J. Appl. Environ. Biol. Sci., 4(7S)302-310, 2014

ISSN: 2090-4274
Journal of Applied Environmental and Biological Sciences
www.textroad.com

Improved Statistical Test Using Shrinkage Covariance Matrix for Identifying Differential Gene Sets

Suryaefiza Karjanto[1], Norazan Mohamed Ramli[1], Rasimah Aripin[2], Nor Azura Md Ghani[1]

[1]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia,
[2]Faculty of Science and Technology, Sunway University, Jalan Universiti, Bandar Sunway, 46150 Petaling Jaya, Selangor, Malaysia,

Received: September 1, 2014
Accepted: November 13, 2014

ABSTRACT

. Microarray technology helps in the identification of new genes in our body. This technology provides the fundamental aspects underlining our life by discovering the genetic causes of differences occurring in the functioning of the human body which is unknown before. For example, a researcher might wish to know the effect of certain treatment by examining the differences in gene activity between treatment and control samples. But the detection of which genes that contribute to certain treatment using statistical test is a problem because the number of samples is smaller than number of variables. Hence, we proposed three methods to help researchers to detect differential gene sets using shrinkage covariance matrix combined withHotelling's T^2 statistic. The performances of the proposed methods were assessed using simulation study. Shrinkage covariance matrix approach shows a promising result for detection of differentially expressed gene sets as compared to other methods.

KEYWORDS: Hotelling'sT^2, gene set analysis, shrinkage covariance matrix

1 INTRODUCTION

Microarray technology is one of the significant achievements in biotechnology history and developed during the second half of the 1990s. An early article defining the application of DNA microarray technology to expression analysis was published in 1995 by Mark Schena and his colleagues at Stanford University [1].In broadest term, microarray technology may be defined as a high-throughput technology to examine the parallel gene expressions levels of thousands of genes at the same time. Precisely, microarray places an orderly arranged of many gene sequences in a grid. The grid that is often used is a glass slide. In general, a single microarray slide may contain thousands of spots. Each spot signifies a single gene and all of them representing the entire set of genes of an organism [2]. The technology has made a novelty discovery since its development and caught many researchers' attention. A number of researchers admit that the breakthrough of this technology is a vital research instrument. The widespread of microarray technology is largely due to its ability to give the quick results, relatively easy to use and precisely perform simultaneous analysis of thousands of genes in a massively parallel manner to researchers in one experiment, hence providing valuable knowledge on gene interaction and function [3].

The challenge of understanding the microarray gene expression has led to the development of new tools in the field of statistics for the analysis of gene expression data such as for the detection of differentially expressed genes between different biological states. Generally, the purpose of differential gene expression studies is to find those genes that produce different expression levels between samples [1]. All cells in the human body contain unique genetic material and the same genes are not active in every cell. Hence, the researchers will understand how these cells function normally and how they are affected when various genes do not perform properly by analyzing which genes are active and inactive in different cell types. For example, a researcher might wish to know the effect of certain treatment by examining the differences in gene activity between treatment and control samples. As a result, the researcher will understand exactly how the treatment affects the

*** Corresponding Author:** Suryaefiza Karjanto, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia,

genes and be able to develop more effective treatments later. Hence, microarray technology helps to further the study based on the gene activity. The gene expression from each sample is measured using microarray and the significantly different gene expression relative to treatment is calculated to conclude the effectiveness of treatment. The identification significantly changed gene study is also known as differential gene expression.

This study is to focus on differential gene expression in gene set analysis [4] and detect the differential gene sets that produce different expression levels between samples. The method is introduced in Section 3 after a description on the properties of Hotelling'sT^2statistic in Section 2. The performance of the proposed method is evaluated in Section 4 through simulation compared with existing methods.

