterizes the type of outlier found previously, based on where it is placed on the plot.

From the experimentation carried out (performed on a i7 octa core machine with 16Gb of RAM in R enviroment (R-Team, 2015)), outliers that are of ”low quality” are found to be very extreme bad outliers, above a certain threshold. On the contrary, the ”mislabeled” type outliers are on the border between the orthogonal type outliers and the bad leverage points.

3.1 Synthetic Dataset

As a first step for our study, we simulated a typical cancer dataset with known outliers as proposed in (Barghash et al., 2016). Each dataset contains two clearly distinguishable sample classes. Abnormal samples do not belong to either class or that are simply mislabeled.

On the rows we have 1000 genes and on the columns 100 samples (50 for each class). The first 900 lines are drawn from the same normal distribution for both classes, the remaining 100 were drawn from different distributions for samples of classes C_1 and C_2, respectively. In addition, samples 10, 15 and 20 of the C_1 class were exchanged with samples 60, 65 and 70 of the C_2 class. Finally the last sample of each class was replaced by one with a different distribution (for example the Poisson distribution). As previously discussed, to detect outliers, samples are hierarchically clustered using Pearson distance and different linkage methods. Subsequently, the CCC and the Silhoette index are used to validate the quality of clustering. The CCC expresses the correlation between the original dissimilarity matrix and the one inferred based on the classification, $CCC \geq 0.8$ denotes a good agreement, whereas $CCC < 0.8$ indicates that the dendrogram is not a good representation of the relationships between objects. Table 1 shows the CCCs corresponding to the various methods, selecting as average linkage the best method for this clustering.

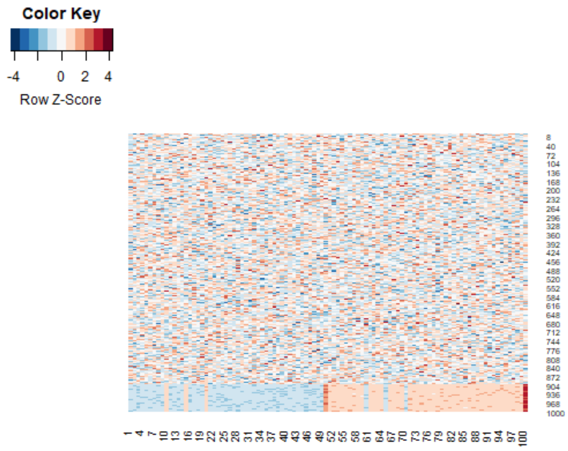

Figure 3: Heatmap of the synthetic dataset.

Table 1: CCC corresponding to the various methods.

Linkage Method	CCC
average	0.92
ward.D2	0.35
complete	0.58
single	0.9
centroid	0.64

Based on the clustering vector and on the set of distances, the algorithm calculates the average dissimilarity of a point x_i to its current class and the lowest dissimilarity of the point to other classes, indicated as $a(x_i)$ and $b(x_i)$, respectively. Formally, for all $x_i \in C_i$ the above dissimilarities are defined as follow:

$$a(x_i) = \frac{1}{|C_i|-1} \sum_{j \in C_i, i \neq j} d(x_i, x_j) \quad \text{and}$$

$$b(x_i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(x_i, x_j).$$

On the other hand, the Silhouette coefficient of a point x_i is defined as

$$s(x_i) = \frac{b(x_i) - a(x_i)}{\max\{a(x_i), b(x_i)\}},$$

$s(x_i)$ ranges between $]-1,1[$ where 1 indicates a better fit to the current cluster and -1 means that the point actually belongs to the other class or a so called neighboring cluster. In fact by definition follows:

- $s(x_i) = 0$ if $|C_i| = 1$,
- $-1 \leq s(x_i) \leq 1$.

It can be observed that the two techniques return the same results, in particular the hierarchical clustering repositions the mislabel outliers in the right cluster,

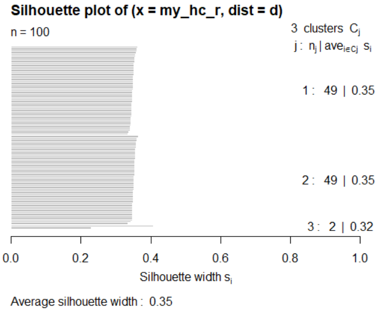

Figure 4: Silhoutette.

