Malaysian Journal of Mathematical Sciences 10(2): 117–135 (2016)

MALAYSIAN JOURNAL OF MATHEMATICAL SCIENCES

Journal homepage: http://einspem.upm.edu.my/journal

Detection of Outliers in Gene Expression Data Using Expressed Robust-t Test

Md. Manzur Rahman Farazi *[1] and A.H.M. Rahmatullah Imon [2]

[1]*Department of Mathematics, Statistics & Computer Science*
Marquette University, Milwaukee, WI 53201, USA
mdmanzurrahman.farazi@marquette.edu
[2]*Department of Mathematical Sciences, Ball State University,*
Muncie, IN 47306, USA

E-mail:
**farazi@juniv.edu*

ABSTRACT

Detection of outliers in gene expression data has drawn a great deal of attention in recent years. Although a variety of outlier detection methods is available in the literature Tomlins et al. (2005) argued that they are not readily applicable to gene expression data. They developed the "cancer outlier profile analysis (COPA)" method to detect cancer genes and outliers. Following their way several methods are proposed in the literature for detecting outliers. Most of these methods are based on t-type tests which are basically nonrobust and hence fail to identify multiple outliers. In this paper we propose a robust version of the t-test that we call expressed robust t (ERT) test. The usefulness of the proposed methods is then investigated by Monte Carlo simulation and real cancer data.

Keywords: Gene expression, Outlier, Cancer outlier profile, Robust statistics.

1. Introduction

Statistical data analysis usually begins with the gathering of observations from a certain population. However, this process of accumulating the data is subject to numerous sources of error. Therefore, the data collected may comprise with some unusually small or large observations, so-called outliers. Determining whether a data set contains one or more outliers is a challenge commonly faced in applied statistics. This is a mostly difficult mission if the properties of the underlying population are not known. However, in many empirical investigations, the assumption that the data come from a particular population is too restrictive or unrealistic. Although outliers are often considered as an error or noise, they may convey important information. Detected outliers are candidates for peculiar data that may otherwise adversely lead to model misspecification, biased parameter estimation and incorrect results. It is therefore important to identify them prior to modeling and analysis. Outlier detection methods have been suggested for numerous applications, such as credit card fraud detection, clinical trials, voting irregularity analysis, data cleansing, network intrusion, severe weather prediction, geographic information systems, athlete performance analysis, and other data-mining tasks [see Hawkins (1980), Barnett and Lewis (1994), Hadi, Imon and Werner (2009)].

The rapid developments of technologies that generate arrays of gene data enable a global view of the transcription levels of hundreds of thousands of genes simultaneously. The outlier detection problem for gene data has its importance but together with the difficulty of high dimensionality. The scarcity of data in high-dimensional space makes each point a relatively good outlier in the view of traditional distance-based definitions. Thus, finding outliers in high dimensional data is more complex. Microarray technology is used in a wide variety of settings for detecting differential gene expression. Classic statistical issues such as appropriate test statistics, sample size, replicate structure, statistical significance, and outlier detection enter into the design and analysis of gene expression studies. Adding to the complexity is the fact that the number of samples I in a microarray experiment is inevitably much less than the number of genes J under investigation and that J is often on the scale of tens of thousands, thus creating a tremendous multiple testing burden.

Fundamental to the task of analyzing gene expression data is the need to identify genes whose patterns of expression differ according to phenotype or experimental condition. Gene expression is a well-coordinated system, and hence measurements on different genes are in general not independent. Given more complete knowledge of the specific interactions and transcriptional controls, it is conceivable that meaningful comparisons between samples can be made by

considering the joint distribution of specific sets of genes. However, the high dimension of gene expression space prohibits a comprehensive exploration, while the fact that our understanding of biological systems is only in its infancy means that in many cases we do not know which relationships are important and should be studied. In current practice, differential expression analysis will therefore at least start with a gene-by-gene approach, ignoring the dependencies between genes. A simple approach is to select genes using a fold-change criterion. This may be the only possibility in cases where no, or very few replicates, are available. An analysis solely based on fold change however does not allow the assessment of significance of expression differences in the presence of biological and experimental variation, which may differ from gene to gene. This is the main reason for using statistical tests to assess differential expression. Generally, one might look at various properties of the distributions of a genes expression levels under different conditions, though most often location parameters of these distributions, such as the mean is considered. Parametric test, such as the t-test, is commonly used. Parametric tests usually have a higher power if the underlying model assumptions, such as normality in the case of the t test, are at least approximately fulfilled. Presence of outliers may often destroy normality pattern so it is essential to identify outliers in gene expression data before any further statistical analysis.