2 Hotelling's T^2 Statistic

The Hotelling'sT^2 is named after Harold Hotelling in 1931, who developed the test statistic as a natural generalization of*t*-statistic. The test statistic develops in multivariate statistic which tests for univariate problems would make use of *t*-statistic. On the contrary, the *t*-statistic disregards for the correlation structure. This classical test statistic solves the univariate procedure problem and takes into account the correlation relationship between data.

The acceptable of this statistic in microarray analysis was due to the characteristics' suitability with the gene expression background of data. This method took into account the multidimensional structure of microarray data. The information for gene interactions was utilized to allow for finding genes whose differential expressions which cannot detectable by univariate methods. The Hotelling'sT^2 statistic gave a prediction rate that is at least as good as univariate procedure including the *t*-test. Furthermore, the test statistic is found to be more sensitive compared to the univariate *t*-statistic for the detection of the gene with certain conditions and summarized the Hotelling'sT^2 to be more efficient [5].

Let *n* represent the number of slides/samples, and *p* was the total number of genes in a gene set. Let X_{ki} be the expression level for gene *i* (where *i*=1, . . . , *p*) of sample *k* (where *k*=1, . . . , *n*) from the treatment group and X_{kj} be the expression level for gene *j* (where *j*=1, . . . , *p*) of sample *k* (where *k*=1, . . . , *n*)from the control group. The expression level vectors for samples *k* from the treatment and control groups can be expressed as $X_i = (X_{k1}, \ldots, X_{ki})^T$and $X_j = (X_{k1}, \ldots, X_{kj})^T$, respectively.The unknown population covariance matrix, $\sum$,was typically estimated by the sample covariance matrix, S_{ij} , for many situations. The sample covariance matrix, S_{ij} was defined as:

$$S_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}\left(X_{ki} - \overline{X}_i\right)\left(X_{kj} - \overline{X}_j\right) \qquad (1)$$

where X_{ki} and X_{kj} is the *k*-th observation of the variable X_i and X_j respectively.The mean, $\overline{X}_i$ was defined as:

$$\overline{X}_i = \frac{1}{n}\sum_{k=1}^{n} X_{ki} \qquad (2)$$

and the $\overline{X}_j$ is the mean for X_{kj}. Suppose we have n_1 and n_2 observations from two groups, such that $n_1 + n_2 = n$. Then, consider testing the null hypothesis that the two groups have equal multivariate means versus the appropriate alternative hypothesis, $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$. The test statistic based on Hotelling'sT^2 was defined as:

$$T^2 = \frac{n_1 n_2}{n}\left(\overline{X}_i - \overline{X}_j\right)' S^{-1}\left(\overline{X}_i - \overline{X}_j\right) \qquad (3)$$

For two subsamples, the pooled sample covariance matrix, *S,* was calculated as:

$$S = \frac{1}{n-2}\left((n_1 - 1)S^{(1)} + (n_2 - 1)S^{(2)}\right) \quad (4)$$

The sub-sample covariance matrix, $S^{(1)}$ and $S^{(2)}$ were defined as in equation (1). The maximum likelihood estimator was employed to obtain the sample covariance matrix. This estimator was unbiased when the number of samples is larger than the number of variables. As a result, the sample covariance matrix in Hotelling's T^2 poses the singularity problem when p is near to n and it is not invertible for p to exceed n. Thus, it will normally cause problem in hypothesis making as the test statistic become unstable.

3 Proposed Shrinkage Covariance Matrix

The proposed methods provide an alternative to estimate covariance matrix using shrinkage method based on the definition of [6, 7, 8, 9]. The approach was adapted to Hotelling'sT^2 and was extended to gene set analysis in microarray study. There were three proposed methods and we referred them as ShrinkA, ShrinkB and ShrinkC for the rest of this study. Generally, the algorithm for the three proposed methods was outlined below:

Step 1: We prepared the data sets with the preprocessing procedure by using suitable normalization and transformation method.

Step 2: We computed the shrinkage target.

Step 3: We searched for the optimal shrinkage intensity by using related definition.