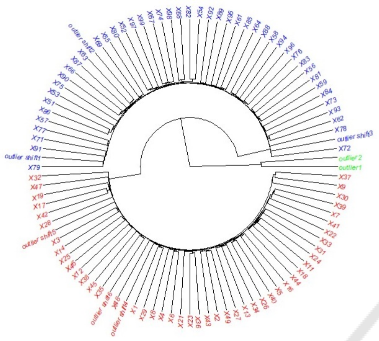

Figure 5: Circular dendrogram.

the abnormal type outliers form a separate cluster, as depicted in Figures 4 and 5.

Using RobPCA for finding outlier in this data, allows to detect more information, as described in Section 2. These results are depicted in Figure 6: in this case, the ”mislabeled” samples are configured as good outliers.

3.2 Real Dataset

The experimentation on real data was carried out on two particular cancer datasets. The first dataset, hereafter denoted as *A*, is characterized by 21120 genes and 593 samples, divided as follows: 591 samples with a Diffuse Large Cell B Lymphoma, a particular type of blood cancer, (DLBCL) and 2 samples with a solid ovarian tumor.

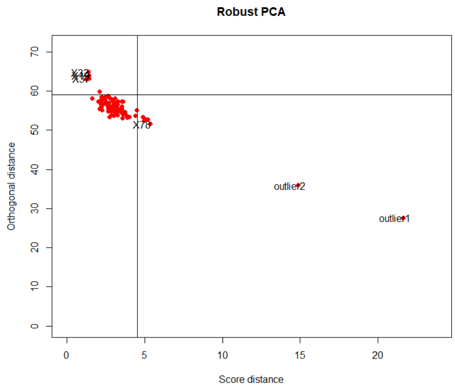

Figure 6: Robust PCA.

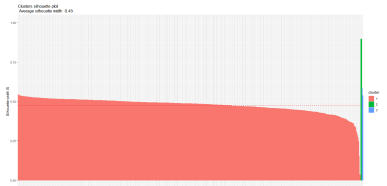

Figure 7: Silhouette Coefficient of Dataset A.

Dataset *B*, instead, is composed by 21120 genes and 597 samples, where 591 are from the same DLBCL dataset and the remaining are equally divided between Follicular Lymphoma (FL), Mantle Cell Lymphoma (MCL) and Burkitt’s Lymphoma (BL).

Raw data [1] are downloaded from Gene Expression Omnibus databases, pre-processed removing backgroud, normalizing and batch effect correction procedures.

The aim of experimental session is twofold: firstly we want to detect if the proposed approach is able to find ovarian outliers from the whole of blood cancer samples giving more robust and detailed results respect to the existing methods in literature. Secondly, we want to stress more the approach proposed when the samples are biologically similar.

3.2.1 Dataset A

The proposed approach provides the following results. Through the Hierarchical Clustering technique we obtain the division of the dataset into 3 clusters. This results in a large cluster with 589 samples and two smaller clusters of cardinality 2, as we can see in the Table 2 and in the Silhouette plot in Figure 7. The first, as expected, containing the two samples with ovarian cancer and the second containing two outliers that the domain experts did not expect.

Table 2: Silhouette Coefficient of Dataset A.

cluster	size	ave.sil.width
1	589	0.47
2	2	0.90
3	2	0.56

Flanking these results by the outcome from Robust PCA we confirm the results obtained from clustering but also we are able to provide information on the classification of the outliers. Figure 8 illustrates

[1] In particular, samples related to DLBCL are associated to GSE10846, GSE132929, GSE23501, GSE34171, GSE87371 and GSE98588; samples of ovarian cancer to GSE9891; whereas the other six samples in Dataset B were randomly choosen from GSE12195, GSE55267, GSE93261, GSE26673 and GSE21452.