We organize this paper in the following way. In section 2, we introduce different concepts regarding gene expression. We also introduce the real data which we analyze later for outlier detection. In section 3 we introduce the concept of outliers, its consequences. We also introduce some commonly used outlier detection techniques here. In section 4, we consider outlier detection methods which are suggested exclusively for gene expression data. The most widely used method for detecting differential gene expression in comparative microarray studies is the two-sample t-statistic. A gene is determined to be significant if the absolute t-value exceeds a certain threshold c, which is usually determined by its corresponding p-value or false discovery rate. Recently, Tomlins et al (2005) introduced the cancer outlier profile analysis (COPA) method for detecting cancer genes which are differentially expressed in a subset of disease samples. Heterogeneous patterns of oncogene activation were observed in the majority of cancer types considered in their studies. Thereafter, several further studies in this direction have been proposed. Tibshirani and Hastie (2007) introduced the outlier sums (OS) method, Wu (2007) proposed the outlier robust t-statistic (ORT), Lian (2008) introduced the maximum ordered subset t-statistics (MOST) for detecting cancer outlier differential gene expression. In this section we investigate the performance of the existing methods and introduced a new robust method by robustifying a t-test that we call expressed robust t (ERT) test. The effectiveness of our proposed method together with

the existing methods is then investigated through a Monte Carlo simulation and also for tumor colon cancer data.

2. Gene Expression and its Pattern

In this section we briefly discuss gene expression and its pattern and introduce the cancer data that we use in our study. Gene expression is the process by which genetic instructions are used to synthesize gene products. These products are usually proteins, which go on to perform essential functions as enzymes, hormones and receptors, for example. Genes that do not code for proteins such as ribosomal RNA or transfer RNA code for functional RNA products. Gene expression analysis is the determination of the pattern of genes expressed at the level of genetic transcription, under specific circumstances or in a specific cell.

For this study we use the data pertaining to the article "Broad patterns of gene expression revealed by clustering of tumor and normal colon tissues probed by oligonucleotide arrays" by Alon et al. (1999). The matrix $http://genomics-pubs.princeton.edu/oncology/affydata/I2000.html$ contains the expression of the 2000 genes with highest minimal intensity across the 62 tissues. The genes are placed in order of descending minimal intensity. Each entry in I2000 is a gene intensity derived from the 20 feature pairs that correspond to the gene on the chip, derived using the filtering process. The data is otherwise unprocessed (for example it has not been normalized by the mean intensity of each experiment). The file $http://genomics-pubs.princeton.edu/oncology/affydata/names.html$ contains the EST number and description of each of the 2000 genes, in an order that corresponds to the order in I2000. Note that some ESTs are repeated which means that they are tiled a number of times on the chip, with different choices of feature sequences. The identity of the 62 tissues is given in file $http://genomics-pubs.princeton.edu/oncology/affydata/tissues.html$. The numbers correspond to patients, a positive sign to a normal tissue, and a negative sign to a tumor tissue. There were 22 tissues from normal sample 40 tissues from tumor sample.

Gene expression in 40 tumor and 22 normal colon tissue samples was taken from Gene expression in 40 tumor and 22 normal colon tissue samples was analyzed with an Affymetrix oligonucleotide array complementary to more than 6,500 human genes. An efficient two-way clustering algorithm was applied to both the genes and the tissues, revealing broad coherent patterns that suggest a high degree of organization underlying gene expression in these tissues. Coregulated families of genes clustered together, as demonstrated for the ribo-

somal proteins. Clustering also separated cancerous from noncancerous tissue and cell lines from in vivo tissues on the basis of subtle distributed patterns of genes even when expression of individual genes varied only slightly between the tissues. For the study the 2,000 genes with highest minimal intensity across the tissues were used. To get an idea about the gene expression in normal sample and tumor sample we construct an index plot of the genes for normal, tumor and all samples which is below. It seems from the figures that tumor samples have higher intensities than normal samples. The ranges of normal genes are 5.82 to 14173.05 whereas for tumor samples these values are 5.89 to 20903.18.

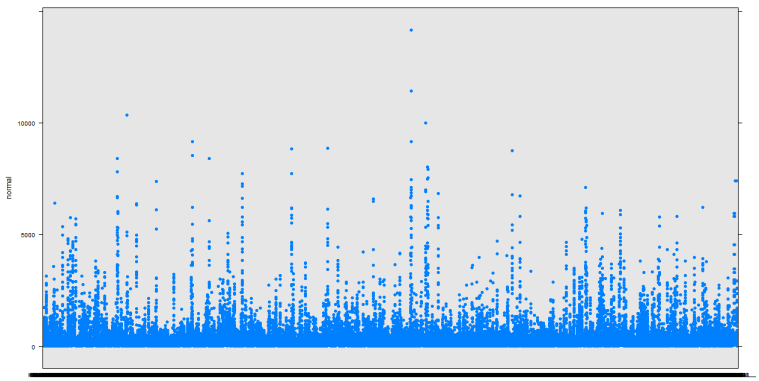

Figure 1: Plot of the gene intensities for normal samples (Horizontal-gene index, vertical intensities)