Step 4: We replaced the sample covariance matrix in Hotelling's T^2 by using the results in **Step 2** and **Step 3**.

Step 5: We calculated theHotelling's T^2 for each of all the gene sets that were measured in datasets.

Step 6: We permuted samples for each gene set and declared as significant gene sets according to the permutation testing.

The proportion of each component in shrinkage estimation was:

$$S_{shrink} = \alpha T_{ij} + (1-\alpha)S_{ij} \quad (5)$$

where shrinkage target, T_{ij} and shrinkage intensity, α was defined as:

$$\alpha = \max\left\{0, \min\left\{\frac{\kappa}{n}, 1\right\}\right\} \quad (6)$$

where κ was a constant and n is the number of samples. The constant κ could be written as:

$$\kappa = \frac{\pi - \rho}{\gamma} \quad (7)$$

where π was the sum of asymptotic variances of the entries of the sample covariance matrix scaled by $\sqrt{n}$. ρ was the sum of asymptotic covariances of the entries of the shrinkage target with the entries of the sample covariance matrix scaled by $\sqrt{n}$. γ was the measurement of the misspecification of the (population) shrinkage

target. If κ were known, we could use κ/n as the shrinkage intensity in practice. Unfortunately, κ is unknown, so we searched for a consistent estimator for κ by $\hat{\kappa}$. This is done by finding consistent estimators for the three estimators π, ρ and γ that is $\hat{\pi}$, $\hat{\rho}$ and $\hat{\gamma}$. The proposed methods ensured the covariance matrix was always a positive definite and well defined. Table 1 showed the shrinkage target and shrinkage intensity for ShrinkA, ShrinkB and ShrinkC.

4 A Simulation Study

In order to evaluate the performance in the shrinkage covariance matrix, simulated data sets were developed by introducing the inter group correlation structure into the simulated data to imitate the multivariate structure in gene set. For a better interpretation of multivariate structure in gene set, the correlation matrix was used. The multivariate normal distribution data was generated using *mvrnorm* function in the *MASS* package. The generated data was assumed as correlation matrix using *rcorrmatrix* function in the *clusterGeneration* package. All programming codes and packages were written in *R* language (http://cran.r-project.org/).