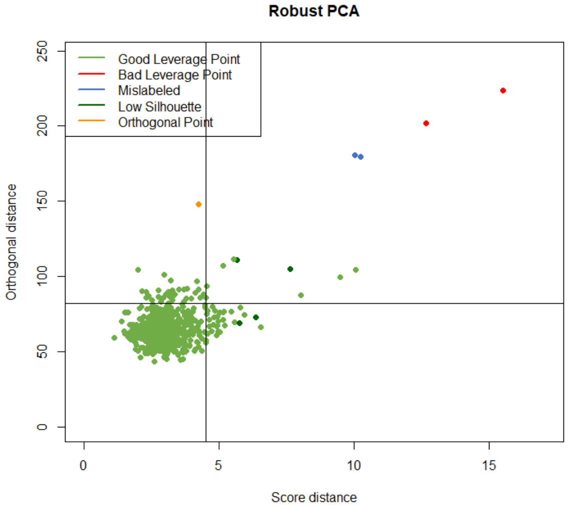

Figure 8: ROBPCA plot of Dataset A.

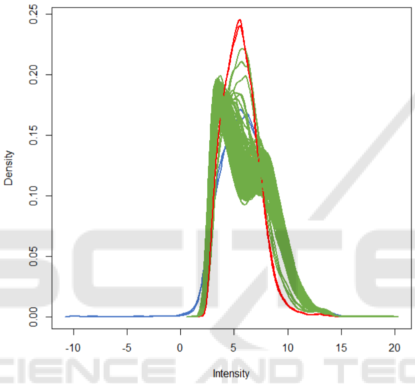

Figure 9: Density Plot per cluster.

the location of outliers: the 4 outliers of clusters 2 and 3 are located in the Bad leverage points quadrant. To investigate better these samples which are differed from most of the data an analysis on the degradation of RNA was performed. This analysis confirms that the 4 outliers correspond to a degradation of the sample, as described below.

On the border with Good Leverage there are observations with a lower silhouette, they indicate that although they are not strictly anomalous samples, they do not fit very well with the data. Only one sample is an orthogonal outlier.

The results produced adopting the proposed approach can be assessed also by using density plots. Particularly, Figure 9 evidences the different distribution between samples labeled according to cluster membership.

On the other hand, Figure 10 illustrates the density of the outliers obtained using RobPCA according to their classification. Mislabeled Samples are depicted in black and red. Abnormal Samples are the samples in blue and gray. In light blue, yellow and purple we have the samples with low silhouette and in green the sample identified as an orthogonal outlier by the Robust PCA. In particular, note the distribution of the sample identified as orthogonal outlier, in green. While all the other have a lower number of genes very expressed than the unexpressed, in this sample the two parts are almost equivalent.

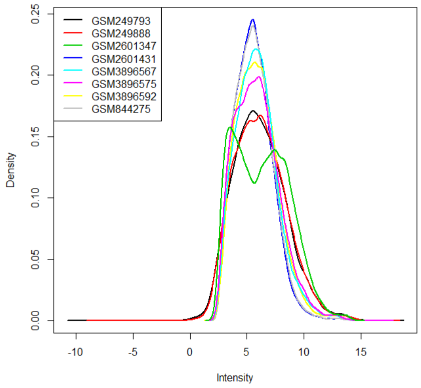

Figure 10: Outliers DensityPlot for each abnormal sample. (GSM is the standard acronym for samples in GEO-Gene Expression Omnibus).

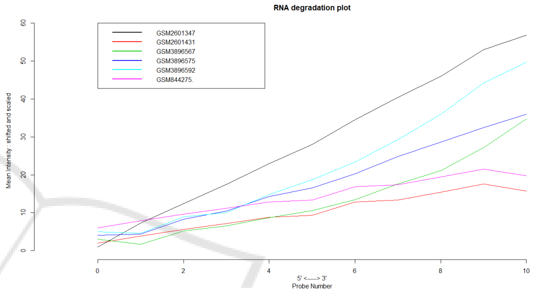

Figure 11: Degradation Plot.

According to this classification, it can be assert that the ovarian samples are correctly detected, whereas the remaining six samples need to be further investigated. To this aim, the quality of the RNA has been examined using a RNA degradation plot.
It can be observed in Figure 11 that the two non-mislabeled outliers are degraded, in fact, the observed trend differs from that of the others and it is not nearly constant. This implies that their information power can be neglected.

Finally, we compared the results provided by the proposed approach with those obtained from the standard ”KS” and ”sum” techniques. Figure 12 reports (as a mechanism of comparison) the Euler-Venn diagram drawn with an on-line tool [2] (Hulsen et al., 2008).