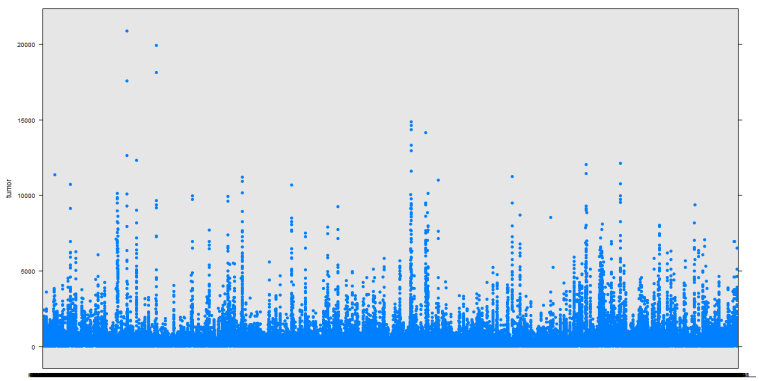

Figure 2: Plot of the gene intensities for tumor samples (Horizontal-gene index, vertical intensities)

The index plots of the genes according to different group indicate that the intensity level in the tumor samples usually much higher than that of normal samples. This indication is clearly a good sign to doubt that there must be

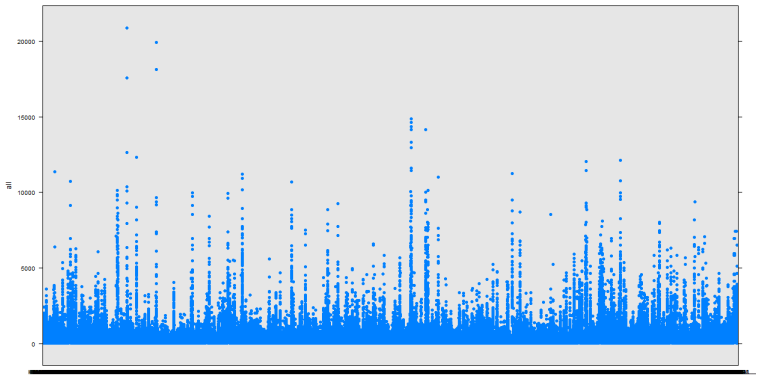

Figure 3: Plot of the gene intensities for all samples (Horizontal-gene index, vertical intensities)

some unusual pattern of heterogeneity in the tumor samples. Our attempt to detect outlier has got a strong background to proceed on.

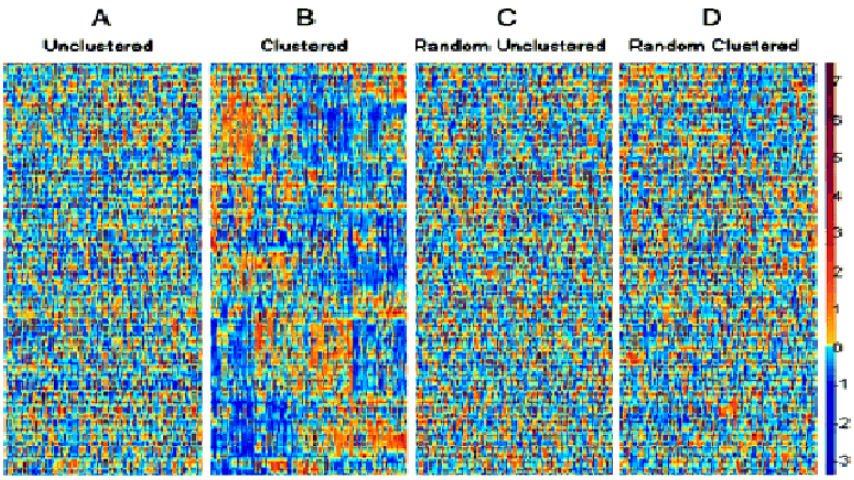

Figure 4: Matrix of gene expression

We have taken Figure 4 from "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays" by A Ulon, a proceedings of the National Academy of Sciences of the United States of America. The vertical axis corresponds to genes, and the horizontal axis to tissues. Each gene was normalized so its average intensity across the tissues is 0, and its SD is 1. The color code used is indicated

in the adjoining scale. (A) Unclustered data set. (B) Clustered data. The 62 tissues are arranged on the vertical axis according to the ordered tree of Figure 4. The 2,000 genes are arranged on the horizontal axis according to their ordered tree. (C) Unclustered randomized data, where the original data set was randomized (the location of each number in the matrix was randomly shifted). (D) Clustered randomized data, subjected to the same clustering algorithm as in B.