Table 1. The shrinkage combinations for ShrinkA, ShrinkB and ShrinkC

Type	Shrinkage Target	Shrinkage Intensity
ShrinkA	$T_{Aij} = \begin{cases} s_{ii} & if\ \ i=j \\ 0 & if\ \ i\neq j \end{cases}$	$\hat{\kappa}_A = \dfrac{\hat{\pi}-\hat{\rho}_A}{\hat{\gamma}_A}$, $\hat{\pi} = \dfrac{1}{n}\sum_{k=1}^{n}\left\{\left(X_{ki}-\bar{X}_i\right)\left(X_{kj}-\bar{X}_j\right)-S_{ij}\right\}^2$ $\hat{\rho}_A = 0$ $\hat{\gamma}_A = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(S_{ij}\right)^2$
ShrinkB	$T_{Bij} = \begin{cases} s_{ii} & if\ \ i=j \\ \sqrt{s_{ii}s_{jj}} & if\ \ i\neq j \end{cases}$	$\hat{\kappa}_B = \dfrac{\hat{\pi}-\hat{\rho}_B}{\hat{\gamma}_B}$, $\hat{\rho}_B = \underbrace{\sum_{i=1}^{p}\hat{\pi}_{ii}}_{on\ diagonal} + \underbrace{\sum_{i=1}^{p}\sum_{j=1,j\neq i}^{p}\frac{1}{2}\left(\sqrt{\frac{S_{jj}}{S_{ii}}}\hat{\vartheta}_{ii,ij}+\sqrt{\frac{S_{ii}}{S_{jj}}}\hat{\vartheta}_{jj,ij}\right)}_{off\ diagonal}$, $\hat{\pi}_{ii} = \frac{1}{n}\sum_{k=1}^{n}\left\{\left(X_{ki}-\bar{X}_i\right)^2-S_{ii}\right\}^2$ $\hat{\vartheta}_{ii,ij} = \frac{1}{n}\sum_{k=1}^{n}\left\{\left(X_{ki}-\bar{X}_i\right)^2-S_{ii}\right\}\left\{\left(X_{ki}-\bar{X}_i\right)\left(X_{kj}-\bar{X}_j\right)-S_{ij}\right\}$ $\hat{\vartheta}_{jj,ij} = \frac{1}{n}\sum_{k=1}^{n}\left\{\left(X_{kj}-\bar{X}_j\right)^2-S_{jj}\right\}\left\{\left(X_{ki}-\bar{X}_{i.}\right)\left(X_{kj}-\bar{X}_j\right)-S_{ij}\right\}$ $\hat{\gamma}_B = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(f_{ij}-S_{ij}\right)^2$ $, f_{ij} = \sqrt{S_{ii}S_{jj}}$
ShrinkC	$T_{Cij} = \begin{cases} s_{ii} & if\ \ i=j \\ \bar{r}\sqrt{s_{ii}s_{jj}} & if\ \ i\neq j \end{cases}$	$\hat{\kappa}_C = \dfrac{\hat{\pi}-\hat{\rho}_C}{\hat{\gamma}_C}$ $\hat{\rho}_C = \hat{\rho}_B$ $\hat{\gamma}_C = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(f_{ij}-S_{ij}\right)^2$ $, f_{ij} = \bar{r}\sqrt{S_{ii}S_{jj}}$

The separation between the two groups measured the difference in the means of the multivariate normal distributions where μ was the vector of gene means and Σ was the covariance matrix of the gene expression on the following density function:

$$fx\left(x_{1,.....,}x_p\right)=\frac{1}{(2\pi)^{p/2}\left|\Sigma\right|^{1/2}}e^{-(x-\mu)'\Sigma^{-1}(x-\mu)/2} \quad (8)$$

The gene set variances were set at one and assumed that the number of samples for both groups is equal. Each case was permutated 10000 times and 100 data sets were generated. The simulated data sets were set to explore the performance of proposed method for two hypotheses: Case 1: No difference (separation) exists between two groups (null hypothesis) and Case 2: There was difference (separation) exists between groups (alternative hypothesis).

The performance of our approach was evaluated by comparing the results with those obtained from two other methods: (1) by using principal component analysis to solve the high dimensionality problem proposed by Kong *et al.* [10] denoted as KPCA, and (2) the Regularized Covariance Matrix Approach (RCMAT) introduced by Yates and Reimers [11]. The RCMAT is quite similar with our proposed methods but the covariance matrix in Hotelling'sT^2 is regularized using the following identity matrix to replace the shrinkage target in equation (5):

$$T_{ij}=\begin{cases}1 & if\ \ i=j\\ 0 & if\ \ i\neq j\end{cases} \quad (9)$$

Since the shrinkage target was penalized to zero and the diagonal to one, consequently information from the covariance matrix is not fully utilized. The shrinkage intensity, α in equation (6) was reduced from 1 towards 0 by increments of 0.01 and the optimum shrinkage intensity would be achieved when the smallest positive eigenvalue was bigger than the reciprocal of the number of genes in the gene set. The optimum intensity would ensure the covariance matrix is a positive definite and invertible. RCMAT and KPCA were comparable with our approach since they were also using Hotelling'sT^2 for testing differentially expressed gene sets.

4.1 Case 1: No difference (separation) exists between two groups (null hypothesis)

A simulation study was performed with parameter combinations under the null hypotheses as display in Table 2. A total of four parameter combinations consisted of two default setting and two changed settings were examined. The default setting used major of axis of separation and no amount of separation at all. For each of the parameter combinations in this simulation study, parameters setting were changed relative to the default setting:

i. Increasing number of variables of 10 and 30;
ii. Increasing number of sample sizes of 10, 20 and 50;
iii. A major axis of variation and;
iv. No amount of separation.