In the first case, our results coincide with those of

[2] web application for the comparison and visualization of biological lists using area-proportional Venn diagrams

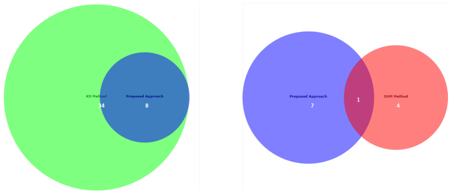

Figure 12: Comparison of approaches for dataset A.

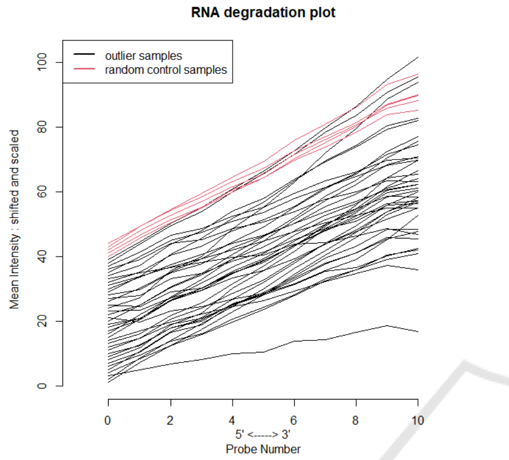

Figure 13: Degradation Plot of anomalous samples found through the KS and SUM techniques.

the ”KS” technique, but are fewer in number. However we believe that our results have more relevance because with ”KS” 34 samples are considered outliers, a high number compared to the nature of the outliers. Normally the number of outliers is very small. The cause probably lies in the origin of the technique which is based on a distance in terms of probability distributions. The second technique, instead, finds different outliers compared to our approach, only one outlier is in common. In Figure13 we show the degradetion plot of all the outliers found with the alternative techniques, with some control samples. We can observe that not all anomalous samples found with these technologies are degraded. This makes us believe our approach is the most reliable.

3.2.2 Dataset B

Differently from previous case, for this dataset the choice of the Pearson distance was necessary because the inserted samples differ little. The Cophenetic coefficient is still high, equal to 0.8647102, then we can consider a good agreement. In this case, through the hierarchical clustering technique we obtain 3 clusters, too. Table 3 reports the obtained results, while Figure 14 draws the related Silhouette plot.

Since the experimental dataset is that of the previous case in which the ovarian tumor samples were replaced with the samples with FL, MCL and BL tumors, the results for the equal part of the dataset are the same. Instead, as it can be observed from the Robust PCA in Figure 15, the samples of type MCL, FL and BL tumors are outliers. In particular, these are Bad Leverage Outliers since they lie in the first quadrant. A different distribution of the FL, MCL and BL tumors samples depicted in red can be observed when compared to the other samples depicted in yellow in Figure 16. The outliers find before are depicted in blue. As it can be observed in Figure 17 and in Figure18, also in this case comparing the results with those of standard techniques, the same consideration as before can be drawn.

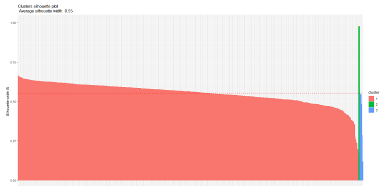

Figure 14: Silhouette Dataset B.

Table 3: Silhouette Coefficient of Dataset B.

cluster	size	ave.sil.width
1	589	0.55
2	2	0.98
3	6	0.42

4 CONCLUSIONS AND FUTURE WORKS

An ensemble mechanism combining Robust PCA and Hierarchical Clustering with opportune distances was proposed to search for abnormalities in gene expression matrices in a more reasonable way. It is configured as an additional tool and allows to derive a pseudo mathematical classification of outlier samples in GEP data focusing on microarray. Moreover since recent works focus only on RNA-seq data (Chen et al., 2020), we will extend our approach to be interchangeable between different platforms such as RNA-seq and GEP derived from Nanostring Technologies. The preliminary experimental results performed using the proposed approach showed that it is possible to make a pseudo-classification of the outliers based on their nature. Future works should be performed to identify the thresholds within which it is possible to associate the mathematically defined outlier to the biological outlier. The obtained results are quite promising and suggest the usefulness of the proposed mechanism as pre-processing for the analysis of datasets that need to be further studied. For example, the proposed mechanism demonstrates to be able to eliminate degraded