3. Outliers and their Identification

Given a data set, outlier detection aims at finding data points which are very different from the remainder.

3.1 Outliers

The term 'outlier' was used in astrophysics to distinguish planets which are 'outlying' in our solar system. This field has received a large attention in the last decades because outliers often represent critical information about an abnormal behavior of the system described by the data. Outliers are also called: event, novelty, anomaly, noise, deviation or exception. However there is no formal definition of an outlier because this intuitive notion varies with the context and the desired characteristics of outliers. In a statistical perspective, Grubbs (1969) defined that "an outlying observation, or outlier, is one that deviates markedly from other members of the sample in which it occurs". Hawkins (1980) defines an outlier as "an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism", while Barnett and Lewis (1994) call an outlier "an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data". They also mentioned "Even before the formal development of statistical method, argument raged over whether, and on what basis, we should discard observations from a set of data on the grounds that they are 'unrepresentative', 'spurious', or 'mavericks' or 'rogues'."

Outliers do not inevitably 'perplex' or 'mislead'; they are not necessarily 'bad' or 'erroneous', and the experimenter may be tempted in some situations not to reject an outlier but to welcome it as an indication of some unexpectedly useful industrial treatment or surprisingly successful agricultural variety. Sometimes it is a matter of subjective judgment on the part of the observer whether or not some observations are genuine members of the main population. If they are contaminants (arising from some other distribution), they may frus-

trate attempts to draw inferences about the original population. Of course, any contaminants which occur in the midst of the data set will not be conspicuous. We call contaminants to be outliers when they appear surprisingly extreme. Outliers may or may not be contaminants; contaminants may or may not be outliers.

Hampel et al. (1986) claim that a routine data set typically contains about 1-10% outliers, and even the highest quality data set cannot be guaranteed free of outliers. One immediate consequence of the presence of outliers is that they may cause apparent non-Normality and the entire classical inferential procedure might breakdown in the presence of outliers.

3.2 Detection of Outliers

Observations arising from large variation of the inherent type are called outliers, while observations subjected to large measurement error or execution errors are termed spurious observations. When we order the sample, the smallest and the largest ordered observations are known as extremes. Whether we declare either of them to be an outlier depends on consideration of how they appear in relation to the postulated model. Extreme values may or may not be outliers. To quote Barnett and Lewis (1994) 'Any outliers, however, are always extreme values in the sample.'

As we mentioned before a large body of literature is outlier detection methods is available. Here we consider those which have most applications in practice.

3.2.1 The three sigma Rule

If we assume a normal distribution, a single value may be considered as an outlier if it falls outside a certain range of the standard deviation. A traditional measure of the 'outlyingness' of an observation with respect to a sample is the ratio between its distance to the sample mean and the sample SD:

$$t_i = \frac{x_i - \bar{x}}{s} \; i = 1, 2, ..., n \tag{1}$$

Observations with $|t| > 3$ are traditionally deemed as suspicious (the three-sigma rule), based on the fact that they would be very unlikely under normality, since $P(|t| > 3) = 0.003$ for a random variable t with a standard normal distribution.

3.2.2 Grubbs' Test

Grubbs (1969) proposed a test to detect outliers in a univariate data set. It is based on the assumption of normality. Grubbs' test is also known as the maximum normed residual test. The Grubbs test statistic is the largest absolute deviation from the sample mean in units of the sample standard deviation. Grubb's Test is a test based on normal distribution, the effects of which are that the data analyzed with this method should have normal distribution. This test should be performed as long as all outliers will be detected. In this test we have two hypotheses: the null-hypothesis (H_0) and the alternative hypothesis (H_1):
H_0: There are no outliers in the data set.
H_1: There is at least one outlier in the data set.

The general formula for Grubbs' Test can be presented as follows:

$$G = \frac{Max|x_i - \bar{x}|}{s} \tag{2}$$

The calculated value of G is compared with the critical value for Grubbs' Test. For the two-sided test, the hypothesis of no outliers is rejected if

$$G > \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{((\alpha/2n),n-2)}}{n-2+t^2_{((\alpha/2n),n-2)}}} \tag{3}$$

with α denoting the critical value of the t distribution with $n-2$ degrees of freedom and a significance level of $\alpha/(2n)$. When the calculated value is higher or lower than the critical value for the chosen statistical significance, then the calculated value can be accepted as an outlier.

3.2.3 Dixon's Q-test

This method was proposed by Dean and Dixon(1951). This test has some restrictions - it is impossible to use this test with a big data set. The Dixon's Q-test is a very simple test for outliers when we suspect that outliers are extreme observations in the data set. Q-test is based on the statistical distribution of 'subrange ratios' of ordered data samples, drawn from the same normal population. Hence, a normal distribution of data is assumed whenever this test is applied.
The test is very simple and it is applied as follows:
1. The n values comprising the set of observations under examination are arranged in ascending order: $x_{(1)} < x_{(2)} < ... < x_{(n)}$.