The above parameter combinations were employed to monitor the performance when the three conditions are applied: $n>p$, $n=p$ and $n<p$. The distribution of p-values was evaluated when no difference exists between the groups in the mean of expression measures of genes in the gene set. In a two-group comparison, each p-value between 0 and 1 was equally likely. The distribution of p-values for ShrinkA, ShrinkB, ShrinkC, RCMAT and KPCA when no difference was detected between the two groups is displayed.All the 100 ranked p-values of the simulation results were displayed using QQ-plot against the uniform distribution.

Table 2. The parameter combinations under null hypothesis

Parameter combinatio	No. of variables	Sample size	Axis of variation	Amount of separatior
1	10	10	Major	0.00
2	10	20	Major	0.00
3	10	50	Major	0.00
4	30	10	Major	0.00
5	30	20	Major	0.00
6	30	50	Major	0.00

4.2 Case 2: There is difference (separation) exists between groups (alternative hypothesis)

We focused on the power of our proposed methods and discovered that our simulation study spans both highly significant to clearly insignificant separations as determined by the mean nominal *p*-value. The simulated data for twelve parameter combinations were generated under alternative hypothesis as summarise in Table 3. Four parameter settings, which included one default setting and three altered settings were studied. For presentation, the default setting used only 20 samples. Specifically, the simulation setup for each of the parameter combinations with three variables were altered relative to the default setting was as follows:

i. Increasing number of variables of 10 and 30;
ii. Number of samples is 20;
iii. Different axis of variation of a major axis of variation and a minor axis of variation and;
iv. Increasing amount of separation between groups of 0.25, 0.50 and 1.00.

Such parameter combinations above were generated to monitor the performance when the two conditions were applied: $n>p$ and $n<p$. Then, all results of the cumulative distribution function of nominal *p*-values were illustrated.

4 RESULTS AND DISCUSSION

The simulation study was performed using four parameter combinations consisted of major of axis of separation and no amount of separation at all as default settings and two set of variables; 10 and 30 and three different sample sizes; 10, 20 and 50 (refer Table 2 for detail explanation of parameter combination). The distribution of *p*-values was evaluated when no difference existed between the groups in the mean of expression measures of genes in the gene set. Table 4 provided a summary of the mean nominal *p*-values of ShrinkA, ShrinkB, ShrinkC, RCMAT and KPCA under no difference (separation) existed between two groups (null hypothesis).

Table 3. The parameter combinations under alternative hypothesis

Parameter combination	No. of variables	Sample size	Axis of variation	Amount of separation
1	10	20	Major	0.25
2	10	20	Minor	0.25
3	10	20	Major	0.50
4	10	20	Minor	0.50
5	10	20	Major	1.00
6	10	20	Minor	1.00
7	30	20	Major	0.25
8	30	20	Minor	0.25
9	30	20	Major	0.50
10	30	20	Minor	0.50
11	30	20	Major	1.00
12	30	20	Minor	1.00

Table 4. Mean nominal *p*-values of ShrinkA, ShrinkB, ShrinkC, RCMAT and KPCA under no difference (separation) exists between groups.