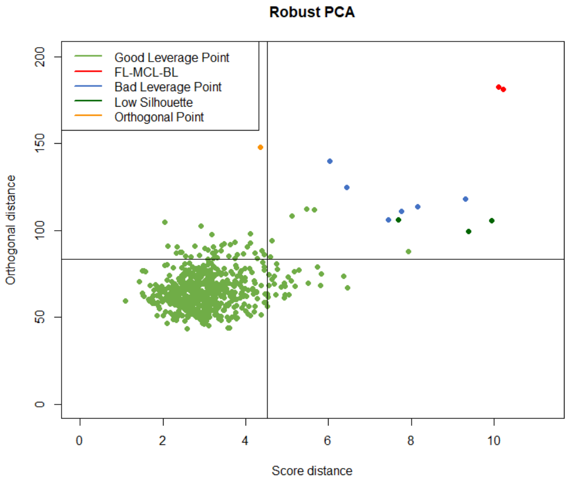

Figure 15: DD-plot Dataset B.

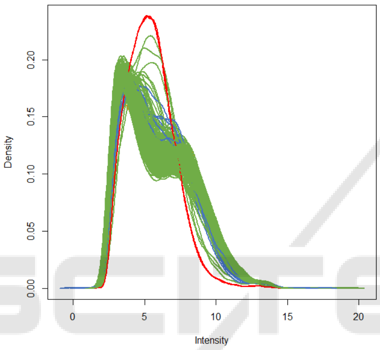

Figure 16: DensityPlot per cluster.

samples, carrying out analyzes on particular samples and possibly reclassifying mislabeled type outliers.

Future research should be devoted to construct a new decision-making model which incorporates the proposed ensemble mechanism as data pre-processing method to identify the anomalies and integrate the anomalies detection tool in the context of microarrays to searching and classifying samples that can generate new biological hypotheses.

ACKNOWLEDGEMENTS

This work was supported in part by the GNCS-INDAM (Gruppo Nazionale per il Calcolo Scientifico of Istituto Nazionale di Alta Matematica) Francesco Severi, P.le Aldo Moro, Roma, Italy.

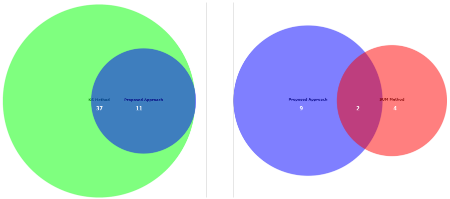

Figure 17: Comparison of approaches for dataset B.

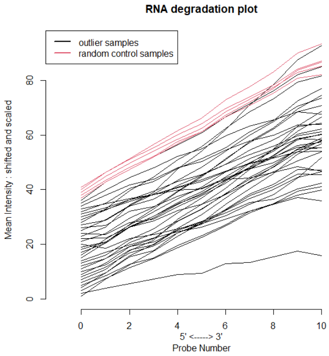

Figure 18: Degradation Plot of anomalous samples found through the KS and SUM techniques.

REFERENCES

Barghash, A., Arslan, T., and Helms, V. (2016). Robust detection of outlier samples and genes in expression datasets. *J. Proteomics Bioinform*, 9(02):38–48.

Chen, X., Zhang, B., and Wang, T. (2020). Robust principal component analysis for accurate outlier sample detection in rna-seq data. *BMC Bioinformatics*, 21, 269.

Croux, C., Filzmoser, P., and Oliveira, M. (2007). Algorithms for projection–pursuit robust principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 87(2):218–225.

Esposito, F., Boccarelli, A., and Del Buono, N. (2020). An nmf-based methodology for selecting biomarkers in the landscape of genes of heterogeneous cancer-associated fibroblast populations. *Bioinformatics and Biology Insights*, 14:117793222090682.

Hubert, M., Rousseeuw, P. J., and Vanden Branden, K. (2005). Robpca: a new approach to robust principal component analysis. *Technometrics*, 47(1):64–79.

Hulsen, T., de Vlieg, J., and Alkema, W. (2008). Biovenn - a web application for the comparison and visualization of biological lists using area-proportional venn diagrams. *BMC Genomics*, 9(1):488.

Paquet, A. and Yang, J. (2010). arrayquality: Assessing array quality on spotted arrays.

R-Team, C. (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Shieh, A. D. and Hung, Y. S. (2009). Detecting outlier samples in microarray data. *Statistical applications in genetics and molecular biology*, 8(1).