2. The statistic Q is a ratio defined as the difference of the suspect value from its nearest one divided by the range of the values. Thus, for testing $x_{(1)}$ or $x_{(n)}$ (as possible outliers) we use the following values:

$$Q = \frac{x_{(2)} - x_{(1)}}{x_{(n)} - x_{(1)}} \; or \; \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}} \tag{4}$$

3. The obtained Q_{obs} value is compared to a critical Q-value (Q_{crit}) found in tables. 4. If $Q_{obs} > Q_{crit}$, then the suspect value can be characterized as an outlier.

3.3 Robust Outlier Detection Methods

In many cases, presence of outliers may make the diagnostic procedure unreliable for which masking and/or swamping can occur. So we need detection methods which are not affected by outliers. The word 'robust' literary means something 'very strong.' So robust statistics are those statistics which do not breakdown easily. The term robustness signifies insensitivity to small deviations from the assumption. That means a robust procedure is nearly as efficient as the classical procedure when classical assumptions hold strictly but is considerably more efficient over all when there is a small departure from them. One objective of robust techniques is to cope with outliers by trying to keep small the effects of their presence. Consequently, we should require resistant estimators. The analogous term used in the literature: resistant statistics.

Here we introduce several statistics which are robust in the presence of outliers. Median and trimmed mean are robust measures of location. For the measure of dispersion we can use the normalized median absolute deviation (MADN). For a set of data the median absolute deviation (MAD) is defined as

$$MAD(x) = Med|x - Med(x)| \tag{5}$$

To make the MAD comparable to the SD in terms of efficiency, we consider the normalized MAD defined as

$$MADN(x) = Mad(x)/0.6745 \tag{6}$$

Two other well-known dispersion estimates are the range defined as

$$R = x_{(n)} - x_{(1)} \tag{7}$$

and the inter-quartile range (IQR) defined as

$$IQR(x) = Q_3 - Q_1 \tag{8}$$

Both of them are based on order statistics; (7) is clearly very sensitive to outliers, while (8) is not.

3.3.1 Robust t like Statistic

Let us now use the robust plug-in technique to obtain a robust t-like statistic by replacing mean by median and SD by the normalized median absolute deviation (MADN). Thus the modified statistic becomes

$$t_i^{'} = \frac{x_i - Median(x)}{MADN(x)} \tag{9}$$

Observations with $|t_i^{'}| > 3$ are identified as outliers.

3.3.2 Test Based on the Interquartile Range

The above-mentioned strategies for identifying outliers are probably most appropriate for symmetric unimodal distributions. If a distribution is skewed, it is recommended to calculate the threshold for outliers from the interquartile distance:

$$Q_1 - 1.5IQR < x_i < Q_3 + 1.5IQR \tag{10}$$

3.3.3 Hampel's Test

In recent years Hampel et al. (1986)'s test for outliers has become very popular in data mining and knowledge discovery. According to this rule an observation is identified as an outlier if

$$x_i - median(x) > 4.5MAD(x) \tag{11}$$

It is interesting to note that Hampel's test is equivalent to robust t test. Recall that according to the robust t test an observation is identified as an outlier according to (9) which yields $x_i - median(x) > 3MADN(x) = 4.4474MAD(x)$

3.4 Masking and Swamping Effects

We often observe that identification methods fail to identify potential outliers or the methods identify cases as outliers which are actually not. In masking it is said that one outlier masks a second outlier, if the second outlier can be considered as an outlier only by itself, but not in the presence of the first outlier. Thus, after the deletion of the first outlier the second instance is emerged as an outlier. Masking occurs when a cluster of outlying observations skews the mean and the covariance estimates toward it, and the resulting distance of the outlying point from the mean is small.

In describing the swamping effect it is said that one outlier swamps a second observation, if the latter can be considered as an outlier only under the presence of the first one. In other words, after the deletion of the first outlier the second observation becomes a non-outlying observation. Swamping occurs when a group of outlying instances skews the mean and the covariance estimates toward it and away from other non-outlying instances, and the resulting distance from these instances to the mean is large, making them look like outliers.

4. Detection of Outliers in Gene Expression Data

A gene expression measurement which differs surprisingly from the other measurements obtained for the same gene on other samples of the same class. The outlying principle assumes that the data, with the possible exception of any outlier, form a sample of a given distribution |here the normal distribution. We will use a reasonable test statistical to decide whether or not the suspect measurement is an outlier.