Parameter combinations	Method				
	Shrink A	Shrink B	Shrink C	RCMAT	KPCA
1	0.4648	0.4666	0.4698	0.4822	0.4874
2	0.4993	0.5316	0.5094	0.5177	0.5221
3	0.5030	0.5104	0.4962	0.4897	0.5063
4	0.5473	0.5624	0.5398	0.5223	0.5363
5	0.5001	0.4793	0.5009	0.5144	0.5100
6	0.4445	0.4766	0.4430	0.4631	0.4910

From above Table 4, the highest mean *p*-value when number of variables was 10 and number of samples was 10 or parameter combination 1 was belong to KPCA with *p*-value 0.4874. When the condition number of variables was 10 with number of samples was 30 and 50 producing $n>p$ condition (parameter combination 2 and parameter combination 3), ShrinkB had the highest mean *p*-value, 0.5316 and 0.5104 respectively and also when number of variables is 30 and number of samples was 10 (parameter combination 4) was 0.5624. In addition, RCMAT had 0.5144 as the highest mean *p*-value when number of variables was 30 and number of samples was 20 (parameter combination 5). The condition with number of variables was 30 and number of samples was 50 (parameter combination 6), the highest mean *p*-value once again belong to KPCA with 0.4910.

Table 5 contained the mean nominal *p*-values of ShrinkA, ShrinkB, ShrinkC, RCMAT and KPCA under alternative hypothesis (refer Table 3 for detail conditions). When number of variables increased from 10 to 30 with fixed number of samples, the mean *p*-value shifted from 0.3261 to 0.3373 for ShrinkA and 0.3362 to 0.3548 for ShrinkC along a major axis of variation and amount of separation is 0.25 (parameter combination 1 to parameter combination 7). On the other word, ShrinkA and ShrinkC exhibited good performance with increasing number of variables with lower mean *p*-value. For same conditions, the ability of detectionwas followed by RCMAT with mean *p*-value shifted from 0.3457 to 0.4213, ShrinkB from 0.3866 to 0.0577 and KPCA from 0.4053 to 0.4430.

For the increased of amount of separation to 0.5 and 1.0 along a major axis of variation when number of variables increased from 10 to 30 with 20 number of samples (parameter combination 3 to parameter combination 9 and parameter combination 5 to parameter combination 11), the same situation was also found which the mean *p*-value of ShrinkA and ShrinkC still lower than other methods. We suggested that the detection power of the ShrinkA and ShrinkC method increased as the amount of separation between two groups increased.

From mean *p*-value, it showed that shrinka easily detected the difference (separation) between groups compared to other methods when axis of variation was changed from major to minor at most of the conditions. Interestingly, we observed that RCMAT and ShrinkC detected the separation easily, in that respective order. For example, when the axis of variation was shifted from major to minor, the mean *p*-value increases from 0.0001 to 0.0551 for shrinka, from 0.0014 to 0.0399 for rcmat, from 0.0040 to 0.0516 for shrinkc, from 0.0192 to 0.0956 for KPCA and from 0.0815 to 0.1519 for shrinkb along the amount of separation was 1.0 (n = 20) with ten variables. (parameter combination 5 to parameter combination 6).

Table 5. Mean nominal *p*-values of ShrinkA, ShrinkB, ShrinkC, RCMAT and KPCA under there is difference (separation) exists between groups.

Parameter combinations	Method				
	Shrink A	Shrink B	Shrink C	RCMAT	KPCA
1	0.3261	0.3866	0.3362	0.3457	0.4053
2	0.4142	0.4738	0.4284	0.4175	0.4764
3	0.3255	0.2656	0.1183	0.1223	0.2395
4	0.2230	0.3547	0.3027	0.2877	0.3459
5	0.0001	0.0815	0.0040	0.0014	0.0192
6	0.0551	0.1519	0.0516	0.0399	0.0956
7	0.3373	0.4577	0.3548	0.4213	0.4430
8	0.4493	0.4766	0.4532	0.4735	0.4669
9	0.1324	0.4506	0.1516	0.0209	0.3912
10	0.3748	0.4697	0.3752	0.3875	0.4395
11	0.0004	0.2575	0.0006	0.0050	0.1560
12	0.0159	0.4125	0.1525	0.1532	0.3336

5 Conclusion

The understanding of biological data has led to the development of new methods in statistics such as [12] and [13]. In this study, we concluded that ShrinkA, ShrinkB, ShrinkC, RCMAT and KPCAmethods produced conservative bias comparative to the expected *p*-value. The deviation from the 45° straight line of q-q plot also was getting larger when number of variables were getting higher than number of samples. However, there was a good agreement between uniform distribution and ShrinkB and KPCA to accept the null hypothesis. On the other word, when no difference (separation) existed between two groups, the ShrinkB and KPCA performed better than other methods.