Here we introduce the existing outlier detection methods for gene expression data. Assuming case-control microarray data were generated for detecting differentially expressed genes consisting of n samples from a normal group and m samples from a cancer group. Let be the expression value for gene $i = (1, 2, ..., p)$ and sample $j = (1, 2, ..., n)$ in the normal group and be the expression value for gene $i = (1, 2, ..., p)$ and sample $j = (1, 2, ..., m)$ in the cancer group. In this study, and without loss of generality, we are only interested in one-sided tests where the activated genes from cancer samples are over-expressed or up-regulated.

4.1 t-statistic

The two-condition t-statistic for gene i is defined by:

$$t_j = \frac{\bar{y}_i - \bar{x}_i}{s_i} \tag{12}$$

where $\bar{y}_i$ is the mean expression value in cancer samples, $\bar{x}_i$ is the mean expression value in normal samples for gene i and s_i is the pooled standard error estimate given by:

$$s_i^2 = \frac{\sum_{1 \le j \le n}(x_{ij} - \bar{x}_i)^2 + \sum_{1 \le j \le m}(y_{ij} - \bar{y}_i)^2}{n + m - 2} \tag{13}$$

The $t-$statistics is powerful when most cancer samples are activated.

4.2 Cancer Outliers Profile Analysis (COPA)

Tomlins et al (2005) defines the COPA statistic as

$$copa_i = \frac{q_r(y_{ij} : 1 \leq j \leq m) - med_i}{mad_i} \tag{14}$$

Where $q_r(.)$ is the r^{th} percentile of the expression data, and med_i is the median expression value for all samples

$$med_i = median((x_{ij} : 1 \leq j \leq n), (y_{ij} : 1 \leq j \leq m))$$

and mad_i is the median absolute deviation of expression values in all samples and is given by:

$$mad_i = 1.4826 * median(((x_{ij} - med_i) : 1 \leq j \leq n), ((y_{ij} - med_i) : 1 \leq j \leq m))$$

The COPA statistic uses a fixed r^{th} sample percentile, which is determined by users.

4.3 Outliers Sums Statistics (OS)

COPA statistic uses a fixed r^{th} sample percentile, which is determined by users. This limitation was overcome by the OS statistic defined by Tibshirani and Hastie (2007) as

$$OS_j = \frac{\sum_{y_{ij} \in R_i}(y_{ij} - med_i)}{mad_i} \tag{15}$$

where

$$R_i = (y_{ij} : y_{ij} > q_{75}((x_{ij} : 1 \leq j \leq n), (y_{ij} : 1 \leq j \leq m)) + IQR((x_{ij} : 1 \leq j \leq n), (y_{ij} : 1 \leq j \leq m)) \tag{16}$$

and IQR(.) is the inter-quantile range of the expression data

$$IQR((x_{ij} : 1 \leq j \leq n), (y_{ij} : 1 \leq j \leq m)) = q_{75}((x_{ij} : 1 \leq j \leq n), (y_{ij} : 1 \leq j \leq m)) - q_{25}((x_{ij} : 1 \leq j \leq n), (y_{ij} : 1 \leq j \leq m))$$

4.4 Outliers Robust t-statistics (ORT)

Wu (2006) modified the OS statistic by proposing the ORT statistic which consists mainly in changing the definition of R_i as:

$$R_i = (y_{ij} : y_{ij} > q_{75}((x_{ij} : 1 \leq j \leq n)) + IQR((x_{ij} : 1 \leq j \leq n))) \tag{17}$$

and replacing med_i in OS by med_{ix}, which is the median expression value in normal samples. Further, mad_i was replaced by

$$mad_i = 1.4826Xmedian(((x_{ij}-med_{ix}) : 1 \leq j \leq n), ((y_{ij}-med_{iy}) : 1 \leq j \leq m))$$

where med_{iy} is the median expression value in cancer samples.
COPA and OS statistics were derived from the t-statistic by replacing the mean and standard errors used in the t-statistic with the median and median absolute deviations, respectively. ORT has been proposed as a more robust statistic that utilizes the absolute difference of each expression value from the median instead of the squared difference of each expression value from the average.

4.5 Maximum Ordered Subset t-statistics (MOST)

Lian (2008) argued that OS and ORT statistics used arbitrary outliers and proposed the MOST statistic which consider all possible values for outlier thresholds. The MOST procedure requires cancer sample expression data be sorted in descending order and the following statistic calculated:

$$MOST_i = max_{1 \leq k \leq m} \left[\frac{\sum_{1 \leq j \leq k}(y_{ij} - med_{ix})}{mad_i'} - \mu_k \right] / \delta_k \tag{18}$$

where μ_k and δ_k are obtained from the order statistics of m samples generated from a standard normal distribution and are used to make different values of the statistic comparable for different values of k.

4.6 The Proposed Outlier Detection Method: Expressed Robust t statistic(ERT)

We observe that most of the outlier detection techniques defined in the previous section contains some non-robust components such as the mean and standard deviations and consequently they may become ineffective in doing their jobs. In our study we propose a new outlier technique modifying one of the existing methods. The proposed technique which we call expressed robust t (ERT) statistic is described below.

We have seen in (12) that the two-condition t-statistic for gene i is defined by: $t_i = \frac{\bar{y}_i - \bar{x}_i}{s_i}$ Since both $\bar{y}_i$,$\bar{x}_i$ and s_i are non-robust, we propose the expressed robust t-statistic as:

$$t_i^{`} = \frac{med_{iy} - med_{ix}}{mad_i} \tag{19}$$

where
$med_{ix} = median[(x_{ij} : 1 \leq j \leq n)]$

$med_{iy} = median[(y_{ij} : 1 \leq j \leq m)]$
$mad_i^{'} = 1.4826Xmedian(((x_{ij} - med_i) : 1 \leq j \leq n), ((y_{ij} - med_i) : 1 \leq j \leq m))$

5. Results on Monte Carlo Study and Cancer Data

Before applying to real data we tested the performance of our newly proposed method by simulation studies. Here we report a Monte Carlo experiment which is designed to assess the performance of the proposed outlier detection methods for gene expression data in comparison with the existing ones. Simulation studies are conducted to compare the performance of newly proposed ERT method with the t-statistic, COPA, OS, ORT and MOST methods. The simulation was conducted in different situations. To test and check the consistency of the test statistic, we generate gene expression for two groups of sample with different sizes in different simulation.

In all simulation we generated $g = 40$ genes. Out of 40 genes we generated 20 genes considering no differences between normal and tumor group. We generated these 20 genes with uniform condition for both groups. Further, we generated another 20 genes with two different situations. To distinguish the two groups, for normal sample and tumor sample we used different ranges. We assume outliers do exist in later 20 genes. The process is done 5 times by changing the number of normal and tumor sample sizes. For the first set of simulation we generated $n = 75$ and $m = 25$ as number of samples from normal and tumor group respectively. For other simulations we chose $(n = 60, m = 40), (n = 55, m = 45), (n = 80, m = 20) and (n = 90, m = 10)$. We applied all the existing methods and our new methods to these simulated data. The results of the number of genes detected as outliers in this simulation experiment are given in Table 1.

Results presented in Table 1 show that among the existing methods, the t test performs well but on a couple of occasions it fail to identify the genuine outliers. The performances of COPA, OS and ORT are not very satisfactory. But the performance of the newly proposed method give better results in comparison with the existing methods and the ERT performs the best. The methods give consistent results over different simulations.

Now we apply the outlier detection methods in cancer data. In the study data we have the data set of intensities of 2,000 genes in 22 normal and 40 tumor colon tissues. The genes chosen are the 2,000 genes with highest minimal intensity across the samples. We try to find out the possible responsible genes

Table 1: Sample table.

Parameter	Simu-1	Simu-2	Simu-3	Simu-4	Simu-5
n	75	60	55	80	90
m	25	40	45	20	10
g	40	40	40	40	40
t	20	13	15	20	20
COPA	6	5	4	5	2
OS	17	15	9	6	4
ORT	7	5	5	2	4
ERT	19	20	20	20	20

for tumor. We believe genes with high expressions might be guilty for tumor. Considering this we applied the existing outlier detection methods and also the newly proposed ERT method to find out the possible genes.

Figure 5 gives a graphical display of the performances of the existing and proposed outluier detection methods with their respective cut-off points. Table 2 demonstrates the number of outliers detected by different methods.