Basically, the real differences between the groups were easier detected with $n>p$ than $n<p$ conditions for all methods with certain axis of variation and amount of separation. Furthermore, from axis of variation's perspective, the differences between the groups with any conditions with a major axis were easier detected rather than a minor axis because of the larger variance. As we expected, the differences between groups for large amount of separation were easier detected compared with smaller amount of separation. All methods were performed consistently according to conditions described earlier but detection ability of ShrinkA and ShrinkC method was higher than other methods across conditions especially along a major axis of variation.

This study discovered the potential of the shrinkage approach to estimate the covariance matrix for microarray data, particularly in comparing gene expression between independent samples. The shrinkage covariance matrix approach showed promising results for testing the differential gene sets expression compared to two established methods. This research provided a new platform and opportunities for further research or studies in microarray-based gene sets and the results were expected to be of interest for further applications in other areas of research with similar data characteristics.

6 Acknowledgement

We would like to extend our appreciation and gratitude to Universiti Teknologi Mara (UiTM) and Research Management Institute of UiTM for supporting this research under the Research Intensive Grant (RIF) with reference number 600-RMI/DANA 5/3/RIF (221/2012).

REFERENCES

1. Schena. M, Shalon. D, Davis. RW and Brown. PO, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray", Science 1995; 270 (5235): 467-470.
2. Babu. MM, "Introduction to microarray data analysis", Computational Genomics: Theory and Application 2004; 225-249.
3. Szabo. A, Boucher. K, Jones. D, Tsodikov. D, Klebanov. LEVB and Yakovlev. AY, "Multivariate exploratory tools for microarray data analysis", Biostatistic 2003; 4(4): 555-567.
4. Dubitzky. W, Granzow. M, Downes. C and Berrar. D, "Introduction to microarray data analysis", A Practical Approach to Microarray Data Analysis 2003; 1-46.
5. Lu. Y, Liu. PY, Xiao. P and Deng. HW, "Hotelling's T^2 multivariate profiling for detecting differential expression in microarrays", Bioinformatics 2005; 21(14): 3105–3113.
6. Schäfer. J and Strimmer. KA, "Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics", Statistical Applications in Genetics and Molecular Biology 2005; 4(1): 32.
7. Ledoit. O and Wolf. M, " Improved estimation of the covariance matrix of stock returns with an application to portfolio selection", Journal of Empirical Finance 2003; 10.5: 603-621.
8. Ledoit. O and Wolf. M, "Honey, I shrunk the sample covariance matrix", The Jurnal of Portfolio Management 2004; 31(1): 110-119.

9. Ledoit. O and Wolf. M, "A well-conditioned estimator for large dimensional covariance matrices", Journal of Multivariate Analysis 2003; 88: 365–411.

10. Yates. PD and Reimers. MA, "RCMAT: A regularized covariance matrix approach to testing gene sets", BMC Bioinformatics 2009; 10: 300.

11. Kong. SW, Pu. WT and Park. PJ, "A multivariate approach for integrating genome-wide expression data and biological knowledge", Bioinformatics 2006; 22(19): 2373-2380.

12. Ullah, R., Zaman, G., & Islam, S. (2013). Stability analysis of a general SIR epidemic model. *VFAST Transactions on Mathematics*, *1*(1).

13. ZEB, A., Zaman, G., MOMANI, S., & ERTÜRK, V. S. (2013). Solution of an SEIR Epidemic Model in Fractional Order. *VFAST Transactions on Mathematics*, *1*(1).