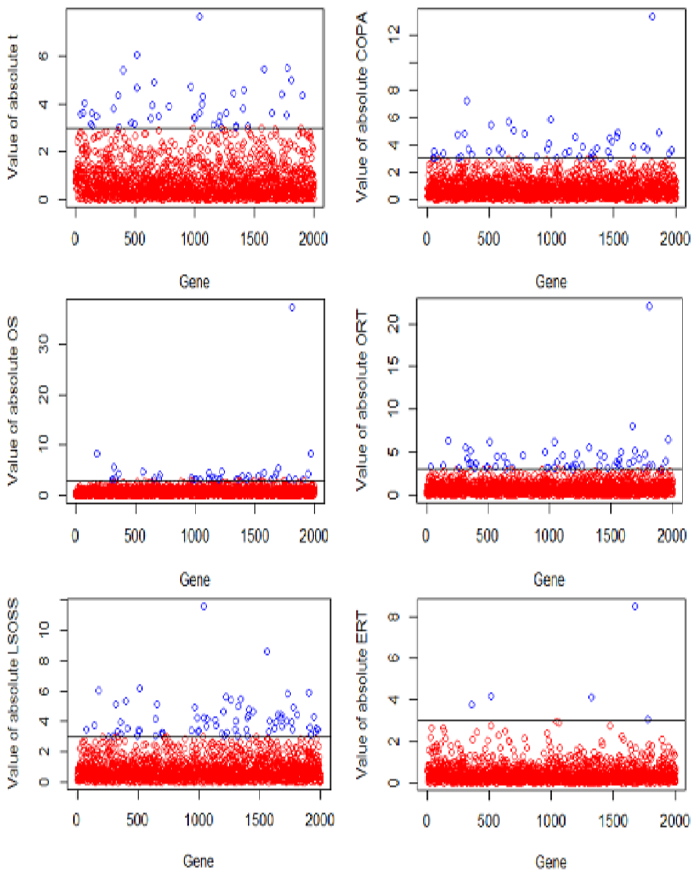

Figure 5: Index plot of genes with cut-off points for different outlier detection

The above figure and table show that the traditional t-test can successfully identify 46 genes as outliers. The COPA, OS, ORT, MOST and LSSOS statistic identify 49, 38, 58 and 78 genes respectively. In turn, our newly proposed statistic ERT can identify 5, 69 and 118 genes. It is worth mentioning that for the real data we do not definitely know which observations are genuine outliers or not so we cannot say which method masks or swamps which observations. Another thing we need to mention here is that the real data contains 2000 genes but in our simulation we consider 40 genes. The reason is even with 40 genes and 100 observations the computation is huge so we did not go for cases like 2000 genes in our simulation experiment.

Table 2: Sample table.

Parameter	Number of Outliers($n = 22, m = 40, g = 2000$)
t-test	46
COPA	49
OS	38
ORT	58
LSSOS	72
ERT	5

6. Conclusion

In our study we propose a new technique for finding outliers in gene expression data. The simulation results suggest that the performance of the proposed ERT statistic outperforms all the existing methods. When evaluating ERT based on tumor cancer data, we studied how many genes among the 2000 genes selected separately by different statistical approaches. The numbers of tumor cancer related genes identified by existing methods were 46, 49, 38, 58, and 78 for the t-statistics, COPA, OS, ORT and MOST respectively. However, our proposed method ERT has identified 5 tumor cancer related genes. Disentanglement the heterogeneous designs of cancer samples is an important goal in medical research, especially for clinical diagnosis and the molecular understanding of cancer mechanisms. The diverse patterns of oncogene activation have been well studied and several useful statistical tools have been proposed. ERT is reasonable model to detect cancer outlier differential gene expression. For each gene, ERT distinguish the expression values of normal and tumor samples. If any gene is expressed heterogeneously in cancer samples, the mean and variance of gene expression values in cancer samples are overemphasized by the classical t-statistic while ERT used the robust statistic Median and MAD in replace of them which gives a reasonable estimate. Our proposed scheme could be useful tool to separate the patterns of tumor cancer with specific gene signatures. In addition, these heterogeneous gene activation patterns may be regarded as the signatures for subtypes of tumor cancer. Thus, the procedure presented could also be useful in detecting and classifying tumor cancer subtypes. Our approach, however, differed from previous studies mainly in that the classification is based on different combinational activation patterns of candidate genes instead of clustering their expression values.

References

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering of tumor and normal colon tissues probed by oligonucleotide arrays, Proceedings of National Academy Science. USA, 96:6745-6750.

Barnett, V. and Lewis, T.B. (1994). *Outliers in Statistical Data*, 2nd ed. New York: Wiley.

Dean, R.B. and Dixon, D. J. (1951). Simplified Statistics for Small Numbers of Observations, *Anal. Chem.*, 23:636-638.

Grubbs, F. E. (1950). Sample criteria for testing outlying observations, *The Annals of Mathematical Statistics*, 21:27-58.

Hadi, A.S., Imon, A.H.M.R. and Werner, M. (2009). Detection of outliers, *Wiley Interdisciplinary Reviews: Computational Statistics*, 1:57-70.

Hampel, F.R., Ronchetti, E.M., Rousseeuw P.J. and Stahel, W. (1986). *Robust Statistics: The Approach Based on Influence Function*, New York: Wiley.

Hawkins D. (1980). *Identification of Outliers*, New York: Chapman and Hall.

Lian H. (2008). MOST: detecting cancer differential gene expression, *Biostatistics*, 9:411-418.

Tibshirani, R. and Hastie, T. (2007). Outlier sums for differential gene expression analysis, *Biostatistics*, 8:2-8.

Tomlins, S. A., Rhodes, D. R. and Perner, S. (2005). Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer, *Science*, 310:644 -648.

Wu, B. (2007). Cancer outlier differential gene expression detection, *Biostatistics*, 8:566-575.